Executing a Reusable Framework for Study-Specific Data Quality Analysis

Kaleigh Wieand[1], Hanieh Razzaghi[1], Kim Dickinson[1], Michael Kahn[2], Jason Roy[3,]
Charles Bailey[1]
Children's Hospital of Philadelphia[1], University of Colorado[2], Rutgers University[3]

Background

Secondary use of electronic health record (EHR) data for research offers the opportunity to conduct research rapidly and generate clinical insights expediently. However, because this data has not been curated for research, it often contains major data quality issues that are uncovered too late in the analytic stage of a research study or project. Every study faces its own unique set of data quality challenges based on the cohort, variables of interest, and analytic methods. Traditional approaches to data quality are often not domain- or use-case specific and there are no standard approaches for assessing data fitness. As a result, most study specific data quality is done ad hoc, which reduces its generalizability and limits reproducibility. Further, the absence of a structured method to interrogate the data can lead to issues appearing late in the analysis or not at all, which can delay timelines and impact the reliability of results. The Kahn framework[1] began to address this problem by identifying broad categories of data quality analyses recommended as a starting point for investigators. And while OHDSI's Characterization tool[2] does offer a structured methodology for analysis, it is limited to producing high-level counts and summary statistics. As a result, the question remains of how to operationalize the Kahn categories in a manner that provides investigators with a clear path forward while allowing for both complex analytic output and flexible customization to suit study-specific needs.

Methods

To bridge this gap, we developed a suite of checks with an underlying reusable framework that can be used across study or research contexts. A group of experts convened to review current progress and provide feedback on potential improvements and additions to the process. After several iterations, we decided on a framework that focuses on three common facets of data quality analysis and allows users to design a given check to fit their specific needs. Users can choose between single or multi-site analysis, exploratory or anomaly detection analysis, and cross-sectional or longitudinal analysis. With these parameters, each check has up to 8 base configurations (Table 1) while still maintaining a customizable data input structure to allow for any cohort, concept set, or variable definition to be examined as part of check execution.

	Single / Multi-Site	Exploratory / Anomaly Detection	Cross-Sectional / Longitudinal
1	Single Site	Exploratory	Cross-Sectional
2	Single Site	Anomaly Detection	Cross-Sectional
3	Multi-Site	Exploratory	Cross-Sectional
4	Multi-Site	Anomaly Detection	Cross-Sectional
5	Single Site	Exploratory	Longitudinal
6	Single Site	Anomaly Detection	Longitudinal
7	Multi-Site	Exploratory	Longitudinal
8	Multi-Site	Anomaly Detection	Longitudinal

Table 1. Eight available base check configurations

The output of a check in tabular format is returned to the user, which can then be taken and analyzed outside of the framework ecosystem. If preferred, the suite is also accompanied by out-of-the-box visualizations built to assist the user in interpreting the results.

Results

The suite currently consists of nine unique modules adapted to be executed against an OMOP database and that cover multiple data quality domains (GitHub). Each module consists of two reusable functions, one to generate tabular output and another to produce graphical visualizations, and a library of metadata documenting every feature. The reusable functions use standard, study-agnostic parameters as outlined by the underlying framework, and configurable CSV files fed into each check give users the freedom to evaluate any variable or domain and quickly update the input when the use case changes. The visualizations that accompany each check, while not required, are designed to reduce workload of the user and simplify the process of interpreting the tabular results. The standardization of graphical output allows for comparability across different study contexts, and the reuse of graph types between modules streamlines the interpretation of results across the whole suite. To demonstrate the process of adjusting a single function

to easily generate different types of output, the function inputs and corresponding graphical output for Base Configuration 5 (Figure 1) and Base Configuration 4 (Figure 2) as seen in Table 1 are included below. By changing the inputs for just three parameters (multi_or_single_site, anomaly_or_exploratory, and time), the user can visualize the underlying data in two unique and informative ways. Additional check-specific parameters also allow for further customization of the results. For example, the domain_tbl parameter reads in a custom CSV file defining each of the domains of interest, meaning the user can easily add or subtract new domains as study needs change without having to alter any of the underlying code.

```
pf_process(cohort = cohort_table,
           multi_or_single_site = 'single',
           anomaly_or_exploratory = 'exploratory',
           time = TRUE,
           time_span = c('2015-01-01', '2022-01-01'),
           time_period = 'year',
           study_name = 'my_study',
           visit_types = c('inpatient', 'outpatient'),
           domain_tbl = read_csv('my_study_domains.csv'),
           visit_type_table = read_csv('my_visit_type_definitions.csv'))
```

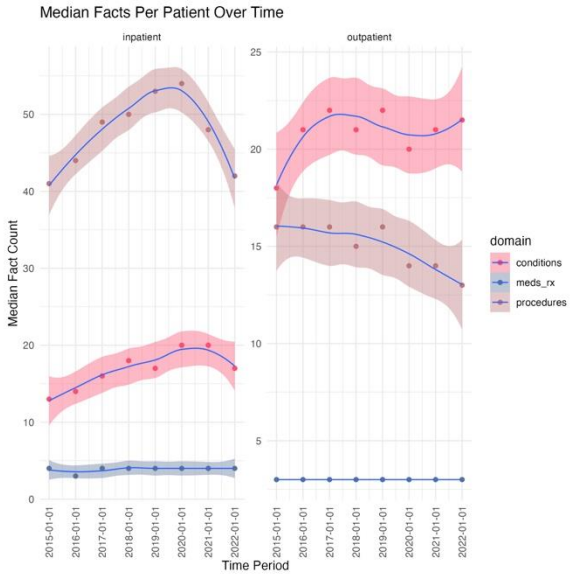

Figure 1. Single Site, Exploratory, Over Time: Function Input & Visualization

```
pf_process(cohort = cohort_table,
           multi_or_single_site = 'multi',
           anomaly_or_exploratory = 'anomaly',
           time = FALSE,
           study_name = 'my_study',
           visit_types = c('inpatient', 'outpatient'),
           domain_tbl = read_csv('my_study_domains.csv'),
           visit_type_table = read_csv('my_visit_type_definitions.csv'))
```

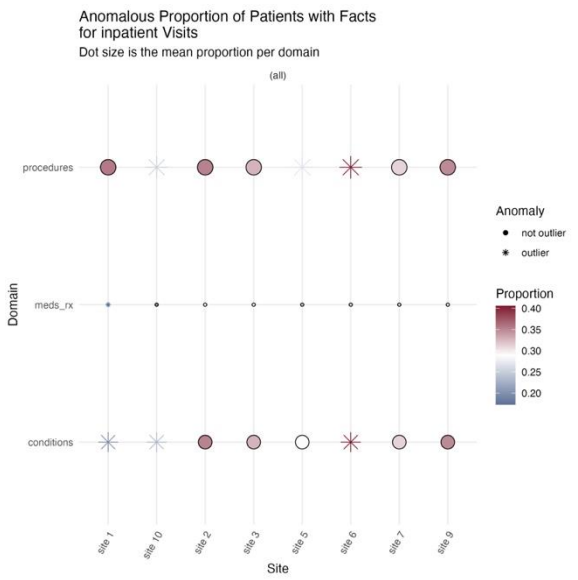

Figure 2. Multi-Site, Anomaly Detection, Static: Function Input & Visualization

Conclusion

We have developed a data quality program that streamlines the process of selecting and executing study-specific evaluations to assess fitness for an intended use. The structured framework, combined with the function inputs' flexibility, facilitates the application of these checks across assorted studies and will allow investigators to easily implement a standard set of data quality analyses. The suite is still growing, with new checks consistently being added to increase the number and types of investigations that are possible.

References

1. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, Estiri H, Goerg C, Holve E, Johnson SG, Liaw ST, Hamilton-Lopez M, Meeker D, Ong TC, Ryan P, Shang N, Weiskopf NG, Weng C, Zozus MN, Schilling L. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC). 2016 Sep 11;4(1):1244. doi: 10.13063/2327-9214.1244. PMID: 27713905; PMCID: PMC5051581
2. Reps J, Ryan P (2024). *Characterization: Characterizations of Cohorts*. R package version 0.2.0, https://github.com/OHDSI/Characterization, https://ohdsi.github.io/Characterization