A Systematic and Sustainable Solution for Assessing Network Data Quality

Kimberley Dickinson[1], Kaleigh Wieand[1], Charles Bailey[1], Hanieh Razzaghi[1]
[1] Applied Clinical Research Center, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania

Background

Performing research on multi-institutional Electronic Health Record (EHR) data can provide insight on a large, diverse patient population, but the quality of the research is highly dependent on the quality of the underlying data. Issues within an institution and discrepancies between institutions can be difficult to uncover and can lead to unexpected or biased results when an analysis is applied uniformly across a patient population. Many networks have developed approaches to assessing DQ, but they are often designed to address specific network concerns.[1, 2, 3] In PEDSnet, a pediatric learning health system with a centralized database used for multidisciplinary healthcare research, we have developed a modular program that augments the capabilities of existing tools such as OHDSI's Data Quality Dashboard (DQD) by not only visualizing findings but incorporating the DQ check process into a larger system of thresholding, communication, and resolution. The process has proven beneficial to our research, but is also designed for reproducibility and scalability. Our DQ program incorporates input from experts who transform local EHR data into an OMOP CDM and data scientists at the Data Coordinating Center (DCC) engaged in scientific research to develop and build checks and to detect and resolve issues.

Methods

We designed our DQ program with the needs of PEDSnet in mind as well as those of the OHDSI community and beyond. The program is highly flexible: we can extend it easily and deploy it across other systems. Prior to developing our DQ system, we identified primary features for reproducibility and sustainability, one of which is the adoption of naming conventions to facilitate defining and tracking metadata. We use the term check type to define a categorization of the purpose of the DQ check. Each combination of check type + check application has a unique identifier used to track metadata. For example, the check type *data cycle changes* computes differences in patient and record counts in a domain between two data versions. The check application *dc_vip_person* within the *data cycles changes* check type measures the difference in the number of persons with at least one inpatient visit between two versions. With this check, we can detect unexpected increases or decreases between data refreshes. To date, we have developed 10 check types and 238 check applications which are described and quantified in Table 1.

Check Type	Check Type Description	Example	Number of Check Applications
Data cycle changes	Computes person counts and row counts between two versions of a dataset	dc_vip_person measures the difference in the number of persons with at least one inpatient admission between versions	72
Vocabulary conformance	Checks present against prescribed vocabularies	vc_co_cid_rows checks whether values in the condition_occurrence.concept_id field are in the allowed vocabularies for diagnosis codes	9
Value set conformance	Checks value sets present against those defined by data model specifications	vs_pd_race_cid_rows checks whether values in the person.race_concept_id fields are in the expected value set for race concepts	6
Unmapped concepts	Computes the proportion of rows not mapped to a non-NULL concept for a field	uc_dr_rows computes the proportion of rows in drug_exposure.dose_unit_concept_id not mapped to a concept_id	15
Completeness of visit facts	Identifies visits with no associated facts	pf_ipvisits_dr_visits computes proportion of inpatient visits with at least one associated drug record	42
Best mapped concepts	Identifies whether concepts are mapped to a preferred level of specificity for the given vocabulary	bmc_rxnorm_dp_rows measures the proportion of drug_exposure.drug_concept_id, limited to prescriptions, that are at least to the specified level of preference in the RxNorm hierarchy	7
Facts over time	Measures change clinical fact volume over a range of time	fot_voml_person measures the difference in the number of people with an outpatient lab in a time increment (e.g. month over month)	64
Domain concordance	Measures the degree of patient overlap meeting two criteria	dcon_ed_visits_conds measures the concordance of patients with an emergency department visit and a diagnosis with an Emergency Header condition_type_concept_id during the visit	9
Facts with associated visit identifier	Computes the proportion of records in each domain that do not have a visit_occurrence_id	mf_visitid_pr_rows finds the number of rows in the procedure_occurrence table that are not associated with a visit_occurrence_id	6
Expected concepts present	Computes the proportion of patients with evidence of a specified concept	ecp_hemoglobin_person computes the number of people with at least one hemoglobin lab divided by the total number of patients with a drug, procedure, and lab record	8

Table 1. DQ check types and check applications

Results

We have integrated check type structure into our overall DQ process, which consists of 3 modular R programs: **library**, **processing**, and **visualization**. Each module operates in the PEDSnet standard R framework, a tool built to facilitate interoperability with remote databases. The output for each module can be directed to a database schema or to csv files, at the user's discretion. There is also a site communication component which is carried out through REDCap forms which are generated and queried in R via REDCap's API. Figure 1 shows an overview of this cyclical process.

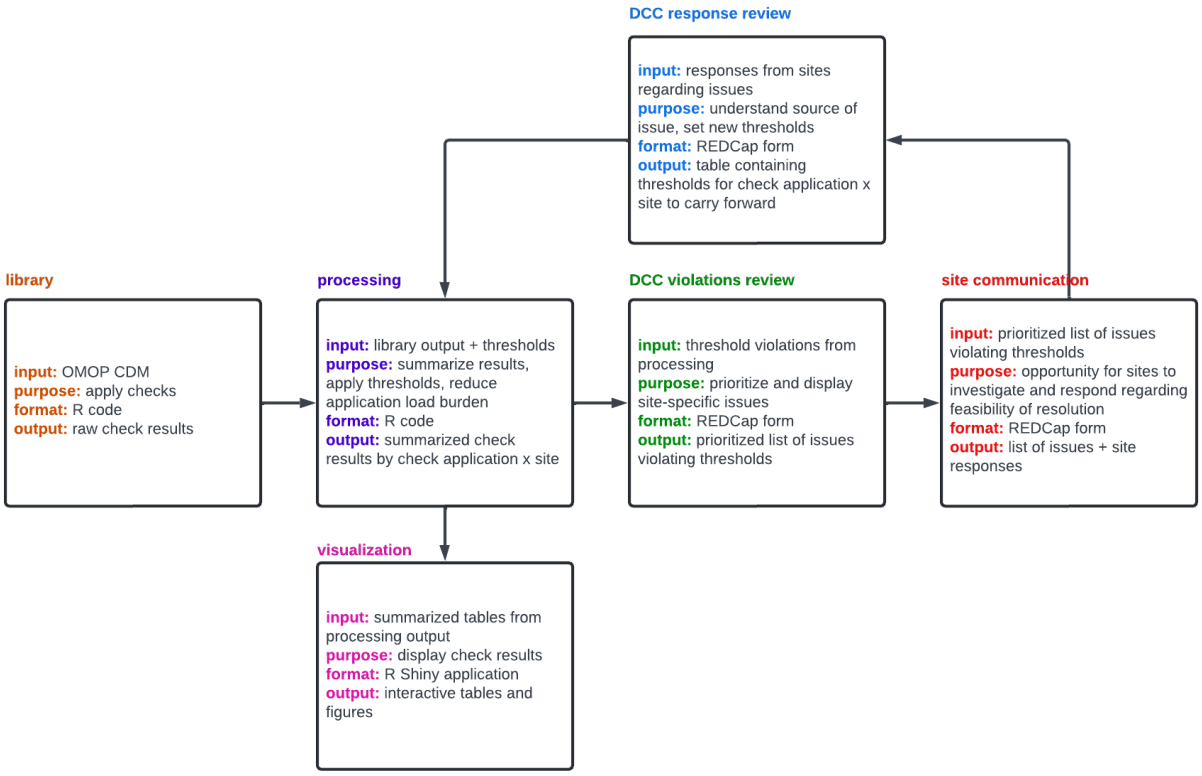

Figure 1. Diagram of PEDSnet DQ process

The **library** module (https://github.com/PEDSnet/dqa_library) queries the OMOP CDM and generates results aggregated by check application and site. The **processing** module (https://github.com/PEDSnet/dqa_processing) reads the library results, applies thresholds, and performs computations to reduce downstream computational time for visualizations.

The **DCC violations review**, **site communication**, and **DCC response review** portion of the process is unique to PEDSnet's DQ infrastructure. Issues flagged in the processing step are imported into a REDCap form and prioritized based on the magnitude and impact on the network. Sites respond, indicating whether the issue's source is known and there is potential for resolution. Responses are reviewed and carried forward into the next round of thresholds, which we can adjust to reflect realistic standards. This feedback loop allows us to limit the issues raised to sites to those that are important and potentially resolvable.

The **visualization** module (https://github.com/PEDSnet/dqa_shiny) reads results from the processing step and displays figures and tables on an R Shiny dashboard. The PEDSnet dashboard is published on a website accessible to all users of our data network, and the code to generate it is accessible on GitHub (https://github.com/PEDSnet/dqa_shiny). It contains network-wide and site-specific visualizations, allowing for comparison across institutions and individual site drill-down. There is a page dedicated to each check type, with visualizations tailored to each check application including bar plots, heat maps and line plots, many of which are interactive. There is dynamic filtering allowing for selection of a site or domain of interest. The dashboard is intended to be an exploratory tool and not a report card. We recognize that unexpected findings, though important to realize, may not be fixable, and it may be appropriate for thresholds to vary across sites depending on their institution's practice patterns and data structures.

Figure 2 is the dashboard page for the *data cycle changes* check type. The heatmap at the top of the page displays each check application by site, with colors indicating magnitude and direction of proportion of change in record count between data the prior and current data submission. Hovering over a tile displays additional metadata. The dropdowns and checkboxes on the left allows the user to filter the content displayed on the pages a site, domain, and check applications of interest.

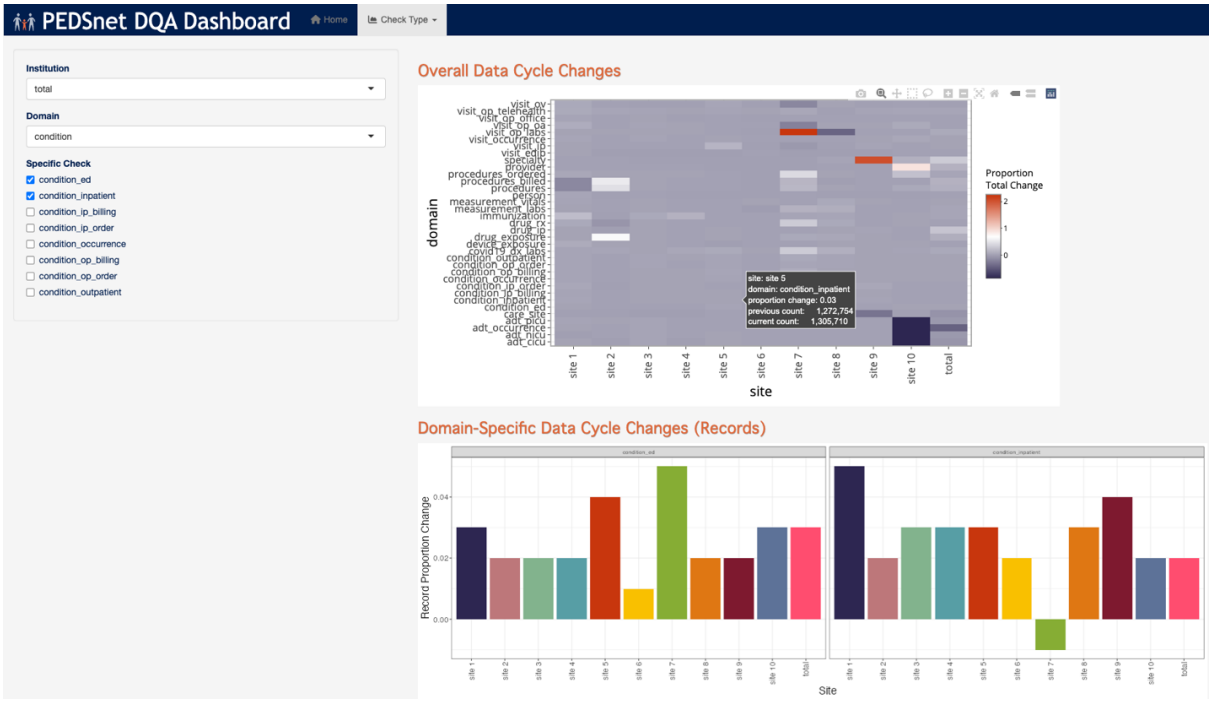

Figure 2. Screenshot of Data Cycles Changes page on the DQ Dashboard

Conclusion

Our data quality program has enhanced our network's ability to detect, diagnose, and resolve issues using a reproducible and sustainable approach. The program has helped us to improve the quality of our data across many domains to make it more research-ready and to provide insight into underlying data intricacies at the outset of new research projects. Our focus on generating the infrastructure in a reproducible manner will allow for other OMOP networks to directly apply our processes, and even for networks using alternative data models to learn from our processes and apply the theoretical framework to their own data.

References

1. Sidky H, Young JC, Girvin AT, Lee E, Shao YR, Hotaling N, Michael S, Wilkins KJ, Setoguchi S, Funk MJ; N3C Consortium. Data quality considerations for evaluating COVID-19 treatments using real world data: learnings from the National COVID Cohort Collaborative (N3C). BMC Med Res Methodol. 2023 Feb 17;23(1):46. doi: 10.1186/s12874-023-01839-2. PMID: 36800930; PMCID: PMC9936475.
2. Abigail E Lewis, Nicole Weiskopf, Zachary B Abrams, Randi Foraker, Albert M Lai, Philip R O Payne, Aditi Gupta, Electronic health record data quality assessment and tools: a systematic review, Journal of the American Medical Informatics Association, Volume 30, Issue 10, October 2023, Pages 1730–1740, https://doi.org/10.1093/jamia/ocad120
3. Bian J, Lyu T, Loiacono A, Viramontes TM, Lipori G, Guo Y, Wu Y, Prosperi M, George TJ, Harle CA, Shenkman EA, Hogan W. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. J Am Med Inform Assoc. 2020 Dec 9;27(12):1999-2010. doi: 10.1093/jamia/ocaa245. PMID: 33166397; PMCID: PMC7727392.