Identifying and Understanding Data Quality Issues in a Pediatric Distributed Research Network

Levon Utidjian, MD, MBI[1], Ritu Khare, PhD[1], Evanette Burrows, MS[1], Greg Schulte[2], Kevin Murphy, BS[1], Sara Deakyne, MPH[2], Richard Hoyt, BS[3], Nandan Patibandla, MS[4], Byron Ruth, BS[1], Aaron Browne, BA[1], Megan Reynolds, BBA[3], Keith Marsolo, PhD[5], Michael Kahn, MD, PhD[2], L. Charles Bailey, MD, PhD[1]

[1]Children's Hospital of Philadelphia, [2]Children's Hospital Colorado, [3]Nationwide Children's Hospital, [4]Boston Children's Hospital, [5]Cincinnati Children's Hospital Medical Center

Background

Collaborations across multiple institutions are essential to achieving adequate cohort sizes in pediatrics research. PEDSnet is a newly established clinical data research network (CDRN) that aggregates electronic health record (EHR) data from 8 of the nation's largest children's hospitals. PEDSnet is part of a much larger network, PCORnet, supported by the Patient-Centered Outcomes Research Institute (PCORI). The goal of PEDSnet is to support a variety of research projects, including comparative effectiveness studies, drug safety studies, descriptive studies, and computable phenotypes.

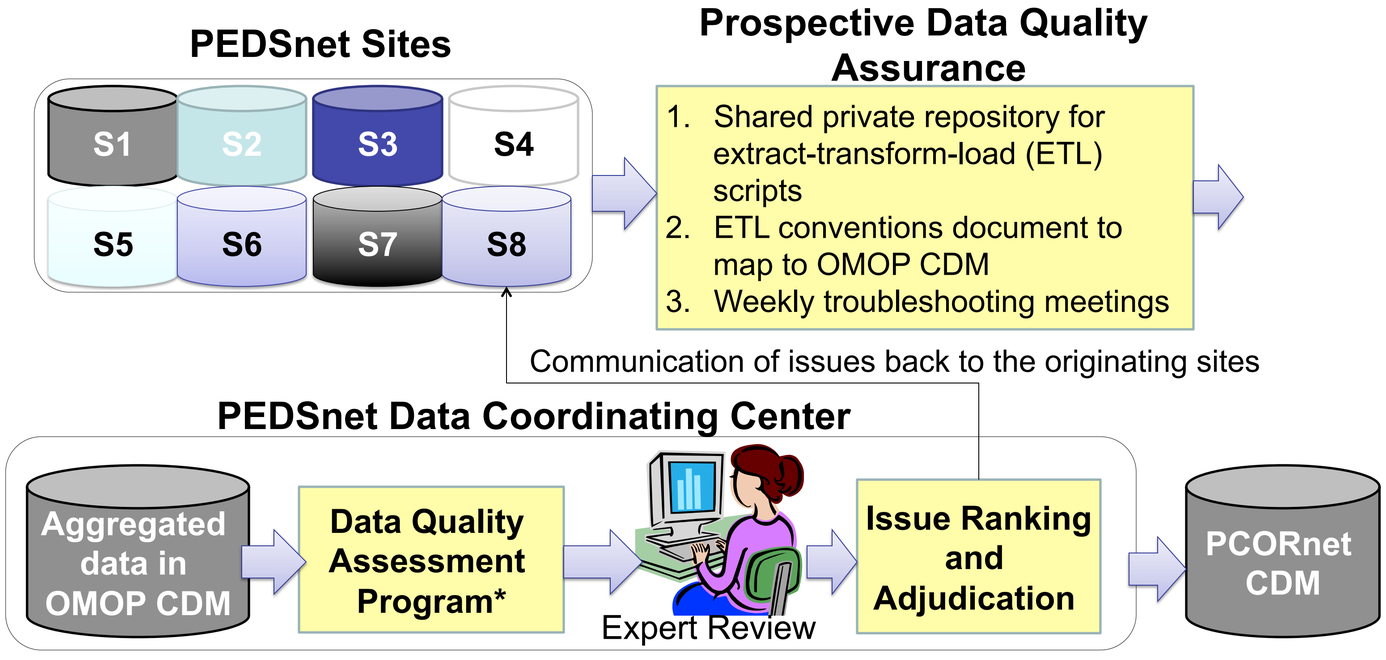
Prominent challenges in building PEDSnet include:

- Lack of EHR data's orientation toward research analytics: unstructured and locally coded data, complex clinical workflows
- Semantic heterogeneity: diverse source systems (5 Epic, 2 Cerner, 1 AllScripts) and vocabularies (ICD9, RxNorm, GPI, CPT-4, etc.)
- Data peculiarities in pediatrics: procedures conducted before a child is born such as prenatal testing or surgery
- Need to supplement the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) with data elements like extra concept identifiers and tables for clinical information like visit payor.

A prerequisite in PEDSnet is to ensure that the network's data is "high quality" but there is a lack of standardized guidelines for conducting data quality assessments (DQA) in large-scale distributed networks.

Methods

PEDSnet Data Quality Assessment Workflow



* DQA Program used scripts to explore data distributions & help identify outliers as in Achilles.

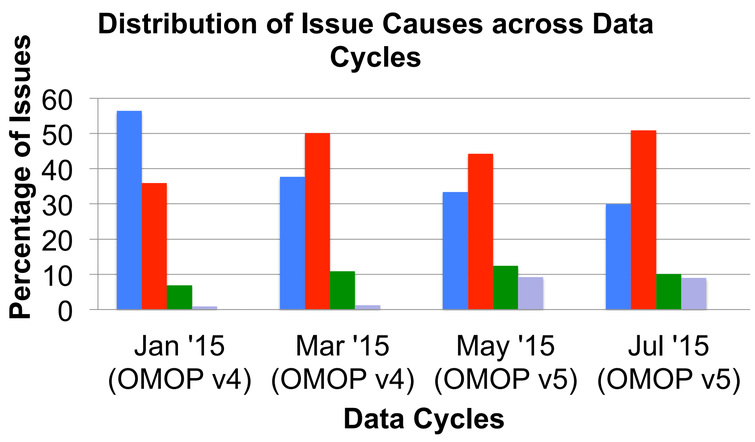
Types of Data Quality Checks implemented in the Program		
Fidelity	degree to which PEDSnet data accurately reflect data from the source systems	e.g. the distributions of gender values do not match between the source system (EHR) and the derived PEDSnet dataset
Consistency	degree to which a specific type of information is recorded in the same way in the different data sources contributing to PEDSnet	value set violations (e.g. incorrect concept identifier used to represent no information in race field) or mapping issues (e.g. a site using internal codes to capture drug information cannot fully map to standard RxNorm codes)
Accuracy	degree to which PEDSnet data correctly reflect the clinical characteristics of patients	e.g. a site having significantly larger ratio for average number of facts (observations, procedures, or drug exposures) per patient
Feasibility	degree to which a given type of information is actually collected and available in PEDSnet	e.g. the recorded gestational age is missing for 70% of patients

Results

After the 4th data cycle, the DQA program identified 591 data quality issues.

Domains with Most Issues	
Drug Exposure	21%
Measurement	13%
Observation	12%
Condition Occurrence	11%
Procedure Occurrence	10%
Person	8%
Visit Occurrence	7%

Most Frequent Issues	
Missing Data	31%
Entity Outliers	15%
Unexpected difference from previous ETL run	9%
Implausible Dates	9%
Temporal Outliers	6%
Unexpected Facts	4%
Unexpected concept identifiers	3%
Numerical Outliers	3%



- ETL: due to an ETL programming error at the site largely due to incompleteness or ambiguity in the conventions document
- Provenance: due to the nature of EHR data, e.g. data entry error, clinical & administrative workflows, true anomaly, etc.
- I2b2 transform: due to limitations of i2b2->PEDSnet transformation scripts
- Non-issue: false alert by the data quality assessment program

Conclusions

We implemented a comprehensive suite of data quality checks with a semi-automatic workflow to conduct DQA in PEDSnet, and made the following contributions:

1. A large-scale data quality assessments in pediatrics on:
   - ~4.7M children (nearly 5% of US child population)
   - Observational data on >97M clinical encounters
2. A real-world account on data quality issues in a CDRN:
   - Identified and tracked the causes of ~600 issues
   - Longitudinal analysis showing trends of causes and domains

Key findings:

- Even after the 4th data cycle, ~74 issues were identified per site.
- The percentage of ETL issues decreased across successive cycles.
- The total number of provenance issues increased across cycles.

Prospective data quality assurance is important but not sufficient for validating data quality in CDRNs. A strong post hoc DQA program is necessary to ensure data quality in CDRNs. This study serves as a data education platform where the experts learn and evolve the data quality program with each iteration.

Future directions:

- Implement advanced data quality checks such as comparisons with external datasets and published results.
- Use the data quality results to determine the analytic fitness of the PEDSnet dataset with respect to a given target research study.

Acknowledgements: This work was supported by PCORI Contract CDRN-1306-01556. The investigators would like to thank the PEDSnet informatics team for their efforts, contributing to the PEDSnet dataset.

Contact: Levon Utidjian, MD, MBI – levon@email.chop.edu