PEDSnet Data Quality
PEDSnet
A National Pediatric Learning Health System (Forrest et al. 2014)

• What is PEDSnet?
  • Clinical Data Research Network (CDRN)
    • Facilitate large-scale research
  • Aggregates pediatric EHR data from multiple institutions

• PCORI Support
  • One of the 13 CDRNs supported by Patient-Centered Outcomes Research Institute (PCORI)
  • Part of a larger network, PCORnet, which combines data from all CDRNs under a common data model
PEDSnet Data Flow

PEDSnet sites (at various locations)

- S1
- S2
- S3
- S4
- S5
- S6
- S7

Extract-transform-load (ETL) data quarterly

- Regular online meetings
- GitHub version control

PEDSnet Common Data Model

PCORnet Common Data Model

Data Quality and Characterization

Query Fulfillment

Data Science

Data Modeling

Study Participation

- ETL specifications
- Data Quality
### PEDSnet DQA - History

- DQA began as a necessity to communicate with sites about problems in their data
  - Initially began as missingness

<table>
<thead>
<tr>
<th>Person Table</th>
<th>Patients</th>
<th>total Pati</th>
<th>Gender</th>
<th>Race</th>
<th>Ethnicity</th>
<th>DOB</th>
<th>Gestational Age</th>
<th>Home ZIP</th>
<th>PCP</th>
<th>Principal Care Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site 1</td>
<td>690,408</td>
<td>14.33%</td>
<td>0.00</td>
<td>0.25</td>
<td>0.27</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Site 2</td>
<td>887,949</td>
<td>18.43%</td>
<td>0.00</td>
<td>0.16</td>
<td>0.23</td>
<td>0.00</td>
<td>0.70</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Site 3</td>
<td>591,689</td>
<td>12.28%</td>
<td>0.00</td>
<td>0.09</td>
<td>0.03</td>
<td>0.00</td>
<td>0.81</td>
<td>0.00</td>
<td>0.02</td>
<td>0.00</td>
</tr>
<tr>
<td>Site 4</td>
<td>642,894</td>
<td>13.34%</td>
<td>0.00</td>
<td>0.25</td>
<td>0.11</td>
<td>0.00</td>
<td>0.72</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Site 5</td>
<td>826,513</td>
<td>17.15%</td>
<td>0.00</td>
<td>0.18</td>
<td>0.06</td>
<td>0.00</td>
<td>0.64</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Site 6</td>
<td>580,762</td>
<td>12.05%</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.80</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Site 7</td>
<td>185,876</td>
<td>3.86%</td>
<td>0.00</td>
<td>0.38</td>
<td>0.47</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Site 8</td>
<td>411,847</td>
<td>8.55%</td>
<td>0.00</td>
<td>0.25</td>
<td>0.10</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.02</td>
<td>0.00</td>
</tr>
<tr>
<td><strong>Total Patients</strong></td>
<td><strong>4,817,938</strong></td>
<td><strong>100.00%</strong></td>
<td><strong>0.01%</strong></td>
<td><strong>29.74%</strong></td>
<td><strong>25.78%</strong></td>
<td><strong>0.00%</strong></td>
<td><strong>79.88%</strong></td>
<td><strong>0.10%</strong></td>
<td><strong>48.13%</strong></td>
<td><strong>0.00%</strong></td>
</tr>
</tbody>
</table>
PEDSnet Approach: Theoretical

- Considerations:
  - Conventions
  - Data Requests
    - e.g., focus on specialty data
  - General standards (e.g., Kahn et al)
  - Anticipating conformance with other research networks (e.g., PCORnet)
  - Anticipating needs for PEDSnet

- Cataloging DQA as a means to create framework for DQA checks
### PEDSnet DQA: Catalog

<table>
<thead>
<tr>
<th>check_type</th>
<th>function</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA-001</td>
<td>value set violation</td>
</tr>
<tr>
<td>AA-002</td>
<td>invalid concept identifier</td>
</tr>
<tr>
<td>AA-003</td>
<td>inconsistency between pk and source value</td>
</tr>
<tr>
<td>AA-004</td>
<td>unexpected fact</td>
</tr>
<tr>
<td>AA-005</td>
<td>illegal vocabulary</td>
</tr>
<tr>
<td>AA-006</td>
<td>inclusion criteria violation</td>
</tr>
<tr>
<td>AA-007</td>
<td>incorrect mapping</td>
</tr>
<tr>
<td>AA-008</td>
<td>illegal concept class</td>
</tr>
<tr>
<td>AA-009</td>
<td>date time inconsistency</td>
</tr>
<tr>
<td>AA-010</td>
<td>invalid format</td>
</tr>
<tr>
<td>BA-001</td>
<td>missing data</td>
</tr>
<tr>
<td>BA-002</td>
<td>no matching concepts</td>
</tr>
<tr>
<td>BA-003</td>
<td>missing expected concept</td>
</tr>
<tr>
<td>BA-004</td>
<td>insufficient facts for visits</td>
</tr>
<tr>
<td>BA-005</td>
<td>insufficient facts for visit types</td>
</tr>
<tr>
<td>CA-001</td>
<td>future event</td>
</tr>
<tr>
<td>CA-002</td>
<td>past event</td>
</tr>
<tr>
<td>CA-003</td>
<td>pre-birth fact</td>
</tr>
<tr>
<td>CA-004</td>
<td>post-death fact</td>
</tr>
<tr>
<td>CA-005</td>
<td>unexpected change in number of records</td>
</tr>
<tr>
<td>CA-006</td>
<td>unexpected change in missingness</td>
</tr>
<tr>
<td>CA-007</td>
<td>entity outliers</td>
</tr>
<tr>
<td>CA-008</td>
<td>temporal outliers</td>
</tr>
<tr>
<td>CA-009</td>
<td>unexpected change in temporal distribution</td>
</tr>
<tr>
<td>CA-010</td>
<td>low record count</td>
</tr>
<tr>
<td>CA-011</td>
<td>implausible numerical values</td>
</tr>
<tr>
<td>CA-012</td>
<td>unexpected distribution</td>
</tr>
<tr>
<td>CA-013</td>
<td>inconsistency in visit types</td>
</tr>
<tr>
<td>CA-014</td>
<td>inconsistent null distribution between</td>
</tr>
<tr>
<td>CA-015</td>
<td>unexpected change in number of fact types</td>
</tr>
<tr>
<td>CA-016</td>
<td>start date after end date</td>
</tr>
<tr>
<td>CB-001</td>
<td>unexpected fact to patient ratio</td>
</tr>
<tr>
<td>CB-002</td>
<td>unexpected most frequent values</td>
</tr>
</tbody>
</table>

- Every check applied to database falls into one of these categories
- CA-001 will check for all future dates as applied to different tables and columns
- 746 checks
<table>
<thead>
<tr>
<th>Framework Term</th>
<th>Check Type</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Conformance: Value (Verification)</strong></td>
<td></td>
</tr>
<tr>
<td>Conformance: Value</td>
<td>Value set violation</td>
</tr>
<tr>
<td></td>
<td>Invalid concept identifier</td>
</tr>
<tr>
<td></td>
<td>Inconsistency between pk and source value</td>
</tr>
<tr>
<td></td>
<td>Unexpected fact</td>
</tr>
<tr>
<td></td>
<td>Invalid format</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Conformance: Relational</td>
<td>Date time inconsistency</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Completeness (Verification)</td>
<td>Missing data</td>
</tr>
<tr>
<td></td>
<td>Missing expected concept</td>
</tr>
<tr>
<td></td>
<td>Insufficient fact for visit types</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Plausibility: Temporal (Verification)</td>
<td></td>
</tr>
<tr>
<td>Plausibility: Temporal</td>
<td>Future event</td>
</tr>
<tr>
<td></td>
<td>Pre-birth fact</td>
</tr>
<tr>
<td></td>
<td>Temporal outliers</td>
</tr>
<tr>
<td></td>
<td>Start date after end date</td>
</tr>
<tr>
<td></td>
<td>Fact date before visit_start_date</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Plausibility: Atemporal (Verification)</td>
<td>Unexpected change in number of records between data cycles</td>
</tr>
<tr>
<td>Plausibility: Atemporal</td>
<td>Unexpected change in missingness of a field between data cycles</td>
</tr>
<tr>
<td></td>
<td>Low record count</td>
</tr>
<tr>
<td></td>
<td>Unexpected distribution</td>
</tr>
<tr>
<td></td>
<td>Inconsistent null distribution between source values and concepts</td>
</tr>
<tr>
<td></td>
<td>Unexpected change in number of fact types between data cycles</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Plausibility: Value (Ver)</td>
<td>Labs outside normal range</td>
</tr>
<tr>
<td>Plausibility: Atemporal (Validation)</td>
<td>Drug is not SCD or more granular</td>
</tr>
<tr>
<td>Plausibility: Value (Ver)</td>
<td>Unexpected fact to patient ratio</td>
</tr>
<tr>
<td>Plausibility: Atemporal (Validation)</td>
<td>Unexpected most frequent values</td>
</tr>
</tbody>
</table>
# PEDSnet DQA: Output in CSV

<table>
<thead>
<tr>
<th>Field</th>
<th>Check Code</th>
<th>Check Type</th>
<th>Finding</th>
</tr>
</thead>
<tbody>
<tr>
<td>admitting_source_value</td>
<td>BA-001</td>
<td>Missing data</td>
<td>94.84%</td>
</tr>
<tr>
<td>care_site_id</td>
<td>CA-007</td>
<td>Identification of entity outliers</td>
<td>Please confirm sites with over 1 million visits (1244)</td>
</tr>
<tr>
<td>death_date,visit_start_date</td>
<td>CA-004</td>
<td>Post-death fact</td>
<td>3021 visits after death</td>
</tr>
<tr>
<td>discharge_to_source_value</td>
<td>BA-001</td>
<td>Missing data</td>
<td>94.79%</td>
</tr>
<tr>
<td>person_id</td>
<td>CA-007</td>
<td>Identification of entity outliers</td>
<td>12345 has over 2K visits?</td>
</tr>
<tr>
<td>provider_id</td>
<td>CA-007</td>
<td>Identification of entity outliers</td>
<td>5555 responsible for most visits</td>
</tr>
<tr>
<td>provider_id</td>
<td>BA-001</td>
<td>Missing Data</td>
<td>1.73%</td>
</tr>
<tr>
<td>visit_concept_id,observation_concept_id</td>
<td>BA-005</td>
<td>Insufficient facts for visit types</td>
<td>30.56% (inpatient visits with no DRG data)</td>
</tr>
<tr>
<td>visit_concept_id,visit_payer_id</td>
<td>BA-005</td>
<td>Insufficient facts for visit types</td>
<td>22.11% (inpatient visits with no insurance data)</td>
</tr>
<tr>
<td>visit_end_date</td>
<td>CA-009</td>
<td>Identification of sudden change in distribution of facts</td>
<td>what do the initial spikes represent?</td>
</tr>
<tr>
<td>visit_occurrence_id</td>
<td>CB-001</td>
<td>Unexpected fact to patient ratio</td>
<td>visit to patient ratio: 37</td>
</tr>
<tr>
<td>visit_source_concept_id</td>
<td>BA-002</td>
<td>No matching concepts</td>
<td>76.31%</td>
</tr>
<tr>
<td>visit_source_value</td>
<td>BA-001</td>
<td>Missing data</td>
<td>0%</td>
</tr>
<tr>
<td>visit_start_date</td>
<td>CA-009</td>
<td>Identification of sudden change in distribution of facts</td>
<td>what do the initial isolated values represent?</td>
</tr>
<tr>
<td>visit_start_date,birth_datetime</td>
<td>CA-003</td>
<td>Pre-birth fact</td>
<td>11872</td>
</tr>
<tr>
<td>visit_start_date,death_date</td>
<td>CA-004</td>
<td>Post-death fact</td>
<td>3733</td>
</tr>
</tbody>
</table>
# PEDSnet Approach: Pragmatic

<table>
<thead>
<tr>
<th>Input</th>
<th>site-level data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output</td>
<td><strong>Level 1</strong>: univariate distributions and figures and graphs</td>
</tr>
<tr>
<td></td>
<td><strong>Level 2</strong>: issues generated from applying catalog of checks</td>
</tr>
<tr>
<td>Program</td>
<td>R, GO, Python</td>
</tr>
<tr>
<td>Format</td>
<td>All on GitHub for list of issues</td>
</tr>
<tr>
<td></td>
<td>Folding in site-by-site comparisons</td>
</tr>
<tr>
<td>Features</td>
<td>Ranking (manual and automated)</td>
</tr>
<tr>
<td></td>
<td>Dynamic feedback and discussion of each issue</td>
</tr>
<tr>
<td></td>
<td>Causes identified and stored over time</td>
</tr>
</tbody>
</table>
Data Quality Assessment

Communication with Sites

- Prepare data quality reports

  - Visual Report
  - GitHub issue report

- Identify causes of issues and propose solutions

  - Ranking

Fact Tables

- drug_exposure
  - High
  - Medium
  - Low
PEDSnet Data Quality Output

DQA: February 2018 (ETLv19): drug_exposure #312

Closed writetoritu opened this issue on Feb 22 · 2 comments

Commented on Feb 22

Description: Unexpected Change In Number Of Records Between Data Cycles
Finding: 58.55%

Added Data Cycle: February 2018 Data Quality Status: new Table: drug_exposure labels on Feb 22

Referenced this issue on Feb 22

DQA Summary: February 2018 (ETLv19) for PEDSnet CDM v2.8.0 #318

Closed writetoritu referenced this issue on Feb 22

Commented on Feb 23

ETL error - could not get script to complete and broke a single union query into two separate queries without accounting for matches which resulted in duplicated records

Added Cause: ETL: programming error Status: solution proposed and removed Status: new labels on Feb 25

Closed this on Feb 25

Commented on Mar 30

Adjusted script and numbers are back in line with expectations
DATA Quality: Evolution of checks and issues

![Graph showing the evolution of data quality checks and issues over data cycles. The graph indicates an increase in data quality issues as the data cycle progresses. There are also markers indicating significant data model changes at specific cycles.](image-url)
Domain-wise Distribution

# Reported Issues

- drug_exposure
- measurement
- condition_occurrence
- observation
- visit_occurrence
- immunization
- procedure_occurrence
- person
- adt_occurrence
- care_site
- fact_relationship
- measurement_orga
- provider
- visit_payer
- location
- death
Check Type Distribution

- temporal outliers
- missing data
- unexpected most frequent values
- unexpected changes in # records
- unexpected changes in # fact types
- invalid concepts
- Others

# Reported Issues

- v2.9 Release
- v3.0 Submission
Site-by-Site Comparison

Current version laboratory data

- Missing/uninformative `measurement_concept_id`; target = <20%
- Missing/uninformative result (value_as_number or value_as_concept_id); target = <20%

<table>
<thead>
<tr>
<th>site</th>
<th>missing_concept_ids</th>
<th>missing_results</th>
</tr>
</thead>
<tbody>
<tr>
<td>site_c</td>
<td>0.00</td>
<td>0.19</td>
</tr>
<tr>
<td>site_d</td>
<td>0.37</td>
<td>0.04</td>
</tr>
<tr>
<td>site_n</td>
<td>0.10</td>
<td>0.13</td>
</tr>
<tr>
<td>site_u</td>
<td>0.00</td>
<td>0.10</td>
</tr>
<tr>
<td>site_v</td>
<td>0.42</td>
<td>0.14</td>
</tr>
<tr>
<td>site_w</td>
<td>0.01</td>
<td>0.31</td>
</tr>
</tbody>
</table>
PEDSnet DQA: Where are we headed?

- More streamlined code
- Interoperability of different DBMS
- Data characterization more extensive – not just reporting on issues
- Queryable data quality
- **Semantic data quality**
## Views On DATA quality

**INFORMATICIST’S VIEW**

- Conforms to data model
- Reflects the source system
- Accuracy:
  - Internal validity
  - Was the data handled correctly?

**RESEARCHER’S VIEW**

- Reflects key clinical facts
- Reflects biology/pathology
- Accuracy:
  - External validity
  - Does the data match what is in the real world?
Data Quality: Ability to Study Rare Conditions

Children with Glomerular Disease
n = 5166

- ... with creatinine and height to calculate egfr
  n =

- ... with at least one visit with nephrologist
  n =

- ... with steroid history available
  n =

- ... with a kidney biopsy
  n =

- ... with urine protein measurement
  n =

- ... with an inpatient and outpatient visit
  n =

lab and anthropometrics
drug history
lab data
specialty data available
procedure data available
continuity of care
Study-specific DQA Index: Quality Measure Study

Data Quality Score in Pediatric Quality Metrics Benchmarking Study

- Index assessed encounters (all), diagnoses (all), drugs (AP, AOM), procedures (SCD, AP)
Study-specific Composite DQA Metric – Abx/Wt

The Data Quality index (DQ index) for a dataset provides a summary of those data quality characteristics closely related to analyzing the impact of antibiotic use on growth trajectories. For a dataset $i$, it is computed as follows:

$$DAI_i = \frac{SDM(abx_i)}{10} + \frac{SDM(steroid_i)}{20} + \frac{SDM(gerd_i)}{40} + 2 \times ht_{-err} + 2 \times wt_{-err} + ht_{-dup} + wt_{-dup}$$

where

$$SDM(x) = (10 \times (x - \mu))^2$$

$abx_i$ = fraction of patients receiving $\geq$1 antibiotic prescription

$steroid_i$ = fraction of patients receiving $\geq$1 systemic steroid prescription

$gerd_i$ = fraction of patients receiving $\geq$1 anti-reflux medication prescription

$ht_{-err}i$ = fraction of height measurements assessed as errors

$wt_{-err}i$ = fraction of weight measurements assessed as errors

$ht_{-dup}i$ = fraction of height measurements assessed as duplicates

$wt_{-dup}i$ = fraction of weight measurements assessed as duplicates
Study-specific Composite DQA Metric – Abx/Wt
Study-specific DQA correlations

- **Steroid Metric** vs. **Antibiotic Metric**: $R = 0.39$
- **Cohort Size** vs. **Antibiotic Metric**: $R = -0.34$
Semantic DQA: Drug Utilization

Proportion of Selected Drugs By Site

- Albuterol among children diagnosed with asthma
- Antiseizure medications among children with seizure disorder
- Antistimulants among children diagnosed with ADHD
- Insulin among children with Type 1 Diabetes
- IV during inpatient stay
- Steroids among children with glomerular disease

site

value
Semantic DQA: Drug Utilization

Clustering of Sites on Drug Utilization as Marker of Semantic Data Quality
Acknowledgments

• Data Coordinating Center
  • Charles Bailey, MD, PhD
  • Connor Callahan, MS
  • Kimberley Dickinson
  • Harris Weinstein
  • Susan Hague
  • Shweta Chavan
• PEDSnet Program Management Office

• Other PEDSnet teams
  • ETL analysts
  • Site Informatics Leads
  • Leadership and Governance
• OHDSI Consortium
• Patients and Families