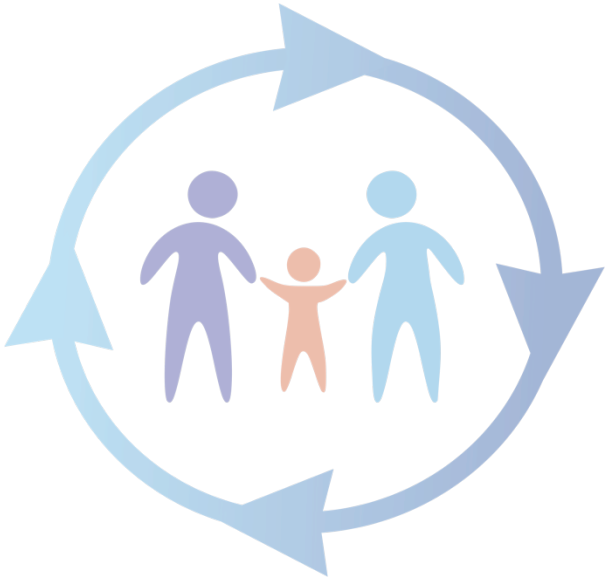
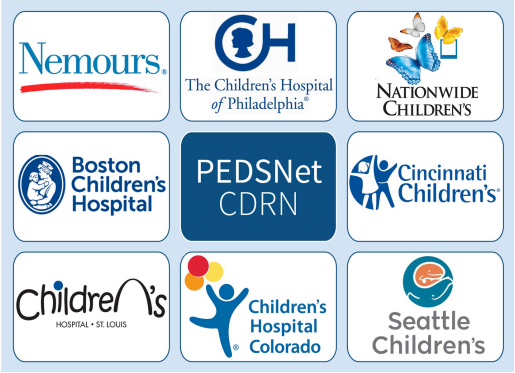
PEDSnet Data Quality



PEDSnet

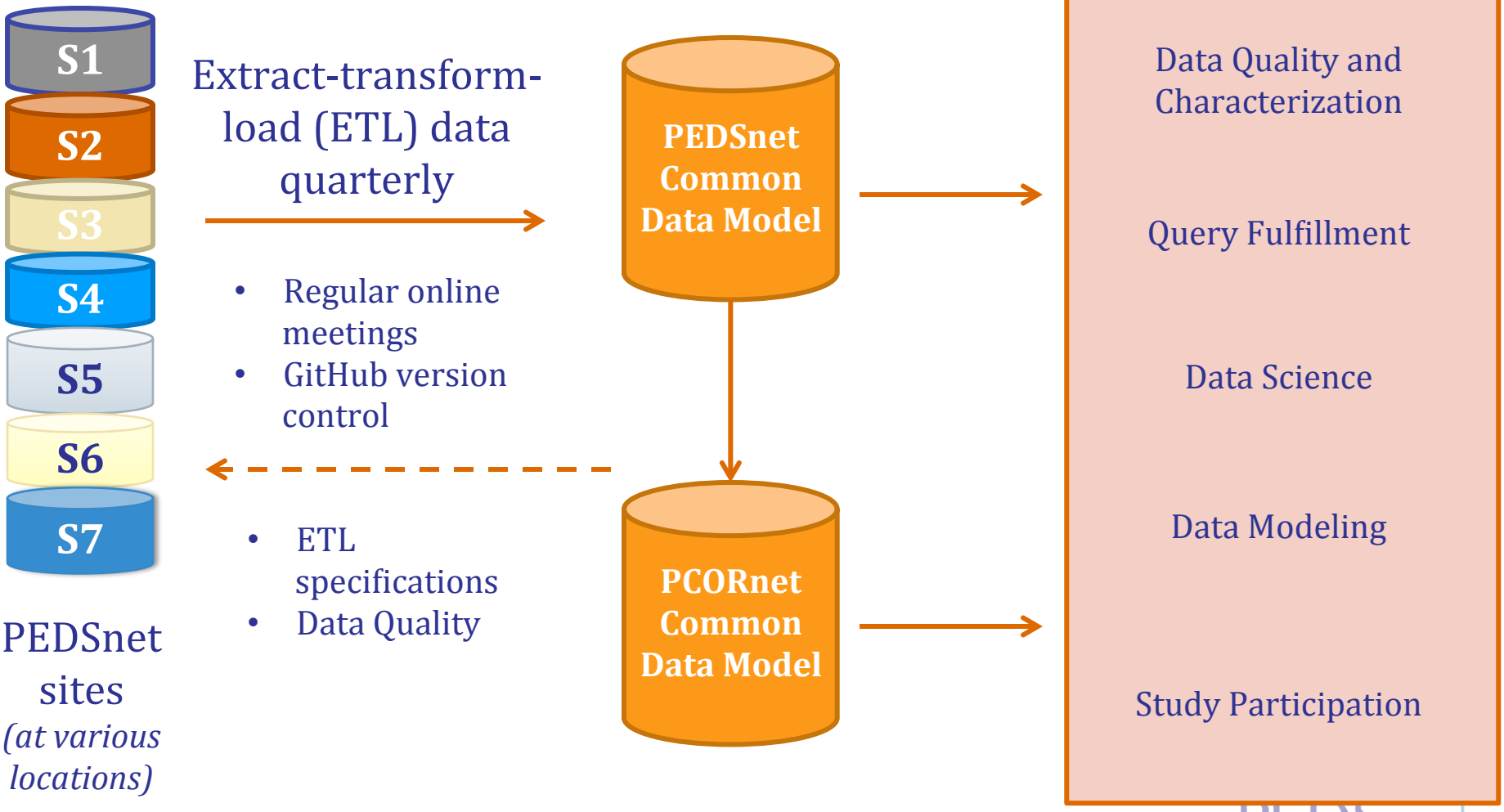
A National Pediatric Learning Health System (Forrest et al. 2014)

- 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
A Pediatric Learning Health System

PEDSnet Data Flow



PEDSnet
A Pediatric Learning Health System

PEDSnet DQA - History

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

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

PEDSnet
A Pediatric Learning Health System

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

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

- 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

PEDSnet
A Pediatric Learning Health System

Framework Term	Check Type	
Conformance: Value (Verification)	Value set violation	Illegal Vocabulary
	Invalid concept identifier	Inclusion Criteria Violation
	Inconsistency between pk and source value	Incorrect mapping
	Unexpected fact	Illegal concept class
	Invalid format	
Conformance: Relational	Date time inconsistency	
Completeness (Verification)	Missing data	No matching concepts
	Missing expected concept	Insufficient facts for visit
	Insufficient fact for visit types	
Plausibility: Temporal (Verification)	Future event	Past event
	Pre-birth fact	Post-death fact
	Temporal outliers	Unexpected change in temporal distributions
	Start date after end date	Max dates are more than 2 weeks apart
	Fact date before visit_start_date	
Plausibility: Atemporal (Verification)	Unexpected change in number of records between data cycles	Unexpected change in missingness of a field between data cycles
	Low record count	Implausible numerical values
	Unexpected distribution	Inconsistency in visit types
	Inconsistent null distribution between source values and concepts	Unexpected change in number of fact types between data cycles
Plausibility: Value (Ver)	Labs outside normal range	Drug is not SCD or more granular
Plausibility: Atemporal (Validation)	Unexpected fact to patient ratio	Unexpected most frequent values

PEDSnet DQA: Output in CSV

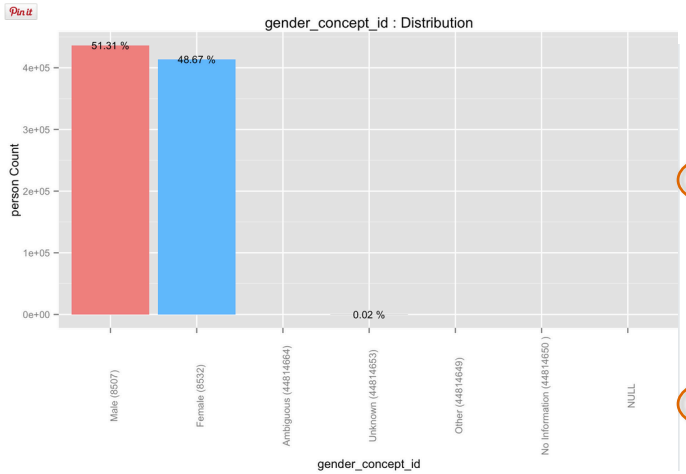
Field	Check Code	Check Type	Finding
admitting_source_value	BA-001	missing data	94.84%
care_site_id	CA-007	Identification of entity outliers	Please confirm sites with over 1 million visits (1244)
death_date,visit_start_date	CA-004	Post-death fact	3021 visits after death
discharge_to_source_value	BA-001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

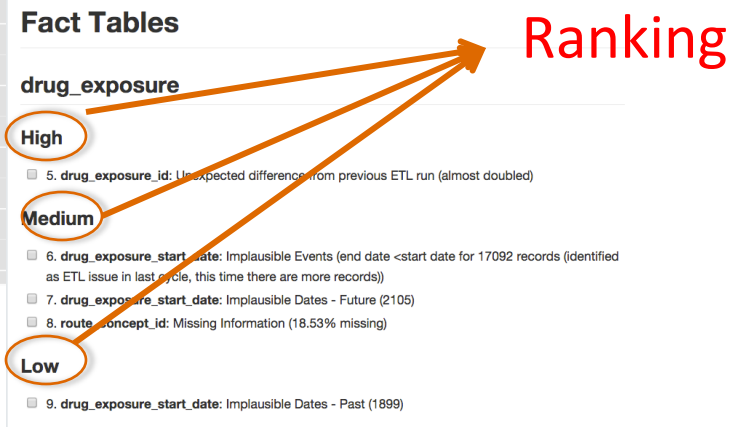
PEDSnet Approach: Pragmatic

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

Data Quality Assessment
Communication with Sites

- Prepare data quality reports





Ranking

Visual Report

GitHub issue report

- Identify causes of issues and propose solutions

PEDSnet
A Pediatric Learning Health System

PEDSnet Data Quality Output

DQA: February 2018 (ETLv19): drug_exposure #312

Closed writetoritu opened this issue on Feb 22 · 2 comments

[redacted] commented on Feb 22

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

[redacted] added **Data Cycle: February 2018** **Data Quality** **Status: new** **Table: drug_exposure** labels on Feb 22

writetoritu referenced this issue on Feb 22

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

15 of 30 tasks complete

[redacted] 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

[redacted] added **Cause: ETL: programming error** **Status: solution proposed** and removed **Status: new** labels on Feb 25

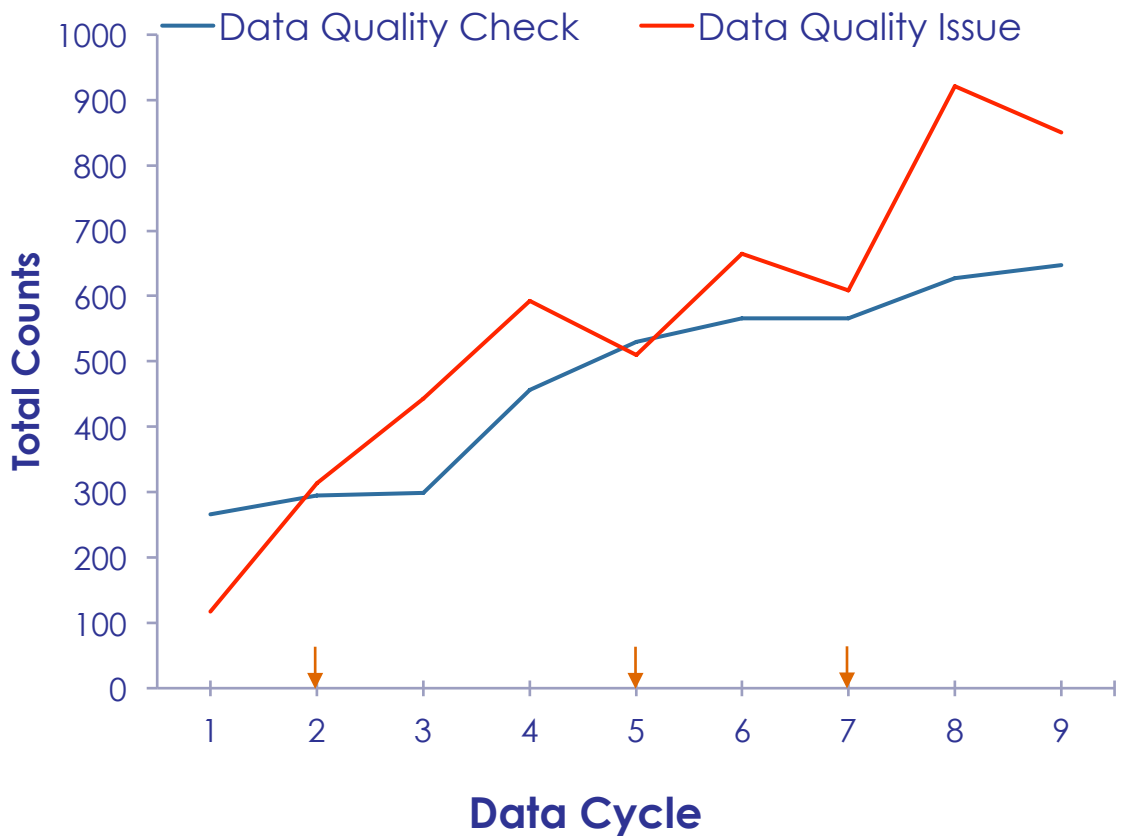
[redacted] closed this on Feb 25

[redacted] commented on Mar 30

adjusted script and numbers are back in line with expectations

PEDSnet
A Pediatric Learning Health System

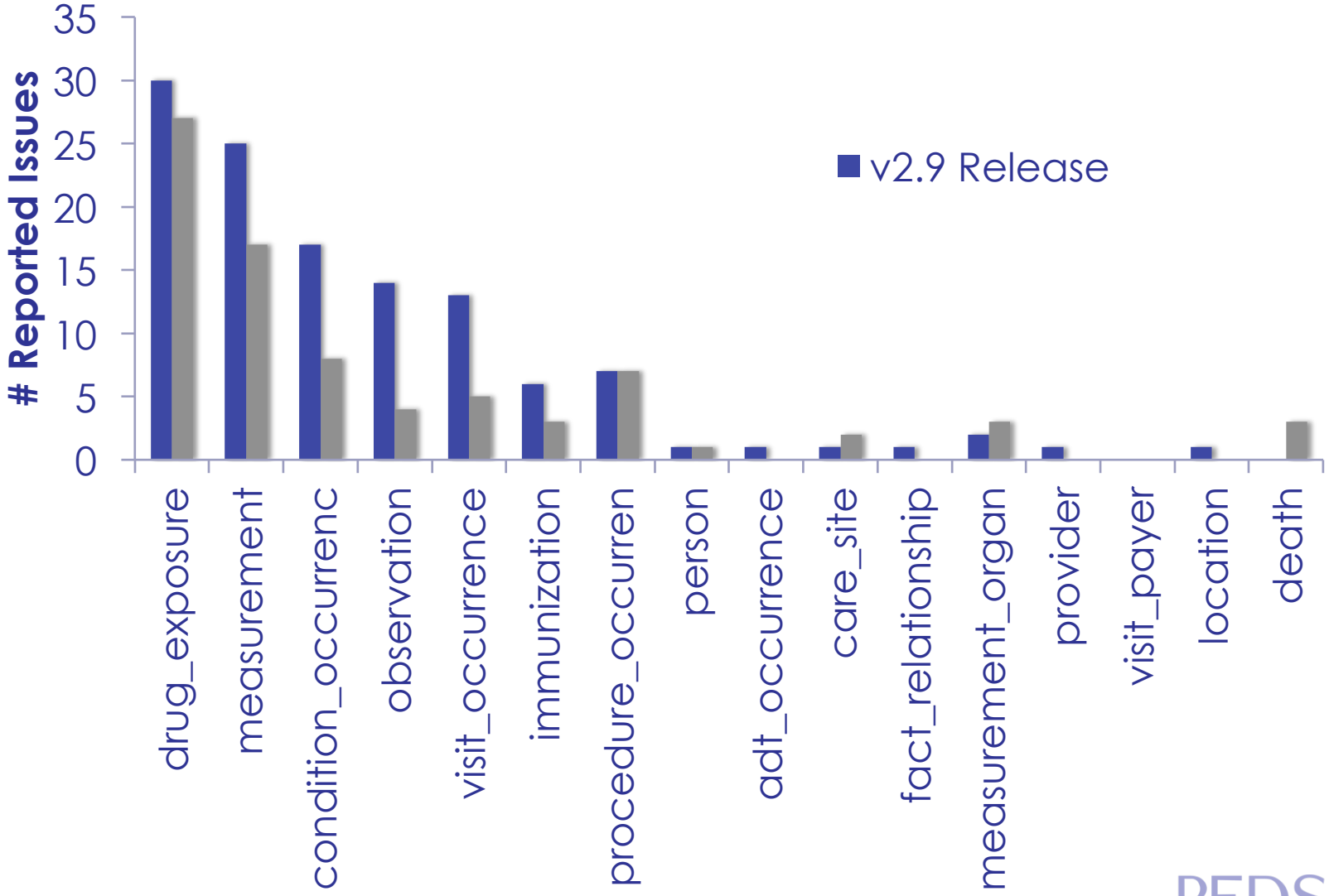
DATA Quality: Evolution of checks and issues



↓ Significant data model change

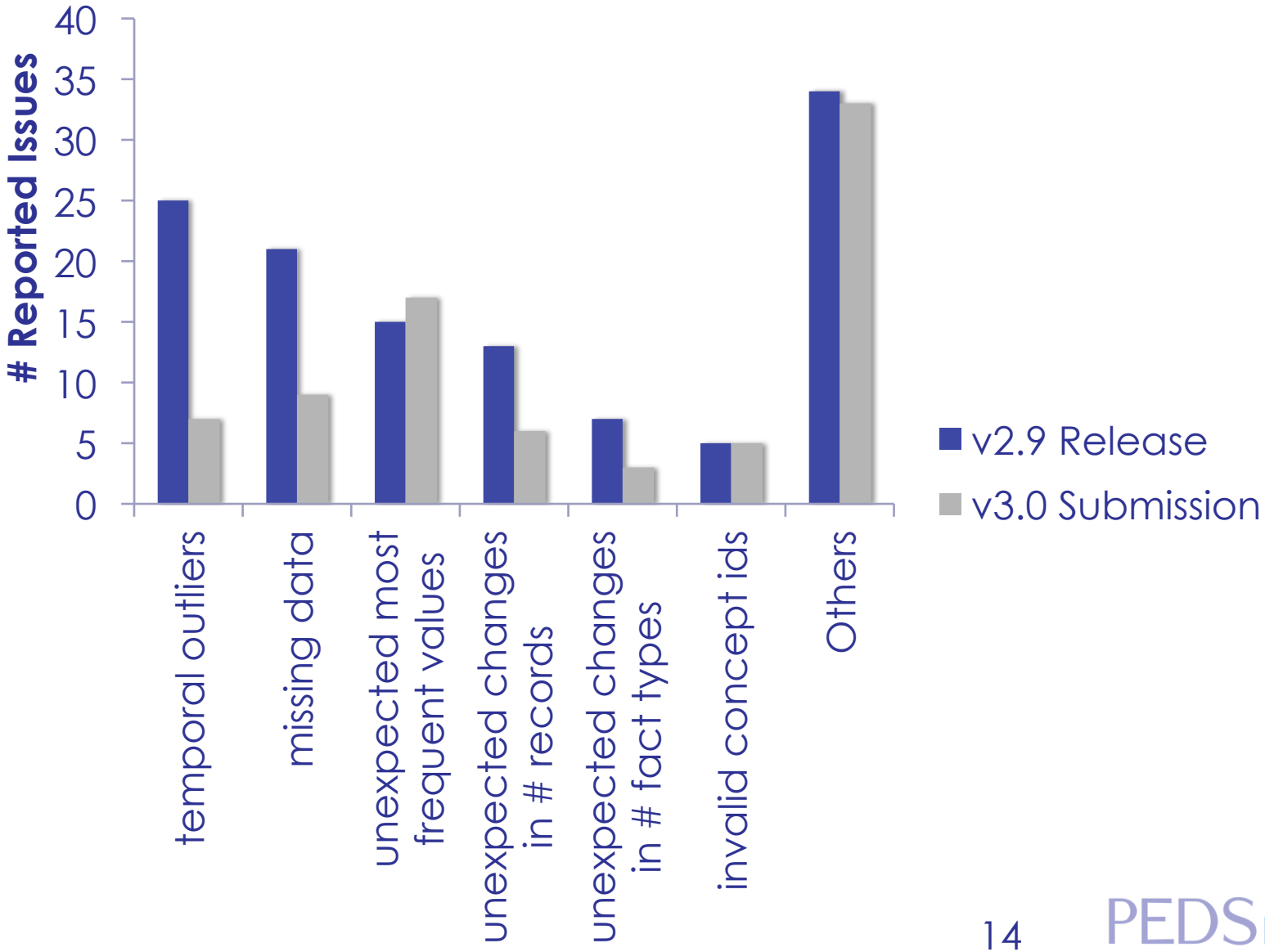
PEDSnet
A Pediatric Learning Health System

Domain-wise Distribution



PEDSnet
A Pediatric Learning Health System

Check Type Distribution



PEDSnet
A Pediatric Learning Health System

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%

site	missing_concept_ids	missing_results
site_c	0.00	0.19
site_d	0.37	0.04
site_n	0.10	0.13
site_u	0.00	0.10
site_v	0.42	0.14
site_w	0.01	0.31

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?

PEDSnet
A Pediatric Learning Health System

Data Quality: Ability to Study Rare Conditions

Children with Glomerular Disease
n = 5166

lab and anthropometrics
… with creatinine and height to calculate egfr
n =

specialty data available
… with at least one visit with nephrologist
n =

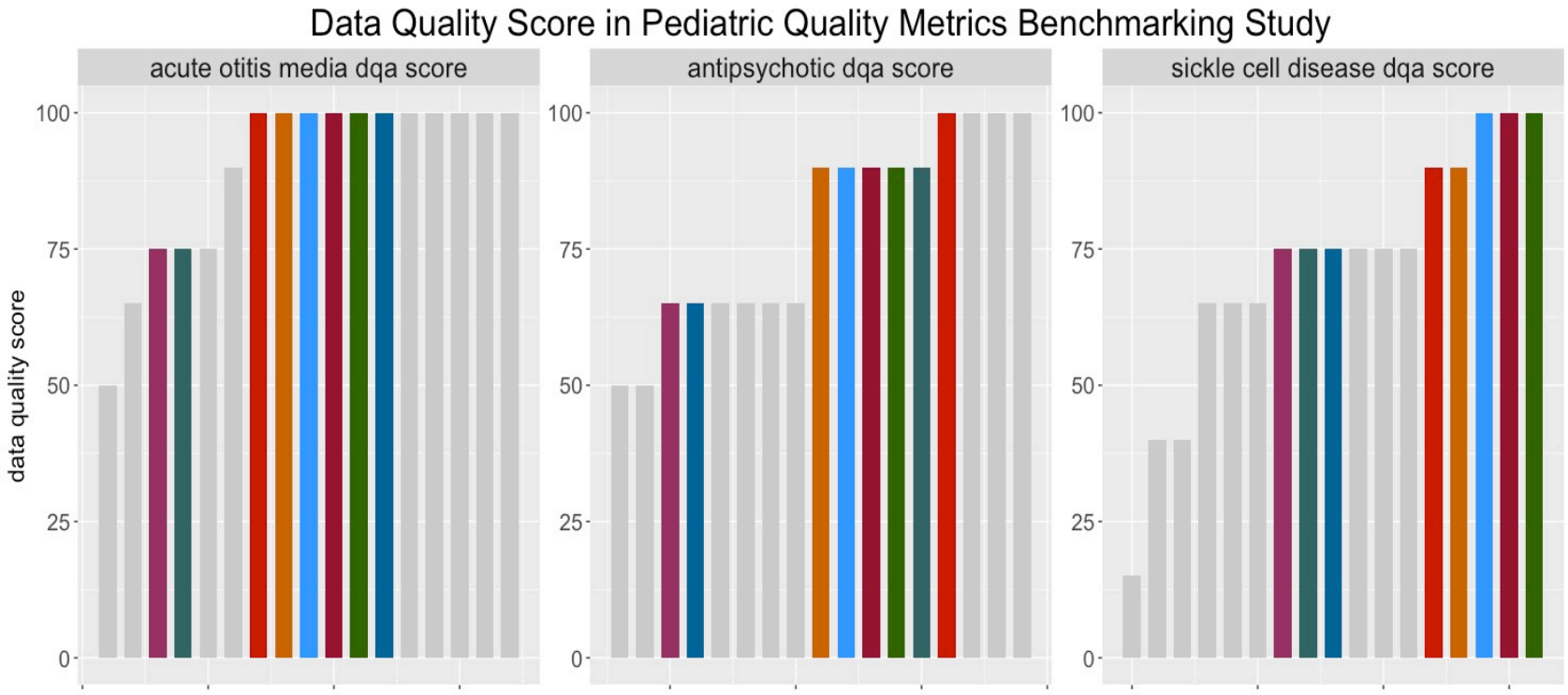
drug history
… with steroid history available
n =

procedure data available
… with a kidney biopsy
n =

lab data
… with urine protein measurement
n =

continuity of care
… with an inpatient and outpatient visit
n =

Study-specific DQA Index: Quality Measure Study



- Index assessed encounters (all), diagnoses (all), drugs (AP, AOM), procedures (SCD, AP)

PEDSnet
A Pediatric Learning Health System

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:

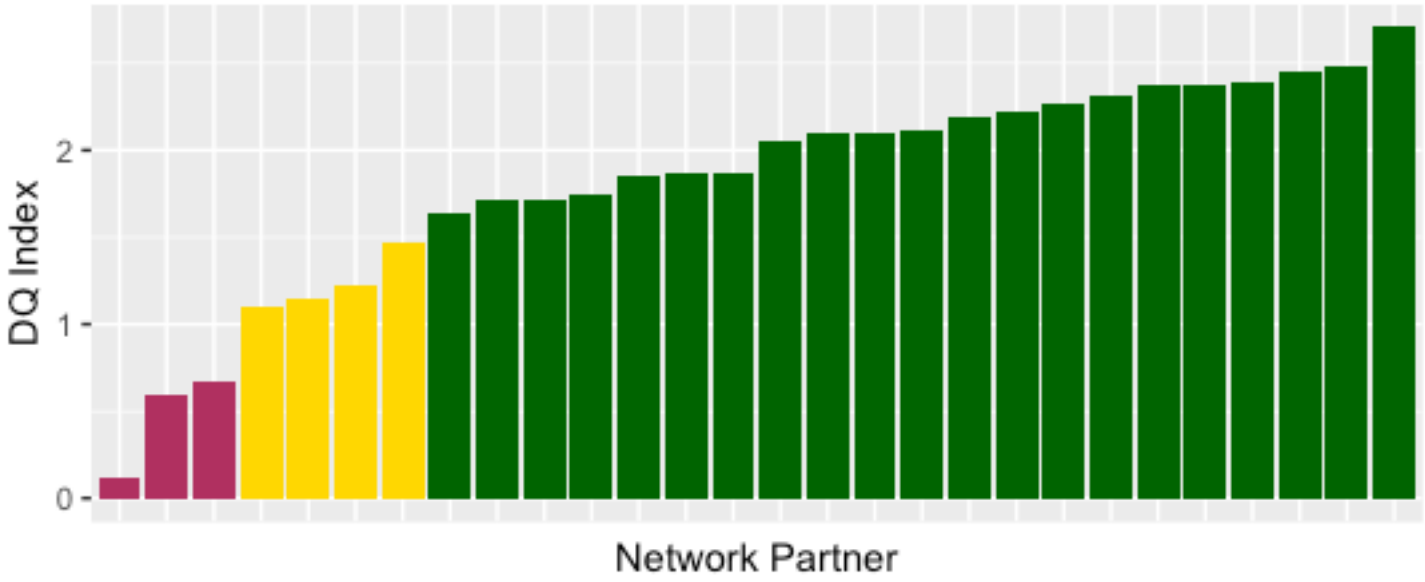
$$DAI_i = \frac{SDM(abx_i)}{10} + \frac{SDM(steroid_i)}{20} + \frac{SDM(gerd_i)}{40} + 2 * ht\_err_i + 2 * wt\_err_i + ht\_dup_i + wt\_dup_i$$

where

$$\begin{aligned}
SDM(x) &= (10 * (x - \overline{x}))^2 \\
abx_i &= \text{fraction of } \textbf{patients} \text{ receiving } \geq 1 \text{ antibiotic prescription} \\
steroid_i &= \text{fraction of } \textbf{patients} \text{ receiving } \geq 1 \text{ systemic steroid prescription} \\
gerd_i &= \text{fraction of } \textbf{patients} \text{ receiving } \geq 1 \text{ anti-reflux medication prescription} \\
ht_err_i &= \text{fraction of height } \textbf{measurements} \text{ assessed as errors} \\
wt_err_i &= \text{fraction of weight } \textbf{measurements} \text{ assessed as errors} \\
ht_dup_i &= \text{fraction of height } \textbf{measurements} \text{ assessed as duplicates} \\
wt_dup_i &= \text{fraction of weight } \textbf{measurements} \text{ assessed as duplicates}
\end{aligned}$$

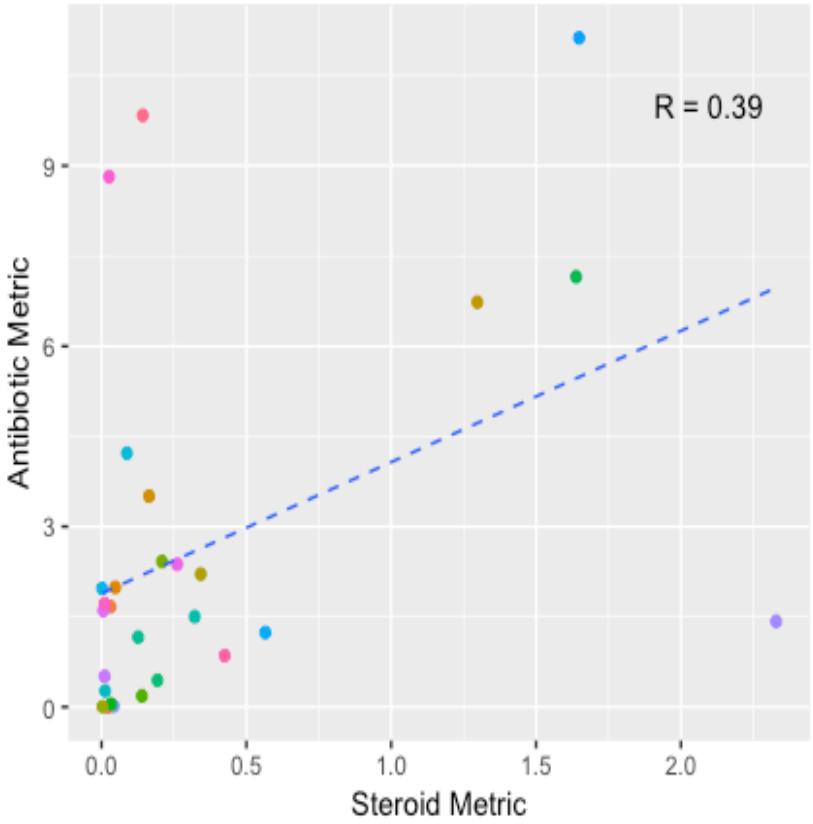
PEDSnet
A Pediatric Learning Health System

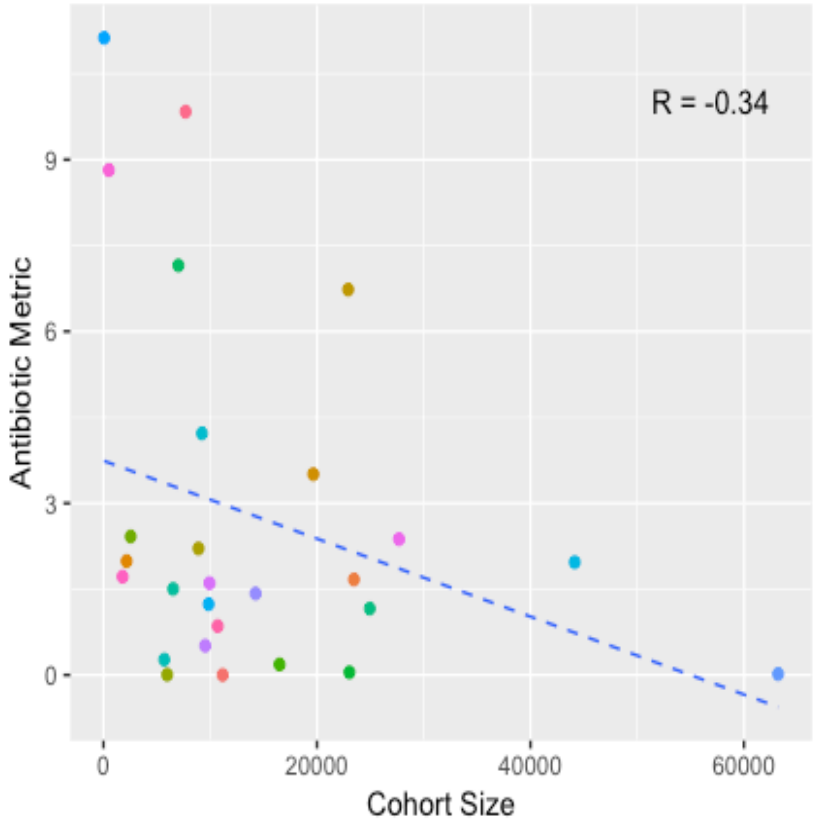
Study-specific Composite DQA Metric – Abx/Wt



PEDSnet
A Pediatric Learning Health System

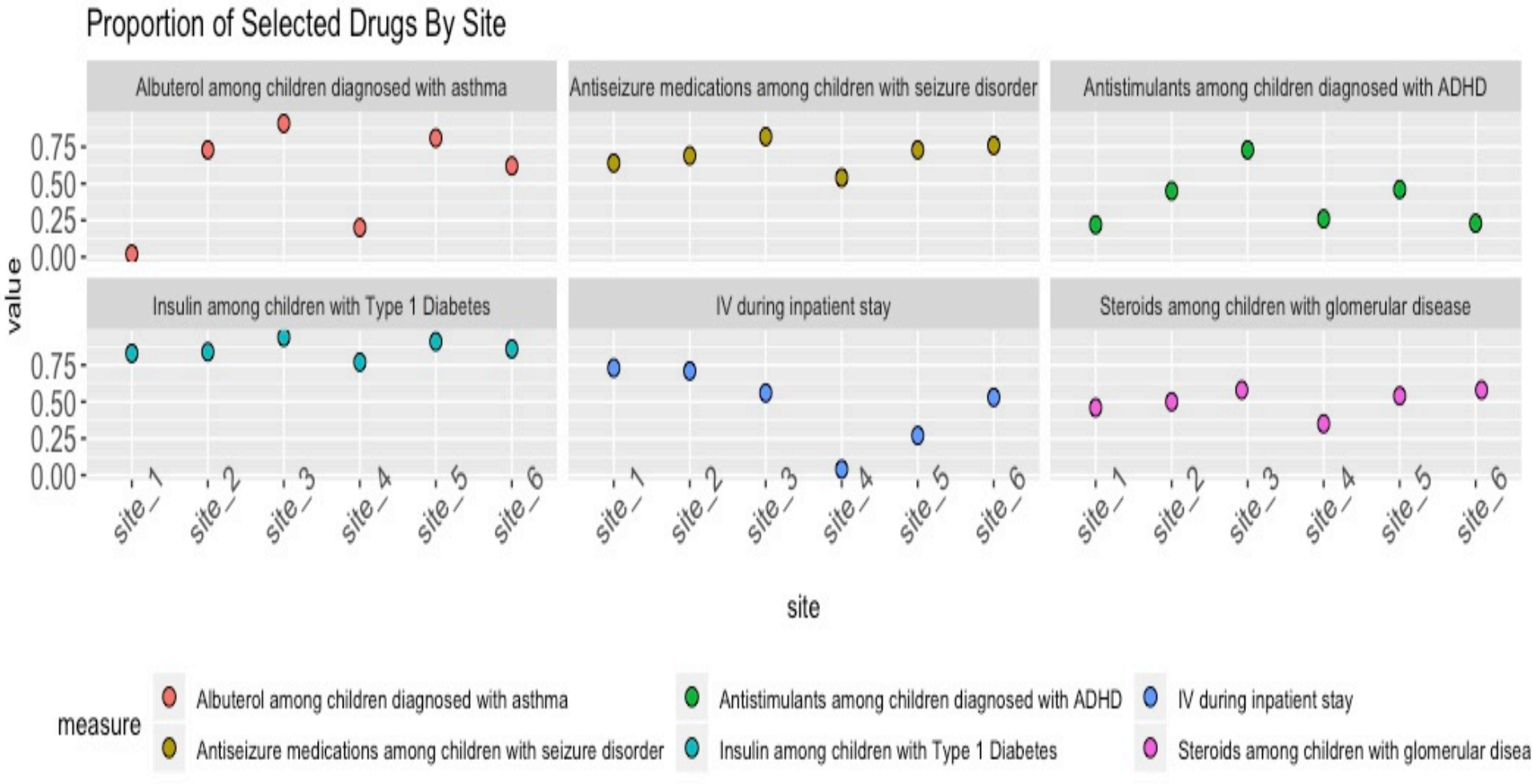
Study-specific DQA correlations





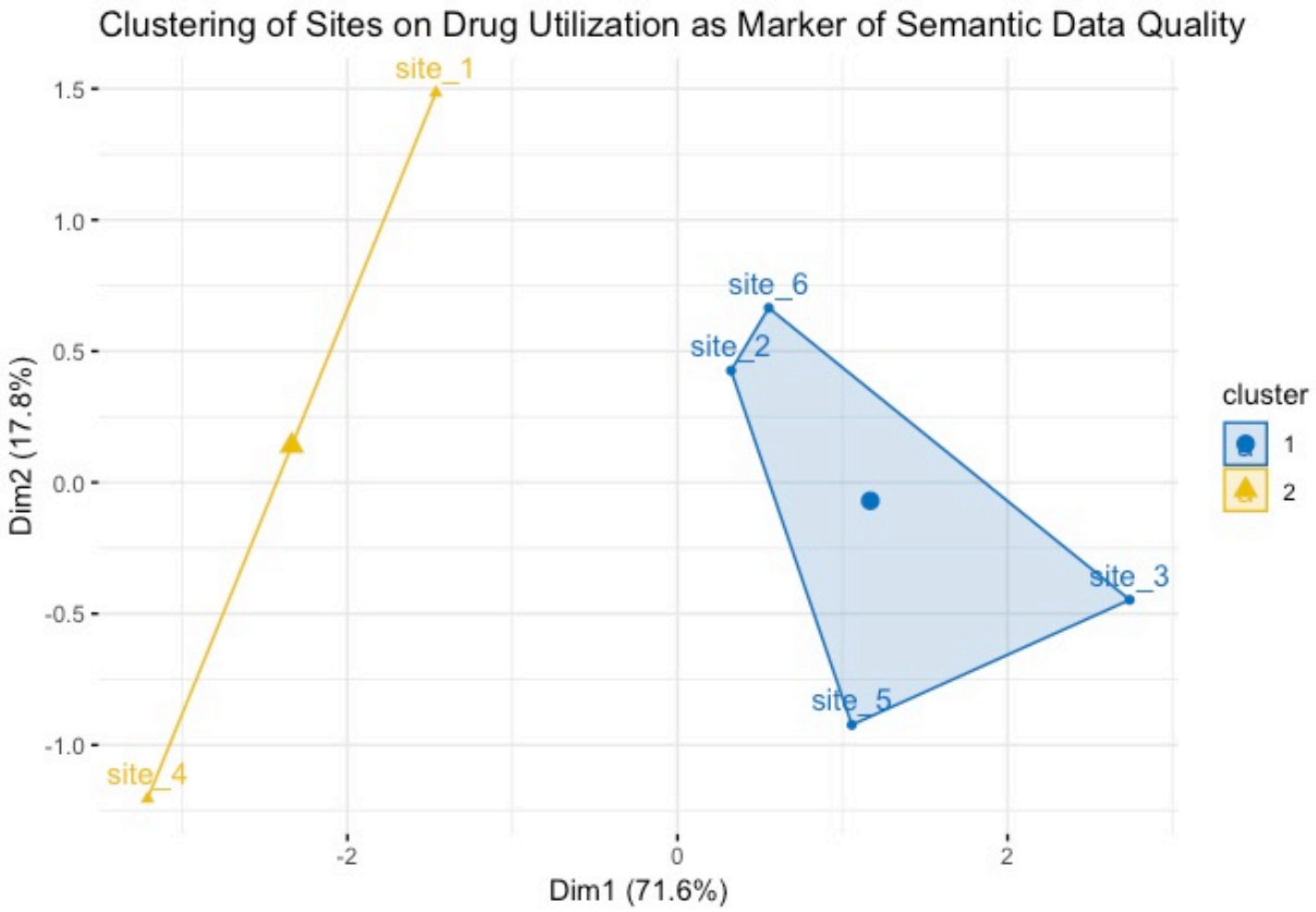
PEDSnet
A Pediatric Learning Health System

Semantic DQA: Drug Utilization



PEDSnet
A Pediatric Learning Health System

Semantic DQA: Drug Utilization



PEDSnet
A Pediatric Learning Health System

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

PEDSnet
A Pediatric Learning Health System