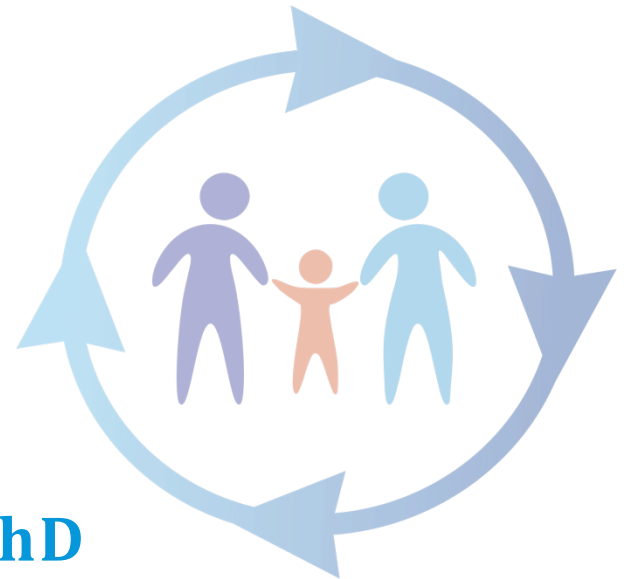


The Data Quality Program in PEDSnet



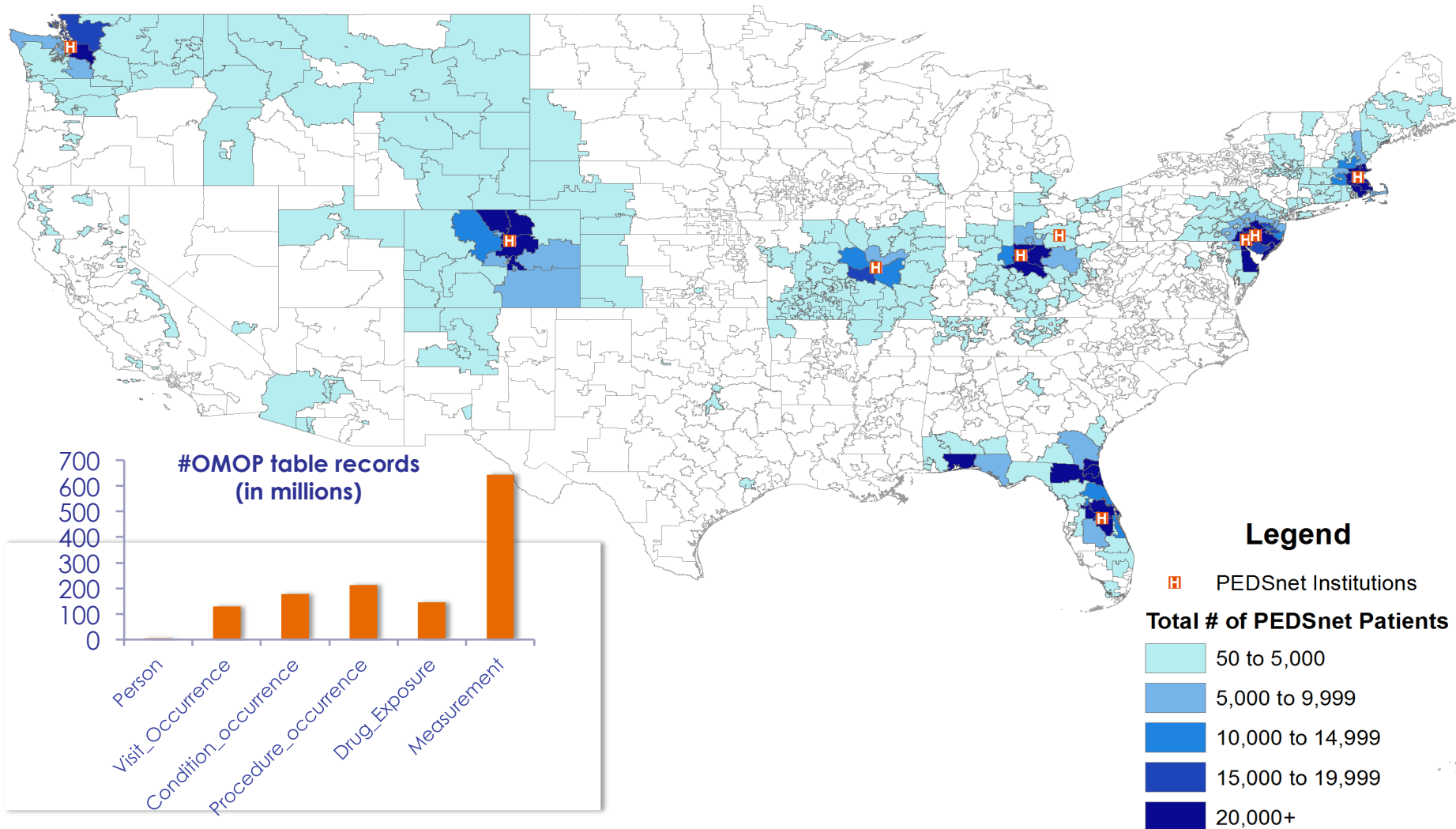
Ritu Khare, PhD

Department of Biomedical and Health Informatics
The Children's Hospital of Philadelphia

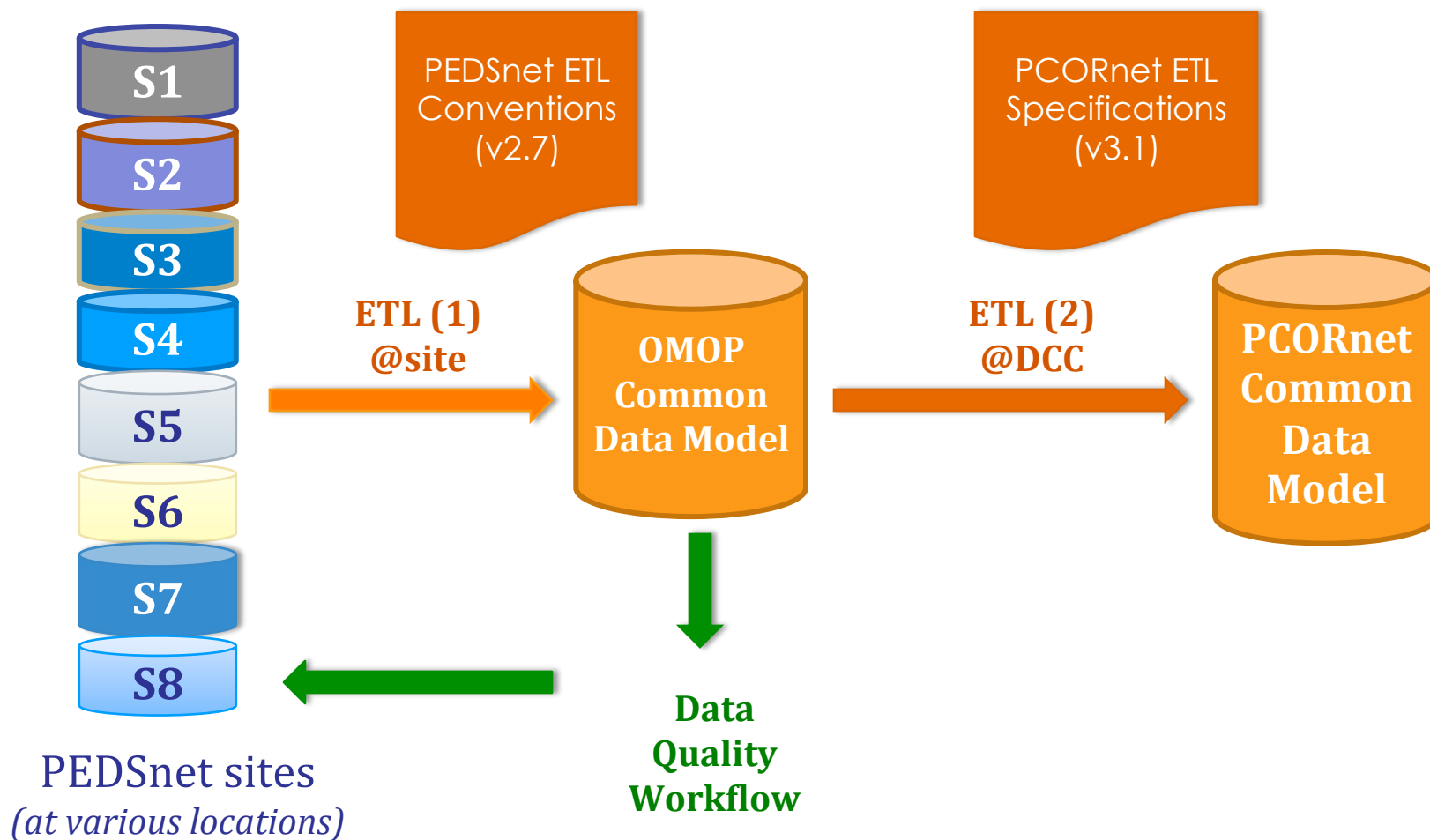
Contents

- PEDSnet Overview
- Data Quality Results (JAMIA)
- Key Data Quality Checks

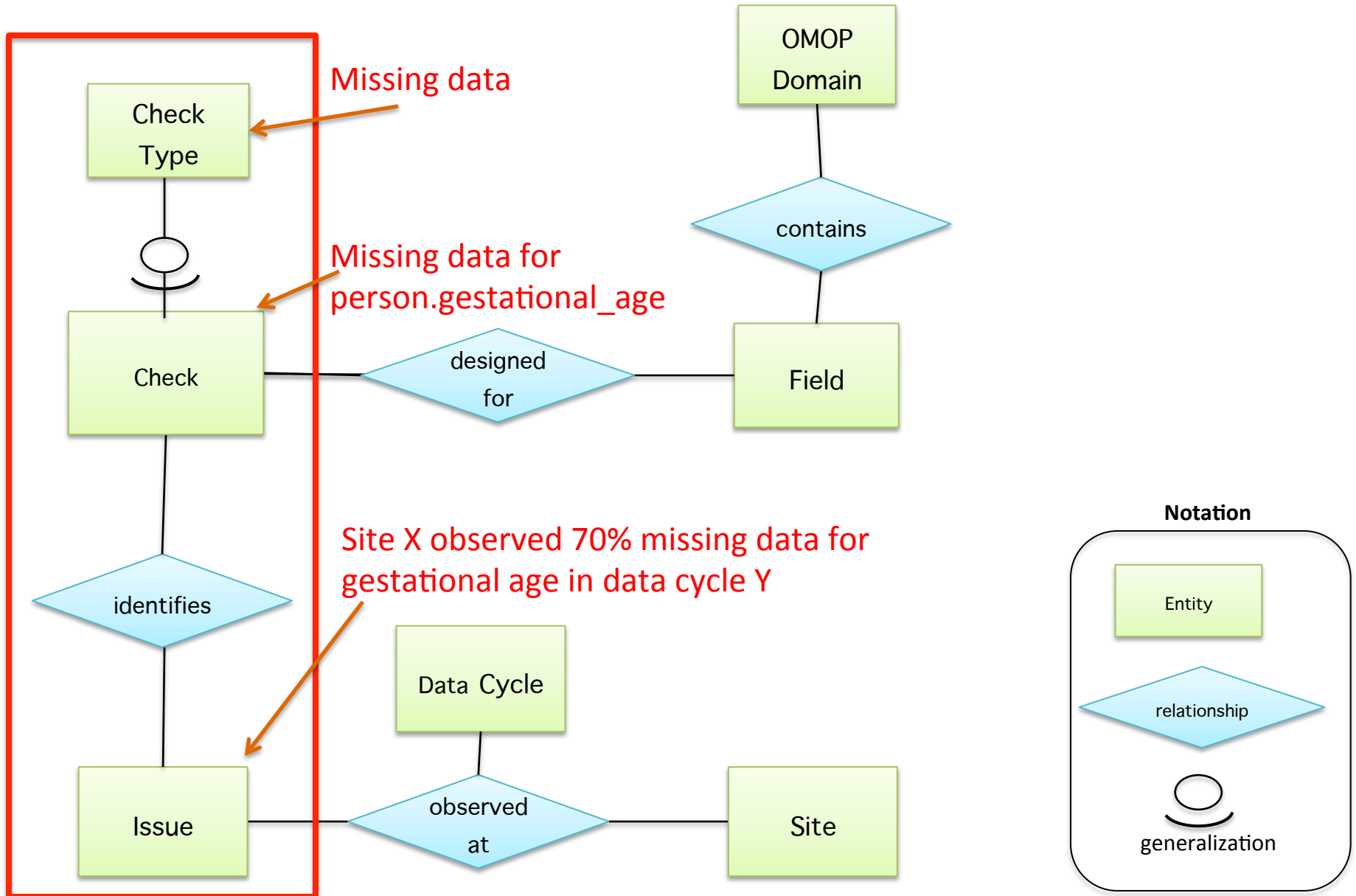
PEDSnet CDRN = 5.86M patients in pediatrics



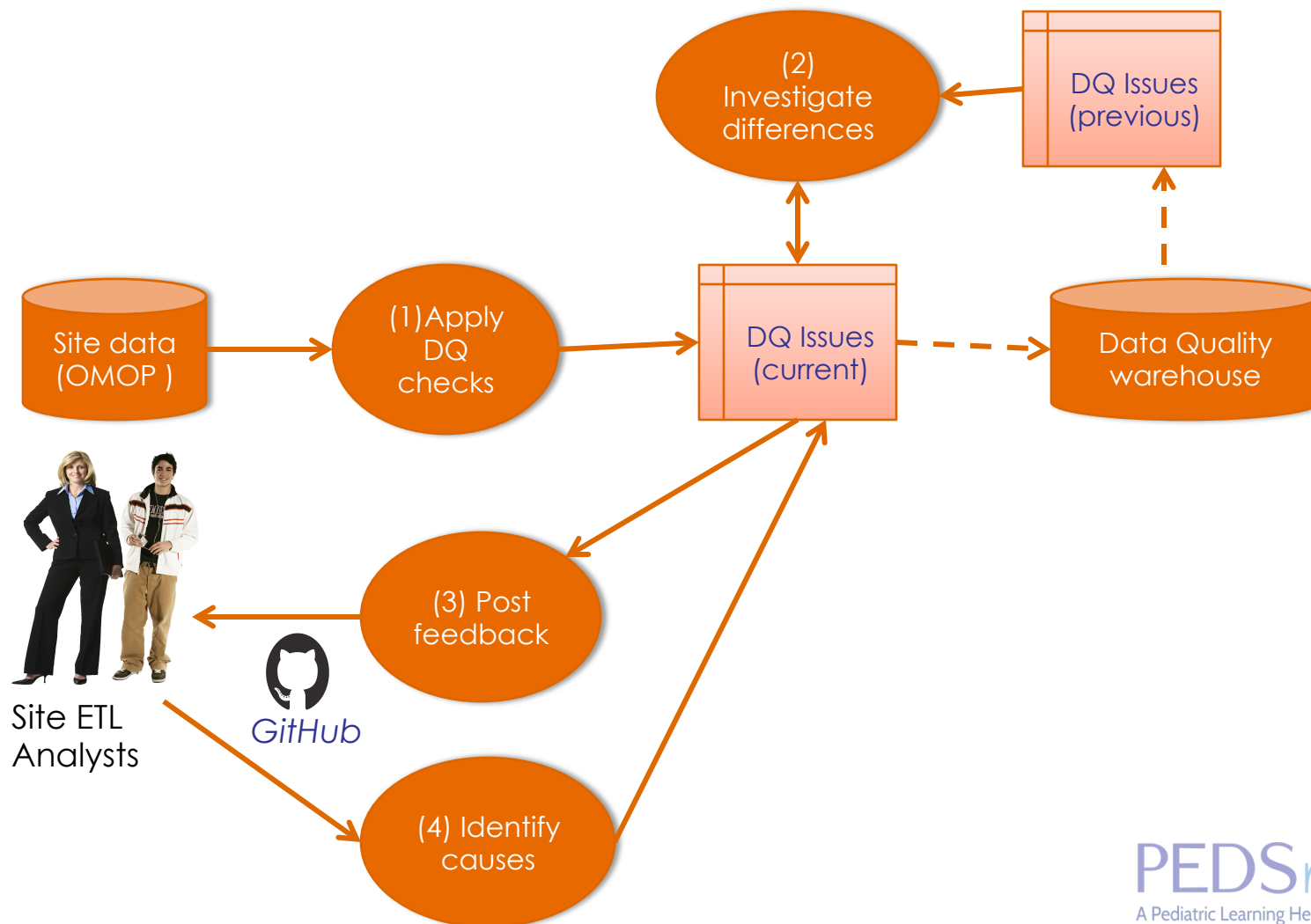
PEDSnet Data Cycle



Data Quality Conceptual Model



PEDSnet Data Quality (DQ) Workflow



GitHub Issue Screenshot

DQA: September 2016 (ETLv11): condition_occurrence/condition_concept_id #289

[Edit](#)[New Issue](#)Closed

opened this issue on Oct 5, 2016 · 1 comment

commented on Oct 5, 2016 • edited by

Description: Unexpected most frequent values

Finding: Shooting pain (concept_id: 4171519) - source values are either: "Achalasia esophagus|530.0"; Need for prophylactic vaccination and inoculation against in|Z23; Need for prophylactic vaccination and inoculation against in|V04.81, none of which should map to shooting pain

added Data Cycle: September 2016 Data Quality Status: new Table: condition_occurrence labels on Oct 5, 2016

commented on Oct 12, 2016

ETL error Fixed by the code fix in issue #292

Assignees

No one—assign yourself

Labels

Cause: ETL: programming er...

Data Cycle: September 2016

Data Quality

Status: solution proposed

Table: condition_occurrence

Projects

None yet

Results

Results of Data Quality

A longitudinal analysis of data quality in a large pediatric data research network

Ritu Khare, Levon Utidjian, Byron J Ruth, Michael G Kahn, Evanette Burrows, Keith Marsolo, Nandan Patibandla, Hanieh Razzaghi, Ryan Colvin, Daksha Ranade ...
Show more

Journal of the American Medical Informatics Association, ocx033,
<https://doi.org/10.1093/jamia/ocx033>

Published: 08 April 2017 **Article history** ▼

Objective: PEDSnet is a clinical data research network (CDRN) that aggregates electronic health record data from multiple children's hospitals to enable large-scale research. Assessing data quality to ensure suitability for conducting research is a key requirement in PEDSnet. This study presents a range of data quality issues identified over a period of 18 months and interprets them to evaluate the research capacity of PEDSnet.

Materials and Methods: Results were generated by a semiautomated data quality assessment workflow. Two investigators reviewed programmatic data quality issues and conducted discussions with the data partners' extract-transform-load analysts to determine the cause for each issue.

Results: The results include a longitudinal summary of 2182 data quality issues identified across 9 data submission cycles. The metadata from the most recent cycle includes annotations for 850 issues: most frequent types, including missing data (>300) and outliers (>100); most complex domains, including medications (>160) and lab measurements (>140); and primary causes, including source data characteristics (83%) and extract-transform-load errors (9%).

Discussion: The longitudinal findings demonstrate the network's evolution from identifying difficulties with aligning the data to a common data model to learning norms in clinical pediatrics and determining research capability.

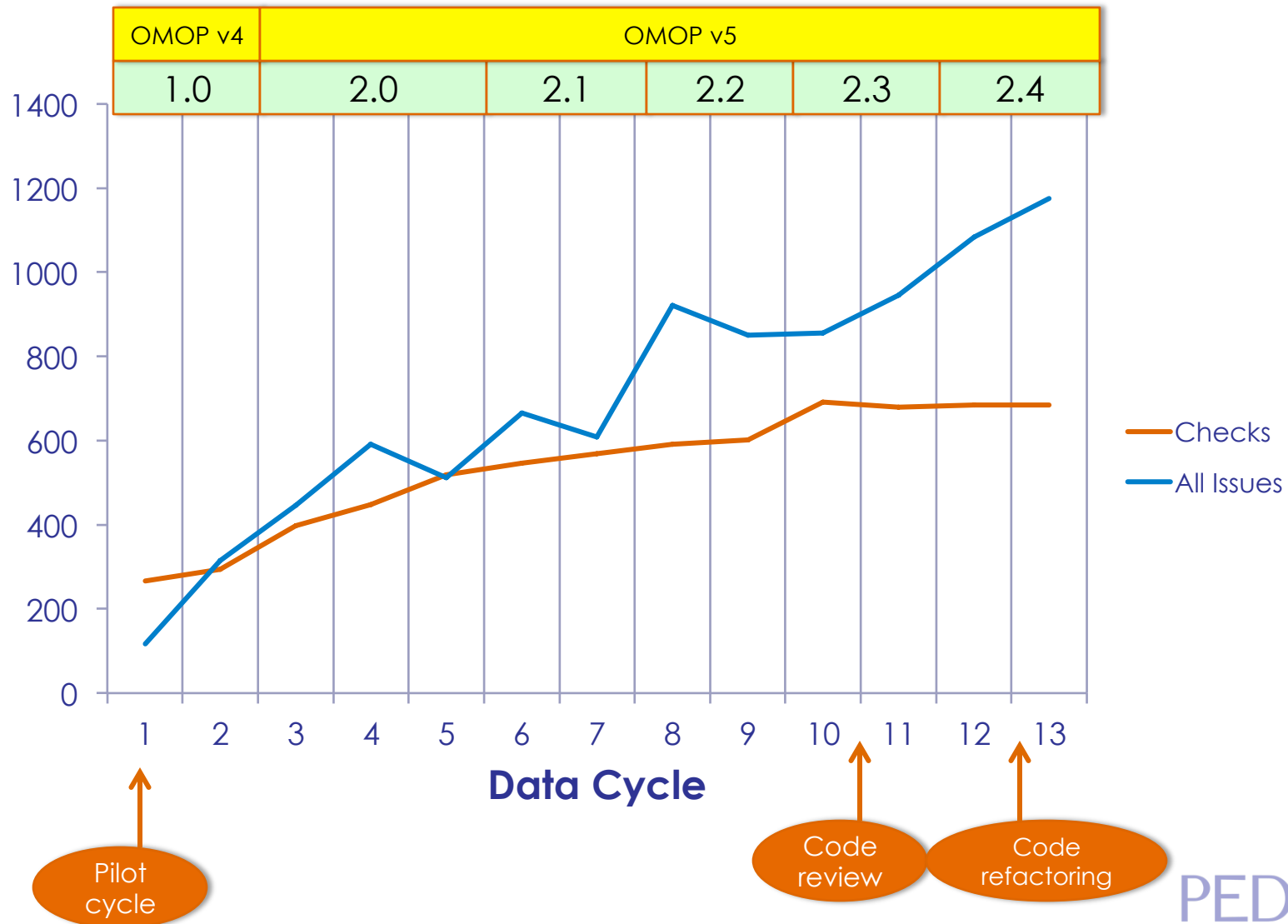
Conclusion: While data quality is recognized as a critical aspect in establishing and utilizing a CDRN, the findings from data quality assessments are largely unpublished. This paper presents a real-world account of studying and interpreting data quality findings in a pediatric CDRN, and the lessons learned could be used by other CDRNs.

Keywords: CDRN, data quality, electronic health record, extract-transform-load, secondary use

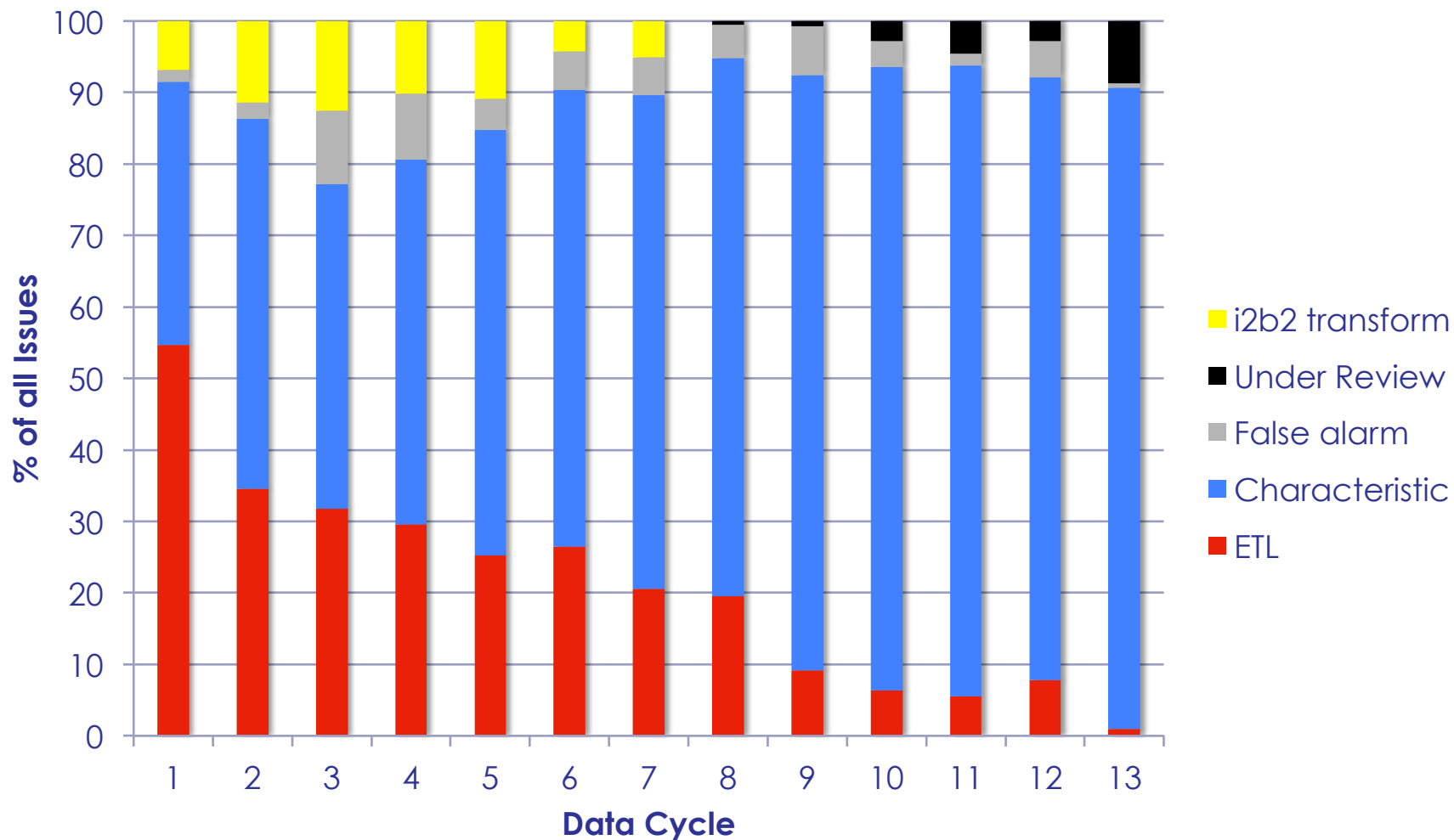
Methods

- PEDSnet Data Quality Warehouse
- January 2015 – March 2017 (13 data cycles)
- Total 9,086 data quality issues and related metadata
 - OMOP domain
 - Field
 - Check type
 - Cause
 - *Identified (all) vs. reported(new) issues*
 - *time to closure (GitHub)*

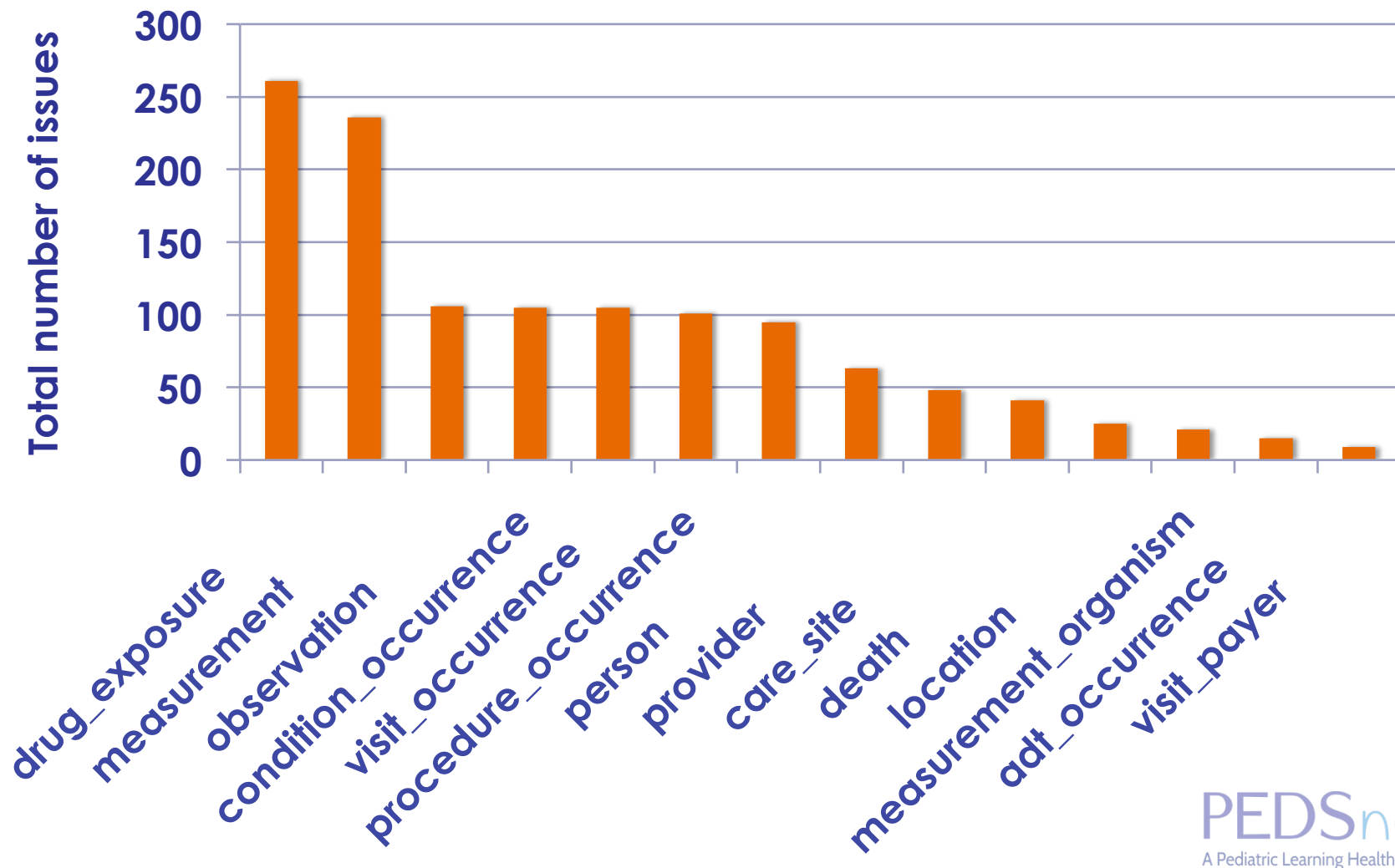
Checks vs. Issues



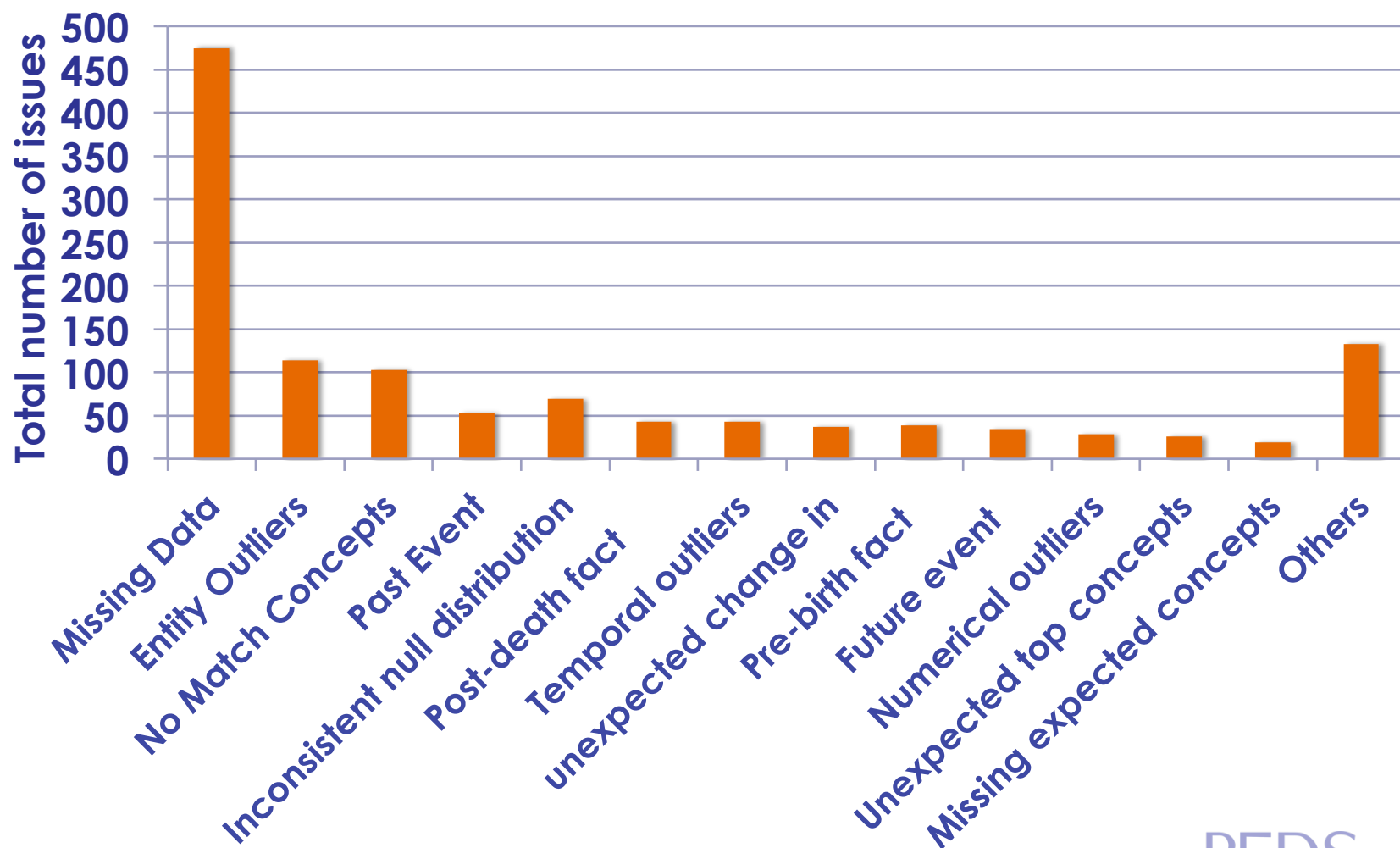
Causes across issues



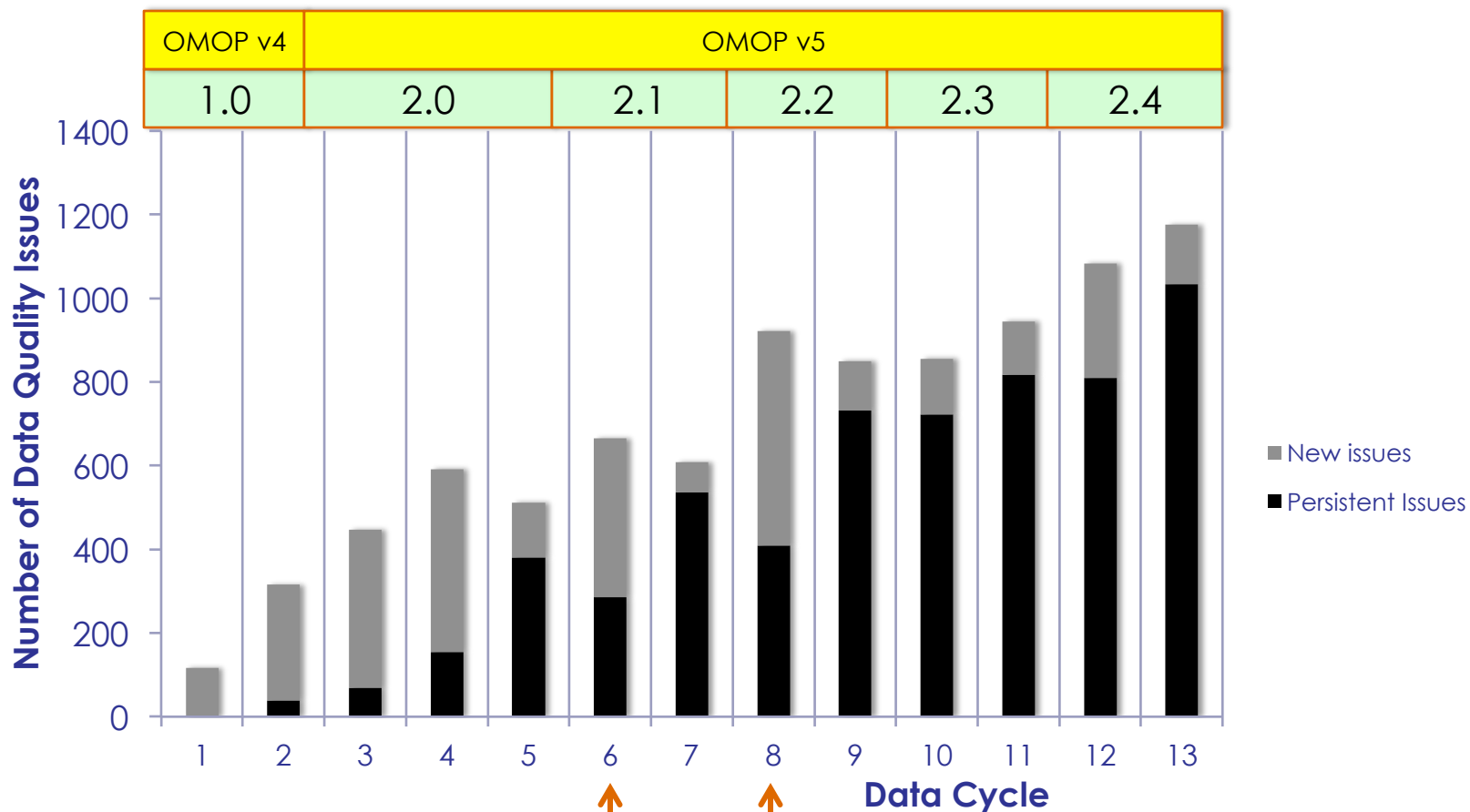
Distribution across Domains (cycle-13)



Distribution across Check Types (cycle-13)

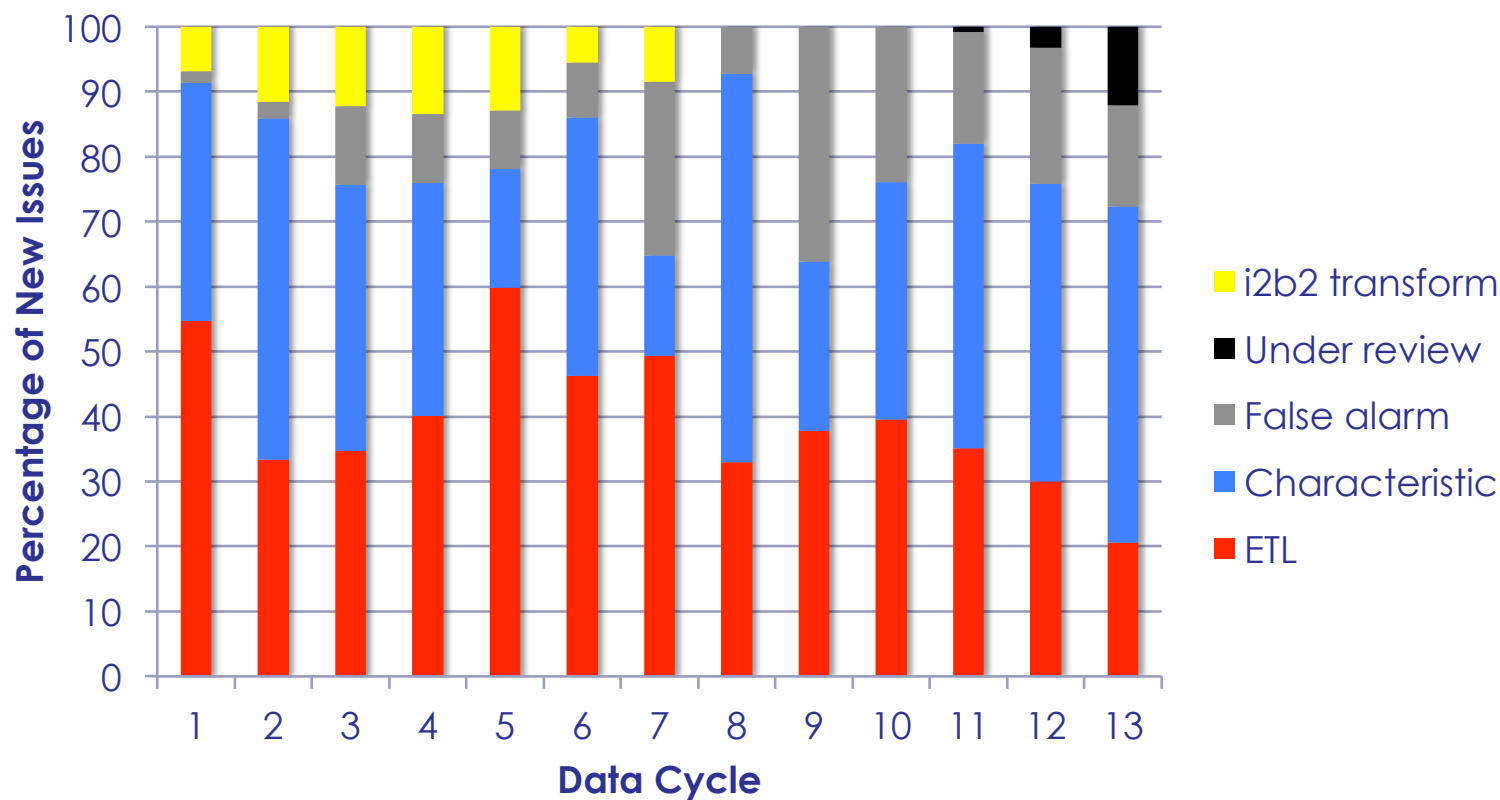


All issues = Persistent + New (*reported*) issues

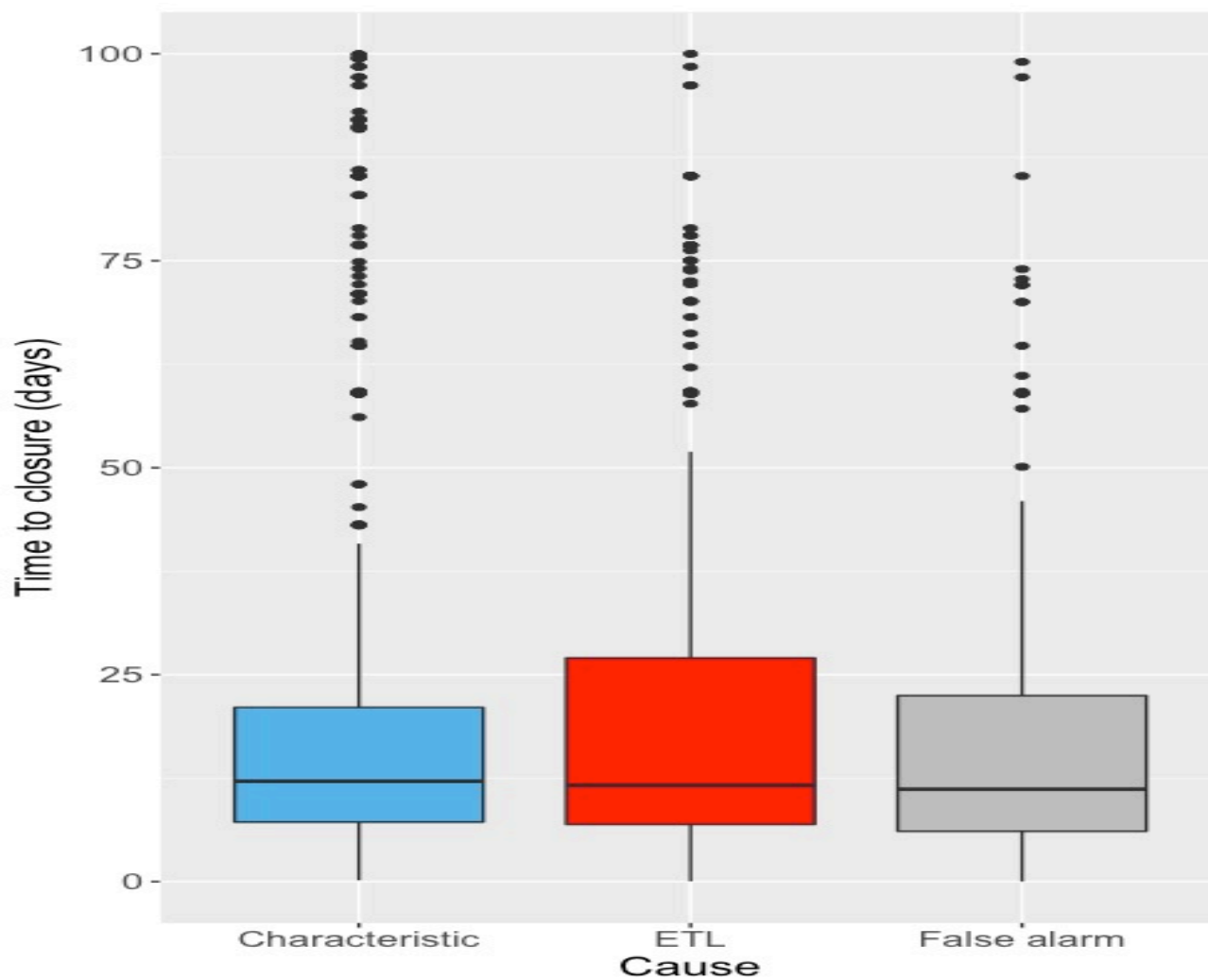


Access to additional
units/data, i2b2-
>OMOP

Causes across new(reported) issues



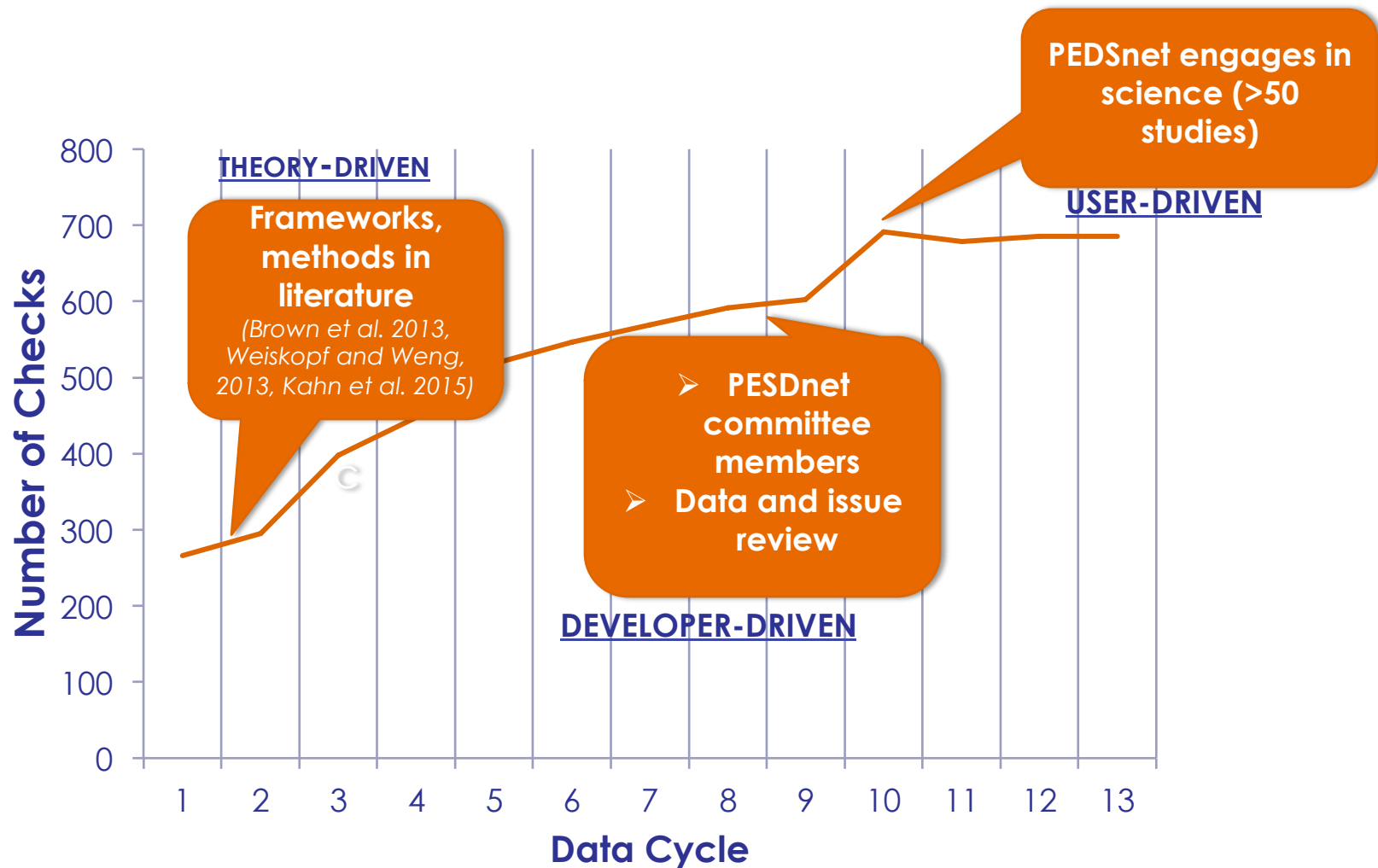
Time to closure of GitHub issues



Results

Data Quality Checks

Evolution of Data Quality Checks



Data Quality Checks in PEDSnet

- Overlap with Achilles Heel
 - Value set violation
 - Invalid concept identifier
 - Illegal vocabulary
 - Missing data
 - No matching concepts
 - Future event
 - Pre-birth fact
 - Post-death fact
 - Start date after end date

Data Quality Checks in PEDSnet

- New Checks (*PEDSnet data committee*)
 - Inclusion criteria violation
 - Date time inconsistency
 - measurement_datetime vs. measurement_date
 - Invalid format
 - procedure_source_value, condition_source_value
 - Unexpected change between data cycles
 - number of records
 - missingness in fields

Data Quality Checks in PEDSnet

- New checks (science queries)
 - Missing expected concept
 - E.g. creatinine labs, nephrology specialty for providers.
 - Insufficient facts for specific visit types
 - E.g. missing DRGs for inpatient admissions
 - Unexpected more frequent values
 - Identify outliers in top conditions and procedures (using cross-site comparison)

Open Questions and Challenges in Check Design

- Design checks for new (unexpected) issues encountered during science queries
 - Determine the combination of fields / tables
 - Determination of thresholds
 - Automatic review of ETL mappings
 - labs, organisms, specialty, route, race, ethnicity, drugs, language, procedure, smoking history
 - 1000s of manually derived mappings
- PEDSnet Data Quality Checks available on GitHub
 - <https://github.com/PEDSnet/Data-Quality-Analysis>

Acknowledgments

- Data Quality Team
 - David Soler
 - Josh Tucker
- PEDSnet data scientists
 - Hanieh Razzaghi
 - Levon Utidjian
- PEDSnet DCC director
 - L. Charles Bailey
- Other PEDSnet teams
 - ETL analysts
 - Site Informatics Leads
 - Leadership and Governance
- PCORnet Governance Committees and DRN OC
- *This work was supported by PCORI Contract CDRN-1306-01556.*
- OHDSI Consortium
- Patients and Families