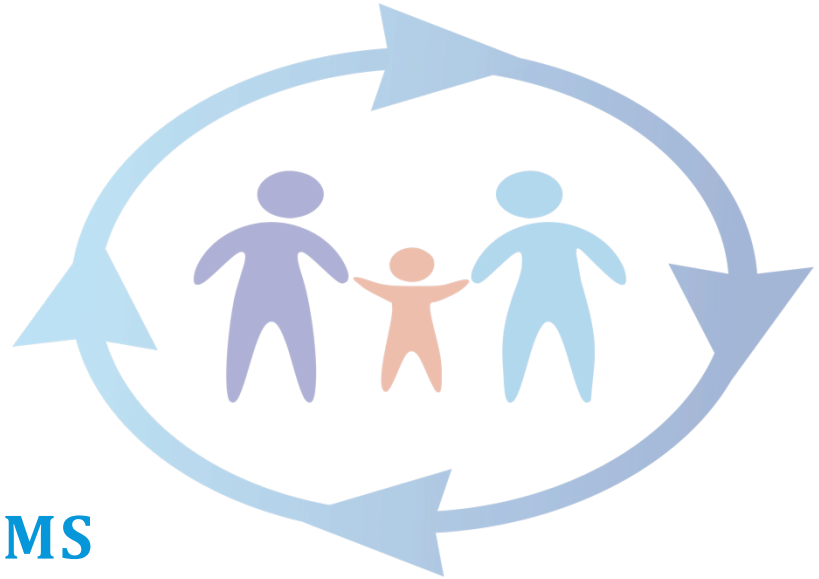
PEDSnet DQA Tutorial



Connor Callahan, MS

The Children's Hospital of Philadelphia

Roadmap

- Background on our process
- Overview of how to run our DQA
- DQA file structure and content
- Example output from our DQA
- Current goals

Background

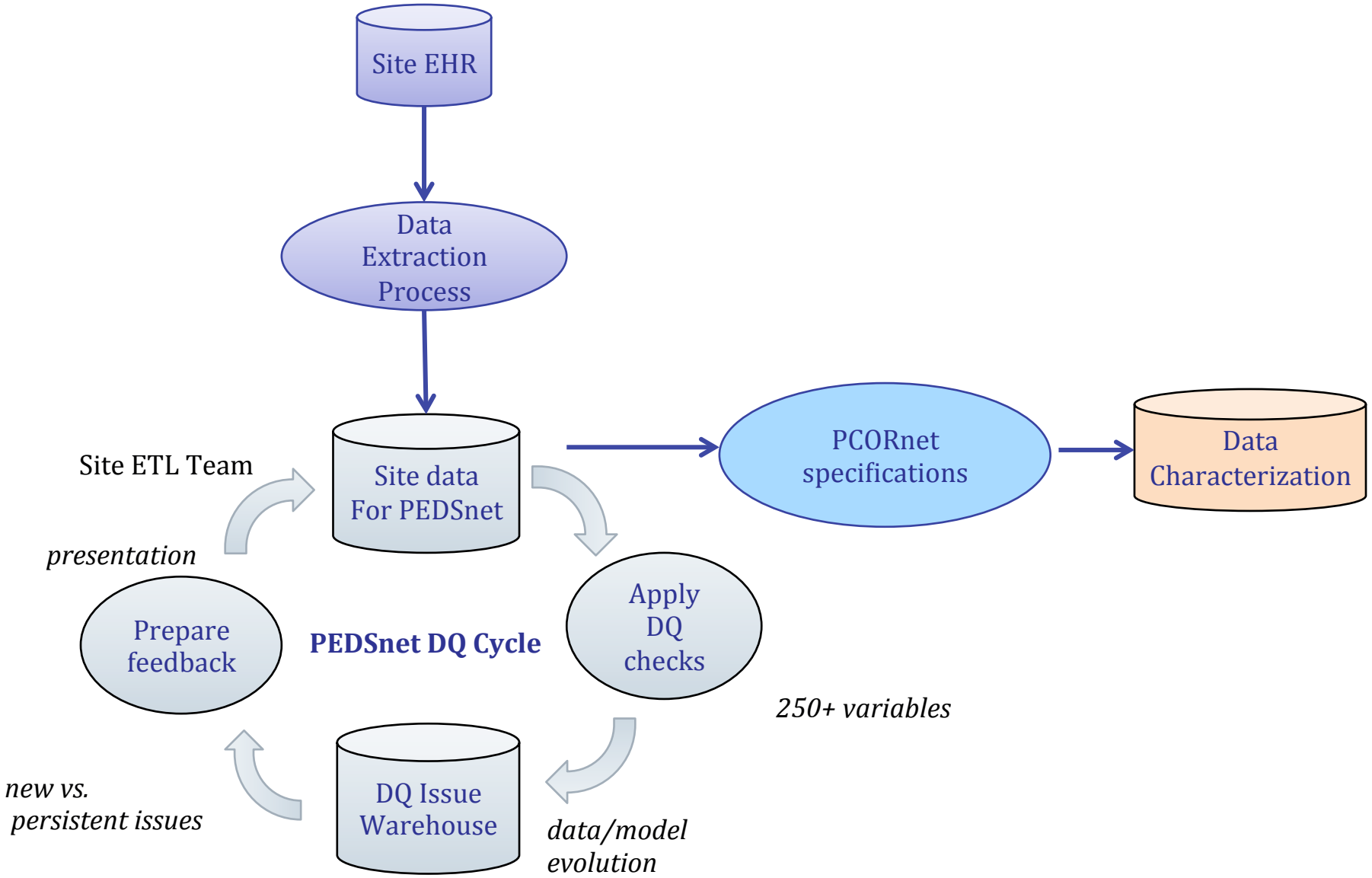
Our Use Case

- PEDSnet has integrated our DQA program into our data pipeline. We view it as essential to data standardization, the evolution of the PEDSnet CDM, and benchmarking goals

- Our institution receives data from our 7 constituent hospitals on a quarterly basis

- We continue to adapt our DQA to our evolving needs

PEDSnet Data Quality Process Overview



PEDSnet
A Pediatric Learning Health System

Current Workflow

- Sites run ETL and data is securely transferred to our site
- We upload that data to our local database and run our DQA toolkit
- Our goal is to create GitHub issues to facilitate a brief, focused conversation with each site's ETL analysts, for each issue
  - Need to avoid issue fatigue

Feedback from DQA

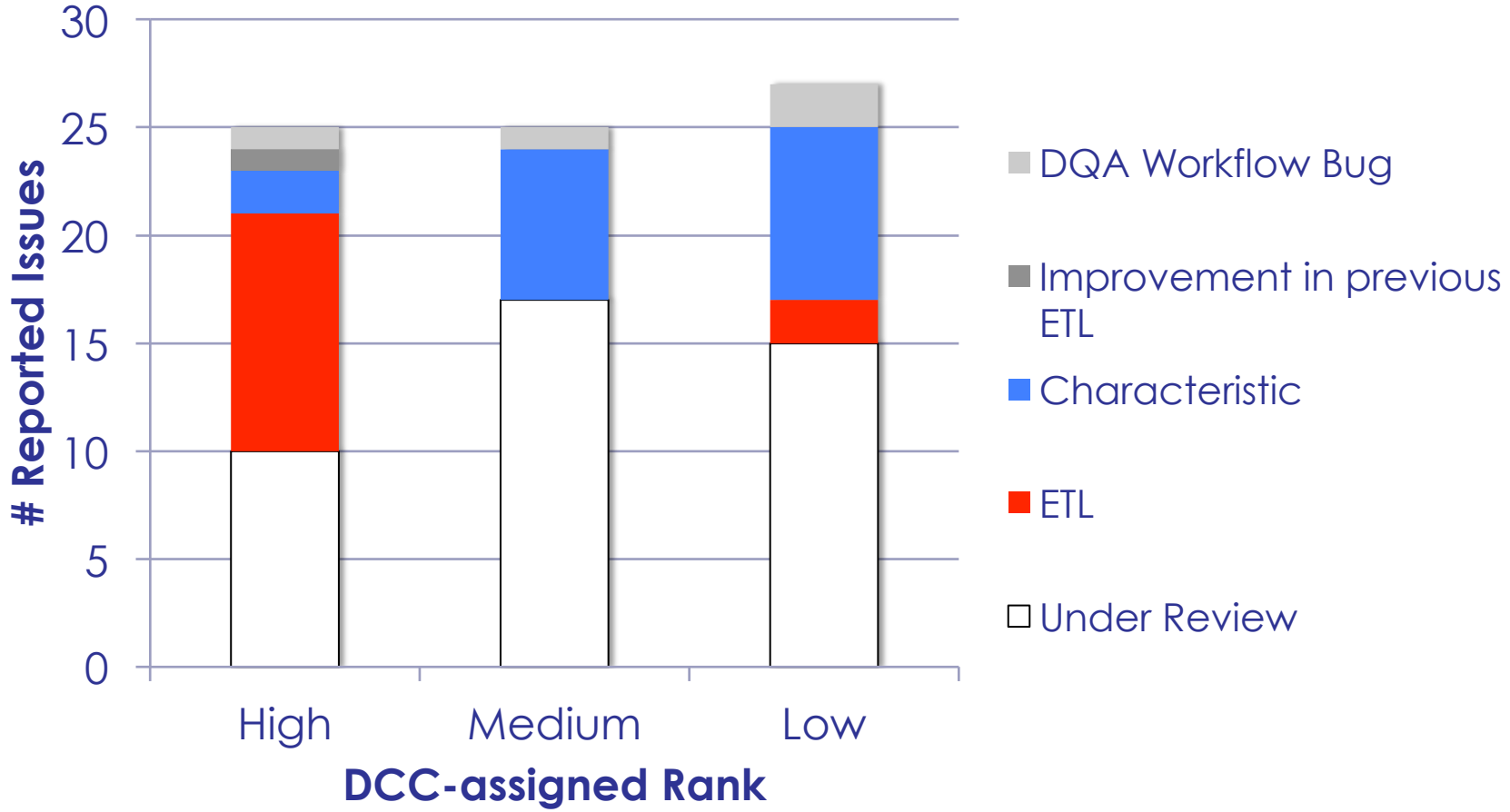
- The cause of each issue is labeled as ETL, persistent (characteristic), an improvement, or a workflow error along with an appropriate cause label

- Rank-order using low, medium, and high importance

- ETL analysts work with the PEDSnet DCC to unravel each issue to see whether ETL can be reworked to improve the result

Rank and Cause Distribution

v3.0 Submission



- High = Major ETL issues
- Medium = characteristic issues + minor ETL issues
- Low = Improvements over previous ETL

PEDSnet
A Pediatric Learning Health System

Running the PEDSnet DQA

Cloning the Repository

- Go to https://github.com/PEDSnet/Data-Quality-Analysis click "clone" to copy the provided link, enter "git clone" and that link in your terminal

- This is a public repository that is updated regularly

- Currently supports Oracle and PostgreSQL databases

Install Necessary Packages

- Below the code links in the GitHub repository is a link of packages that need to be installed for the DQA to work

- Our DQA utilizes an R package developed by Dr. Charles Bailey at CHOP called ‘Argos’ which streamlines our database connections using a .json file

- The package extensively uses the tidyverse via the ‘dplyr’ package

Prepare the Configuration Files

- From the top directory, go to ‘Resources’ and open ‘PEDSnet_config.yml’ and add information relating to your database connection, output directory, CDM version, etc.

- Create an argos directory (preferably a hidden .argos directory) and add your argos configuration .json
  - This is outlined on the repository

Run the DQA

- Set your working directory to the top directory of the package

- Source the 'Run_DQA.R' file in this directory
  - Can run a single report with generateSingleReport()
  - Can run all reports with runDQA()

- Any errors or printed information will be recorded in the file 'dqa.log' in the top directory

DQA Structure & Content

Data Directory

- The data directory is split into three subdirectories by use case
  - The PreviousDataSummary folder contains data from the previous data cycle that is used in comparison checks
    - -i.e. did the total number of inpatient visits change significantly?
  - The DQACatalog folder contains an inventory of check information that is read in every time a check is run
    - -i.e. threshold values and a brief description of the check for the ETL analyst
  - The ValueSets directory contains table-specific information for checks
    - i.e. what filtering conditions should be used or what are the accepted procedure_type_concept_ids

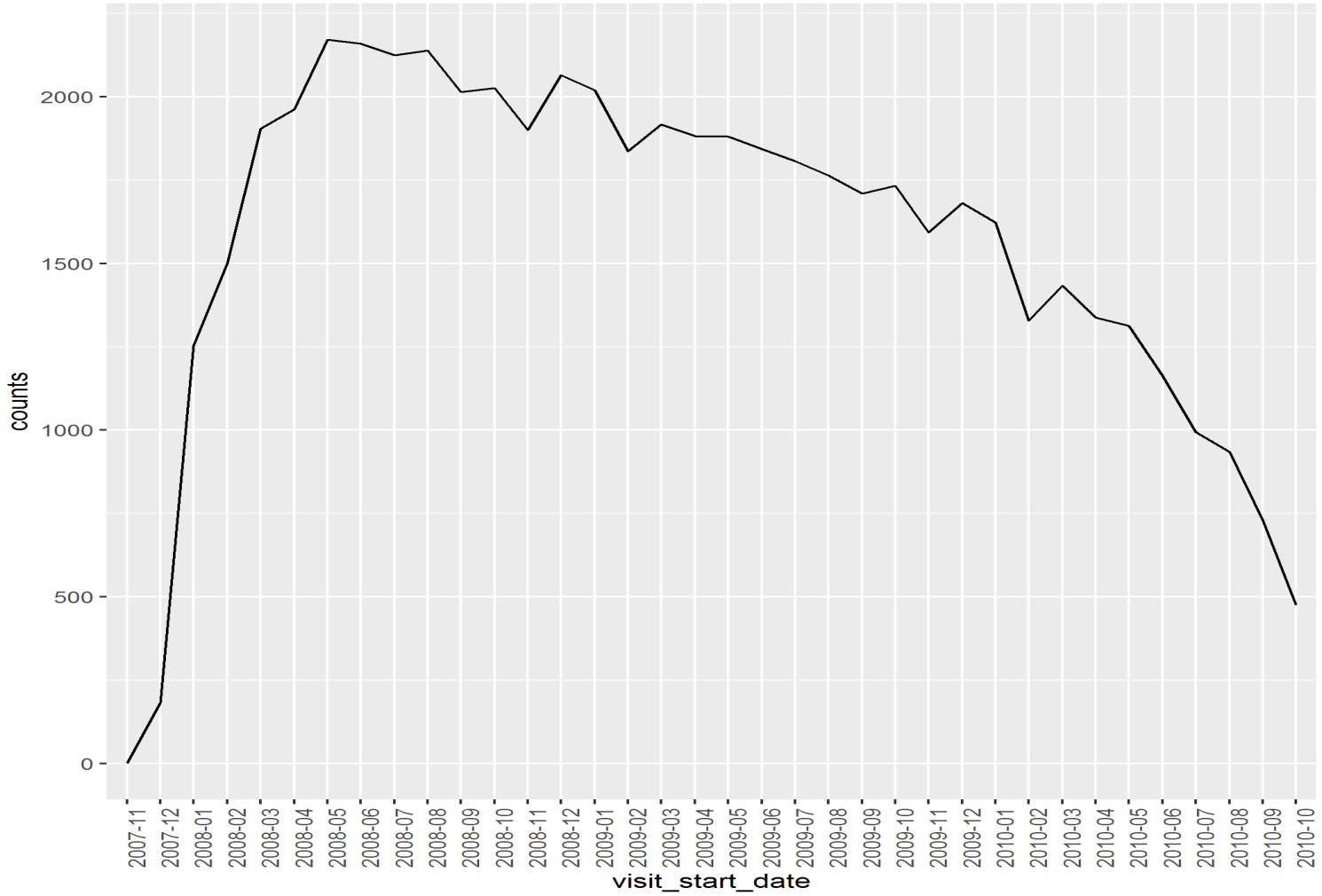
Library

- The library directory contains a script for each of the checks that our DQA implements

- This directory also includes functions for database operations in the script “PerformDatabaseOperations.R” which executes queries to produce specific data frames used in checks and reports
  - By putting generalized queries in this file, we avoid updating each check

- Various plots are also produced using functions from the “CreatePlots.R” file

Example Plot: Visit_start_date Monthly Count



PEDSnet
A Pediatric Learning Health System

Main Directory

- The main directory is split into two levels
  - Level 1 focuses on univariate checks
  - Level 2 focuses on checks that utilize multiple tables with foreign keys

- Within each level is a series of report functions that implement all checks around a given table

Example Output

Results

- Results directory will now have the DQA output spread through four separate subdirectories

- Data directory includes summary information used for the DQA process such as ‘total_counts.csv’ as well as records such as ‘top_inpatient_drugs.csv’ which are used for issue generation

- Images directory contains all images generated while the report directory contains summary markdowns, including those images

Issues

- The ‘issues’ directory includes a list of issues found by the DQA that for each domain

- Issues are copied and summarized in GitHub using a Go Language program that can recognize if an issue should be reposted or updated based on GitHub labeling

- We want to discuss all new issues at hand while avoiding fatigue with sites by reposting resolved or characteristic issues

Examples

- Illegal vocabulary in condition table:
  - https://github.com/PEDSnet/CHOP/issues/758
- Unexpected drop in visit payer records between data cycles:
  - https://github.com/PEDSnet/CHOP/issues/752
- Pre-Birth visits:
  - https://github.com/PEDSnet/CHOP/issues/690
- Unexpected Change in number of outpatient visits:
  - https://github.com/PEDSnet/CHOP/issues/731
- Temporal Outliers due to flu season:
  - https://github.com/PEDSnet/CHOP/issues/636

Current Goals

DQA Continues to Evolve

- Each data cycle (quarter), our data quality package adapts to our needs and standards
  - Improve checks based on stakeholder feedback
  - Add new fields and reports based on information added to PEDSnet CDM
  - Design new checks based on ongoing findings in our data

Expanding DQA Scope

- We are currently looking to add metadata collection to our data quality process
- This will involve recording information from each check each time the program is run
- This could provide further insight into how our DQA is performing overall, how sites have evolved over time based on feedback, and open avenues for statistical investigation of data quality

Acknowledgments

- Data Quality Team
  - Ritu Khare , PhD
- PEDSnet data scientists
  - Hanieh Razzaghi , MPH
  - Levon Utidjian, MD
- PEDSnet DCC director
  - Charles Bailey, MD, PhD
- 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

PEDSnet
A Pediatric Learning Health System