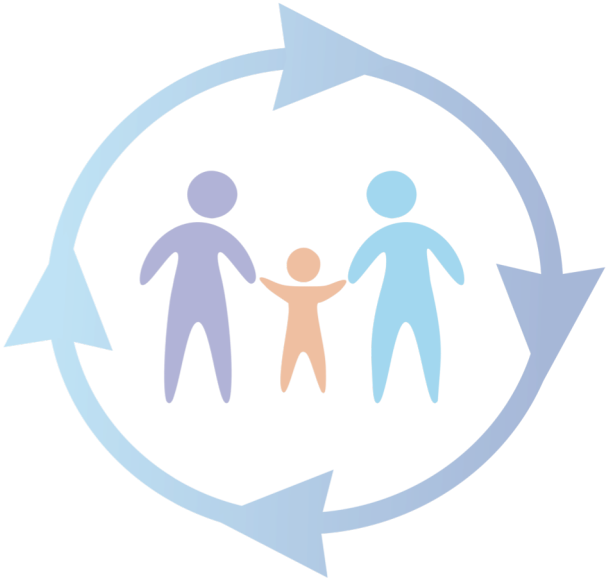
PEDSnet Data Science Training



5/14/2019

Overview

- Web sessions

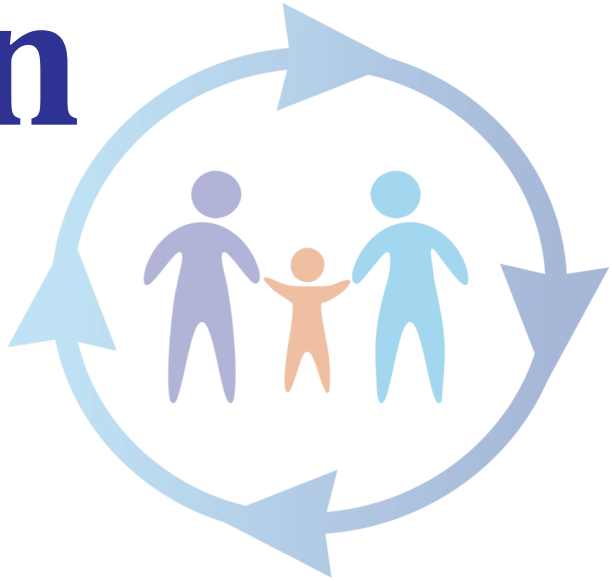
Data Model & Vocabularies	April 2019
Data Request Fulfillment	May 2019
Web Application	June 2019
PEDSnet Standard R Framework	July 2019

- Hands-on workshop – August 01-02, 2019

- Use of network data and scientific projects

PEDSnet
A Pediatric Learning Health System

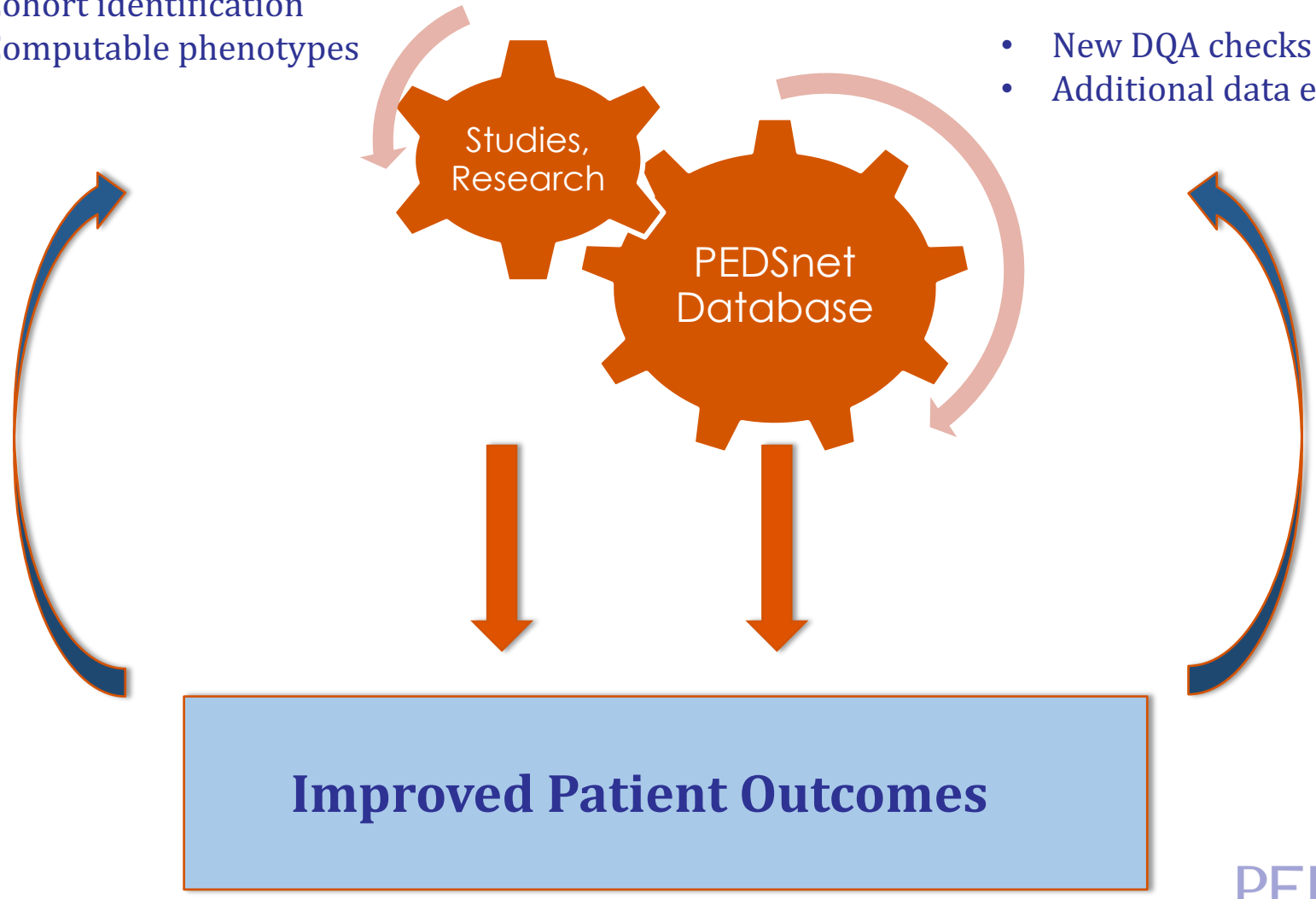
Introduction



PEDSnet as a learning network

- Codeset development
- Cohort identification
- Computable phenotypes

- New DQA checks
- Additional data elements



PEDSnet
A Pediatric Learning Health System

Past Work

Type	Examples
Government	• Anthropometric measurements • Serve as learning health network in large training grants (P50) and research mentorship grants (K12)
Academic	• Asthma and BMI association • Dosing of croup during inpatient hospitalization • Numerous computable phenotypes
Industry	• Trial feasibility • Trial recruitment efforts
Linkage	• Dispensing (outside claims) and prescribing (PEDSnet) for PCORnet abx study

PEDSnet Data Request Process

External collaborator completes “front door” request

Administrative review by PMO, Data, and Research Committees

Prioritization by Research Committee

Administrative processes (e.g., budget, scope of work, contracting, etc)

Study query and remaining administrative processes

PEDSnet
A Pediatric Learning Health System

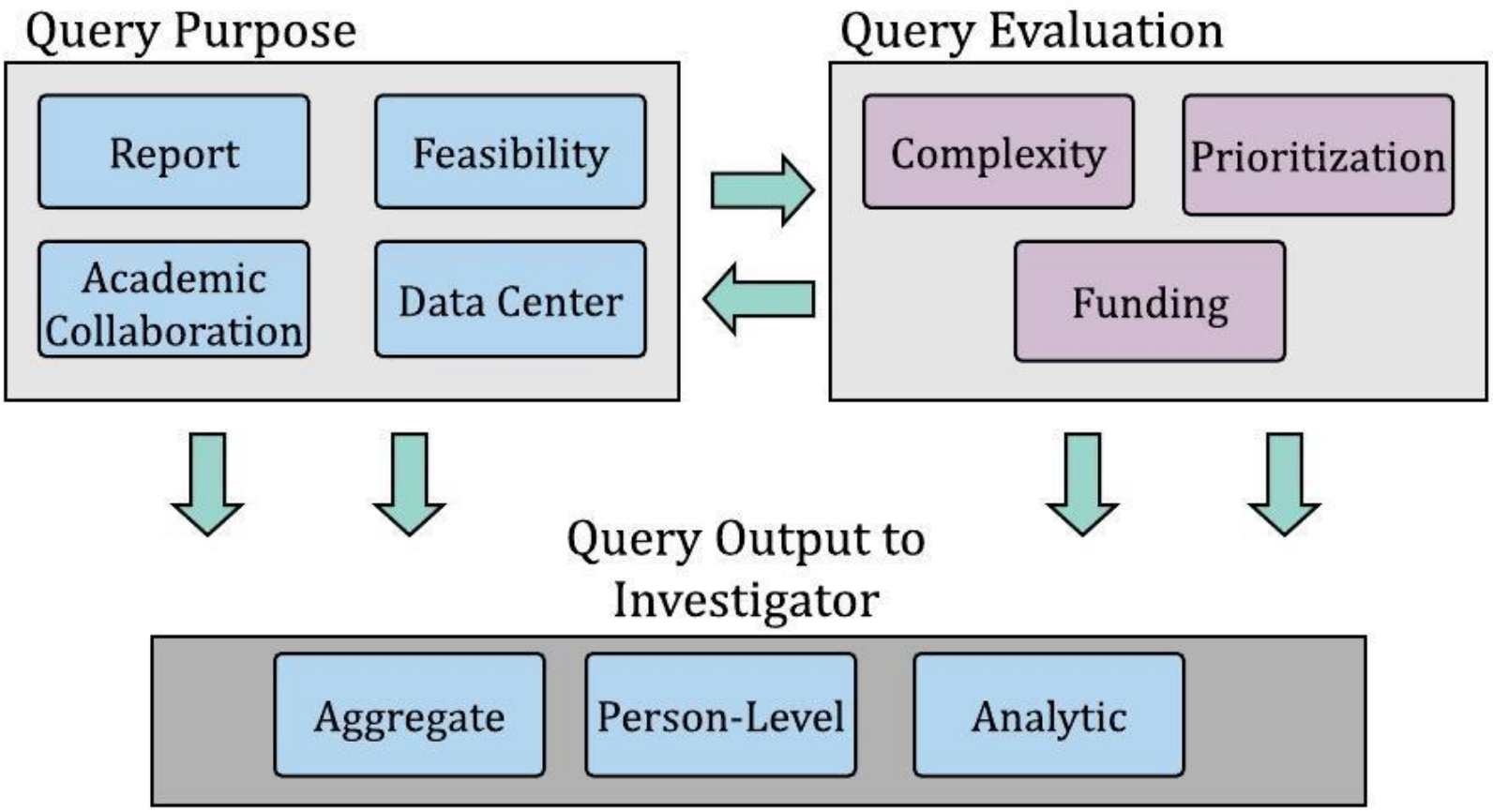
Session Overview / Objectives

- Data request types and how they relate to output
- Specification Form
  - Purpose
  - Major sections
- Major steps in the query process
- Aggregate vs Person-Level Query
- Tools necessary for good data science
  - DCC-specific software review (Atlassian)
- PEDSnet risk reduction standards

PEDSnet
A Pediatric Learning Health System

Types of Data Requests

Overview



PEDSnet
A Pediatric Learning Health System

Overview

Consideration	Definition
Query Purpose	• Determines the role of the analytic team • Sponsor plays a big role in level of formality and collaboration
Query Evaluation	• Usually determined in collaboration with Research Committee (RC) • Outcome of this step usually determines timeline and engagement with sponsor and RC
Query Output	• Output can be somewhat related to purpose but are two separate considerations • Sometimes one large study can require two different kinds of output • Difference between output to DCC data science team vs. output to investigator

Query Purpose

Feasibility	Report	Academic Collaboration	Data Center
• Usually performed before funding is available • Can be very quick or complicated • Usually aggregate-level output	• Can be population report, trial feasibility assessment, trial facilitated recruitment • Varies in scope and sponsor-dependent	• Purpose is usually for a paper • High level of input and direction from data scientists • Less formal than a report	• Less analytic work and more focus on data integration • PPRL work as an example

PEDSnet
A Pediatric Learning Health System

Query Evaluation

Complexity

- Dependent on number of variables, level of DQ, number of derivations, number of codesets, etc
- Requires high level of input from data scientists

Funding

- Can be government, academic, industry
- Highest level of project formality from industry

Prioritization

- Usually set by Research Committee
- Combination of multiple considerations and related to overall goal and direction of network

Query Output

Aggregate

- Examples include counts of patients with certain conditions, drug distributions, health utilization, etc
- Ranges in complexity

Analytic

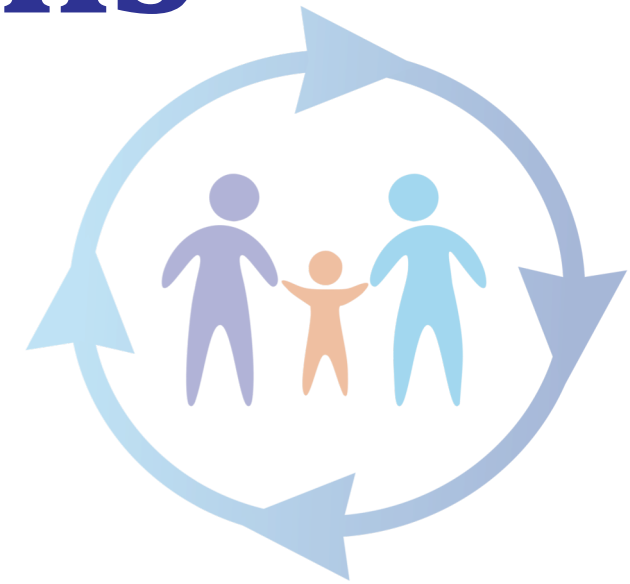
- Output is usually also aggregate but requires a level of analysis (e.g., graphs, statistical tests, etc)
- Usually in combination with aggregate

Person-Level

- Release of datasets
- Usually outside analytic team
- Data scientists still should serve as data experts throughout project
- Output resembles data model

PEDSnet
A Pediatric Learning Health System

Specifications Form



Purpose of Specification Form

- Fosters collaboration with investigators
- Serves as communication tool and guide throughout query process
- Provides documentation for agreed upon terms of query
- Contains all major elements of data query requirements
- Standardizes approach to queries across studies
- ***NOTE:*** *This document can serve as a space that points to other files in the repo (e.g., if logic rules or stored somewhere else)*

Specification Form: Table of Contents

PEDSnet
A Pediatric Learning Health System

PEDSnet
A Pediatric Learning Health System

Major Sections of Form

Section	Summary
Requestor Information	• PI and organizational information (e.g., contact information) • Brief description / overview of study (title, funding status, aims, etc)
Query Description	• Characteristics for cohort inclusion (inclusion/exclusion criteria) • Description of other variables in study)
Variable Identification	• Mapping of study variables to domains in data model
Logic Rules for Query	• Written rules for how to use the data to derive the variables of interest
Output Structure	• Separate sections for person-level datasets and aggregate output • Attempts to distinguish fields in data model vs derivations for study
Codeset Identification	• All codesets used in study should be listed in this section • Resembles the vocabulary.concept tables
Index	• Example of study variables and logic rules to demonstrate difference

Data Variable Identification (Example in Index)

	Cohort Identification	Other Variables
Demographics	• Children between the ages of 5-10 years old	• Race, ethnicity, and gender
Visits	• Outpatient, inpatient, and ED. Only face to face visits.	• All visits for the identified cohort
Diagnoses	• Asthma	• Medically complex diagnoses (for flagging) --- see codeset for diagnoses • All inpatient discharge diagnoses • All respiratory-related outpatient and ED visits --- see codeset
Procedures	• None	• Chest x-rays
Medications	• None	• Bronchodilators, steroids • Antipsychotics, mood stabilizers, antidepressants
Immunizations	• None	• None
Labs/Measurement	• None	• All heights, weights, BMI, and BMI z-scores available for patient • PFT measurements • Blood tests for viral and bacterial infections
Death	• None	• None
Provider	• None	• All interactions with a pulmonary specialist
Visit Payer	• None	• None
Care Site	• None	• All interactions with a pulmonary specialist
Other	• None	• None

- lists **only data** variables in each domain
- logic should not be in this section
- `other variables` is purposefully broad
- a study variable may be a combination of data variables + logic

PEDSnet
A Pediatric Learning Health System

Logic Rules for Query (Example in Index)

Logic

Rules for Cohort Inclusion

- Children should receive 2+ visits of asthma between the ages of 5-10 years old.
- Here, a visit is counted as a face to face visits associated with an asthma diagnosis. Once a patient is included, we pull all their face to face visits outside of the age window for inclusion. Face to face visits should exclude lab and radiology visits.
- Pull both office visits and problem list diagnoses, as long as these do not occur on the same day
- The index visit for a patient is the date of their first asthma diagnosis for inclusion.

Rules for Cohort Exclusion

- If patients have multiple visits for asthma but do not meet the requirement of 2+ face to face visits, they should still be excluded.
- A patient is excluded if 2+ asthma diagnoses are outside of the age window of 5-10 years, defined as the day of a patient's fifth birthday until the first day of the patient's eleventh birthday.

Rules for Dataset Creation

- Create a complexity flag for each patient for each body system. The flag should indicate if a patient has 2+ visits for that body system in the window of interest. A yearly complexity grid should be created for each year a patient is in the cohort.

Other Rules

- Diagnoses for HIV and pregnancy-related diagnoses will need to be excluded from the cohort.

Output Structure

	PEDSnet Variables	Derived Variables
Patient Cohort Table		
Visits		
Diagnoses		
Procedures		
Medications		
Immunizations		
Labs/Measurement		
Death		
Provider		
Visit Payer		
Care Site		
Code Lookup Table		
Other		

- ***Patient Cohort Table*** provides metadata for the included patients in cohort
  - **usually** a person-level table
  - can flag major variables of interest (e.g., number of visits, indicator for outcome of interest, etc)
  - usually the only table in dataset that requires derivations
- **Other tables should be similar to structure of data model tables**

PEDSnet
A Pediatric Learning Health System

Codeset Identification

Domain	Vocabulary	Concepts	Codeset	Notes

- Purpose is to create codesets that map to standardized vocabularies in PEDSnet data model
- Usually too complex and cumbersome to keep in word document --- can instead point to directory in repo containing codesets (as well as name of the file)
- Table structure can contain codesets or serve as shell to list where full codesets can be found

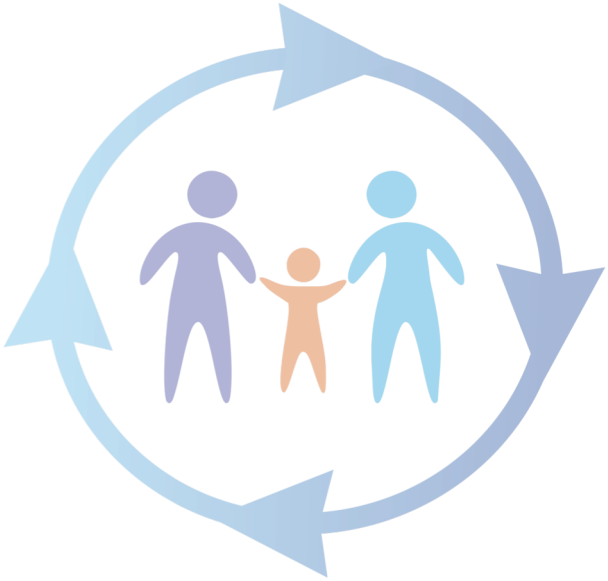
Completing Specification Form

Section	Responsible Party	Notes
Requestor Information	• Can be DCC, project manager, or study PI	• Usually the simplest part of form • Mostly administrative
Query Description	• Study PI	• Study investigator summarizes in research framework
Variable Identification	• First attempt by study PI	• Will usually need help / guidance from DCC • Serves as good education for the investigator, re: semantics of data model
Logic Rules for Query	• Both DCC and study team	• DCC leads with questions and guides conversation • Very important to document • Highly collaborative process
Output Structure	• DCC	• Study team can provide input but DCC leads this discussion
Codeset Identification	• DCC as default but low-effort queries may require study team to lead	• Usually a very time consuming part of the query development process • Most of the time the DCC will lead with input from study team – though there are exceptions

Managing Investigator Expectations

- Investigators may have additional requests after they receive the data or continue to develop their study question
- The specification form is useful to help set expectations for what the study data output will be before data is delivered
- Common last minute issues have included:
  - Additional variables not originally requested
  - Patient counts for attrition table or study flow diagram for eventual study publication
- Asking investigators to commit to what was requested in the specification form can help mitigate such requests.

Study Conduct and Query Process



Regulatory Considerations

- Engage with PEDSnet PMO early to determine regulatory needs
- DCC default data privacy:
  - date shifting
  - ID replacement
  - no source values
  - obscuring site names
- Difference in regulatory needs based on output (aggregate vs person-level) and data recipient (PEDSnet vs non-PEDSnet)
- PEDSnet has a data governance framework
- Determine level of IRB approval

Study Startup and Checklist

- Contract
- Budget
- Regulatory and governance decisions
  - DUA
  - RUD
  - IRB
- RC approval and prioritization
- Institutional Official sign-off on final dataset

PEDSnet
A Pediatric Learning Health System

"Data Governance Grid"

PEDSnet Data Sharing Procedure				
Data Set Type	**Required Procedures**	**Data Recipient**		
		PEDSnet Member[1]	**PCORnet Distributed Research Network Operations Center (DRN OC)[2]**	**Other**
Aggregate Counts for Feasibility Analyses	Human Subjects Review	None	None	None
	Legal Review	None[3]	None	None[3]
	Network Participation Approval	None	None	None
	Institutional Participation Approval	None	None	None
Aggregate Counts for Publishable Research	Human Subjects Review	None	None	None
	Legal Review	None[3]	None	None[3]
	Network Participation Approval	Executive Committee Approval	Executive Committee Approval	Executive Committee Approval
	Institutional Participation Approval	None	None	None
De-identified person-level (safe harbor method of de-identification only) **OR** **Limited Data Set with allowable LDS HIPAA identifiers EXCEPT actual dates; may include obscured dates[6]**	Human Subjects Review	Not Human Subjects Research Determination[4]	Not Human Subjects Research Determination[4]	Not Human Subjects Research Determination[4]
	Legal Review	PEDSnet DUA and RUD	Data Release Agreement [per PEDSnet, not PCORnet SOP] [PCORnet DSA between PCORnet and PEDSnet sites][5]	PEDSnet DUA and RUD
	Network Participation Approval	Executive Committee Approval	Executive Committee Approval	Executive Committee Approval
	Institutional Participation Approval	Prospective Site PI Approval	Prospective Site PI Approval	Prospective Site PI Approval

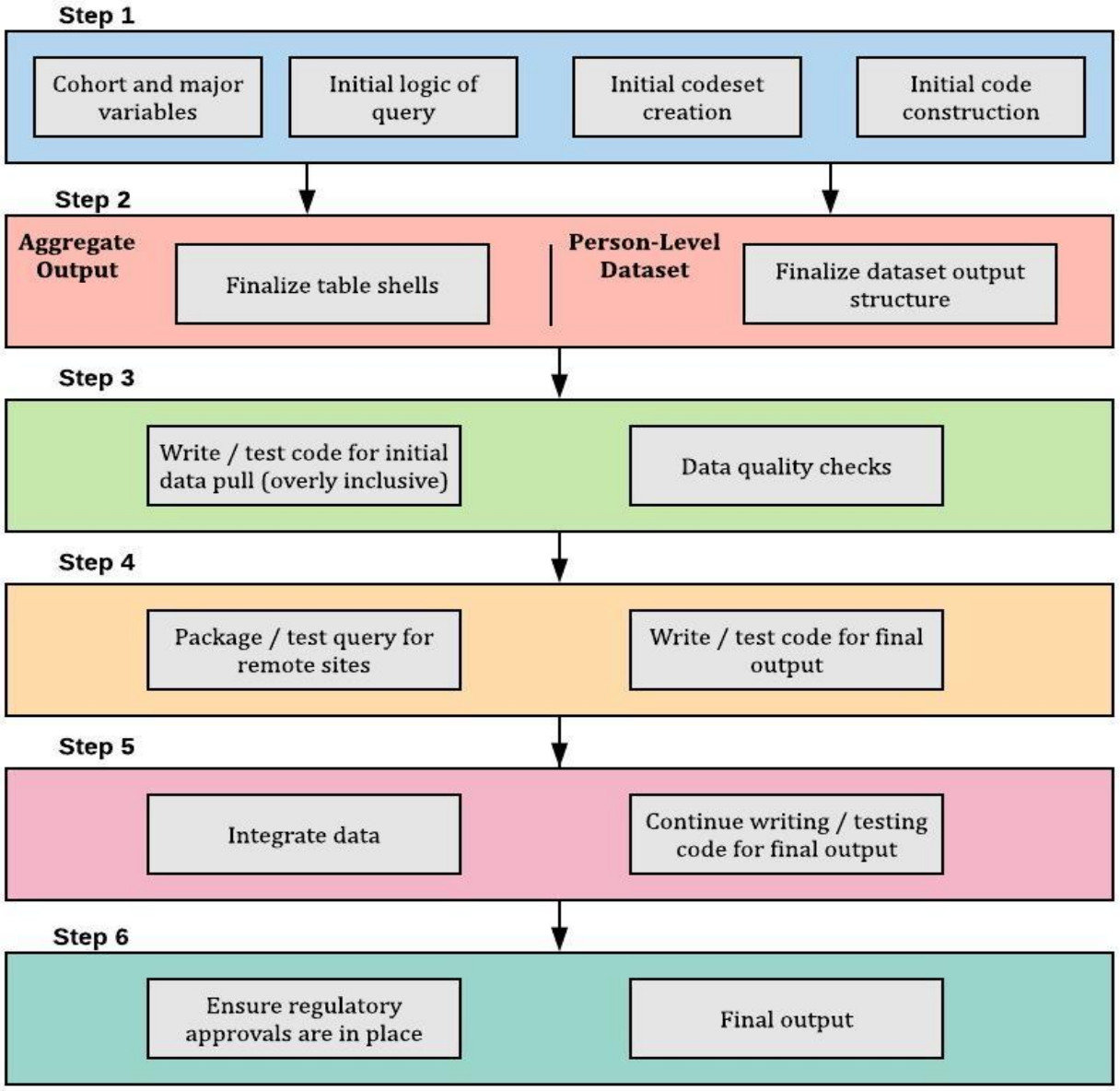
"Data Governance Grid"

A Pediatric Learning Health System

Limited Data Set with Actual Dates	Human Subjects Review	IRB Determination/Approval required by PEDSnet site -If NHSR/Exempt: no further review -If HSR: IRB approval with IRB reliance for sites providing data (NPRA MRA or SMART IRB MRA)	IRB Determination/Approval required by PEDSnet site -If NHSR/Exempt: no further review -If HSR: IRB approval with IRB reliance for sites providing data (NPRA MRA or SMART IRB MRA)	IRB Determination/Approval required by PEDSnet site -If NHSR/Exempt: no further review -If HSR: IRB approval with IRB reliance for sites providing data (NPRA MRA or SMART IRB MRA)
	Legal Review	PEDSnet DUA and RUD	Data Release Agreement [per PEDSnet, not PCORnet SOP] [PCORnet DSA between PCORnet and PEDSnet sites][5]	PEDSnet DUA and RUD
	Network Participation Approval	EC Approval	EC Approval	EC Approval
	Institutional Participation Approval	Prospective Site PI Approval	Prospective Site PI Approval	Prospective Site PI Approval
Data Set with PHI Direct Identifiers	**Human Subjects Review**	IRB Approval required by PEDSnet site with IRB reliance for sites providing data (PEDSnet MRA or SMART IRB MRA)[7]	IRB Approval required by PEDSnet site with IRB reliance for sites providing data (PEDSnet MRA or SMART IRB MRA)[7]	IRB Approval required by PEDSnet site with IRB reliance for sites providing data (PEDSnet MRA or SMART IRB MRA)[7]
	Legal Review	Consent and Authorization (may or may not also have a written assurance); OR Waiver of authorization with written assurance, BAA (if required per institution), PEDSnet DUA and RUD	Consent and Authorization (may or may not also have a written assurance); OR Waiver of authorization with written assurance; BAA (if required per institution), PCORnet DUA[PCORnet DSA between PCORnet and PEDSnet sites]	Consent and Authorization (may or may not also have a written assurance); OR Waiver of authorization with written assurance, BAA (if required per institution), PEDSnet DUA and RUD
	Network Participation Approval	EC Approval	EC Approval	EC Approval
	Institutional Participation Approval	Prospective Site PI Approval	Prospective Site PI Approval	Prospective Site PI Approval

PEDSnet
A Pediatric Learning Health System

Process



PEDSnet
A Pediatric Learning Health System

Key Questions

Development:

- *What question is the study team trying to answer?*
- *What are the data domains required to answer questions?*
- *How should data model elements be mapped to the table outputs of interest?*

Determine complexity:

- *How many codesets are required? How difficult are the codesets to create?*
- *Are there any derivations required?*
- *How many cohorts need to be created?*
- *What analyses must be performed?*

Process and Output:

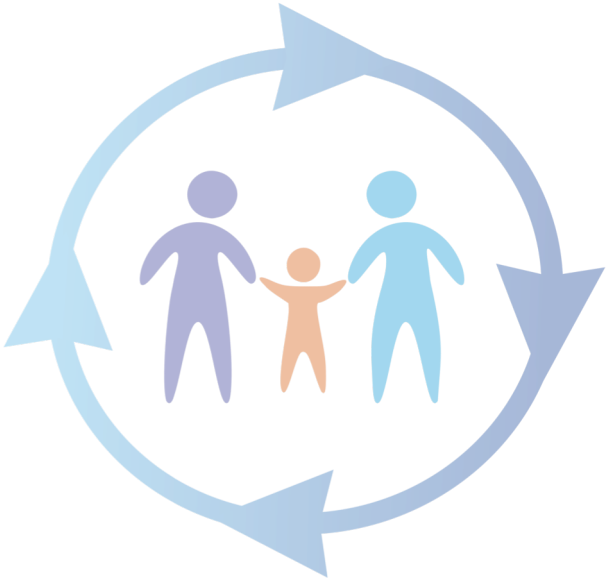
- *How can the cohort of interest be narrowed?*
- *What are the intermediate tables that need to be developed?*
- *How can the data be represented in a way to ease computation of output?*

PEDSnet
A Pediatric Learning Health System

Use Intermediate Datasets!

- Narrows population to cohort of interest
- Always save in study-specific schema; source of truth
- Allows copying to local data frames for creation of tables and figures
- Contains major variables of interest and allows for flexibility in using data
- Can more easily assess data quality of output

Aggregate Output



Common Aggregate Data Output

Table Type	Definition / Example
Table 1	• Demographics • Age/Sex/Race Distributions • Provides high level distribution of question of interest
Attrition Table	• Shows how cohort is narrowed down • Starts with large 'n' and ends with a very specific cohort of interest
Incidence / Prevalence	• Rate of selected conditions or observations of interest
Utilization	• For a selected cohort, the drug / procedure utilization patterns (e.g., median use)
Health System Level	• Computes number of visits, specialty visits, selected utilization • Usually normalized to person-time

Considerations

- Attrition
- Codesets are everything!
- What are the things they want to know
- Small cell suppression

Example: GLEAN manuscript tables

	Full Population N (%) or median (IQR)	**Sub-cohort: nephrotic syndrome, minimal change, FSGS or membranous** N (%) or median (IQR)
N	6657	3315
Age in years at diagnosis		
<2	273 (4%)	217 (7%)
2-4	1233 (19%)	912 (28%)
5-9	2028 (30%)	921 (28%)
10-14	1616 (24%)	679 (20%)
15-19	1360 (20%)	524 (16%)
20+	147 (2%)	62 (2%)
Year of first diagnosis		
<2009	1308 (20%)	712 (21%)
2009-2011	1843 (28%)	937 (28%)
2012-2014	1529 (23%)	738 (22%)
2015-2017	1801 (27%)	842 (25%)
Follow-up time since diagnosis (years)	3.3 (1.10, 6.42)	3.6 (1.33, 6.91)
# of nephrology visits (per person-year)	2.3 (0.5, 7.3)	2.6 (0.5, 6.9)
# of hospitalizations (per person-year)	0.5 (0.1, 1.8)	0.4 (0, 1.6)
CKD stage within ± 30 days of diagnosis		
1	2077 (31%)	1186 (36%)
2	968 (15%)	380 (11%)
3	681 (10%)	248 (7%)
4	304 (5%)	92 (3%)
5	390 (6%)	128 (4%)
CKD stage at last follow-up		
1	1212 (18%)	674 (20%)
2	713 (11%)	332 (10%)
3	392 (6%)	188 (6%)
4	148 (2%)	67 (2%)
5	366 (5%)	189 (6%)

- two cohorts
- mostly a health system follow-up table
- CKD stage is a derived variable
- specialty visits and hospitalizations need clear definitions
- codesets relatively straightforward

PEDSnet
A Pediatric Learning Health System

GLEAN

DEVELOPMENT	Example Considerations
What question is the study team trying to answer?	• How many children meet the computable phenotype for glomerular disease? • What are some basic distributions of drug utilization and health seeking behaviors of the cohort?
What are the data domains required to answer the questions?	• person, condition, visits, care site, provider, measurement
How should data model elements be mapped to the table outputs of interest?	• CKD stage is a derivation of eGFR – need age, creatinine, sex, height (how should nearest height be calculated? what is the limit?) • Need to determine if follow-up should be determined by first and last in-person visit or any interaction with the health system • `nephrology visits` can be either care site or provider specialty and need to figure out if inpatient or OA visits apply

GLEAN

Complexity	Example Considerations
How many codesets are required? How difficult are the codesets to create?	• nephrology visits, hospitalizations, diagnosis codes (flagging subtypes), creatinine, demographics height, visit type
What are the data domains required to answer the questions?	• person, condition, visits, care site, provider, measurement
How many cohorts need to be created?	• two for the total computation – one for the full and one subcohort
What analyses must be performed?	• computing person years • need to understand how to compute follow-up and what to do with right censor

GLEAN

Process and Output	Example Considerations
How can the cohort of interest be narrowed?	• subcohorts should be created from the full GLEAN cohort
What are the intermediate tables that need to be developed?	• table with all the categorical level output • person-level with flags for output that requires computation (e.g., medians, person-level) ****there is no "right" way to do this – could have created it as all person-level*
How can the data be represented in a way to ease computation of output?	• intermediate tables help in pulling the measure we want and then doing the computations • categorical variables are already counts and easy to compute • table shells already provided by GLEAN investigator

PEDSnet
A Pediatric Learning Health System

GLEAN: Intermediate Tables

	category integer	total_pts integer	cohort text	concept_of_interest text	site text
206	2018	75	nephrotic_pts	year_first_diag	pedsnet
207	12	76	nephrotic_pts	age_first_diag	pedsnet
208	16	76	case_pts	age_first_diag	bch
209	4	76	case_pts	age_first_diag	bch
210	2005	77	nephrotic_pts	year_first_diag	pedsnet
211	2006	79	case_pts	year_first_diag	bch

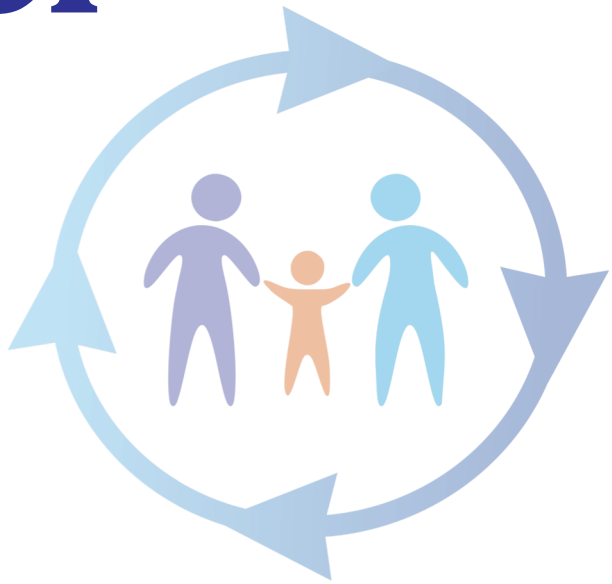
categorical variables

	person_id integer	outcome numeric	cohort text	measure text	site text
1	1	693	case_pts	follow_up_in_days	pedsnet
2	1	1	case_pts	hospitalizations	pedsnet
3	2	20	nephrotic_pts	neph_visits	pedsnet
4	2	1	nephrotic_pts	hospitalizations	pedsnet
5	2	1	case_pts	hospitalizations	pedsnet
6	2	20	case_pts	neph_visits	pedsnet
7	2	1642	nephrotic_pts	follow_up_in_days	pedsnet
8	2	1642	case_pts	follow_up_in_days	pedsnet
9	3	1660	case_pts	follow_up_in_days	pedsnet

person-level output

PEDSnet
A Pediatric Learning Health System

Person-Level Output



Considerations

Key Consideration	Description
Determining complexity	• determined by number of codesets, derivations, number of variables; difficulty in deriving cohort; requirement for computation at DCC (e.g., matching, complexity algorithm)
Engaging researchers	• most researchers are not prepared for level of complexity and component of 'messy data' • ask key questions early
Balancing requirements for analysis with data privacy	• **request for: *give me all the data* not feasible** • **minimum amount of data to complete study while providing researchers sufficient data for sub- and sensitivity analyses** • **no source values**
Determining level of DQ and data investigation	• clinical data inherently 'messy' • level of funding and priority usually determines effort for DQ investigations

PEDSnet
A Pediatric Learning Health System

Determining Output

- one table that usually contains meta-study data (e.g., demographics, major exposure/outcome variables, derivations, etc)
- try to limit derivations
- keep structure of tables similar to PEDSnet data model
- provide a primary key where possible
- logic for deriving query should be documented early

Study Example: Obesity and Asthma

- study that examined relationship between obesity and subsequent development of asthma
- index visit was first visit in system with no asthma diagnosis
- case patients matched to a control patient
- study contained a mix of derived and CDM variables
- provided lookup table to investigators to look up *concept_id's*

PEDSnet
A Pediatric Learning Health System

Study Example Output

Table	Content
cohort table (derived table)	• all patients in cohort • metadata table: • person_id, matched_patient, payer, demographic info, site, age at index visit, date of index visit, BMI at index visit, obesity flag, total visit count, and flag for case or control • PK is person_id
condition_ occurrence and drug_exposure	• only conditions and drugs related to exposure, output, and confounding variables
visit_occurrence	• all visits for patients in cohort
measurement	• all BMI's for patients in cohort
lookup table	• provides a study-specific concept table to crosswalk concept_id to concept_name

ASTHMA STUDY

DEVELOPMENT	Example Considerations
What question is the study team trying to answer?	• What is the association of asthma onset with obesity? • What are the factors that might confound this association?
What are the data domains required to answer the questions?	• person, condition, visits, drugs, visit payer, care site, provider, measurement
How should data model elements be mapped to the table outputs of interest?	• study team wants to determine obesity – need to define obesity from anthropometrics (z-scores) and determine cut-offs • follow-up period was only through 2016 so need to take that into consideration in all data tables • need to define exclusionary diagnoses (e.g., cystic fibrosis)

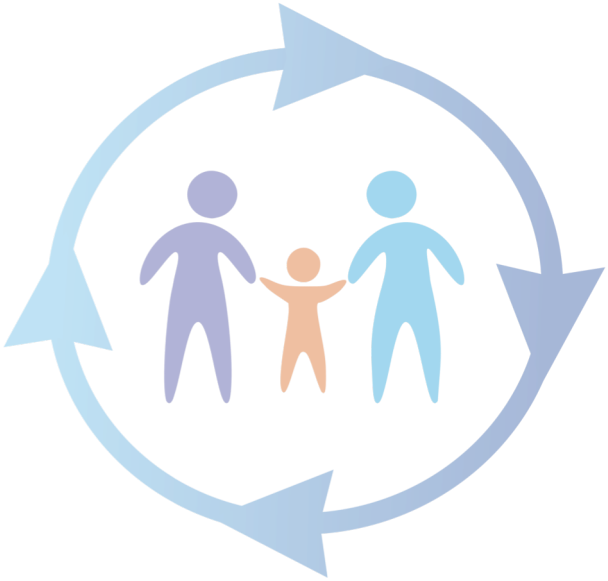
ASTHMA STUDY

Complexity	Example Considerations
How many codesets are required? How difficult are the codesets to create?	• need to balance asking for all patient data (large cohort in this case) with what is necessary to answer the research question • study team needed narrow list of measurements, drugs, and conditions
What are the data domains required to answer the questions?	• person, condition, visits, measurement, visit payer
How many cohorts need to be created?	• this study required cases and controls but not different output for each – just a flag indicating which is which
What analyses must be performed?	• study team will use this as a cohort study; need to define exposures and outcomes very clearly • provide flags for computations – need to negotiate with investigator (we flagged cases vs controls)

ASTHMA OUTPUT

Process and Output	Example Considerations
How can the cohort of interest be narrowed?	• narrow output for cohorts based on specific codesets • exclusionary criteria can be applied to cohort on DCC end – do not need to create overly inclusive cohort
What are the intermediate tables that need to be developed?	• tables that are returned to DCC can contain all domains and the broadeset level codeset (e.g., all steroids instead of just inhaled corticosteroids) • can develop distributions and play around with codesets as needed
How can the data be represented in a way to ease computation of output?	• create metadata table with some outcomes of interest flagged – race/gender, index visit, age at index visit, case/cohort assignment, etc • fact tables limited to facts of interest for cohort

Good Data / Coding Practices



Strategies for Good Coding Practices

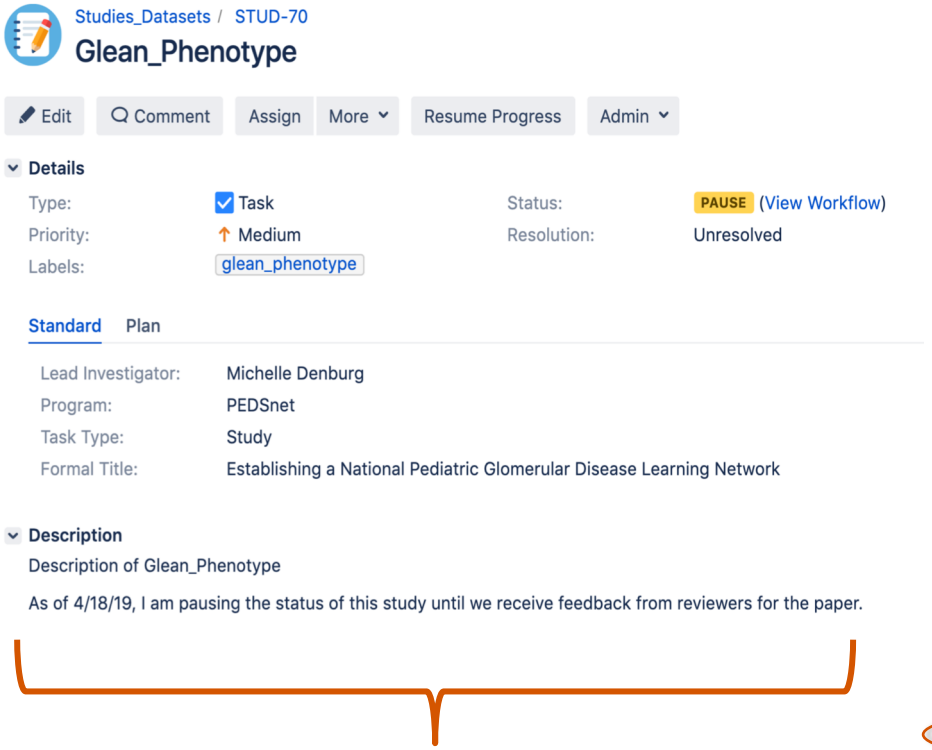
- Task and project management
  - tracks projects
  - one central location for communication about studies
  - reporting to different entities
- Version control
  - track as code evolves – commit often!
  - facilitate collaboration and code review
  - availability of previous versions and backup
- Wiki pages
  - general knowledge management
  - record for meetings
  - information sharing

Task management (Jira)

- Track individual assignments
- Background information on study all in one place
- Collaborate through comments
- Organizes tasks and subtasks
- Holds attachments
- Links to development branch

Task Management: JIRA

TASK (Parent) LEVEL

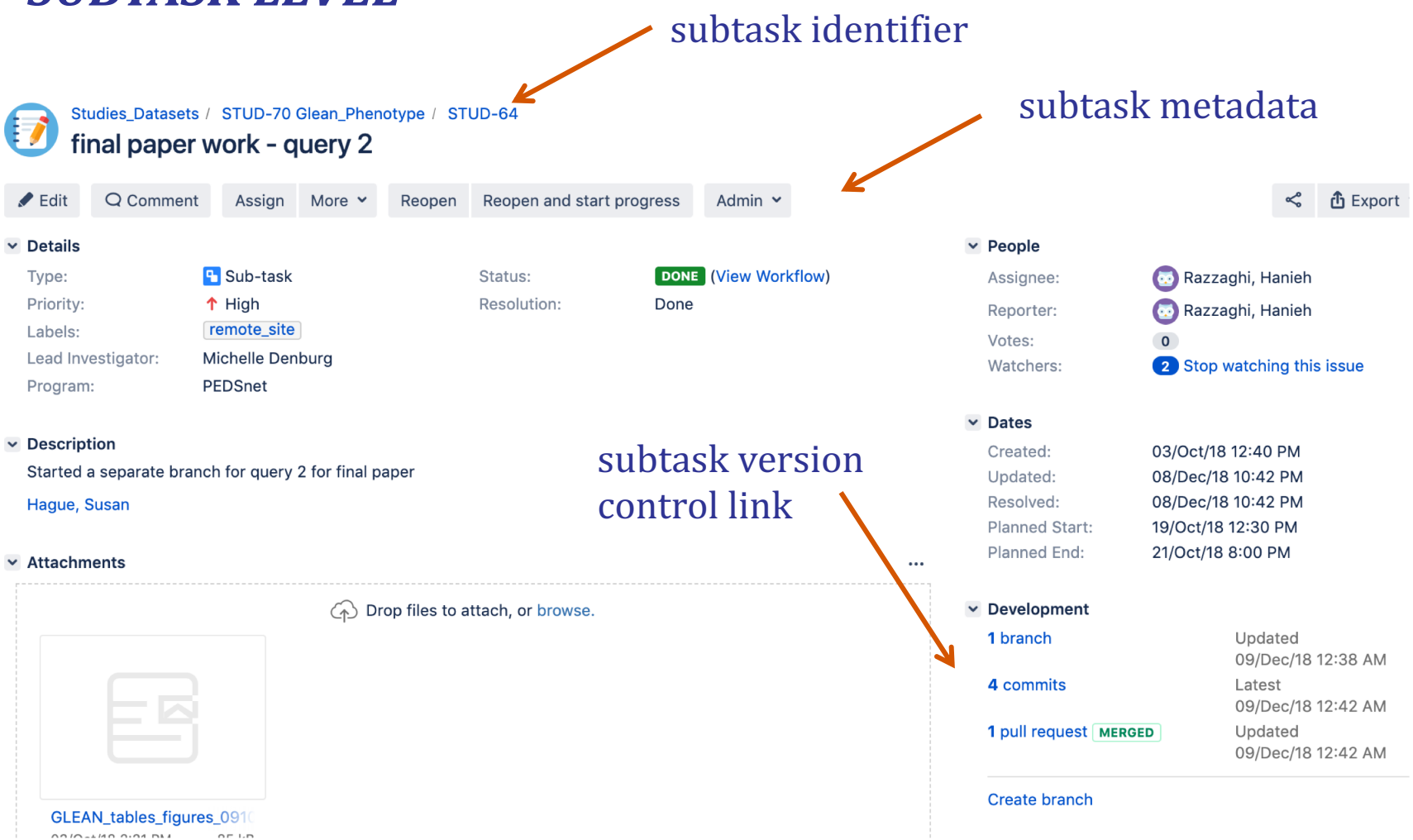


Task Metadata

Subtask List

Task Management: JIRA

SUBTASK LEVEL



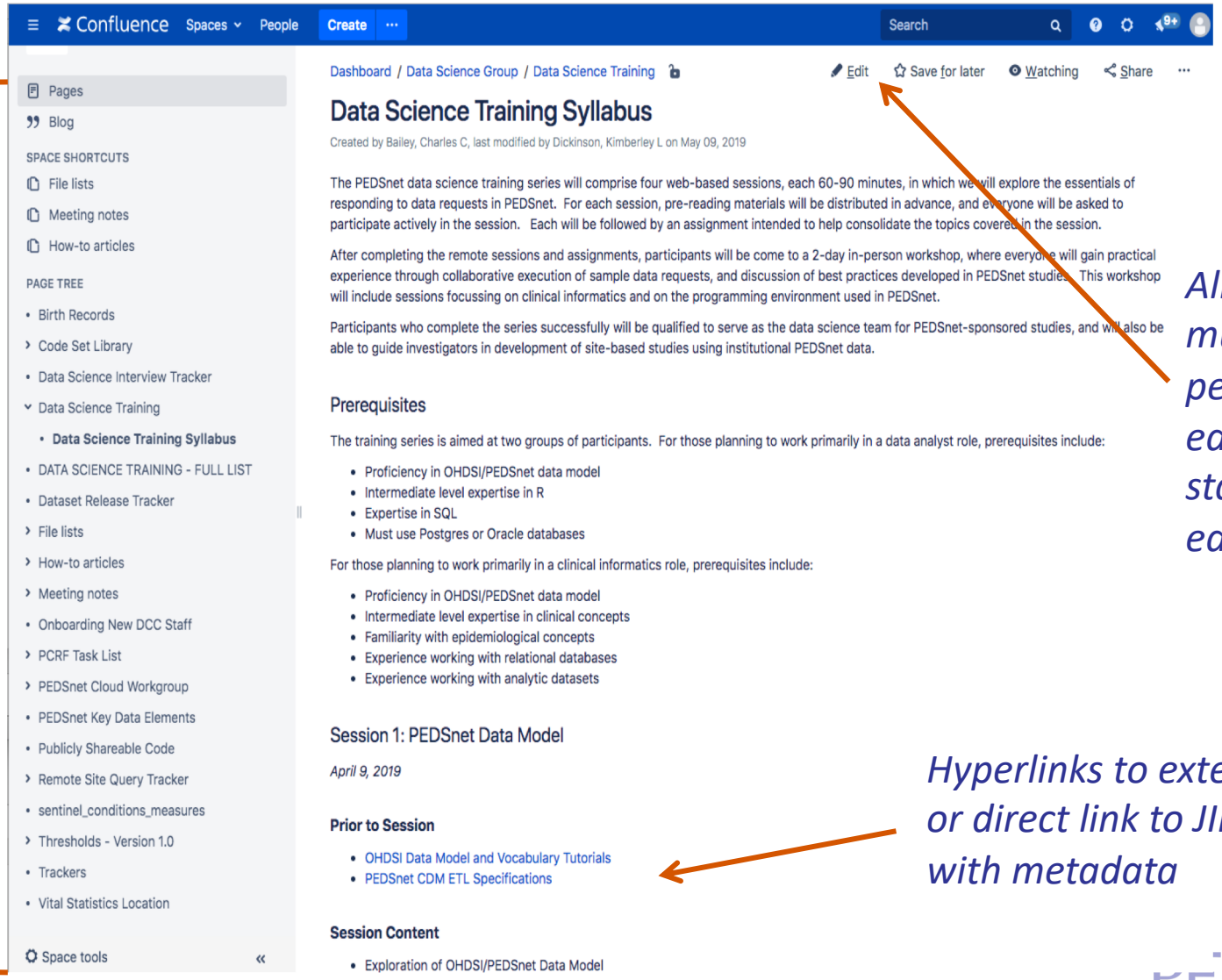
PEDSnet
A Pediatric Learning Health System

Wiki for documentation (Confluence)

- Documentation of processes
- Recording meeting minutes
- Sharing to-do lists
- Creating shared space for best practices
- Central location for documentation and shared communication for organizational management

Confluence

Exports to PDF and Word



Menu of Pages: Includes meeting minutes, interview trackers, coding practices, etc

Allows for multiple people to edit and status of each edit

Hyperlinks to external pages or direct link to JIRA tickets with metadata

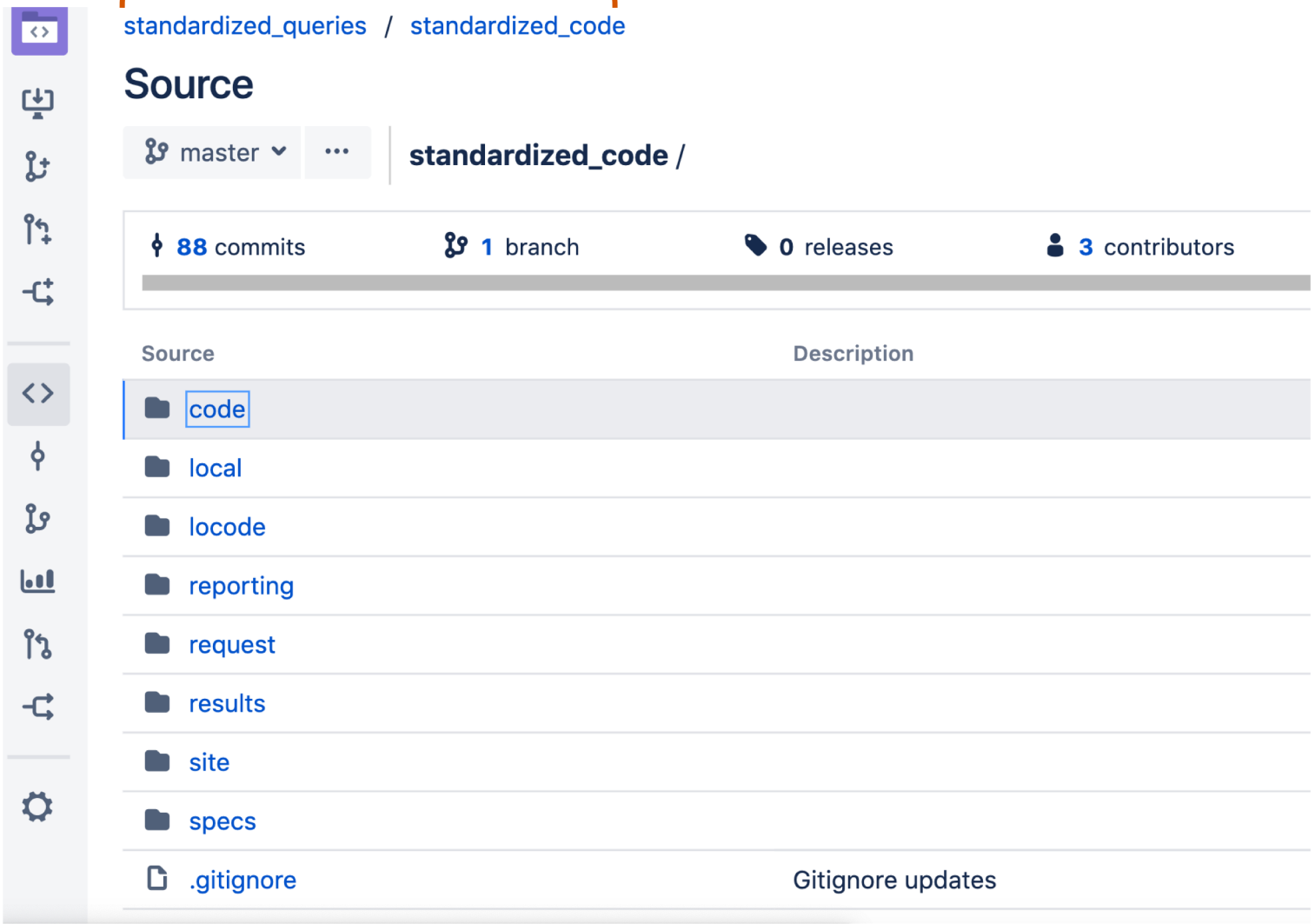
PEDSnet
A Pediatric Learning Health System

Version Control (BitBucket & GitHub)

- Central place to store code
- Eases collaboration
- Branching for different tasks
- Easier to track bugs and document fixed

- BitBucket for research projects
- GitHub for 'publically' shareable code

BitBucket



Clone

Create Branch

Create PR

Create Fork

Source Code

Commits

Branches

Graphs

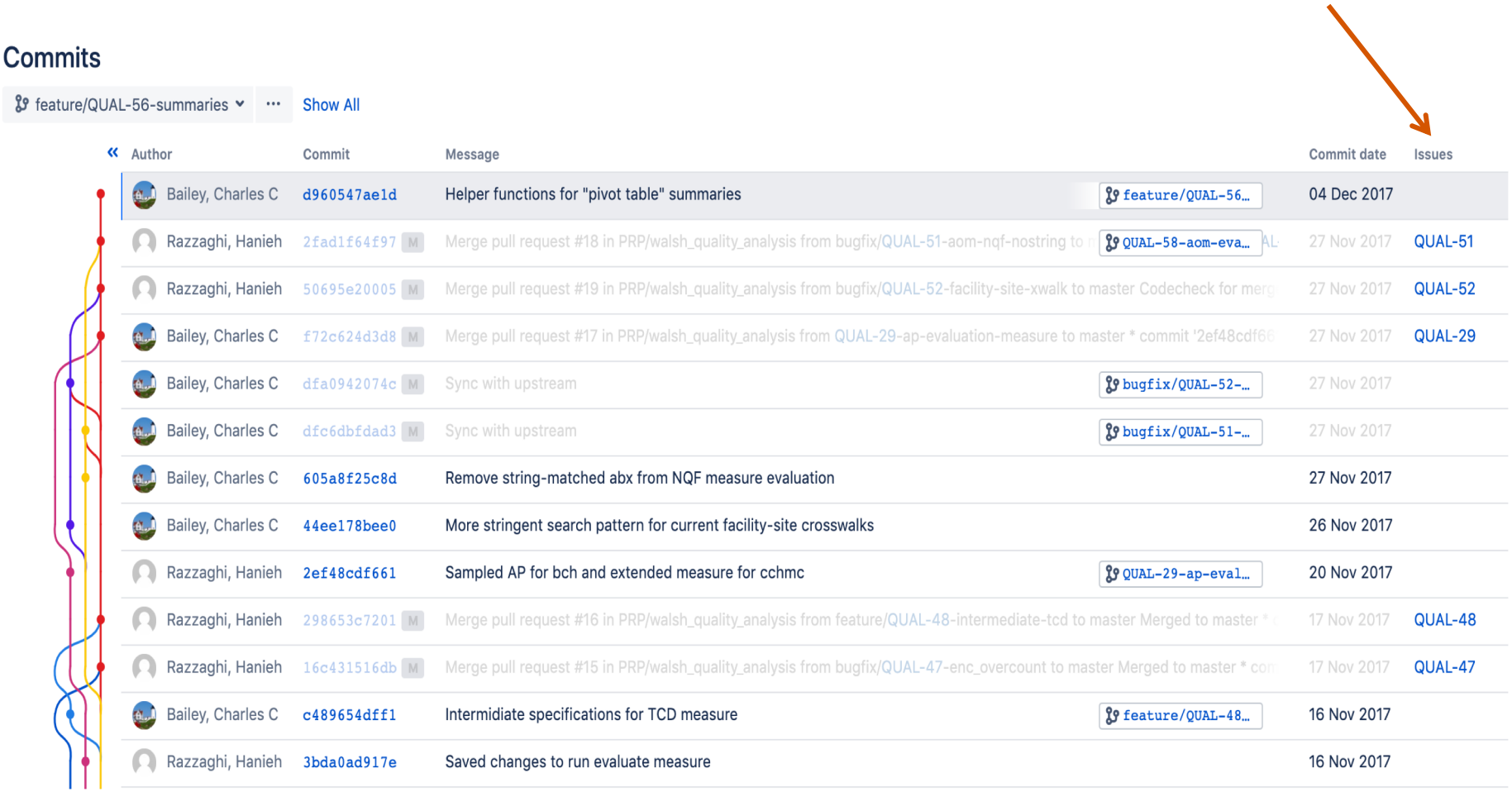
Pull Requests

Forks

PEDSnet
A Pediatric Learning Health System

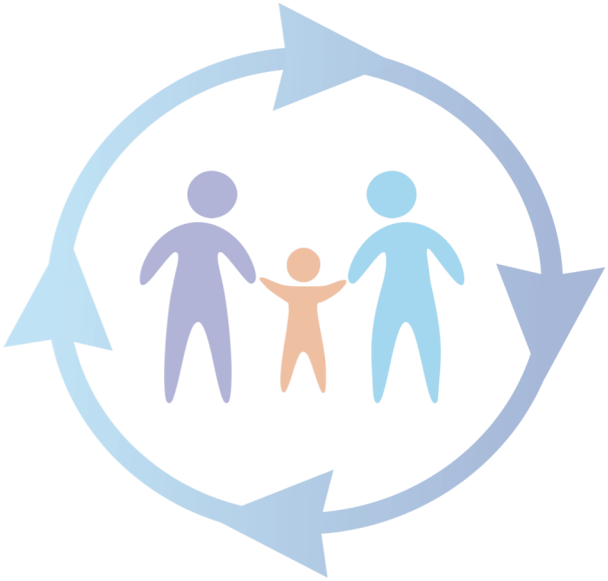
BitBucket

Links back to JIRA issue



PEDSnet
A Pediatric Learning Health System

Data Privacy: Risk Reduction Standards



Reducing Identity Risk

Direct reidentification

- Matching data to widely available information that identifies an individual
- Risk minimization primarily through network architecture

Indirect reidentification

- Matching unique pattern in data for an individual to the same pattern in another data source (*e.g.* claims, local EHR, different study)
- Risk minimization primarily during dataset construction

Regulatory guideposts

Common Rule

- Standard is readily identifiable
- Safe data handling may balance identity risk (for secondary use only)
- PEDSnet generally operates as NHSR, exempt/expedited

HIPAA Privacy Rule

- Standard is bright line list of identifiers
- Level of PHI determines administrative protections
- PEDSnet generally operates using limited datasets

PEDSnet
A Pediatric Learning Health System

Standard Risk Reduction I

Embedded identities

- Avoid disclosure of source values
- Review and redact when needed

Institutional risk

- One-time labels for sites, clinics, providers, etc.

PEDSnet
A Pediatric Learning Health System

Practices – Aggregate Data

- Small cell suppression/aggregation
- Masked names
  - Anonymous labels
  - Substitute labels (e.g. specialty)
- Avoid intersections
- Responsible Use of Data (RUD) agreement

PEDSnet
A Pediatric Learning Health System

Standard Risk Reduction II

Date anchors	• One-year shift window • Age transformation
Dataset linkage	• One-time IDs
Sensitive Data	• HIV status • Pregnancy • Mental health

PEDSnet
A Pediatric Learning Health System

Practices – Individual-level Data

Cohort formation	• Eligibility criteria • Sensitivity analyses
Minimum Data	• Data element selection • Data element structure • Elimination of sensitive data
Limited Distribution	• Study team • RUD

Lessidentification

PEDSnet
A Pediatric Learning Health System

PEDSnet::Lessidentify

Intervention	Purpose
Single-use IDs	Not linkable across datasets
Date shifts/age transform	Reduces risk due to anchor events
Value replacement	Avoids transferring sensitive values while allowing recognition of similarity
Value redaction	Removal of data that aren't needed and cause risk
Noise introduction	Perturbing values (esp. dates) reduces risk due to intervals

Lessidentify Operation

Before Transform

person_id	visit_occurrence_id	birth_date	visit_start_date	site
1858690	13544482	1/10/14	9/9/14	nemours
1858690	22942467	1/10/14	3/12/15	nemours
3652592	107562095	10/17/06	7/10/12	chop
3652592	138972307	10/17/06	7/03/12	chop
3652592	112806972	10/17/06	→ 5/8/15	chop

3125 days

IDs replaced but consistent (e.g., 1869660 = 8026, with two unique visits)

Dates replaced, but ***intervals are preserved and assigned uniquely***

Site names replaced

After Transform

person_id	visit_occurrence_id	birth_date	visit_start_date	site
8044	50233	12/29/13	8/26/14	site_63359
8044	49589	12/29/13	1/27/15	site_63359
20297	51683	10/05/06	6/28/12	site_62489
20297	51468	10/05/06	6/22/12	site_62489
20297	51439	10/05/06	→ 4/25/15	site_62489

3125 days

PEDSnet
A Pediatric Learning Health System

Configuration

Date window	Columns to preserve, replace, and drop	Mappings saved and can be rerun across multiple tables
• Anchor events can be limited to defined window	• Defaults for common data models • Aliasing of columns with connected values	• Retained for future reidentification or linking

PEDSnet
A Pediatric Learning Health System

Example config file

```
{
    "cdm_type": "PCORnet",
    "output_tag": "__scrubbed",
    "load_maps_from": "ap_maps.json",
    "save_maps_to": "ap_maps_v2.json",
    "window_days": "366",
    "preserve": "patid",
    "min_date": "2015-01-01",
    "max_date": "2015-12-31",
    "date_threshold_action": "retry",
    "preserve": "qr/^((?!person_id|visit_occurrence_id|provider_id).)*_id/"
}
```