Welcome to the Sentinel Innovation Center Webinar Series

The webinar will begin momentarily

- Please visit www.sentinelinitiative.org for recordings of past sessions and details on upcoming webinars.

- Note: closed-captioning for today's webinar will be available on the recording posted at the link above.

Sentinel

Data Curation in PCORnet®: Lessons Learned and Implications for Regulatory Decision-Making

pcornet®
The National Patient-Centered Clinical Research Network

Keith Marsolo, PhD

Associate Professor

Department of Population Health Sciences

Duke Clinical Research Institute

Duke University School of Medicine

Disclosures

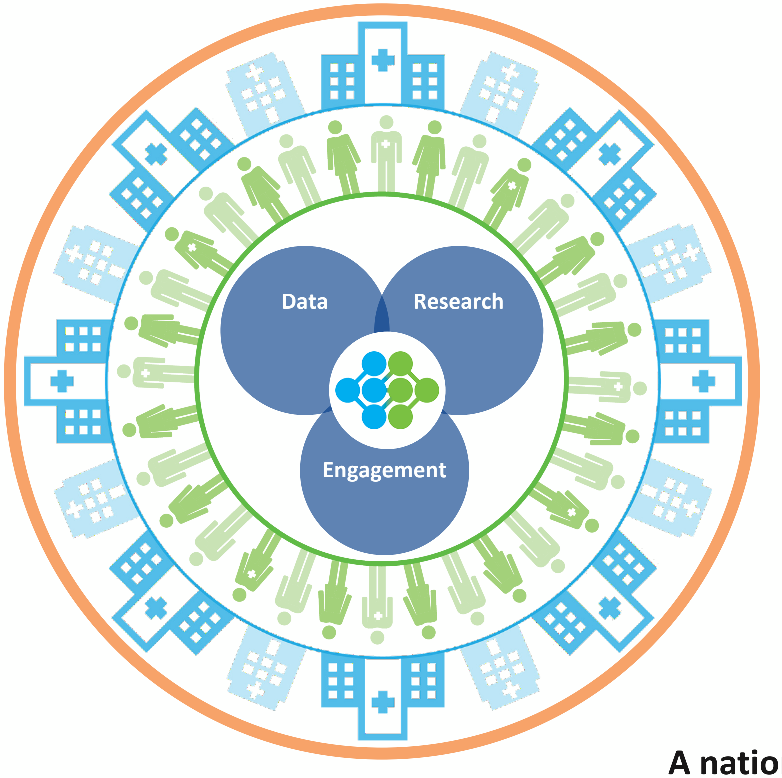
- Consulting support from Novartis
- Investigator on research contracts from Amgen & Bayer
- Co-inventor – Hive Networks, Inc.

- Duke University is part of the Coordinating Center for PCORnet®, the National Patient-Centered Research Network. PCORnet® has been developed with funding from the Patient-Centered Outcomes Research Institute® (PCORI®). Duke University's participation in PCORnet® is funded through PCORI® Award (CC2-Duke-2016).
- The statements presented in this work are solely the responsibility of the author(s) and do not necessarily represent the views of other organizations participating in, collaborating with, or funding PCORnet® or of the Patient-Centered Outcomes Research Institute® (PCORI®).

Goals

- Describe current practices and lessons learned from efforts to assess data quality and dataset suitability within the National Patient-Centered Clinical Research Network (PCORnet®)
- Discuss implications for the use of EHR data more broadly to support regulatory decision-making

PCORnet is a "network of networks" that harnesses the power of partnerships

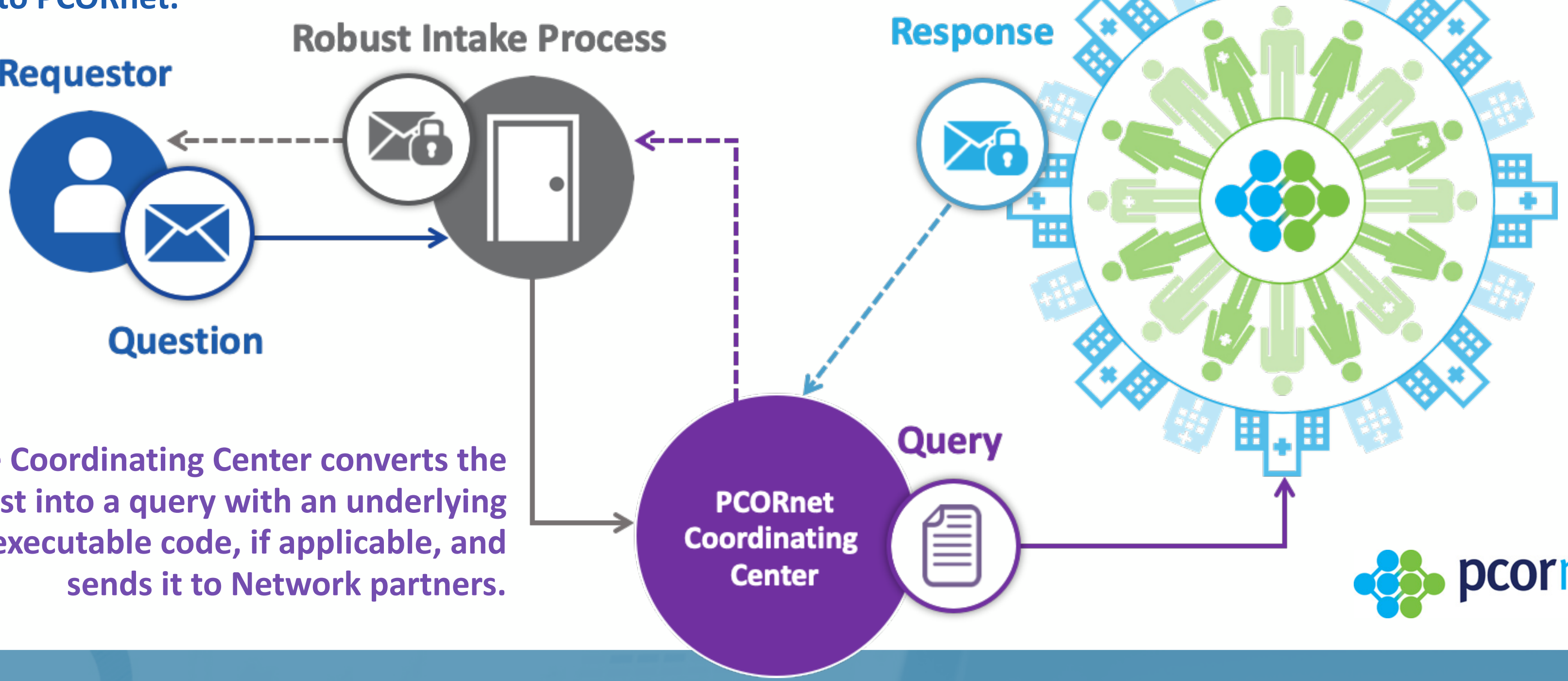


Clinical Research Networks (CRNs) + Health Plan Research Networks (HPRNs) + Patient Partners + Coordinating Center = A national infrastructure for people-centered clinical research

pcornet®

A secure infrastructure to make real-world data accessible

PCORnet was developed with a secure and streamlined infrastructure that offers researchers a simple process for querying the accessible data and deriving efficient insights.



The Requestor sends a question to PCORnet.

PCORnet Leadership reviews the question and consults with Requestor about next steps.

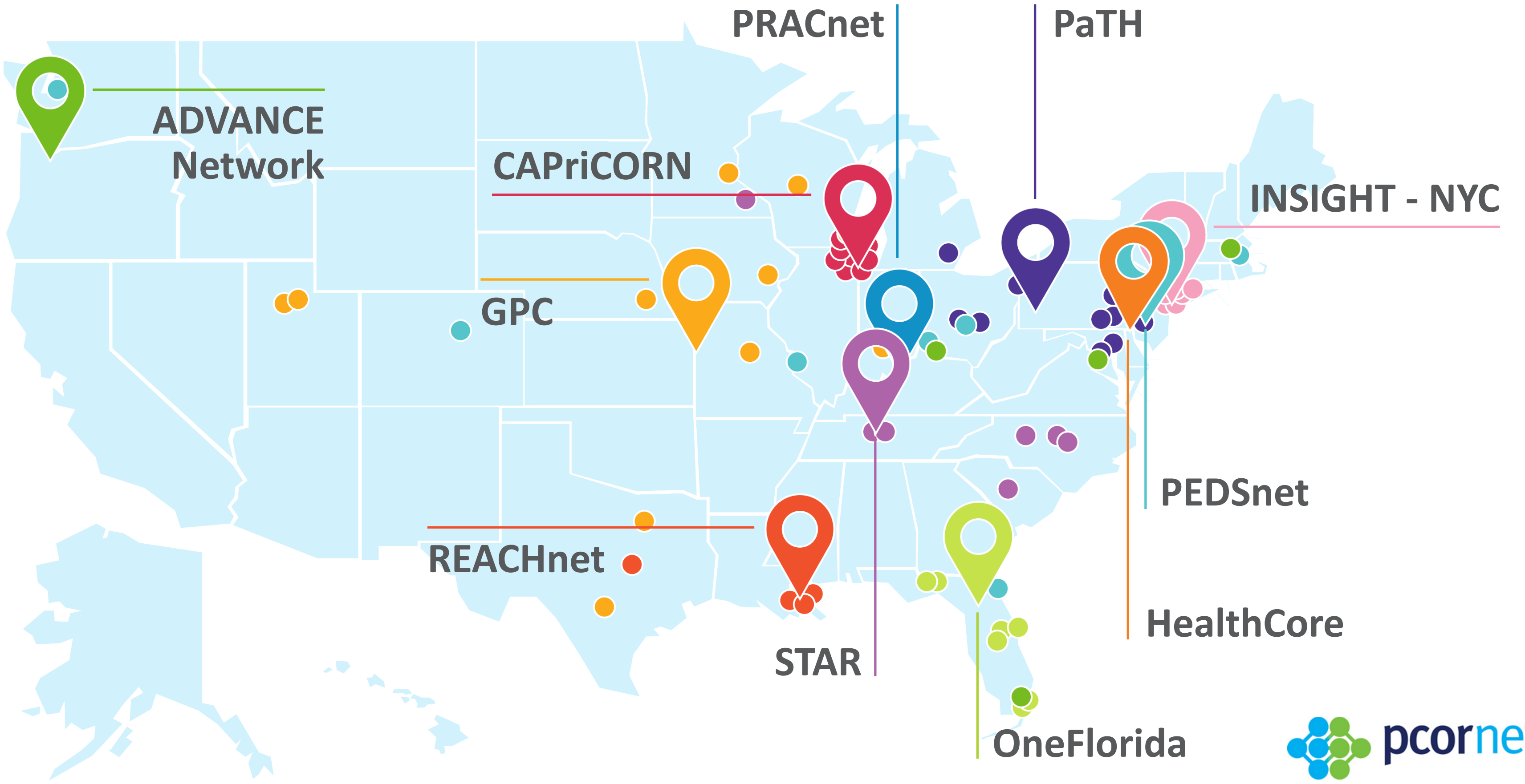
Network partners review the query and provide a response, which is sent back through the Coordinating Center and to the Requestor.

The Coordinating Center converts the request into a query with an underlying executable code, if applicable, and sends it to Network partners.

pcornet®

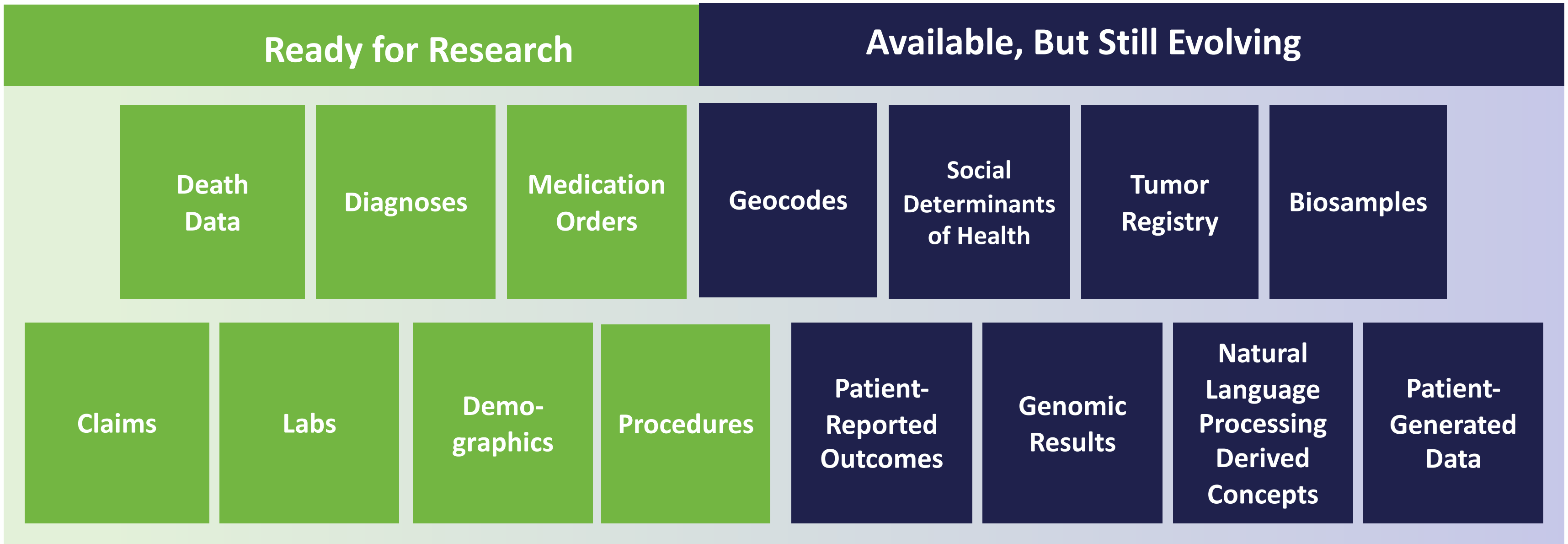
PCORnet CRNs & HPRNs

The PCORnet solution starts with real-world data. PCORnet-partnered CRNs and HPRNs can help users conduct research more efficiently. Users can access data from everyday medical encounters from more than 66 million people across the United States.



pcornet

Domains within the PCORnet Common Data Model



Ready for Research

Death Data

Diagnoses

Medication Orders

Claims

Labs

Demo-graphics

Procedures

Available, But Still Evolving

Geocodes

Social Determinants of Health

Tumor Registry

Biosamples

Patient-Reported Outcomes

Genomic Results

Natural Language Processing Derived Concepts

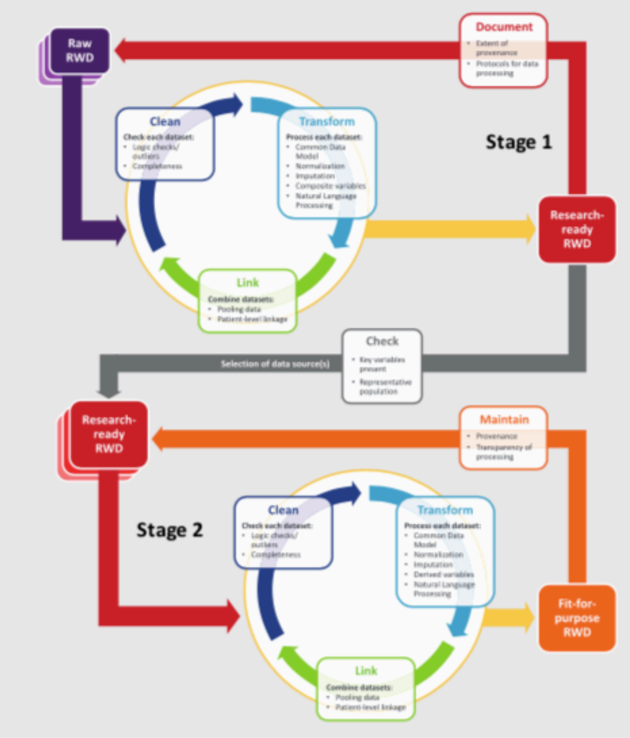
Patient-Generated Data

Data available from several Clinical Research Networks, in the PCORnet Common Data Model and ready for use in research.

Data available at some Clinical Research Networks, may or may not be in the PCORnet Common Data Model and require additional work for use in research.

Moving from raw data to fit-for-purpose

- PCORnet follows a two-stage process to assess suitability
  - **Foundational** curation – establish a baseline level of data quality
  - **Study-specific** – ensure data are fit-for-purpose for a given study or analysis
- Foundational data curation is not static – view as a **continuous learning cycle**
  - Continuous assessment of performance
  - Close gap between foundational and study-specific – add new data checks based on study findings



pcornet®

https://healthpolicy.duke.edu/sites/default/files/atoms/files/characterizing_rwd.pdf

FDA definition of fit-for-purpose

- In order to determine the suitability of RWD for regulatory decision-making, **FDA will assess the relevance and reliability of the source and its specific elements**. This assessment will be used to determine whether the RWD source(s) and the proposed analysis can generate evidence that is sufficiently robust to be used for a given regulatory purpose.

Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices
https://www.fda.gov/media/99447/download

Relevance

- The RWD contain sufficient detail to capture the use of the device, exposures, and the outcomes of interest in the appropriate population (i.e. **the data apply to the question at hand**);
- The data elements available for analysis are capable of addressing the specified question when valid and appropriate analytical methods are applied (i.e. **the data are amenable to sound clinical and statistical analysis**); and
- The RWD and RWE they provide are **interpretable using informed clinical/scientific judgment**

Reliability

- Data accrual
  - Relates to how the data are collected (e.g., operational manual, data element definitions, methods of aggregation, etc.)
- Data assurance
  - Quality control standards to ensure data and analyses are reliable and trustworthy (e.g., registry best practices)
- RWD sources are not necessarily expected to fulfill all characteristics of reliability

How does the PCORnet data curation process relate to the FDA definition?

- Relevance
- Reliability – data accrual
- Reliability – data assurance

How does the PCORnet data curation process relate to the FDA definition?

- Relevance
- Reliability – data accrual
- Reliability – data assurance

Foundational curation is mostly focused here (with some aspects of accrual & relevance)

How does the PCORnet data curation process relate to the FDA definition?

- Relevance

Study-specific characterization is targeted here

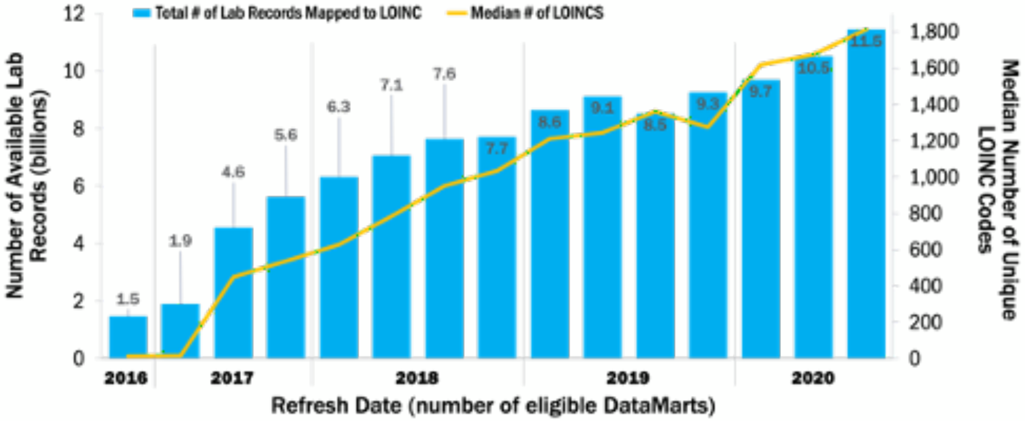
- Reliability – data accrual

- Reliability – data assurance

Foundational curation is mostly focused here (with some aspects of accrual & relevance)

Why is foundational curation focused more on data assurance?

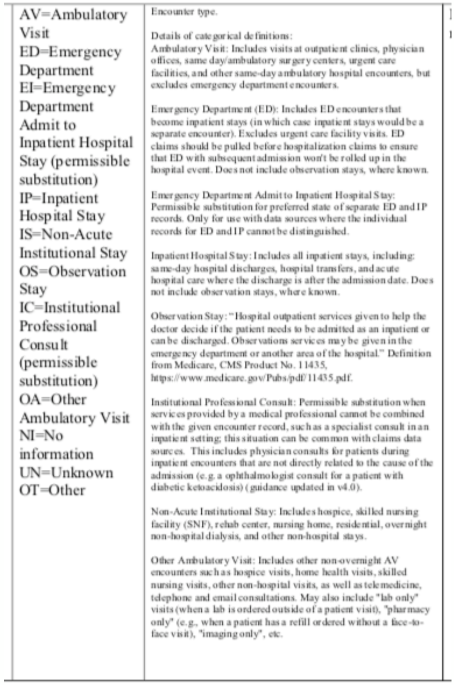
- Many EHR domains are being harmonized / standardized for the first time
- Given volume of data, it is overwhelming to both harmonize and assess fitness for specific study questions / populations at the same time



Eligible DataMarts: PCORnet 2.0 DataMarts that include EHR data and populate the LAB_RESULT_CM table and were approved prior to August 3, 2020. DataMart Refreshes: The refreshes displayed here are the first and third refreshes in previous cycles and every refresh in the current cycle. Other Notes: Each column indicates the number of available laboratory results across the network, in billions. The line shows the median number of unique LOINC codes within a DataMart. We see an increase from a median of 14 LOINC codes in Nov 2016 to well over 1,600 codes in March 2020.

Harmonization examples - Encounter type

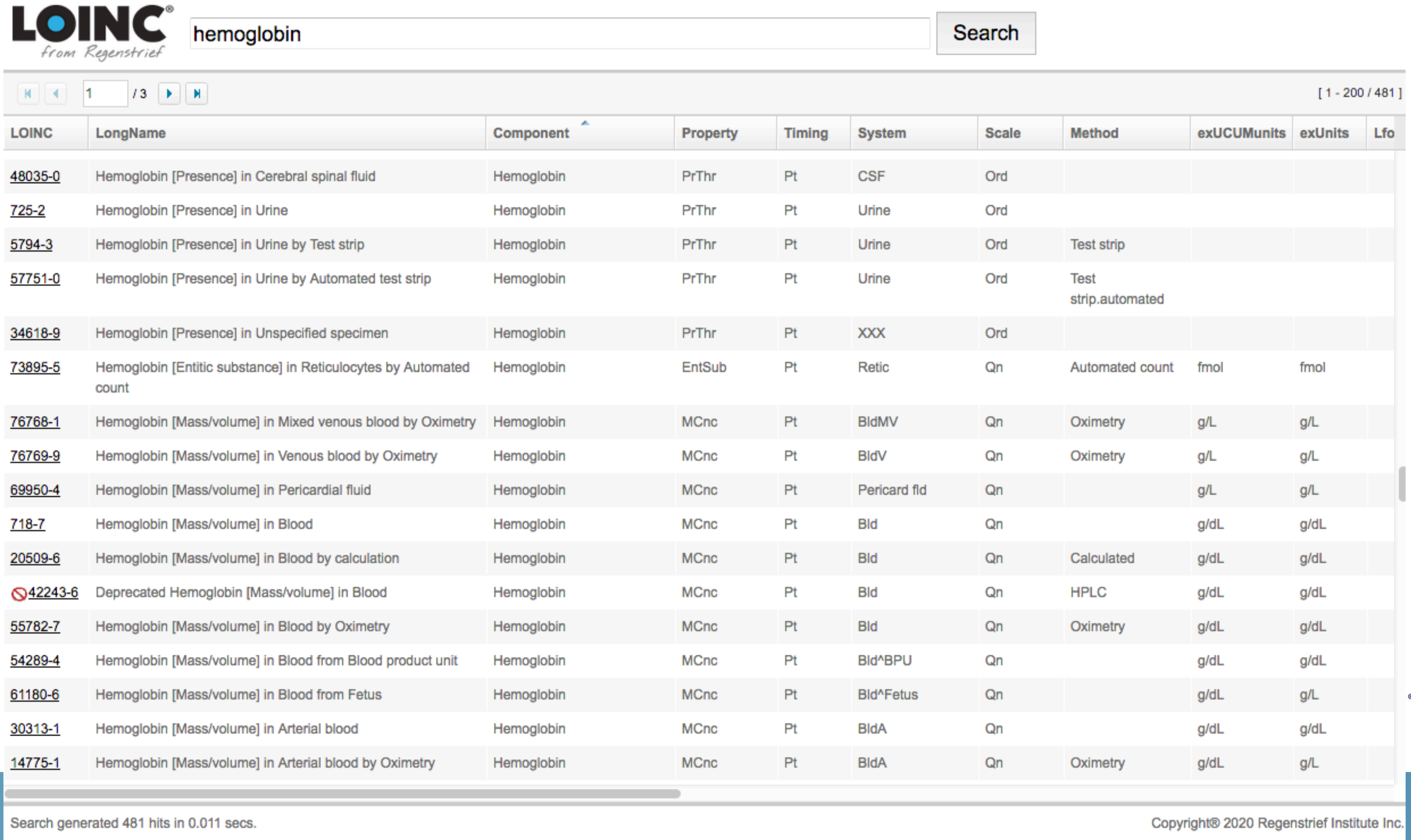
- REGISTRATION
- EMPTY
- LAB REQUISITION
- INITIAL CONSULT
- ANTI-COAG VISIT
- PROCEDURE VISIT
- OFFICE VISIT
- CONSENT FORM
- SCREENING FORM
- EXTERNAL HOSPITAL ADMISSION
- LETTER (OUT)
- REFILL
- IMMUNIZATION
- HISTORY
- RESEARCH ENCOUNTER
- REFERRAL
- ORDERS ONLY
- RX REFILL AUTHORIZE
- MEDS ONLY (WEB)
- MEDS VOID (WEB)
- RESOLUTE PROFESSIONAL BILLING HOSPITAL PROF FEE
- EPISODE CHANGES
- ANCILLARY ORDERS
- PHARMACY VISIT
- BPA
- ROUTINE PRENATAL
- INITIAL PRENATAL
- OPHTH OFFICE VISIT
- ABSTRACT
- WALK-IN
- TREATMENT PLAN
- ALLIED HEALTH
- NURSE ONLY
- SOCIAL WORK
- NUTRITION
- PHYSICAL THERAPY
- OCCUPATIONAL THERAPY
- SPEECH THERAPY
- ROADMAP
- CASE MANAGEMENT
- EDUCATION
- SURGICAL H&P
- CLINICAL SUPPORT
- MEDS ONLY / E - PRESCRIBE
- PFT ONLY
- TRANSPLANT PRE-EVALUATION
- TRANSPLANT EVALUATION
- TRANSPLANT FOLLOW-UP
- TRANSPLANT RESULTS ENTRY
- IMMUNOTHERAPY
- ALLERGY TESTING
- SPECIMEN COLLECTION
- AUTO RELEASE ORDERS
- URODYNAMIC TESTING
- PRE-NATAL
- CONSULT CHECKLIST
- BOWEL MANAGEMENT
- CARE CONFERENCE
- INTAKE/TRIAGE
- VNS REPROGRAM/SHUTOFF
- CLINICAL NOTE
- GENETICS
- PASTORAL
- THERAPY VISIT
- INTAKE - NEW PATIENT
- HIM SCANS
- PRE-VISIT PLANNING
- TRANSCRIBED ORDERS
- SCHOOL TEACHER/INTERVENTION
- CHILD LIFE
- THERAPY PROGRESS SUMMARY
- BRONCHOSCOPY REQUEST
- HEMONC SOCIAL WORK
- AUD CONSULT
- OPH CONSULT
- ALG CONSULT
- UROLOGY COMPLEX INTAKE
- RESPIRATORY THERAPY
- HOSPITAL ENCOUNTER
- UPDATE
- PCP/CLINIC CHANGE
- WAIT LIST
- CLERICAL ORDERS
- MOTHER BABY LINK
- LACTATION ENCOUNTER
- CANCELED
- APPOINTMENT
- SURGERY
- ANESTHESIA
- ANESTHESIA EVENT
- UNMERGE
- HEALTH MAINTENANCE LETTER
- PATIENT EMAIL
- E-VISIT
- MOBILE ORDER ONLY
- QUESTIONNAIRE SERIES SUBMISSION
- PATIENT OUTREACH
- CONTACT MOVED
- NURSE TRIAGE
- E-CONSULT
- E-CONSULT COMMUNITY ORDER
- TELEMEDICINE
- EXTERNAL CONTACT
- OPHTH EXAM
- HOSPICE ADMISSION
- HOME HEALTH ADMISSION
- HOME CARE VISIT
- HOME CARE UPDATE
- PATIENT WEB UPDATE
- COMMUNITY ORDERS
- COMMITTEE REVIEW
- POST MORTEM DOCUMENTATION
- BILLING ENCOUNTER
- HOSPITAL
- CONFIDENTIAL
- OPH TESTING
- EDUCATOR
- VOICE CLINIC
- TELEPHONE
- EEG
- EXERCISE
- CARDIOLOGY TESTING
- PUMP/CGM INITIATION ORDERS
- MED TAPER SCHEDULE
- GENETIC COUNSELOR
- NEONATOLOGY TESTING
- CARE CONFERENCE - PATIENT/FAMILY PRESENT
- HOME VISIT - PALLIATIVE CARE
- ABUSE REPORTING
- CARE COORDINATOR
- SPECIAL NEEDS SUMMARY
- EARLY INTERVENTION
- HI NEURODEVELOPMENTAL CLINIC TRACKING
- INFUSION ORDERS
- ENT CLINIC VISITS
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES CLINIC
- VPI - ENT/SPEECH
- INTAKE
- HVMC PLANNING
- PRE-OP PHYSICAL
- PLAN OF CARE
- ENT INPATIENT VISIT
- HOSPITAL TO HOSPITAL TRANSFER
- DEVELOPMENTAL TESTING
- BIOETHICS CONSULT
- ENDO STIM TESTING
- HIM INTERFACE CREATED
- SURGICAL SITE INFECTION
- DERM PATCH TESTING
- INTAKE CONSULT
- ADEC INTAKE
- CPST-PSY ENCOUNTER
- ECONSULT TELEMEDICINE



AV=Ambulatory Visit
ED=Emergency Department
EI=Emergency Department Admit to Inpatient Hospital Stay (permissible substitution)
IP=Inpatient Hospital Stay
IS=Non-Acute Institutional Stay
OS=Observation Stay
IC=Institutional Professional Consult (permissible substitution)
OA=Other Ambulatory Visit
NI=No information
UN=Unknown
OT=Other
	Encounter type.
	Details of categorical definitions:
	Ambulatory Visit: Includes visits at outpatient clinics, physician offices, same day/ambulatory surgery centers, urgent care facilities, and other same-day ambulatory hospital encounters, but excludes emergency department encounters.
	Emergency Department (ED): Includes ED encounters that become inpatient stays (in which case inpatient stays would be a separate encounter). Excludes urgent care facility visits. ED claims should be pulled before hospitalization claims to ensure that ED with subsequent admission won't be rolled up in the hospital event. Does not include observation stays, where known.
	Emergency Department Admit to Inpatient Hospital Stay: Permissible substitution for preferred state of separate ED and IP records. Only for use with data sources where the individual records for ED and IP cannot be distinguished.
	Inpatient Hospital Stay: Includes all inpatient stays, including: same-day hospital discharges, hospital transfers, and acute hospital care where the discharge is after the admission date. Does not include observation stays, where known.
	Observation Stay: "Hospital outpatient services given to help the doctor decide if the patient needs to be admitted as an inpatient or can be discharged. Observations services may be given in the emergency department or another area of the hospital." Definition from Medicare, CMS Product No. 11435, https://www.medicare.gov/Pubs/pdf/11435.pdf.
	Institutional Professional Consult: Permissible substitution when services provided by a medical professional cannot be combined with the given encounter record, such as a specialist consult in an inpatient setting; this situation can be common with claims data sources. This includes physician consults for patients during inpatient encounters that are not directly related to the cause of the admission (e.g. a ophthalmologist consult for a patient with diabetic ketoacidosis) (guidance updated in v4.0).
	Non-Acute Institutional Stay: Includes hospice, skilled nursing facility (SNF), rehab center, nursing home, residential, overnight non-hospital dialysis, and other non-hospital stays.
	Other Ambulatory Visit: Includes other non-overnight AV encounters such as hospice visits, home health visits, skilled nursing visits, other non-hospital visits, as well as telemedicine, telephone and email consultations. May also include "lab only" visits (when a lab is ordered outside of a patient visit), "pharmacy only" (e.g., when a patient has a refill ordered without a face-to-face visit), "imaging only", etc.

pcornet®

Harmonization examples - Lab results



LOINC	LongName	Component	Property	Timing	System	Scale	Method	exUCUMunits	exUnits	Lfo
48035-0	Hemoglobin [Presence] in Cerebral spinal fluid	Hemoglobin	PrThr	Pt	CSF	Ord				
725-2	Hemoglobin [Presence] in Urine	Hemoglobin	PrThr	Pt	Urine	Ord				
5794-3	Hemoglobin [Presence] in Urine by Test strip	Hemoglobin	PrThr	Pt	Urine	Ord	Test strip			
57751-0	Hemoglobin [Presence] in Urine by Automated test strip	Hemoglobin	PrThr	Pt	Urine	Ord	Test strip.automated			
34618-9	Hemoglobin [Presence] in Unspecified specimen	Hemoglobin	PrThr	Pt	XXX	Ord				
73895-5	Hemoglobin [Entitic substance] in Reticulocytes by Automated count	Hemoglobin	EntSub	Pt	Retic	Qn	Automated count	fmol	fmol	
76768-1	Hemoglobin [Mass/volume] in Mixed venous blood by Oximetry	Hemoglobin	MCnc	Pt	BldMV	Qn	Oximetry	g/L	g/L	
76769-9	Hemoglobin [Mass/volume] in Venous blood by Oximetry	Hemoglobin	MCnc	Pt	BldV	Qn	Oximetry	g/L	g/L	
69950-4	Hemoglobin [Mass/volume] in Pericardial fluid	Hemoglobin	MCnc	Pt	Pericard fld	Qn		g/L	g/L	
718-7	Hemoglobin [Mass/volume] in Blood	Hemoglobin	MCnc	Pt	Bld	Qn		g/dL	g/dL	
20509-6	Hemoglobin [Mass/volume] in Blood by calculation	Hemoglobin	MCnc	Pt	Bld	Qn	Calculated	g/dL	g/dL	
🚫42243-6	Deprecated Hemoglobin [Mass/volume] in Blood	Hemoglobin	MCnc	Pt	Bld	Qn	HPLC	g/dL	g/dL	
55782-7	Hemoglobin [Mass/volume] in Blood by Oximetry	Hemoglobin	MCnc	Pt	Bld	Qn	Oximetry	g/dL	g/dL	
54289-4	Hemoglobin [Mass/volume] in Blood from Blood product unit	Hemoglobin	MCnc	Pt	Bld^BPU	Qn		g/dL	g/dL	
61180-6	Hemoglobin [Mass/volume] in Blood from Fetus	Hemoglobin	MCnc	Pt	Bld^Fetus	Qn		g/dL	g/L	
30313-1	Hemoglobin [Mass/volume] in Arterial blood	Hemoglobin	MCnc	Pt	BldA	Qn		g/dL	g/dL	
14775-1	Hemoglobin [Mass/volume] in Arterial blood by Oximetry	Hemoglobin	MCnc	Pt	BldA	Qn	Oximetry	g/dL	g/L	

Search generated 481 hits in 0.011 secs.

Copyright® 2020 Regenstrief Institute Inc.

Designing foundational data checks

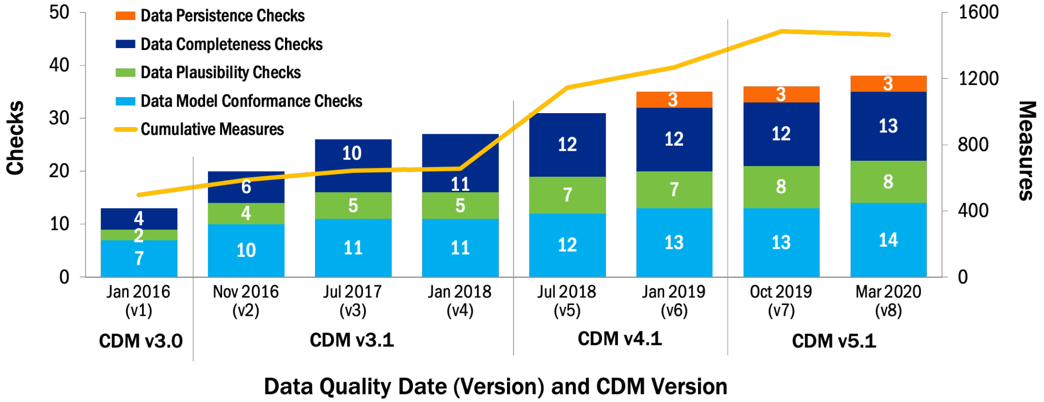
- Do the records conform to the structure/format of the CDM?
- Are records internally consistent (e.g., specimen source is valid for selected LOINC code)?
- If data are to be used in an analysis, are all necessary fields populated?
- Do the values make sense?

- Must keep in mind:
  - Some fraction of the data will always be “dirty” – no errors is usually a problem
  - EHRs change over time – older data (before ~ 2014) are less standardized
  - Need to allow for variation in population / practice patterns

  - Factors can help determine what checks are required, and what are optional

PCORnet foundational data checks

- **Conformance** — Data adhere to the format of the CDM
  - *Fields do not contain values outside of the CDM specification*
- **Completeness** — Values appear where we expect them
  - *Diagnosis codes have an associated diagnosis type (e.g., ICD-9, ICD-10, SNOMED)*
- **Plausibility** — Values that appear make sense
  - *Less than 5% of records are associated with a future date*
- **Persistence** — Patients / records do not disappear between refreshes
  - *Less than a 5% decrease in the number of patients or records in a CDM table between refreshes*



Growth in foundational data quality checks over time. Checks: Rules such as “Values must conform to CDM specifications.” Measures: The number of CDM tables and/or fields affected by the checks. Includes data from PCORnet Data Curation team.

pcornet®

PCORnet data checks - Conformance

Type	Check	Description	Cycle Added
Required	DC 1.01	Required tables not present	1
	DC 1.02	Expected tables not populated	1
	DC 1.03	Required fields not present	1
	DC 1.04	Fields do not conform to CDM specifications	1
	DC 1.05	Tables have primary key definition errors	1
	DC 1.06	Fields contain values outside of CDM spec.	1
	DC 1.07	Fields have non-permissible missing values	1
	DC 1.08	Tables contain orphan PATIDs	1
	DC 1.09	Tables contain orphan ENCOUNTERIDs	2
	DC 1.10	Replication errors between ENCOUNTER, DIAGNOSIS & PROCEDURES	2
	DC 1.11	More than 5% of encounters assigned to 1 patient	3
	DC 1.12	Tables contain orphan PROVIDERIDs	5
	DC 1.13	More than 5% of ICD, CPT, LOINC, RXCUI, or NDC codes do not conform to the expected length or content	6
	DC 1.14	Patients in the DEMOGRAPHIC table are not in the HASH_TOKEN table	8

PCORnet data checks - Plausibility

Type	Check	Description	Cycle Added
Investigative	DC 2.01	More than 5% of records have future dates	1
	DC 2.02	More than 10% of records fall into high/low categories for selected variables	1
	DC 2.03	More than 5% of patients have illogical date relationships	2
	DC 2.04	Average number encounters per visit is > 2.0 for IP, EI, or ED encounters	2
	DC 2.05	More than 5% of lab results have inappropriate specimen source [for selected tests]	3
	DC 2.06	Median lab results are statistical outliers [for selected tests]	5
	DC 2.07	Average number of principal diagnoses per encounter is above threshold (2.0 for IP & EI)	5
	DC2.08	The monthly volume of encounter, diagnosis, procedure, vital, prescribing, or laboratory records is an outlier.	7

PCORnet data checks - Completeness

Type	Check	Description	Cycle Added
Investigative	DC 3.01	Average # of diagnoses with known diagnosis type per encounter is below threshold	1
	DC 3.02	Average # of procedures with known procedure type per encounter is below threshold	1
	DC 3.03	More than 10% of records have missing/unknown values for selected fields	1
Required	DC 3.04	Less than 50% of patients with encounters have DIAGNOSIS records	2
	DC 3.05	Less than 50% of patients with encounters have PROCEDURES records	2
Investigative	DC 3.06	More than 10% of IP & EI encounters with a diagnosis are missing principal diagnosis	2
	DC 3.07	DX, PX, & encounter records in AV, ED, EI, IP setting are <75% complete 3 months prior to current month	3
	DC 3.08	Less than 80% of prescribing orders mapped to a Tier 1 RXCUI (encodes ingredient, strength, & dose form)	3
	DC 3.09	Less than 80% of lab results mapped to LOINC	3
	DC 3.10	Less than 80% of quantitative lab results specify the normal range	3
	DC 3.11	Vital, Rx, Lab records are <75% complete 3 months prior to current month	4
	DC 3.12	Less than 80% of quantitative lab results mapped to LOINC specify SPECIMEN_SOURCE & RESULT_UNIT	5
	DC 3.13	The percentage of patients with selected lab tests is below threshold	8

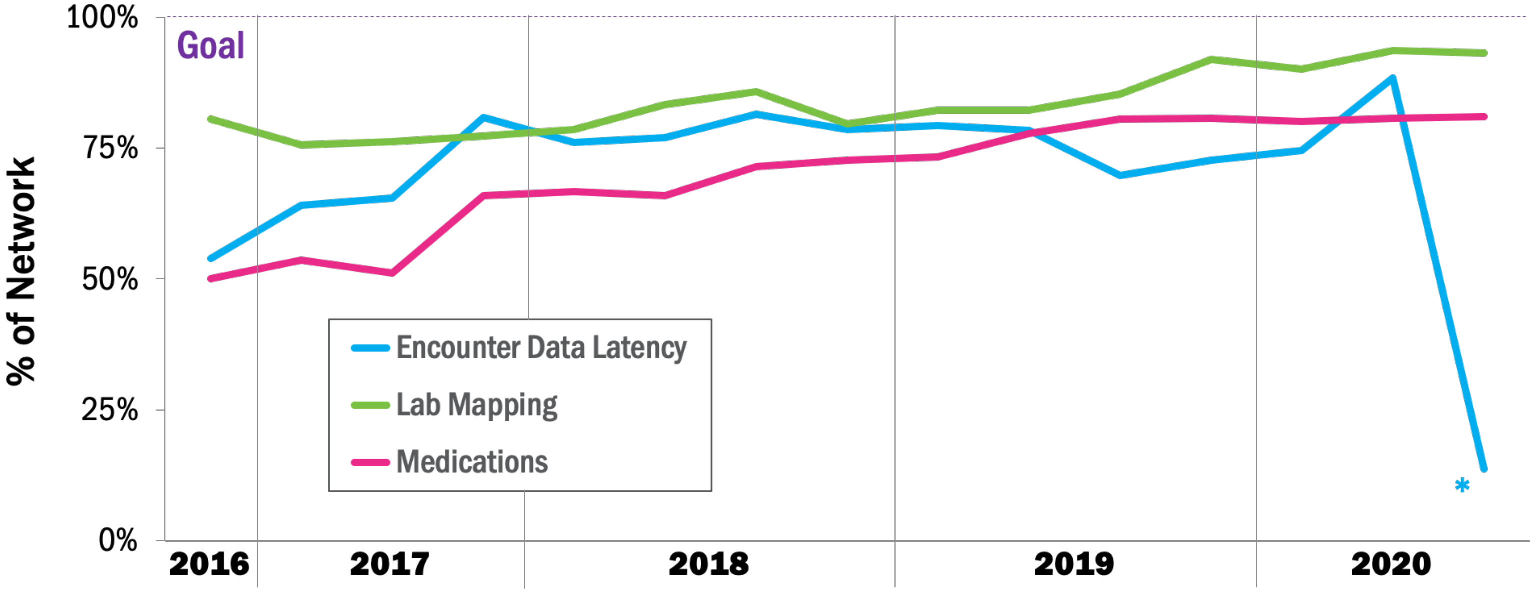
PCORnet data checks - Persistence

Type	Check	Description	Cycle Added
Investigative	DC 4.01	More than a 5% decrease in the number of patients or records in a CDM table	6
	DC 4.02	More than a 5% decrease in the number of patients with diagnosis, procedures, labs or prescriptions during an ambulatory (AV), emergency department (ED), or inpatient (IP) encounter.	6
	DC 4.03	More than a 5% decrease in the number of records for ICD9 or ICD10 diagnosis or procedure codes or CPT/HCPCS procedure codes.	6

Causes of data check failures

- Non-remediable
  - Population characteristics
  - Source system limitation - data does not exist and/or system artifact
- Remediable
  - Problem mapping to reference terminology / CDM value set
  - Source system limitation - data not in system available to datamart team
  - Issue introduced by extract-transformation-load process
- Not all checks will be broadly remediable; some sites may not be able to improve their performance

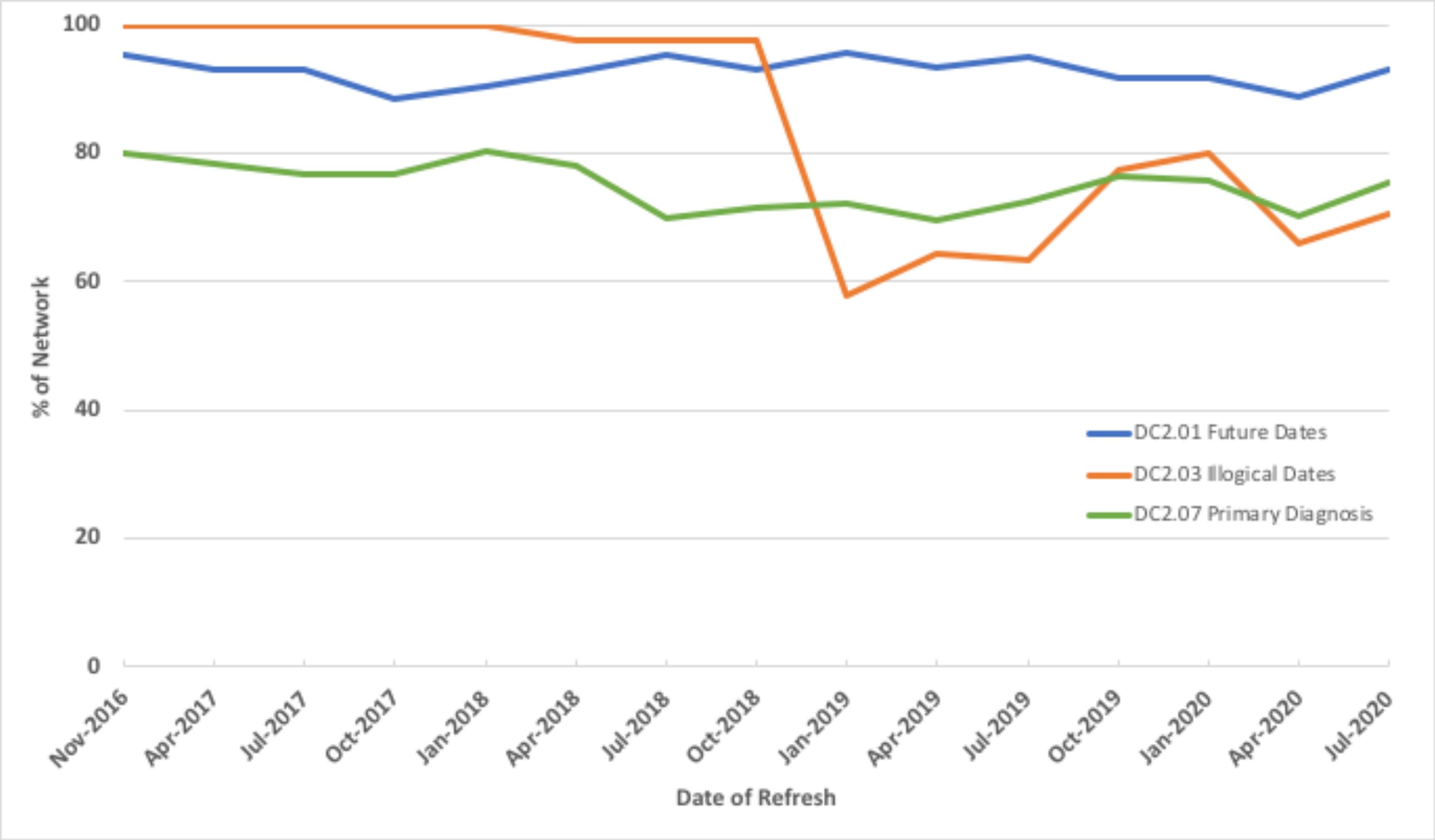
Key foundational data checks



** This is an artifact of the COVID-19 pandemic, because the latency calculations compares April 2020 counts to average volume. At most institutions, volumes in more recent months are now closer-to-normal so next measurement point at July 2020 should be more typical.* *

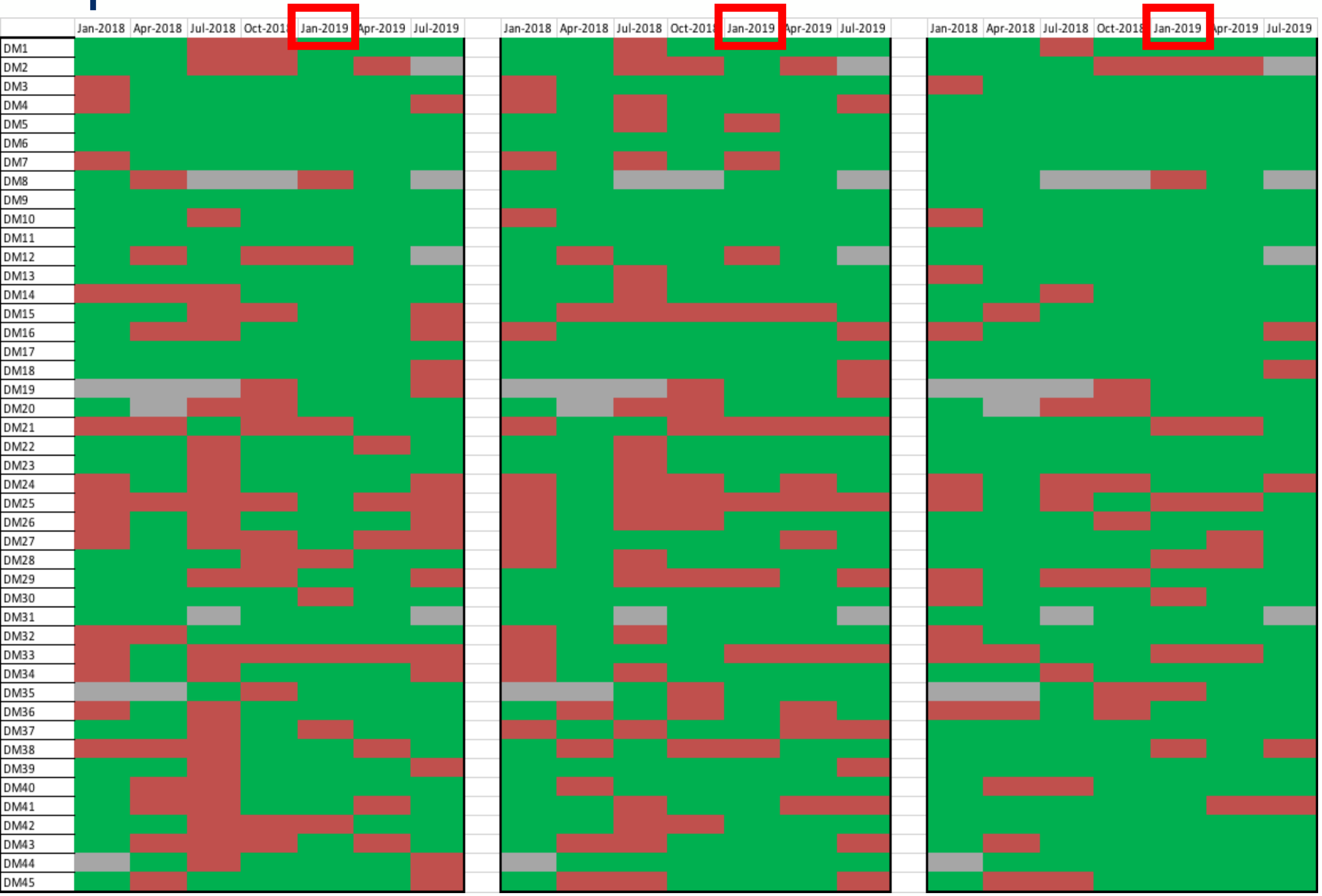
Eligible DataMarts: PCORnet 2.0 DataMarts that include EHR data and were approved prior to August 3, 2020. Data latency is also limited to DataMarts that do not use date obfuscation and include inpatient, ambulatory, and/or emergency department encounters. Since the denominator varies by metric it is not displayed on the X-axis. DataMart refreshes: The refreshes displayed here are the first and third refreshes in previous cycles and every refresh in the current cycle. Other notes: Data latency is measured as the difference in months between the month when the data curation query was executed and the most recent month in which encounter data were ≥75% complete. Lab mapping is the percentage of DataMarts that map at least 80% of their lab records to LOINC. Medications is the percentage of DataMarts that map at least 80% of their Prescribing records to the preferred RXNORM codes.

Results of selected completeness measures



ornet®

Data persistence



DC4.01 DC4.02 DC4.03

Persistence Measures	
DC 4.01	More than a 5% decrease in the number of patients or records in a CDM table
DC 4.02	More than a 5% decrease in the number of patients with diagnosis, procedures, labs or prescriptions during an ambulatory (AV), emergency department (ED), or inpatient (IP) encounter.
DC 4.03	More than a 5% decrease in the number of records for ICD9 or ICD10 diagnosis or procedure codes or CPT/HCPCS procedure codes.

🟩	Pass
🟥	Fail
⬜	Not approved; No refresh

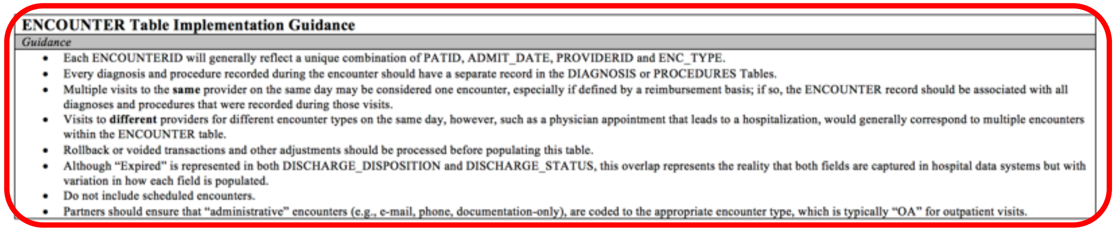
□ First refresh check officially introduced

Curation as a learning process

- Findings from curation influencing the CDM
- Study findings influencing curation

Impact of Data Curation on the CDM

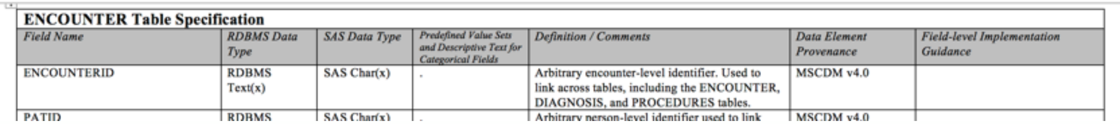
- Curation surfaced instances where there is ambiguity in the CDM specification
  - CDM is silent on the issue – *what to do if date of death is completely unknown?*
  - Unexpected complexity in source data – *how to separate race & ethnicity if captured in a single field?*
- Developed Implementation Guidance (IG) to reduce variability & improve downstream analytics



ENCOUNTER Table Implementation Guidance

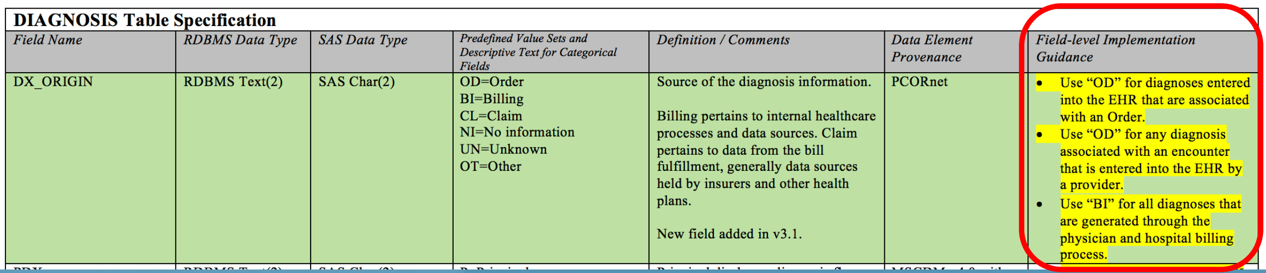
Guidance

- Each ENCOUNTERID will generally reflect a unique combination of PATID, ADMIT_DATE, PROVIDERID and ENC_TYPE.
- Every diagnosis and procedure recorded during the encounter should have a separate record in the DIAGNOSIS or PROCEDURES Tables.
- Multiple visits to the **same** provider on the same day may be considered one encounter, especially if defined by a reimbursement basis; if so, the ENCOUNTER record should be associated with all diagnoses and procedures that were recorded during those visits.
- Visits to **different** providers for different encounter types on the same day, however, such as a physician appointment that leads to a hospitalization, would generally correspond to multiple encounters within the ENCOUNTER table.
- Rollback or voided transactions and other adjustments should be processed before populating this table.
- Although "Expired" is represented in both DISCHARGE_DISPOSITION and DISCHARGE_STATUS, this overlap represents the reality that both fields are captured in hospital data systems but with variation in how each field is populated.
- Do not include scheduled encounters.
- Partners should ensure that "administrative" encounters (e.g., e-mail, phone, documentation-only), are coded to the appropriate encounter type, which is typically "OA" for outpatient visits.



ENCOUNTER Table Specification

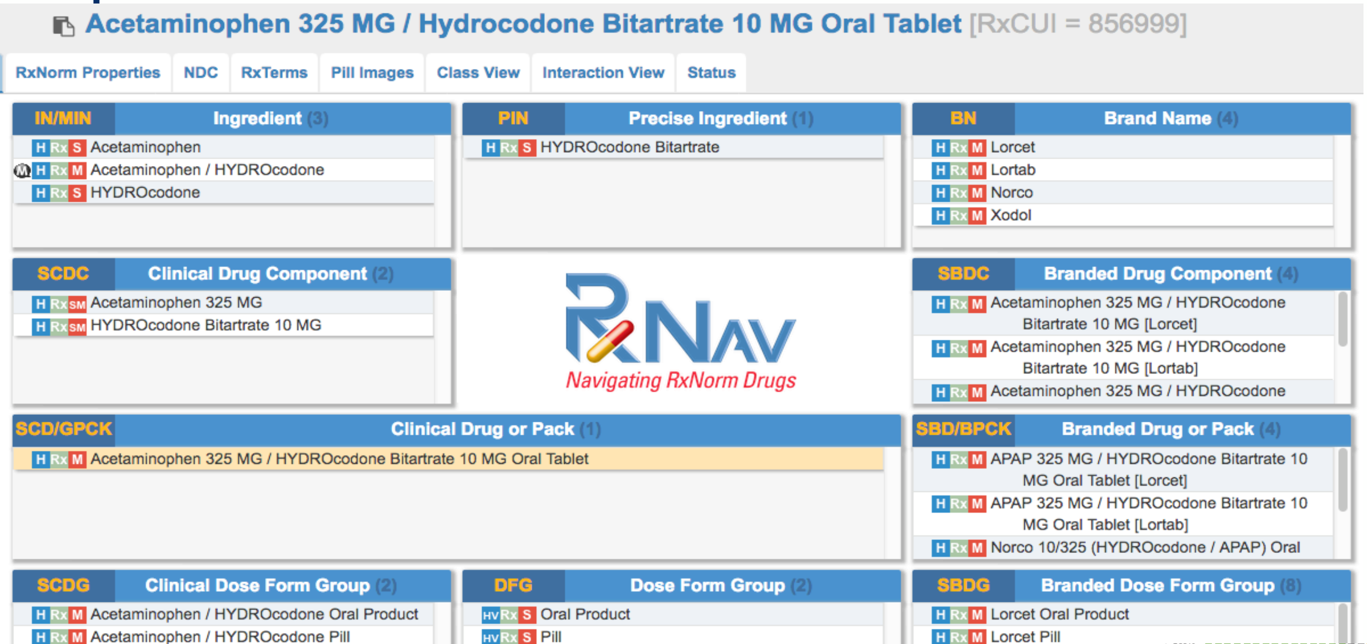
Field Name	*RDBMS Data Type*	*SAS Data Type*	*Predefined Value Sets and Descriptive Text for Categorical Fields*	*Definition / Comments*	*Data Element Provenance*	*Field-level Implementation Guidance*
ENCOUNTERID	RDBMS Text(x)	SAS Char(x)	.	Arbitrary encounter-level identifier. Used to link across tables, including the ENCOUNTER, DIAGNOSIS, and PROCEDURES tables.	MSCDM v4.0	
PATID	RDBMS	SAS Char(x)	.	Arbitrary person-level identifier used to link	MSCDM v4.0	



DIAGNOSIS Table Specification

Field Name	*RDBMS Data Type*	*SAS Data Type*	*Predefined Value Sets and Descriptive Text for Categorical Fields*	*Definition / Comments*	*Data Element Provenance*	*Field-level Implementation Guidance*
DX_ORIGIN	RDBMS Text(2)	SAS Char(2)	OD=Order BI=Billing CL=Claim NI=No information UN=Unknown OT=Other	Source of the diagnosis information. Billing pertains to internal healthcare processes and data sources. Claim pertains to data from the bill fulfillment, generally data sources held by insurers and other health plans. New field added in v3.1.	PCORnet	• Use "OD" for diagnoses entered into the EHR that are associated with an Order. • Use "OD" for any diagnosis associated with an encounter that is entered into the EHR by a provider. • Use "BI" for all diagnoses that are generated through the physician and hospital billing process.

Impact of Studies – Prescribing



Acetaminophen 325 MG / Hydrocodone Bitartrate 10 MG Oral Tablet [RxCUI = 856999]

IN/MIN Ingredient (3)
- H Rx S Acetaminophen
- H Rx M Acetaminophen / HYDROcodone
- H Rx S HYDROcodone

PIN Precise Ingredient (1)
- H Rx S HYDROcodone Bitartrate

BN Brand Name (4)
- H Rx M Lorcet
- H Rx M Lortab
- H Rx M Norco
- H Rx M Xodol

SCDC Clinical Drug Component (2)
- H Rx SM Acetaminophen 325 MG
- H Rx SM HYDROcodone Bitartrate 10 MG

SBDC Branded Drug Component (4)
- H Rx M Acetaminophen 325 MG / HYDROcodone Bitartrate 10 MG [Lorcet]
- H Rx M Acetaminophen 325 MG / HYDROcodone Bitartrate 10 MG [Lortab]
- H Rx M Acetaminophen 325 MG / HYDROcodone

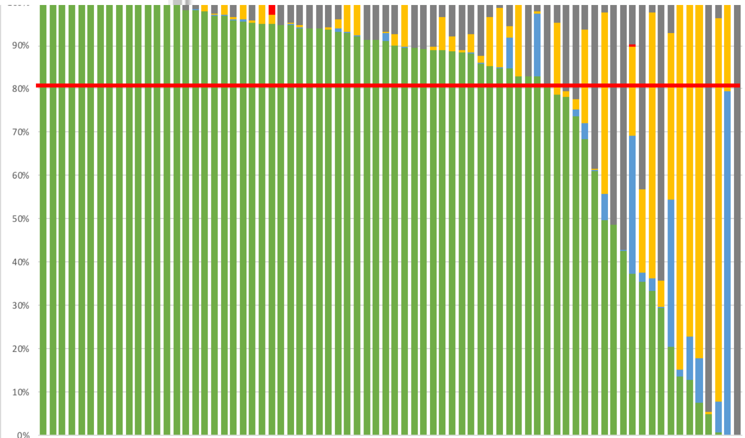
SCD/GPCK Clinical Drug or Pack (1)
- H Rx M Acetaminophen 325 MG / HYDROcodone Bitartrate 10 MG Oral Tablet

SBD/BPCK Branded Drug or Pack (4)
- H Rx M APAP 325 MG / HYDROcodone Bitartrate 10 MG Oral Tablet [Lorcet]
- H Rx M APAP 325 MG / HYDROcodone Bitartrate 10 MG Oral Tablet [Lortab]
- H Rx M Norco 10/325 (HYDROcodone / APAP) Oral

SCDG Clinical Dose Form Group (2)
- H Rx M Acetaminophen / HYDROcodone Oral Product
- H Rx M Acetaminophen / HYDROcodone Pill

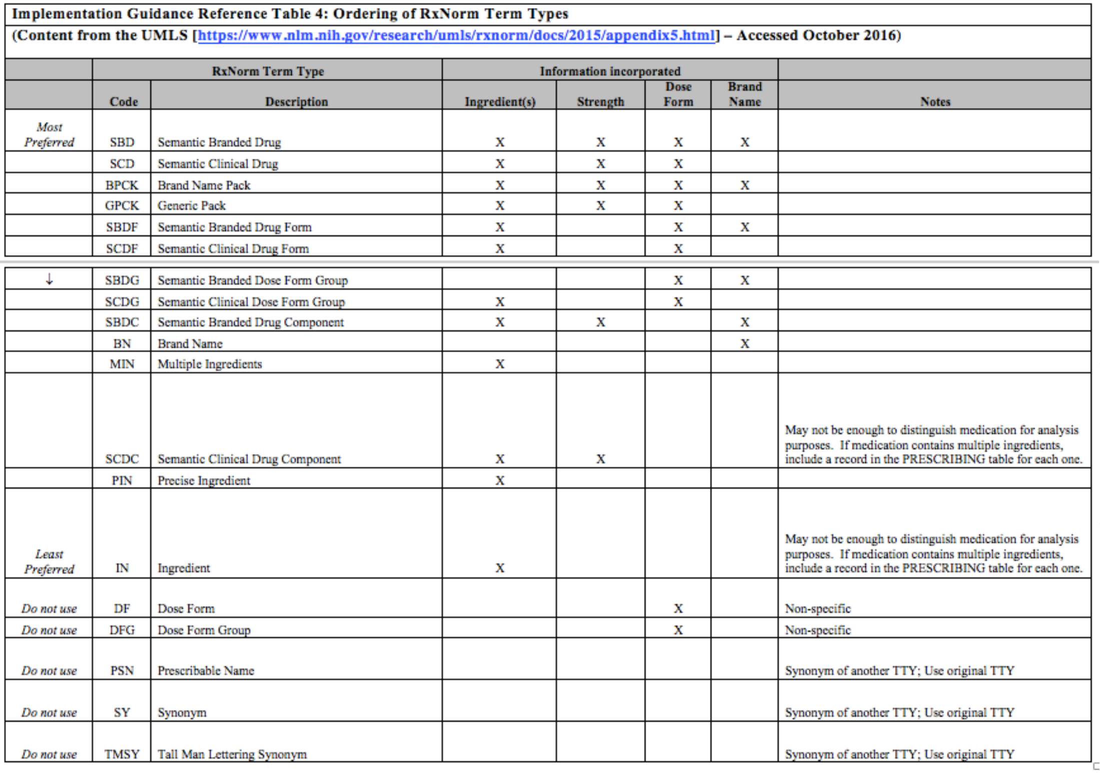
DFG Dose Form Group (2)
- HV Rx S Oral Product
- HV Rx S Pill

SBDG Branded Dose Form Group (8)
- H Rx M Lorcet Oral Product
- H Rx M Lorcet Pill
- H Rx M Lortab Oral Product
- H Rx M Lortab Pill
- H Rx M Norco Oral Product



Impact of Studies – Prescribing (2)

Variability in prescribing data led to updates in IG



Implementation Guidance Reference Table 4: Ordering of RxNorm Term Types
(Content from the UMLS [https://www.nlm.nih.gov/research/umls/rxnorm/docs/2015/appendix5.html] – Accessed October 2016)

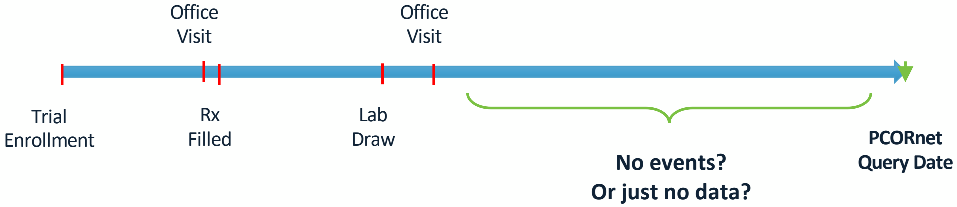
	RxNorm Term Type		Information incorporated				
	Code	**Description**	**Ingredient(s)**	**Strength**	**Dose Form**	**Brand Name**	**Notes**
Most Preferred	SBD	Semantic Branded Drug	X	X	X	X	
	SCD	Semantic Clinical Drug	X	X	X		
	BPCK	Brand Name Pack	X	X	X	X	
	GPCK	Generic Pack	X	X	X		
	SBDF	Semantic Branded Drug Form	X		X	X	
	SCDF	Semantic Clinical Drug Form	X		X		
↓	SBDG	Semantic Branded Dose Form Group			X	X	
	SCDG	Semantic Clinical Dose Form Group	X		X		
	SBDC	Semantic Branded Drug Component	X	X		X	
	BN	Brand Name				X	
	MIN	Multiple Ingredients	X				
	SCDC	Semantic Clinical Drug Component	X	X			May not be enough to distinguish medication for analysis purposes. If medication contains multiple ingredients, include a record in the PRESCRIBING table for each one.
	PIN	Precise Ingredient	X				
Least Preferred	IN	Ingredient	X				May not be enough to distinguish medication for analysis purposes. If medication contains multiple ingredients, include a record in the PRESCRIBING table for each one.
Do not use	DF	Dose Form			X		Non-specific
Do not use	DFG	Dose Form Group			X		Non-specific
Do not use	PSN	Prescribable Name					Synonym of another TTY; Use original TTY
Do not use	SY	Synonym					Synonym of another TTY; Use original TTY
Do not use	TMSY	Tall Man Lettering Synonym					Synonym of another TTY; Use original TTY

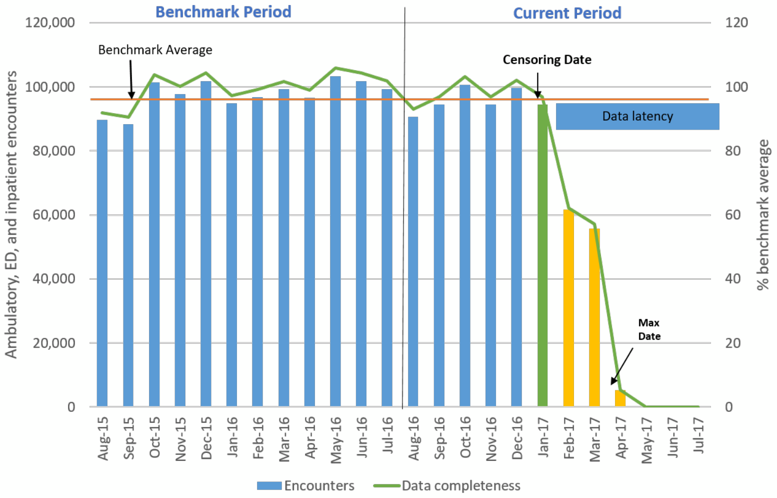
Variability in implementation led to further clarifications of the IG

- **Do NOT assign a CUI that contains more information than is supported by the source data.** For instance, medication orders that only reference a generic medication should not be assigned a branded CUI unless there is a 1:1 relationship between the brand and the generic.
- While SBD is the most preferred of the RxNorm Term Types, **we expect that the one most likely to be present in EHR data will be SCD.** Do NOT assign multiple SBD codes to a single medication order in an attempt to represent all possible branded medications.
- Medications with approved formulations should have an RXCUI that can adequately represent all ingredients with a single code (e.g., SBD, SCD, MIN). **Partners should contact the DRN OC if they run across examples of medications with approved formulations that cannot be represented by a single code**.

Impact of studies – Data latency

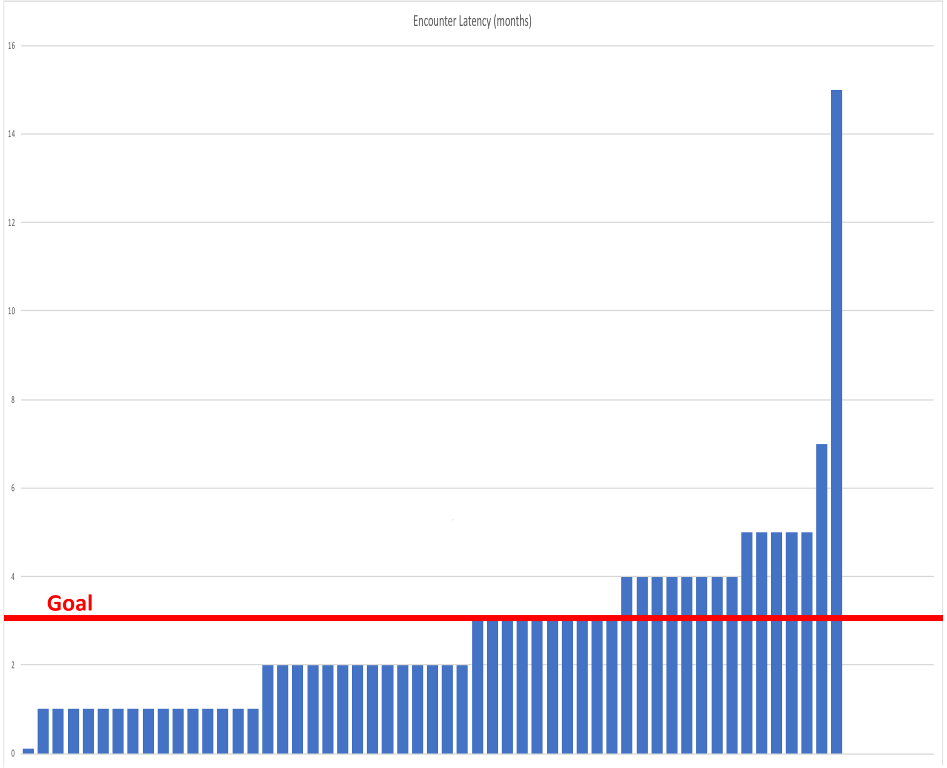
- Latency / completeness of data



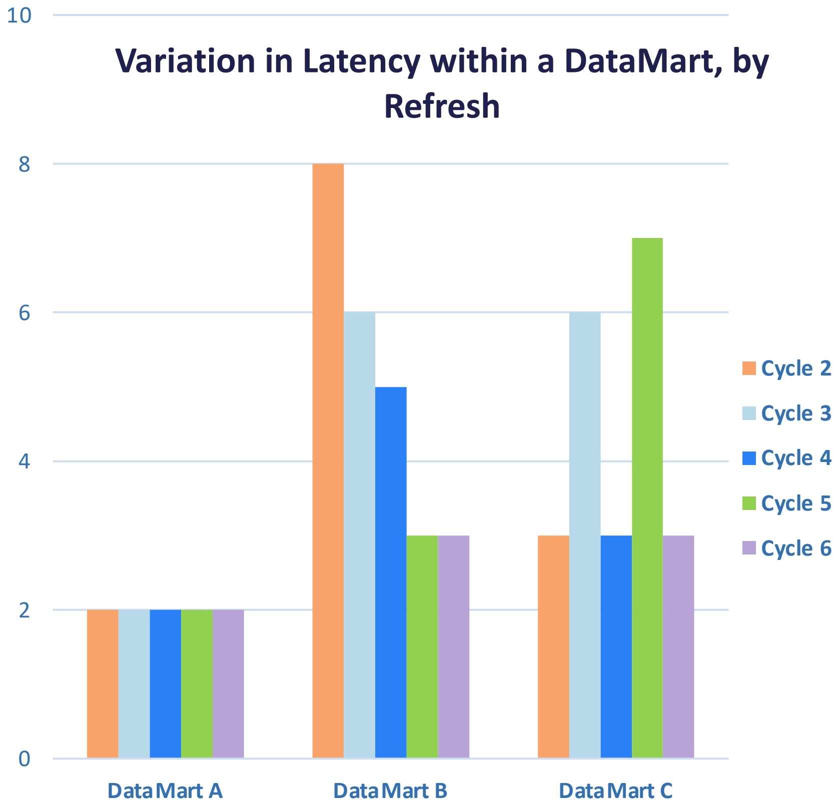


- Questions:
  - *"How complete & up-to-date are the data?" (DSMB)*
  - *"What's the data censoring date for participants?" (Statistician)*
- Developed latency calculation & incorporated into data curation

Latency results (pre-COVID)



Eligible DataMarts: PCORnet 2.0 DataMarts which include inpatient, ambulatory, and/or Emergency Department encounters and do not use date obfuscation



Future work

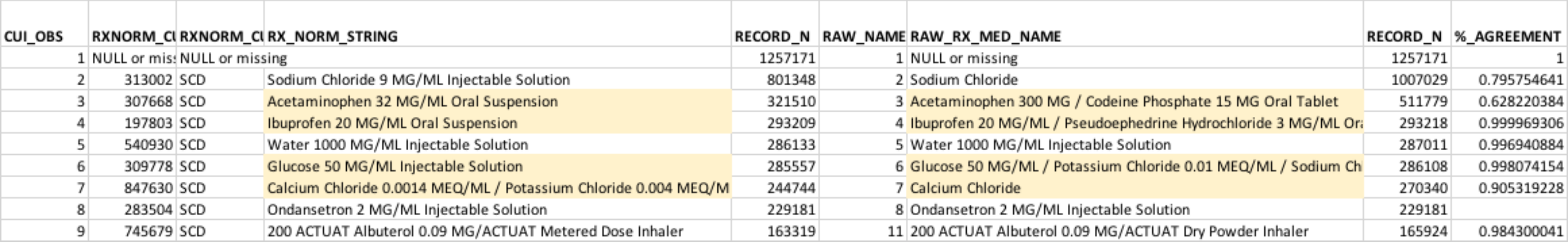
- Assessment of source-to-CDM mappings
- Closing of the gap between foundational and study-specific curation

Assessment of source-to-CDM mappings

- Certain domains within the EHR are not captured in the same terminology used for analysis / data sharing (e.g., RxNorm for medications & LOINC for laboratory results)
- Existing data checks can assess whether CDM records are internally consistent (e.g., specimen source is appropriate for given LOINC code)
- Less capable of determining whether the CDM record is truly reflective of what is in the source (e.g., was the right RxNORM code selected in the first place?)

Assessment of source-to-CDM mappings

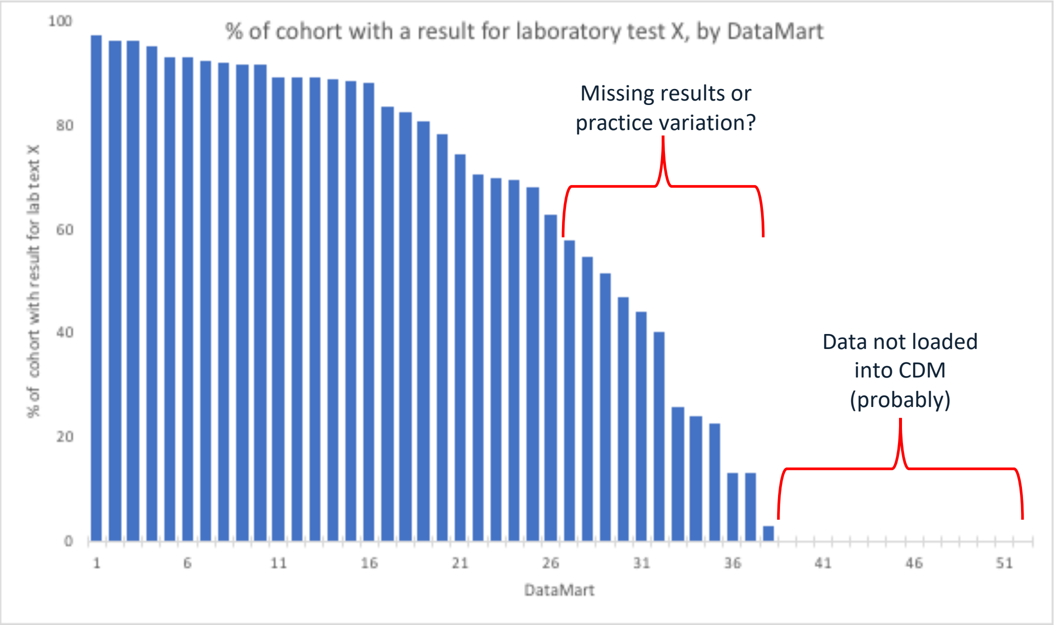
- Many CDMs contain “raw” text fields that store information about a record as it existed in the source system
- Develop procedures to compare the raw and encoded values & flag potential issues



CUI_OBS	RXNORM_CI	RXNORM_CI	RX_NORM_STRING	RECORD_N	RAW_NAME	RAW_RX_MED_NAME	RECORD_N	%_AGREEMENT
1	NULL or mis	NULL or missing		1257171	1	NULL or missing	1257171	1
2	313002	SCD	Sodium Chloride 9 MG/ML Injectable Solution	801348	2	Sodium Chloride	1007029	0.795754641
3	307668	SCD	Acetaminophen 32 MG/ML Oral Suspension	321510	3	Acetaminophen 300 MG / Codeine Phosphate 15 MG Oral Tablet	511779	0.628220384
4	197803	SCD	Ibuprofen 20 MG/ML Oral Suspension	293209	4	Ibuprofen 20 MG/ML / Pseudoephedrine Hydrochloride 3 MG/ML Ora	293218	0.999969306
5	540930	SCD	Water 1000 MG/ML Injectable Solution	286133	5	Water 1000 MG/ML Injectable Solution	287011	0.996940884
6	309778	SCD	Glucose 50 MG/ML Injectable Solution	285557	6	Glucose 50 MG/ML / Potassium Chloride 0.01 MEQ/ML / Sodium Ch	286108	0.998074154
7	847630	SCD	Calcium Chloride 0.0014 MEQ/ML / Potassium Chloride 0.004 MEQ/M	244744	7	Calcium Chloride	270340	0.905319228
8	283504	SCD	Ondansetron 2 MG/ML Injectable Solution	229181	8	Ondansetron 2 MG/ML Injectable Solution	229181	
9	745679	SCD	200 ACTUAT Albuterol 0.09 MG/ACTUAT Metered Dose Inhaler	163319	11	200 ACTUAT Albuterol 0.09 MG/ACTUAT Dry Powder Inhaler	165924	0.984300041

Closing of the gap between foundational and study-specific curation

- **Study-specific curation**: Identify potential quality concerns for key variables within a given study population
- Determine whether issues are related to the data or reflect normal practice variation



Current efforts – Lab, Dx & Px Groups

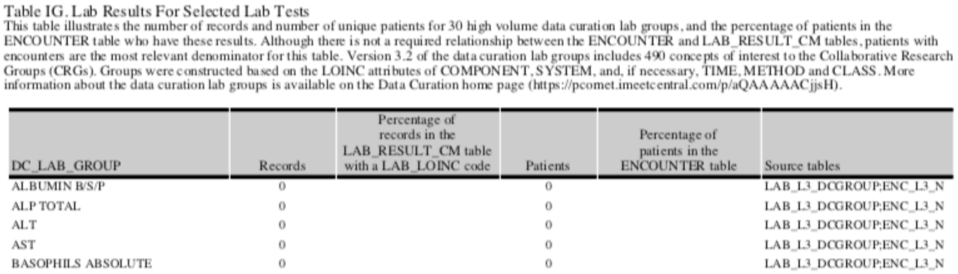


Table IG. Lab Results For Selected Lab Tests

This table illustrates the number of records and number of unique patients for 30 high volume data curation lab groups, and the percentage of patients in the ENCOUNTER table who have these results. Although there is not a required relationship between the ENCOUNTER and LAB_RESULT_CM tables, patients with encounters are the most relevant denominator for this table. Version 3.2 of the data curation lab groups includes 490 concepts of interest to the Collaborative Research Groups (CRGs). Groups were constructed based on the LOINC attributes of COMPONENT, SYSTEM, and, if necessary, TIME, METHOD and CLASS. More information about the data curation lab groups is available on the Data Curation home page (https://pcornet.imeetcentral.com/p/aQAAAAACjjsH).

DC_LAB_GROUP	Records	Percentage of records in the LAB_RESULT_CM table with a LAB_LOINC code	Patients	Percentage of patients in the ENCOUNTER table	Source tables
ALBUMIN B/S/P	0		0		LAB_L3_DCGROUP;ENC_L3_N
ALP TOTAL	0		0		LAB_L3_DCGROUP;ENC_L3_N
ALT	0		0		LAB_L3_DCGROUP;ENC_L3_N
AST	0		0		LAB_L3_DCGROUP;ENC_L3_N
BASOPHILS ABSOLUTE	0		0		LAB_L3_DCGROUP;ENC_L3_N

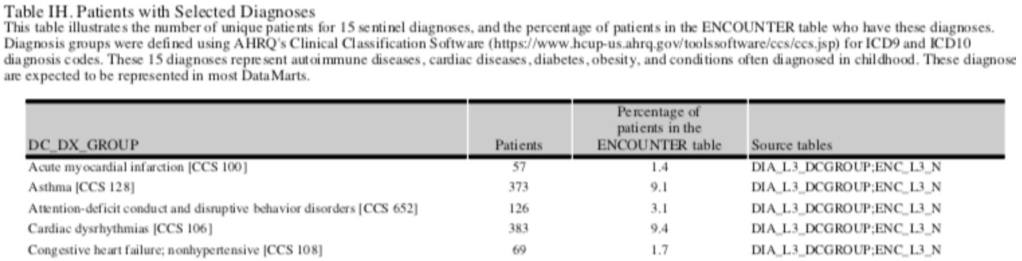


Table IH. Patients with Selected Diagnoses

This table illustrates the number of unique patients for 15 sentinel diagnoses, and the percentage of patients in the ENCOUNTER table who have these diagnoses. Diagnosis groups were defined using AHRQ's Clinical Classification Software (https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp) for ICD9 and ICD10 diagnosis codes. These 15 diagnoses represent autoimmune diseases, cardiac diseases, diabetes, obesity, and conditions often diagnosed in childhood. These diagnose are expected to be represented in most DataMarts.

DC_DX_GROUP	Patients	Percentage of patients in the ENCOUNTER table	Source tables
Acute myocardial infarction [CCS 100]	57	1.4	DIA_L3_DCGROUP;ENC_L3_N
Asthma [CCS 128]	373	9.1	DIA_L3_DCGROUP;ENC_L3_N
Attention-deficit conduct and disruptive behavior disorders [CCS 652]	126	3.1	DIA_L3_DCGROUP;ENC_L3_N
Cardiac dysrhythmias [CCS 106]	383	9.4	DIA_L3_DCGROUP;ENC_L3_N
Congestive heart failure; nonhypertensive [CCS 108]	69	1.7	DIA_L3_DCGROUP;ENC_L3_N

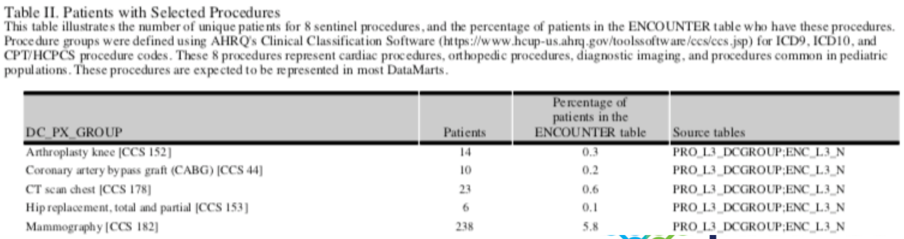


Table II. Patients with Selected Procedures

This table illustrates the number of unique patients for 8 sentinel procedures, and the percentage of patients in the ENCOUNTER table who have these procedures. Procedure groups were defined using AHRQ's Clinical Classification Software (https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp) for ICD9, ICD10, and CPT/HCPCS procedure codes. These 8 procedures represent cardiac procedures, orthopedic procedures, diagnostic imaging, and procedures common in pediatric populations. These procedures are expected to be represented in most DataMarts.

DC_PX_GROUP	Patients	Percentage of patients in the ENCOUNTER table	Source tables
Arthroplasty knee [CCS 152]	14	0.3	PRO_L3_DCGROUP;ENC_L3_N
Coronary artery bypass graft (CABG) [CCS 44]	10	0.2	PRO_L3_DCGROUP;ENC_L3_N
CT scan chest [CCS 178]	23	0.6	PRO_L3_DCGROUP;ENC_L3_N
Hip replacement, total and partial [CCS 153]	6	0.1	PRO_L3_DCGROUP;ENC_L3_N
Mammography [CCS 182]	238	5.8	PRO_L3_DCGROUP;ENC_L3_N

How to interpret these results?

- Absence of expected concepts likely indicates a problem

- Determining whether a given percentage is difficult, given size of dataset

- Proposed solution – create “population reports”
  - For a series of conditions, define co-morbidities, events, medications and labs of interest
  - Generate statistics across time & care settings
  - Benchmark & compare across centers to determine outliers

Summary

- Issues discussed here are inherent to EHR data – they are not specific to PCORnet!
- Data curation is a process for continuous improvement – both methods and quality
- Will need to continue to develop & share best practices around fitness-for-use assessments & how they translate to FDA guidance
- Have spent years understanding the pitfalls of working with administrative claims – will take time to develop that knowledge around EHR data

Questions?

- Acknowledgements
  - Laura Qualls
  - Darcy Louzao
  - Sujung Choi
  - Tom Phillips
  - Brad Hammill
  - Alli Haufler
  - Katie Arnold
  - Michelle Smerek
  - Lauren Cohen
  - Lesley Curtis
  - Adrian Hernandez