Sentinel

Identification and mitigation of structured electronic health record source data mapping issues

Keith Marsolo, PhD
Associate Professor
Department of Population Health Sciences
Duke Clinical Research Institute
Duke University School of Medicine
January 23, 2023

Project motivation

Harmonization of EHR data

- Many EHR data domains (e.g., medication orders, laboratory results) are not captured in standard formats
- To use these data for research or public health surveillance, must first harmonize to a reference standard – need procedures to verify the transformation is accurate

Assessing completeness of EHR data

- Within a health system, "EHR data" may come from a single enterprise EHR with multiple modules, or from many different systems (e.g., EHR, lab, billing, etc.), which may have changed over time (e.g., Cardiology procedures ordered through System X from 2017-2021, but billed through System Y from 2017-2019 and System Z from 2020-2021)
- Given that health systems have access to multiple streams of EHR data (e.g., clinician-entered information, orders, billing data, claims submitted to health plans, etc.) – can we look at these different streams to help determine if we have a "complete" set of EHR data?

Initial examples shown within these slides are taken from the National Patient-Centered Clinical Research Network (PCORnet®), but the same challenges exist regardless of the source

Harmonization example - representing a medication in RxNorm

	RxNorm Term Type	Information encoded				Example medication representation
	Description	**Ingredient(s)**	**Strength**	**Dose Form**	**Brand Name**	**Original string - Augmentin XR 12 HR 1000 MG Extended Oral Release Tablet**
Most Granular	Semantic Branded Drug	X	X	X	X	Augmentin XR 12 HR 1000 MG Extended Release Oral Tablet
	Semantic Clinical Drug	X	X	X		12 HR Amoxicillin 1000 MG / Clavulanate 62.5 MG Extended Release Oral Tablet
	Brand Name Pack	X	X	X	X	N/A
	Generic Pack	X	X	X		N/A
	Semantic Branded Drug Form	X		X	X	Amoxicillin / Clavulanate Extended Release Oral Tablet [Augmentin]
	Semantic Clinical Drug Form	X		X		Amoxicillin / Clavulanate Extended Release Oral Tablet
↓	Semantic Branded Dose Form Group*			X	X	Augmentin Oral Product; Augmentin Pill (Requires two records)
	Semantic Clinical Dose Form Group*	X		X		Amoxicillin / Clavulanate Oral Product; Amoxicillin / Clavulanate Pill (Requires two records)
	Semantic Branded Drug Component	X	X		X	Amoxicillin 1000 MG / Clavulanate 62.5 MG [Augmentin]
	Brand Name				X	Augmentin
	Multiple Ingredients	X				Amoxicillin / Clavulanate
	Semantic Clinical Drug Component*	X	X			Amoxicillin 1000 MG; Clavulanate 62.5 MG (Requires two records)
	Precise Ingredient	X				N/A
Least Granular	Ingredient*	X				Amoxicillin; Clavulanate (Requires two records)
Non-specific	Dose Form			X		Extended Release Oral Tablet
	Dose Form Group*			X		Oral Product; Pill (Requires two records)
	Prescribable Name					
	Synonym					
	Tall Man Lettering Synonym					

* Denotes term types that require multiple records to represent multi-ingredient medications

Within the PCORnet Common Data Model, medication orders and administrations (at most sites) are coded using RxNorm

RxNorm is an interoperability standard maintained by the National Library of Medicine that provides normalized names for medications (hence, *RxNorm*). It can represent medications at various levels of granularity

Even if Sentinel leverages a different standard to represent EHR-based medications, data partners may still need to transform data to/from RxNorm

PCORnet has defined a set of preferred "tiers" for the different RxNorm Term Types

	RxNorm Term Type	Information encoded				Example medication representation	Tier
	Description	Ingredient(s)	Strength	Dose Form	Brand Name	Original string - Augmentin XR 12 HR 1000 MG Extended Oral Release Tablet	
Most Granular	Semantic Branded Drug	X	X	X	X	Augmentin XR 12 HR 1000 MG Extended Release Oral Tablet	Tier 1 (most preferred) These terms types encode the maximum amount of information
	Semantic Clinical Drug	X	X	X		12 HR Amoxicillin 1000 MG / Clavulanate 62.5 MG Extended Release Oral Tablet	
	Brand Name Pack	X	X	X	X	N/A	
	Generic Pack	X	X	X		N/A	
	Semantic Branded Drug Form	X		X	X	Amoxicillin / Clavulanate Extended Release Oral Tablet [Augmentin]	Tier 2
	Semantic Clinical Drug Form	X		X		Amoxicillin / Clavulanate Extended Release Oral Tablet	
↓	Semantic Branded Dose Form Group*			X	X	Augmentin Oral Product; Augmentin Pill (Requires two records)	
	Semantic Clinical Dose Form Group*	X		X		Amoxicillin / Clavulanate Oral Product; Amoxicillin / Clavulanate Pill (Requires two records)	
	Semantic Branded Drug Component	X	X		X	Amoxicillin 1000 MG / Clavulanate 62.5 MG [Augmentin]	
	Brand Name				X	Augmentin	
	Multiple Ingredients	X				Amoxicillin / Clavulanate	
	Semantic Clinical Drug Component*	X	X			Amoxicillin 1000 MG; Clavulanate 62.5 MG (Requires two records)	Tier 3
	Precise Ingredient	X				N/A	
Least Granular	Ingredient*	X				Amoxicillin; Clavulanate (Requires two records)	
Non-specific	Dose Form			X		Extended Release Oral Tablet	Tier 4 (Do not use)
	Dose Form Group*			X		Oral Product; Pill (Requires two records)	
	Prescribable Name						
	Synonym						
	Tall Man Lettering Synonym						

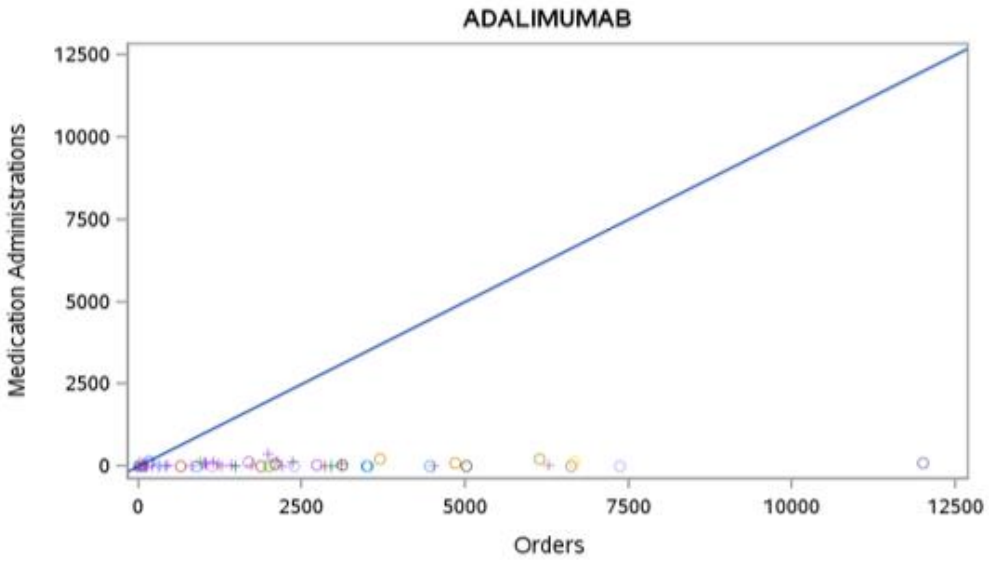
* Denotes term types that require multiple records to represent multi-ingredient medications

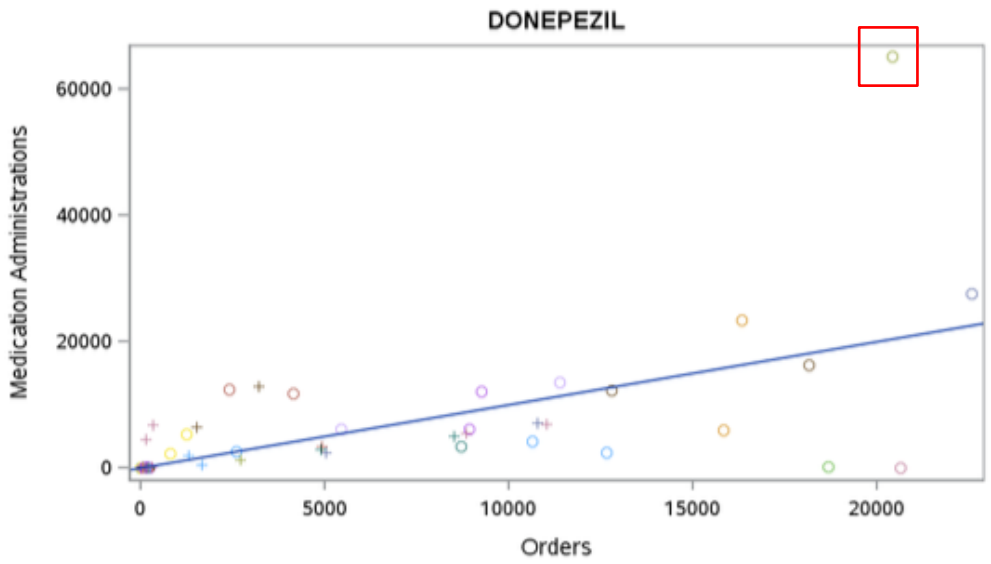
Example harmonization issue – medication mapping

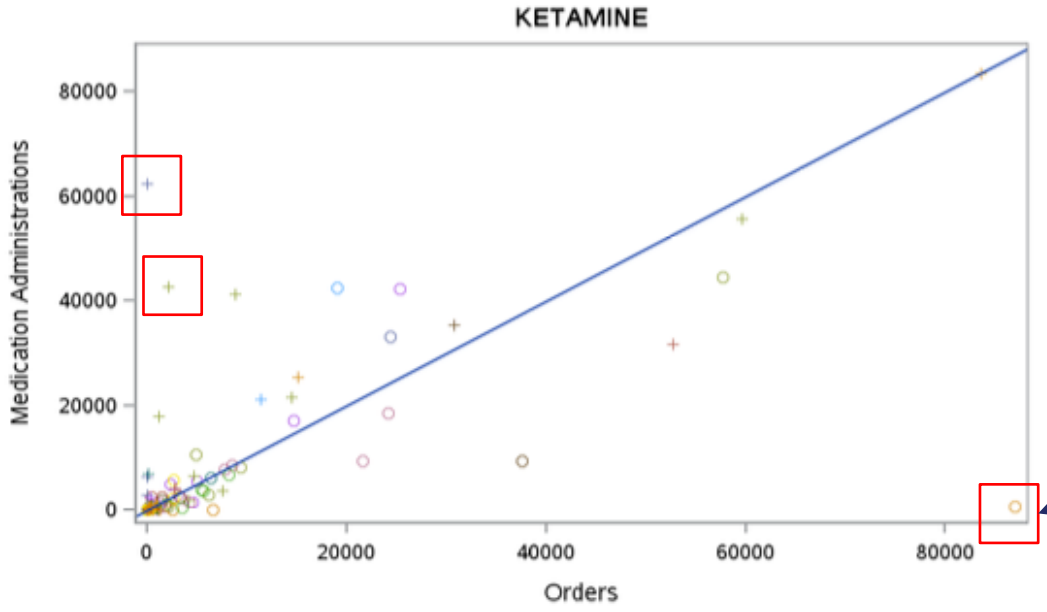
Highest-volume medication records by RxNorm code				Highest-volume medication records by name (within the EHR)			Percent Agreement
Rank based on Code	**RxNorm Code**	**Medication name (derived from RxNorm code)**	**Record Count by Code**	**Rank based on Name**	**Medication name (from EHR)**	**Record Count by Name**	
1	Null or missing		1257171	1	Null or missing	1257171	100%
2	313002	Sodium Chloride 9 MG/ML Injectable Solution	801348	2	Sodium Chloride	1007029	79.6%
3	307668	Acetaminophen 32 MG/ML Oral Suspension	321510	3	Acetaminophen 300MG / Codeine Phosphate 15 MG Oral Tablet	511779	
4	197803	Ibuprofen 20 MG/ML Oral Suspension	293209	4	Ibuprofen 20 MG/ML / Pseudoephedrine Hydrochloride 3 MG/ML Oral Suspension	293218	
5	540930	Water 1000 MG/ML Injectable Solution	286133	5	Water 1000 MG/ML Injectable Solution	287011	99.6%
6	309778	Glucose 50 MG/ML Injectable Solution	285557	6	Glucose 50 MG/ML / Potassium Chloride 0.01 MEQ/ML / Sodium Chloride 0.0342 MEQ/ML Injectable Solution	286108	99.8%

Yellow highlighting indicates a discordance in medications (e.g., RxNorm code represents a single ingredient in RxNorm vs. multi-ingredient order within the EHR)

Potential completeness issue – differences in rates based on provenance (orders vs. medication administrations)







Red boxes indicate potential outlier

Project goals and objectives

Objective 1 (data harmonization): Assess the mapping and harmonization of structured electronic health record (EHR) data to reference terminologies for laboratory results, medication orders and administrations (inpatient and outpatient) & characterize the severity of issues that are uncovered that would impact their use in public health surveillance or research analyses

Objective 2 (data completeness): Develop metrics (queries) to allow for comparisons across complementary data domains; Specifically look at profiles of records across time, care setting, and provenance for a given condition to identify issues with data completeness.

Objective 1: Methods to assess mapping of structured EHR data to reference terminologies

General approach:

- Develop queries to assess mapping of medication orders, medication administrations and laboratory tests – limited to the top 200 by volume
- For each medication / lab, generate statistics on all the different combinations within the structured fields and "raw" source fields
  - Example - for a given medication name, summarize the number of records/patients for associated RxNorm codes, dose units, dose forms, as well as the corresponding "raw" fields
- Queries are written for the PCORnet Common Data Model (CDM), but can be repurposed to other CDMs

RAW Medication Name	RxNorm Code	CDM Dose Unit	RAW Dose Unit	Number of Records	Number of Patients
CALCIUM CARBONATE 300 MG (750 MG) CHEWABLE TABLET	1044532	Other		12	2
	1044532	Other	mg of elemental	13	11
	1044532	Other	mg of salt	50564	14817
	1044532	Other	tablet	1	1
	1484737	Other		3	2
	1484737	Other	mg of elemental	4	3
	1484737	Other	mg of salt	51092	14887
	1484737	Other	tablet	2	2

Example statistics for Dose Unit for a single medication

Objective 1: Evaluation (ongoing)

Project will evaluate the following:

- Number of medication codes/ laboratory tests associated with more than one name within the EHR and vice versa
- Concordance between lab name / medication name (brand and/or ingredient) within the EHR and that derived from the associated code
- Concordance between discrete fields (e.g., lab result unit, medication dose, etc.) and those associated with the associated LOINC / RxNorm code
- Use results to characterize issues by severity (e.g., LOINC code mis-match, combination medication represented by single-ingredient RxNorm code, generic medication represented by brand name, etc.)

Severity	Example issue	Rationale
Critical	(1) Lab test mismatch (incorrect LOINC code) (2) Multi-ingredient drug uses single ingredient RxNorm code (3) Single ingredient drug uses multi-ingredient RxNorm code	(1-3) The LOINC/RxNorm codes that are assigned to these records are incorrect and would not actually represent the test result or exposure to the specified medication.
Major	(1) Ingredient-level RxNorm code utilized when more granular available (single-ingredient drugs only) (2) More granular RxNorm code used than supported by the data	(1) The ingredient is correct, but the other metadata is missing, meaning those records may be excluded if the drug has forms that are not part of an analysis (i.e., topical creams). (2) This example is the inverse – records that should have been excluded were included.
Moderate	(1) Generic medication uses brand name RxNorm code (2) Brand name medication uses a generic-level RxNorm code	(1) Any study that looking for the use of a specific brand of medication will include extra records. (2) Studies that are looking at the use of a specific branded medication will miss records.
Minor	(1) Distribution of lab results is an outlier for a given LOINC.	(1) The test may be only used on specific populations (e.g., inpatients), which may bias results.

Objective 1: mapping between RAW medication name & RxNorm code

Number of RxNorm code associated with a given RAW name (PRESCRIBING) – *Assumption is median ~1*

Site	Median (Min, Max)
DP1	1 (1, 1)
DP2	1 (1, 2)
DP3	1 (1, 2)

Number of RxNorm codes associated with a given RAW name (MED_ADMIN) – *Assumption is median ~1*

Site	Median (Min, Max)
DP1	1 (1, 2)
DP2	1 (1, 2)
DP3	1 (1, 1)

Examples: RAW medication names mapped to >1 RxNorm code (PRESCRIBING)

Site	Raw RX Medication Name	RxNorm code	Name Associated with RxNorm code
DP2	NALOXONE 0.4 MG/ML INJECTION SOLUTION		
DP2	NALOXONE 0.4 MG/ML INJECTION SOLUTION	1191222	NALOXONE HYDROCHLORIDE 0.4 MG/ML INJECTABLE SOLUTION
DP3	EPHEDRINE	1116233	ephedrine hydrochloride 30 MG/ML Injectable Solution
DP3	ePHEDrine	1116294	1 ML ephedrine sulfate 50 MG/ML Injection

- The records with a "blank" RxNorm code would not show up in analyses.
- Records from DP3 use an overly-specific RxNorm code, given information available in RAW name.

Objective 1: mapping between RxNorm codes and RAW name

Number of RAW names associated with a given RxNorm code (PRESCRIBING) – *Assumption is median ~1*

Site	Median (Min, Max)
DP1	1 (1, 5)
DP2	1 (0, 2)
DP3	1 (1, 3)

Number of RAW names associated with a given RxNorm code (MED_ADMIN) – *Assumption is median ~1*

Site	Median (Min, Max)
DP1	1 (1, 13)
DP2	1 (1, 3)
DP3	1 (1, 4)

RxNorm code mapped to 0, >1 RAW name (Subset of PRESCRIBING)

Site	RxNorm Code	Name associated with RxNorm Code	Raw RX Medication Name	Ramification
DP1	142436	FENTANYL CITRATE	FENTANYL CITRATE 0.05 MG/ML IJ SOLN	Likely okay, though RxNorm code only specifies the ingredient
DP1	142436	FENTANYL CITRATE	FENTANYL CITRATE INJ 50 MCG/ML CUSTOM AMP/VIAL	Likely okay, though RxNorm code only specifies the ingredient
DP2			PROPOFOL INTRAOP INFUSION	No RxNorm code specified
DP3	1665050	CEFAZOLIN 1000 MG INJECTION	CEFAZOLIN	RxNorm code is overly specific
DP3	1665060	CEFAZOLIN 2000 MG INJECTION	CEFAZOLIN 2 GRAM/100 ML IN DEXTROSE	RxNorm code differs from raw name
DP3	1665060	CEFAZOLIN 2000 MG INJECTION	ceFAZolin	RxNorm code is overly specific

RxNorm code mapped to 0, >1 RAW name (Subset of MED_ADMIN)

Site	RxNorm code	Name associated with RxNorm Code	Raw Medication Administrated Name	
DP1	313002	SODIUM CHLORIDE 9 MG/ML INJECTABLE SOLUTION	AMPICILLIN-SULBACTAM 3 G IV MBP	Mapping error, unless second record created for ampicillin-sulbactam
DP1	313002	SODIUM CHLORIDE 9 MG/ML INJECTABLE SOLUTION	BOLUS IV FLUID <masked>	Likely okay
DP1	313002	SODIUM CHLORIDE 9 MG/ML INJECTABLE SOLUTION	CALCIUM CHLORIDE 4 GRAMS IN 500 ML 0.9% NACL MIXTURE FOR CVVHD <masked>	Mapping error, unless second record created for calcium chloride
DP1	313002	SODIUM CHLORIDE 9 MG/ML INJECTABLE SOLUTION	CEFEPIME 2 G EXTENDED INFUSION MBP <masked>	Mapping error, unless second record created for cefepime
DP3	161	ACETAMINOPHEN	5653989	Raw name not present, no way to assess
DP3	161	ACETAMINOPHEN	954132	Raw name not present, no way to assess

Objective 1: concordance between ingredients

Match between “cleaned” RAW ingredient & ingredient from specified in RxNorm code (MED_ADMIN)

Site	Match /N (Pct.)
DP1	103/120 (85.8%)
DP2	121/122 (99.2%)
DP3	0/200 (0%)

Most flagged errors in DP1 may be caused by a single inpatient mixture medication being represented by 2 or more records in the CDM. Examples are provided on the next slide.

Error in DP2 is due to a multi-ingredient medication listed as a single-ingredient record.

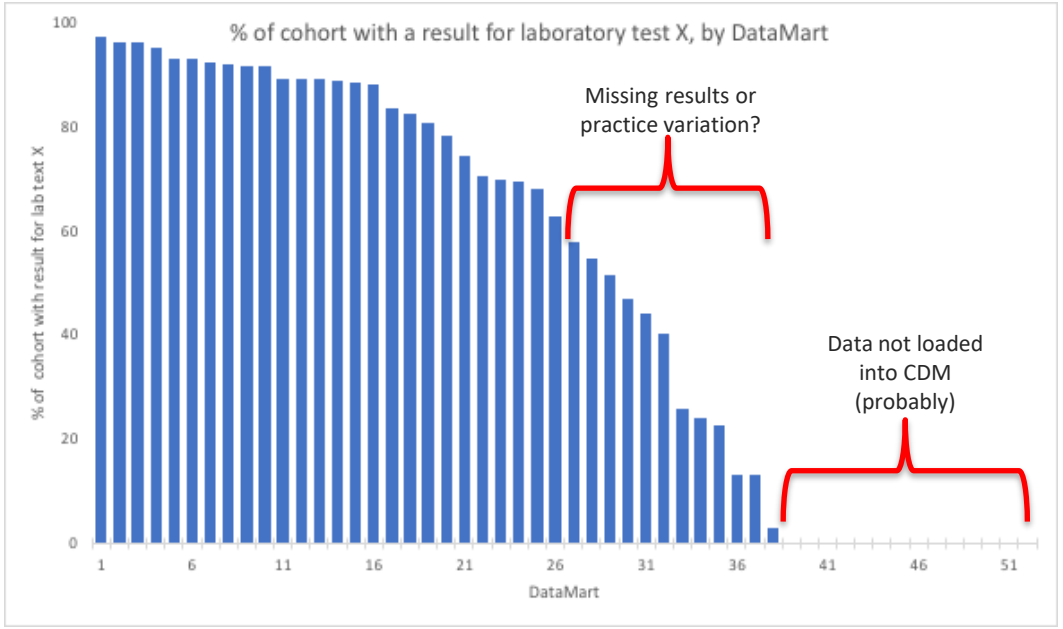
DP3 does not populate the RAW fields with text in their MED_ADMIN table. They use numbers instead, making analysis impossible.

Potential mismatches between ingredient (MED_ADMIN)

Site	Raw Medication Name (Cleaned)	Raw Medication Name (original)	Name associated with RxNorm Code (Cleaned)	Name associated with RxNorm Code ((original)	Notes
DP1	IV BUILDER PEDS CONTINUOUS	IV BUILDER PEDS CONTINUOUS <masked>	WATER	sterile water	These records are likely fine, though it is hard to determine for sure based on the raw medication name.
DP1	NICU LVP INFUSION BUILDER	NICU LVP INFUSION BUILDER <masked>	WATER	sterile water	
DP1	NUTRITION ADULT - FLUSH SCHEDULED	NUTRITION ADULT - FLUSH SCHEDULED <masked>	WATER	water	
DP1	CEFEPIME	CEFEPIME 2 G EXTENDED INFUSION MBP <masked>	SODIUM CHLORIDE	sodium chloride 9 MG/ML Injectable Solution	May be okay if additional record(s) for the active ingredient is present. Mapping error if not.
DP1	CEFTRIAXONE	CEFTRIAXONE 1 G IV MBP	SODIUM CHLORIDE	sodium chloride 9 MG/ML Injectable Solution	
DP1	DEXTROSE	DEXTROSE 10% AND NACL INFUSION	SODIUM CHLORIDE	sodium chloride 234 MG/ML Injectable Solution	
DP1	AMPICILLIN/SULBACTAM	AMPICILLIN-SULBACTAM 3 G IV MBP	SODIUM CHLORIDE	sodium chloride 9 MG/ML Injectable Solution	
DP1	KETAMINE	KETAMINE INFUSION <masked>	SODIUM CHLORIDE	sodium chloride 9 MG/ML Injectable Solution	
DP1	PIPERACILLIN/TAZOBACTAM	PIPERACILLIN SOD-TAZOBACTAM 3.375 G IV MBP	SODIUM CHLORIDE	sodium chloride 9 MG/ML Injectable Solution	
DP1	SODIUM PHOSPHATE	SODIUM PHOSPHATE INFUSION CVVH <masked>	DEXTROSE	150 ML glucose 50 MG/ML Injection	Appears to be mapping error, though glucose could also be delivered as part of CVVH therapy.
DP1	ZZ IMS TEMPLATE	ZZ IMS TEMPLATE	METOPROLOL TARTRATE	metoprolol tartrate 25 MG Oral Tablet	Appears to be mapping error
DP1	TPN ADULT CONTINUOUS	TPN ADULT CONTINUOUS <masked>	WATER	Water 1000 MG/ML Injectable Solution	These are all components that would be included in total parenteral nutrition (TPN), so may not be an error.
DP1	TPN ADULT CYCLIC	TPN ADULT CYCLIC <masked>	DEXTROSE	glucose 700 MG/ML Injectable Solution	
DP1	TPN NEONATAL CONTINUOUS	TPN NEONATAL CONTINUOUS <masked>	DEXTROSE	glucose 700 MG/ML Injectable Solution	
DP1	TPN PEDS CONTINUOUS	TPN PEDS CONTINUOUS <masked>	DEXTROSE	glucose 700 MG/ML Injectable Solution	
DP1	TPN WITHOUT H2 ANTAGONIST	TPN WITHOUT H2 ANTAGONIST <masked>	CYSTEINE	10 ML cysteine hydrochloride 50 MG/ML Injection	
DP2	PRENATAL VITAMIN	PRENATAL VITAMIN WITH CALCIUM NO.72-IRON 27 MG-FOLIC ACID 1 MG TABLET	IRON	iron	Multi-ingredient medication listed as single ingredient. Would need additional records for calcium and folic acid to not be an error.

Objective 2: Standardized metrics to generate comparisons based on provenance (ongoing)

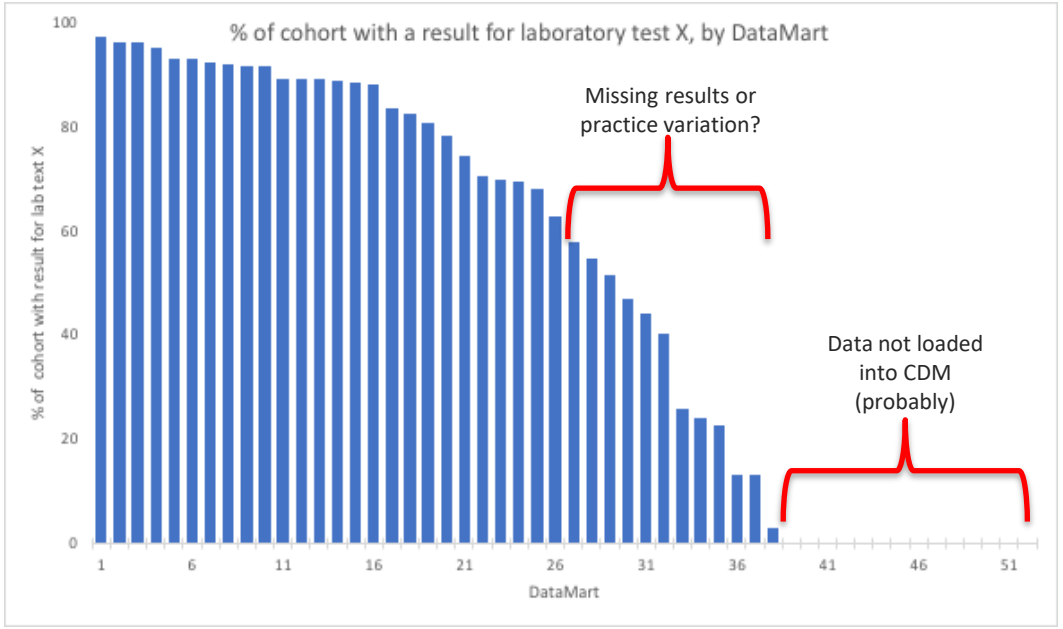
- In this example, the absence of records likely indicates a problem
- Determining whether a given percentage is “correct” can be difficult without additional context
- To address this, will develop queries that allow the comparison of records based on time, care setting and provenance to identify potential issues with data completeness



Objective 2: Standardized metrics to generate comparisons based on provenance (ongoing)

Approach:

- Work with FDA to select 4 conditions of interest
- Within each condition, work with FDA to define associated diagnoses, procedures, medications and laboratory tests (up to 8 in each category)
- Distribute query package to partner sites to generate summary statistics stratified by:
  - Year (2016-2021)
  - Encounter type (e.g., outpatient, inpatient, emergency department, other)
  - Provenance (e.g., clinician-entered information, EHR billing)
- Focus of analysis will be within-DataMart comparisons, though cross-DataMart comparisons can also be informative.



Conditions and concepts (1)

Chronic Hepatitis C (Hep C)	Chronic Kidney Disease (CKD)	
Diagnoses: • Hepatic decompensation • Hepatitis B • HIV **Procedures:** • Liver ultrasonography • Liver biopsy • Hepatitis A/B vaccination **Medications:** • Interferon • Nucleoside Analogue (Antiviral) • NS5A Inhibitor **Labs:** • Albumin, serum • Bilirubin • HCV RNA • INR	**Diagnoses:** • Anemia • Dyslipidemia • Hyperkalemia • Metabolic acidosis • Pericarditis • Thyroid dysfunction • Uremic bleeding • Uremic neuropathy **Medications:** • ESAs • Iron • ACEi • ARBs • Beta blockers • Loop diuretics • Amlodipine • Sodium/glucose cotransporter-2 inhibitors	**Procedures:** • Abdominal imaging (ultrasound, CT, MRI, x-ray) • Dialysis • Renal arteriography / venography • Voiding cystourethrography Pyelography • Kidney biopsy **Labs:** • Albumin, serum • Albumin, urine • Blood urea nitrogen (BUN) • Creatinine, serum • Creatinine, urine • eGFR • Ferritin • Hematocrit (HCT) • Hemoglobin • Parathyroid hormone • Protein, urine

Conditions and concepts (2)

Chronic obstructive pulmonary disease (COPD)

Diagnoses:
- Asthma
- Congestive Heart Failure
- Coronary Artery Disease
- Depression
- Diabetes
- Environmental exposure
- Lung Cancer
- Osteoporosis
- Pulmonary heart disease
- Sleep disorders

Medications:
- Beta adrenergic agonists
- Muscarinic antagonists
- Oral / inhaled glucocorticoids
- O2 supplementation
- First line antibiotics
- Macrolides
- Fluoroquinolones
- Antipseudomonal penicillin
- Cephalosporins
- Aminoglycoside

Procedures:
- Chest X-Ray
- Chest CT
- Pulmonary Function Test
- Pulse oximetry
- Ventilation

Labs:
- alpha-1 antitrypsin (ATT)
- PaO2
- PaCO2
- pH
- Bicarbonate
- Hemoglobin
- Leukocytes
- Platelets
- Sputum gram stain and culture

Pulmonary Function Test Results:
- FEV1
- FEV1/FVC

Pulmonary Arterial Hypertension (PAH)

Diagnoses:
- Congenital heart disease
- HIV
- Rheumatoid arthritis
- Scleroderma
- Systemic lupus erythematosus

Procedures:
- Echocardiogram
- Electrocardiogram (ECG)
- Cardiac catheterization
- Ventilation perfusion scan (VQ)

Medications:
- Prostacyclins
- Endothelin Receptor Antagonists
- sCG Stimulator
- PDE5 inhibitors

Labs:
- ALP
- ALT
- Bilirubin
- BNP
- BUN
- CO2
- Calcium
- Chloride
- Creatinine, serum
- Glucose
- Hematocrit (HCT)
- Human Chorionic Gonadotropin
- Platelets
- Potassium
- Sodium
- Thyrotropin

Pulmonary Function Test Results:
- Total Lung Capacity
- Diffusion capacity (DLCO)

Background info on Data Partners

- Large academic medical centers
- All currently use Epic as their enterprise EHR (but all have switched from other systems in the last 10 years)
- Data provenance for diagnoses and procedures
  - Both DP1 and DP2 populate clinician-entered (ordered) and billed *diagnoses*
  - DP2 populates both ordered and billed *procedures*
  - DP3 only populates billed *diagnoses* and *procedures*
  - None of the 3 DPs populate the other provenance types in the PCORnet Common Data Model (e.g., Claims or Derived [e.g., derived through natural language processing])
- PCORnet provenance statistics by data domain

Diagnoses	Number of Partners	Procedures	Number of Partners
Orders only	19	Orders only	16
Billing only	18	Billing only	20
Both	23	Both	24

Structure of results

- Summarize each condition by co-morbidity, procedure, medication, and lab, stratified by encounter type and provenance at each Data Partner by year
  - Encounter types
    - Telehealth is grouped into the "Other" encounter category (with Observation Stays and Institutional Consults)
    - Within the Epic EHR, ED visits that lead to an inpatient admission are treated as a single encounter - in the PCORnet CDM, these are represented by the IP/EI encounter type
    - OA = other ambulatory – lab-only visits, refill encounters, etc.
  - Provenance values:
    - OD – Order / clinician-entered (i.e., entered into EHR)
    - BI – EHR billing
    - CL – Claim (not used in any of the responding DPs)
    - DR – Derived (not used in any of the responding DPs)
    - Missing
- Note – to remove readability, have removed rows that are not focus of discussion in some tables

Example finding – drop in Encounter type

Table 20: CKD Diagnosis Information at Data Partner 2												
	2016		2017		2018		2019		2020		2021	
Description	**N**	**Percent**	**N**	**Percent**	**N**	**Percent**	**N**	**Percent**	**N**	**Percent**	**N**	**Percent**
Number of Unique Patients	12297		13558		14617		16626		16600		20772	
Co-morbidity: Dyslipidemia	7662	62.3%	8523	62.9%	9371	64.1%	10880	65.4%	11163	67.3%	14446	69.6%
By encounter type	.		.		.		.		.		.	
AV - Ambulatory	6444	84.1%	7053	82.8%	7716	82.3%	9082	83.5%	9146	81.9%	12304	85.2%
ED – Emergency Dept	800	10.4%	49	0.6%	48	0.5%	61	0.6%	99	0.9%	341	2.4%
IP/EI – Emergency to Inpatient	3019	39.4%	3511	41.2%	3931	41.9%	4326	39.8%	4079	36.5%	4620	32.0%
OA – Other Ambulatory	533	7.0%	978	11.5%	1246	13.3%	1679	15.4%	2042	18.3%	2960	20.5%
Other	514	6.7%	560	6.6%	534	5.7%	642	5.9%	2338	20.9%	1537	10.6%
Missing	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
By provenance	.		.		.		.		.		.	
OD – Clinician / Ordered	2891	37.7%	2098	24.6%	2319	24.7%	2474	22.7%	3807	34.1%	5134	35.5%
BI - EHR Billing	7596	99.1%	8402	98.6%	9205	98.2%	10690	98.3%	10818	96.9%	14006	97.0%

Drop in the ED encounter type may be due to uptake of IP/EI encounter type

Example finding – provenance rates differ by co-morbidity

Table 19: CKD Diagnosis Information at Data Partner 1

	2016		2017		2018		2019		2020		2021	
	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
Number of Unique Patients	10279		10974		11896		12614		12674		13007	
Anemia	5210		5690		6283		6582		6609		6999	
By provenance	.		.		.		.		.		.	
OD	3292	63.2%	3935	69.2%	4299	68.4%	4446	67.5%	4409	66.7%	4660	66.6%
BI	4934	94.7%	5397	94.9%	5973	95.1%	6329	96.2%	6287	95.1%	6713	95.9%
Pericarditis	394		420		464		526		613		651	
By provenance	.		.		.		.		.		.	
OD	83	21.1%	83	19.8%	91	19.6%	105	20.0%	82	13.4%	125	19.2%
BI	388	98.5%	418	99.5%	458	98.7%	522	99.2%	607	99.0%	643	98.8%
Uremic Bleeding	778		901		858		887		809		828	
By provenance	.		.		.		.		.		.	
OD	597	76.7%	693	76.9%	664	77.4%	680	76.7%	584	72.2%	594	71.7%
BI	707	90.9%	804	89.2%	787	91.7%	810	91.3%	750	92.7%	762	92.0%
Metabolic Acidosis	987		1207		1276		1596		1783		1868	
By provenance	.		.		.		.		.		.	
OD	37	3.7%	41	3.4%	43	3.4%	41	2.6%	59	3.3%	64	3.4%
BI	973	98.6%	1192	98.8%	1260	98.7%	1573	98.6%	1760	98.7%	1845	98.8%

Very different rates of capture within the ”OD” provenance tag. While there appears to be high levels of completeness for billing data, if Data Partners do not have access to those data, there could be significant missingness.

Example finding – lower than expected rates of completeness for billed diagnoses

Table 22: COPD Diagnosis Information at Data Partner 1

	2016		2017		2018		2019		2020		2021	
	N	%	N	%	N	%	N	%	N	%	N	%
Number of Unique Patients	8776		8893		9454		10159		9267		9217	
Asthma	3140	35.8%	2904	32.7%	2187	23.1%	1985	19.5%	1771	19.1%	1637	17.8%
By provenance	.		.		.		.		.		.	
OD	1477	47.0%	1318	45.4%	1222	55.9%	1243	62.6%	1124	63.5%	992	60.6%
BI	3028	96.4%	2818	97.0%	2074	94.8%	1865	94.0%	1630	92.0%	1515	92.5%
Missing	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%

Table 23: COPD Diagnosis Information at Data Partner 2

	2016		2017		2018		2019		2020		2021	
	N	%	N	%	N	%	N	%	N	%	N	%
Number of Unique Patients	9556		10095		10519		10901		10271		11192	
Asthma	2003	21.0%	1823	18.1%	1349	12.8%	1342	12.3%	1269	12.4%	1495	13.4%
By provenance	.		.		.		.		.		.	
OD	367	18.3%	463	25.4%	508	37.7%	520	38.7%	601	47.4%	663	44.3%
BI	1960	97.9%	1753	96.2%	1236	91.6%	1214	90.5%	1105	87.1%	1320	88.3%
Missing	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%

Rates for the "BI" provenance value are lower than usual for DPs 1 and 2 (88-92% instead of ~99%) – may be missing patients, which could have ramifications for billing-only DPs like DP3.

Example finding – variable rates of provenance for procedure data

Table 32: CKD Procedure Information at Data Partner 2

	2016		2017		2018		2019		2020		2021	
	N	%	N	%	N	%	N	%	N	%	N	%
Number of Unique Patients	12297		13558		14617		16626		16600		20772	
Biopsy, Kidney	145		156		140		171		157		199	
By provenance	.		.		.		.		.		.	
OD	91	63%	101	65%	73	52%	92	54%	79	50%	110	55%
BI	123	85%	131	84%	121	86%	145	85%	135	86%	172	86%
Missing	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%
Dialysis	1249		1295		1257		1409		1377		1437	
By provenance	.		.		.		.		.		.	
OD	1071	86%	1071	83%	1063	84%	1207	86%	1198	87%	1224	85%
BI	1241	99%	1290	99%	1251	99%	1396	99%	1370	99%	1422	99%
Missing	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%
Pyelography	240		275		283		296		306		405	
By provenance	.		.		.		.		.		.	
OD	182	76%	218	79%	230	81%	229	77%	230	75%	322	79%
BI	192	80%	215	78%	202	71%	231	78%	230	75%	302	75%
Missing	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%

Table 38: PAH Procedure Information at Data Partner 2

	2016		2017		2018		2019		2020		2021	
	N	%	N	%	N	%	N	%	N	%	N	%
Number of Unique Patients	1370		1493		2208		2485		2257		2476	
Echocardiograph	940		1060		1571		1759		1497		1670	67%
By provenance	.		.		.		.		.		.	
OD	587	62%	671	63%	1069	68%	1153	66%	1015	68%	1071	64%
BI	760	81%	832	79%	1151	73%	1256	71%	1101	74%	1261	76%
Missing	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%
Ventilation–Perfusion (VQ) Scan	152		204		277		270		211		240	
By provenance	.		.		.		.		.		.	
OD	40	26%	97	48%	118	43%	109	40%	83	39%	108	45%
BI	150	99%	204	100%	275	99%	267	99%	210	99%	238	99%
Missing	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%

Rates for "BI" are low for certain procedures, but not others. Some records may be missing for DP2; alternatively, DPs with only billing data may be missing records.

Example finding – comparing ratio of administrations (MED_ADMIN) to orders (PRESCRIBING)

Percent ratio of CKD meds (MED_ADMIN / PRESCRIBING) at Data Partner 2						
	2016	2017	2018	2019	2020	2021
ACE Inhibitor	35.9%	31.8%	31.1%	28.5%	24.7%	22.0%
Amlodipine	42.3%	42.3%	39.0%	37.1%	34.3%	31.2%
ARBs	27.4%	27.7%	26.3%	25.8%	23.1%	22.4%
ESAs	94.3%	127.4%	118.5%	109.2%	96.5%	92.8%
Iron	97.4%	87.1%	52.4%	64.1%	43.0%	27.3%
Loop Diuretics	54.1%	53.7%	54.2%	53.2%	51.4%	49.7%
Non-specific Beta Blockers	48.0%	47.6%	47.1%	44.7%	40.4%	39.2%
SGL2T Inhibitors	0.0%	3.7%	3.8%	1.2%	12.2%	16.6%

Expect rates to be relatively stable over time.

Big changes may indicate dropped records, though it could also be influenced by changes in practice patterns (e.g., first dose of a med no longer administered during a clinic visit).

Example finding – potential missing medication records

Percent ratio of COPD meds (MED_ADMIN / PRESCRIBING) at Data Partner 2						
	2016	2017	2018	2019	2020	2021
Aminoglycoside	53%	61%	39%	24%	15%	16%
Anti-pseudomonal Penicillin	4210%	1733%	453%	798%	5746%	9313%
BA Agonists	41%	41%	41%	41%	33%	32%
Cephalosporins	312%	455%	937%	1118%	1259%	318%
First Line Antibiotics	28%	24%	27%	27%	27%	26%

- Extreme differences are likely due to missed records (e.g., not loaded into CDM, utilizing codes that were not part of the query *[same query codes used for both tables, however]*).
- Comparing % of patients receiving the medications between DP1 & DP2, MED_ADMIN percentages look to be closer in line than PRESCRIBING.

Table 48: COPD Medication Prescription Information at Data Partner 1

	2016		2017		2018		2019		2020		2021	
	N	%	N	%	N	%	N	%	N	%	N	%
Number of Unique Patients	8776		8893		9454		10159		9267		9217	
Anti-pseudomonal Penicillin	648	7%	938	11%	893	10%	830	8%	645	7%	593	6%
Cephalosporins	1420	16%	1443	16%	1498	16%	1872	18%	1709	18%	1624	18%

Table 49: COPD Medication Prescription Information at Data Partner 2

	2016		2017		2018		2019		2020		2021	
	N	%	N	%	N	%	N	%	N	%	N	%
Number of Unique Patients	9556		10095		10519		10901		10271		11192	
Anti-pseudomonal Penicillin	21	0%	48	0%	196	2%	97	1%	13	0%	8	0%
Cephalosporins	249	3%	181	2%	97	1%	90	1%	66	1%	281	3%

Table 60: COPD Medication Administration Information at Data Partner 1

	2016		2017		2018		2019		2020		2021	
	N	%	N	%	N	%	N	%	N	%	N	%
Number of Unique Patients	8776		8893		9454		10159		9267		9217	
Anti-pseudomonal Penicillin	629	7%	911	10%	872	9%	808	8%	624	7%	578	6%
Cephalosporins	1366	16%	1404	16%	1460	15%	1823	18%	1678	18%	1585	17%

Table 61: COPD Medication Administration Information at Data Partner 2

	2016		2017		2018		2019		2020		2021	
	N	%	N	%	N	%	N	%	N	%	N	%
Number of Unique Patients	9556		10095		10519		10901		10271		11192	
Anti-pseudomonal Penicillin	884	9%	832	8%	888	8%	774	7%	747	7%	745	7%
Cephalosporins	778	8%	824	8%	909	9%	1006	9%	831	8%	894	8%

Example finding – changes in encounter types for labs

Table 77: PAH Lab Information at Data Partner 3

	2016		2017		2018		2019		2020		2021	
Description	N	%	N	%	N	%	N	%	N	%	N	%
Number of Unique Patients	787		1189		2454		2737		2743		3286	
ALP	587		991		1990		2225		2184		2711	
By encounter type	.		.		.		.		.		.	
AV	0	0%	63	6%	927	47%	1437	65%	1518	70%	2163	80%
ED	0	0%	23	2%	253	13%	388	17%	413	19%	510	19%
IP/EI	0	0%	74	8%	856	43%	1243	56%	1268	58%	1561	58%
OA	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%
Other	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%
Missing	587	100%	961	97%	1346	68%	783	35%	524	24%	106	4%
By provenance	.		.		.		.		.		.	
OD	587	100%	991	100%	1983	100%	2207	99%	2158	99%	2707	100%
Missing	0	0%	0	0%	7	0%	18	1%	29	1%	53	2%
Creatinine, Serum	737		1161		2400		2668		2635		3185	
By encounter type	.		.		.		.		.		.	
AV	0	0%	145	13%	1591	66%	2251	84%	2311	88%	3014	95%
ED	0	0%	43	4%	463	19%	679	25%	680	26%	852	27%
IP/EI	0	0%	99	89%	1233	51%	1721	65%	1712	65%	2060	65%
OA	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%
Other	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%
Missing	737	100%	1151	99%	1924	80%	1149	43%	846	32%	193	6%
By provenance	.		.		.		.		.		.	
OD	737	100%	1161	100%	2391	100%	2647	99%	2606	99%	3182	100%
Missing	0	0%	0	0%	9	0%	21	1%	32	1%	60	2%

Provenance is less of an issue with labs (all tend to be OD), but information on encounter type can inform analytical strategy (e.g., looking for records by date instead of by encounter type).

Example finding – examining rates of test completion

Table 75: PAH Lab Information at Data Partner 1

	2017		2019		2021	
	N	%	N	%	N	%
Number of Unique Patients	986		1438		1401	
ALP	797	81%	1163	81%	1151	82%
ALT	798	81%	1165	81%	1153	82%
Bilirubin	800	81%	1168	81%	1155	82%
BNP	263	27%	527	37%	518	37%
BUN	882	90%	1270	88%	1252	89%
Calcium	876	89%	1272	89%	1252	89%
Chloride BSP	880	89%	1271	89%	1252	89%
Creatinine Blood	86	9%	170	12%	151	11%
Creatinine Serum	881	89%	1277	899%	1255	90%
Glucose	886	90%	1274	89%	1260	90%
HCG	0	0%	0	0%	0	0%
HCG Presence	0	0%	0	0%	0	0%
HCG Urine Presence	0	0%	0	0%	0	0%
HCG Urine PT	0	0%	0	0%	0	0%
HCG Urine Time	0	0%	0	0%	0	0%
Hematocrit	857	87%	1239	86%	1217	87%
Platelets	848	86%	1238	86%	1216	87%
Potassium	880	89%	1271	88%	1252	89%
Sodium	880	89%	1271	88%	1252	89%
Total CO2	881	89%	1270	88%	1252	89%
TSH	519	53%	760	53%	745	53%

Table 77: PAH Lab Information at Data Partner 3

	2017		2019		2021	
	N	%	N	%	N	%
Number of Unique Patients	1189		2737		3286	
ALP	991	83%	2225	81%	2711	83%
ALT	963	81%	2228	81%	2709	82%
Bilirubin	978	82%	2212	81%	2701	82%
BNP	245	21%	196	7%	264	8%
BUN	1065	90%	2554	93%	3057	93%
Calcium	994	84%	2542	93%	3036	92%
Chloride BSP	921	78%	2464	90%	2986	91%
Creatinine Blood	128	11%	321	12%	317	10%
Creatinine Serum	1161	98%	2668	98%	3185	97%
Glucose	972	82%	2473	90%	2986	91%
HCG	7	1%	48	2%	76	2%
HCG Presence	25	2%	27	1%	28	1%
HCG Urine Presence	29	2%	37	1%	55	2%
HCG Urine PT	0	0%	0	0%	0	0%
HCG Urine Time	0	0%	0	0%	0	0%
Hematocrit	1139	96%	2591	95%	3090	94%
Platelets	1138	96%	2600	95%	3089	94%
Potassium	1145	96%	2644	97%	3172	97%
Sodium	1146	96%	2638	96%	3167	96%
Total CO2	408	34%	200	7%	228	7%
TSH	818	69%	1685	62%	2029	62%

Percentages can help identify missing labs (if 0%), or if there are differences in labs that should be ordered together as part of a panel (e.g., Hematocrit & Platelets)

Summary

- While initial queries were limited to a handful of Data Partners, results surfaced several findings
- Mapping issues exist in CDM data – assessing harmonization can help identify mapping errors, *but need to set expectations that Data Partners properly populate RAW fields*
- Having access to multiple streams of data provenance can help raise potential data issues, *but only if they are present in the CDM*
- Many Data Partners do not populate all data streams, so efforts could be made to raise awareness of their potential importance
- In the meantime, findings about potential gaps can help study teams as they assess data with single streams of provenance (e.g., billing only, or orders only)
- Condition-specific breakdowns can also help inform analysis plans by surfacing changes in rates of encounter types over time

Project Team

Duke

- Keith Marsolo
- Lesley Curtis
- Larry Hill
- Jennifer Xu
- Gretchen Sanders
- Laura Qualls
- Yinghong Zhang
- Tom Phillips

Harvard Pilgrim Healthcare Institute

- Judy Maro
- Kevin Coughlin
- Daniel Kiernan
- Christine Draper

Additional PCORnet Data Partners

- Alanna Chamberlain (Mayo Clinic)
- Jiang Bian (U of Florida)

FDA

- Sara Dutcher
- Jose Hernandez
- Monique Falconer

Questions?