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

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.
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.

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

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

The Requestor's question is received by PCORnet Leadership, who review the question and consult with the Requestor about next steps.

The Coordinating Center converts the request into a query and sends it to Network partners for review.

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

- ADVANCE Network
- CAPriCORN
- GPC
- REACHnet
- PRACnet
- PaTH
- INSIGHT - NYC
- PEDSnet
- HealthCore
- OneFlorida
- STAR
- HealthCore
- HealthCore
## Domains within the PCORnet Common Data Model

<table>
<thead>
<tr>
<th>Ready for Research</th>
<th>Available, But Still Evolving</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death Data</td>
<td>Geocodes</td>
</tr>
<tr>
<td>Diagnoses</td>
<td>Social Determinants of Health</td>
</tr>
<tr>
<td>Medication Orders</td>
<td>Tumor Registry</td>
</tr>
<tr>
<td>Claims</td>
<td>Biosamples</td>
</tr>
<tr>
<td>Labs</td>
<td>Patient-Reported Outcomes</td>
</tr>
<tr>
<td>Demographics</td>
<td>Genomic Results</td>
</tr>
<tr>
<td>Procedures</td>
<td>Natural Language Processing Derived Concepts</td>
</tr>
</tbody>
</table>

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.

[pcornet logo]
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

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

- **Reliability – data assurance**
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

- 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-ADOPPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOUNTER
- EB PLANNING
- FEES/VOICE
- HEPATOBLASTOMA LIVER TRANSPLANT FOLLOW UP
- PRE-ADOPTION ENCOURAGE
## Harmonization examples - Lab results

<table>
<thead>
<tr>
<th>LOINC</th>
<th>LongName</th>
<th>Component</th>
<th>Property</th>
<th>Timing</th>
<th>System</th>
<th>Scale</th>
<th>Method</th>
<th>xUCUMunits</th>
<th>exUnits</th>
<th>Lfo</th>
</tr>
</thead>
<tbody>
<tr>
<td>48035-0</td>
<td>Hemoglobin [Presence] in Cerebral spinal fluid</td>
<td>Hemoglobin</td>
<td>PrThr</td>
<td>Pt</td>
<td>CSF</td>
<td>Ord</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>725-2</td>
<td>Hemoglobin [Presence] in Urine</td>
<td>Hemoglobin</td>
<td>PrThr</td>
<td>Pt</td>
<td>Urine</td>
<td>Ord</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5794-3</td>
<td>Hemoglobin [Presence] in Urine by Test strip</td>
<td>Hemoglobin</td>
<td>PrThr</td>
<td>Pt</td>
<td>Urine</td>
<td>Ord</td>
<td>Test strip</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>57751-0</td>
<td>Hemoglobin [Presence] in Urine by Automated test strip</td>
<td>Hemoglobin</td>
<td>PrThr</td>
<td>Pt</td>
<td>Urine</td>
<td>Ord</td>
<td>Test strip.automated</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34618-9</td>
<td>Hemoglobin [Presence] in Unspecified specimen</td>
<td>Hemoglobin</td>
<td>PrThr</td>
<td>Pt</td>
<td>XXX</td>
<td>Ord</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>73995-5</td>
<td>Hemoglobin [Entitic substance] in Reticulocytes by Automated count</td>
<td>Hemoglobin</td>
<td>EntSub</td>
<td>Pt</td>
<td>Retic</td>
<td>Qn</td>
<td>Automated count</td>
<td>fmol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>76766-1</td>
<td>Hemoglobin [Mass/volume] in Mixed venous blood by Oximetry</td>
<td>Hemoglobin</td>
<td>MCnc</td>
<td>Pt</td>
<td>BlodMV</td>
<td>Qn</td>
<td>Oximetry</td>
<td>g/L</td>
<td>g/L</td>
<td></td>
</tr>
<tr>
<td>76768-9</td>
<td>Hemoglobin [Mass/volume] in Venous blood by Oximetry</td>
<td>Hemoglobin</td>
<td>MCnc</td>
<td>Pt</td>
<td>BlodV</td>
<td>Qn</td>
<td>Oximetry</td>
<td>g/L</td>
<td>g/L</td>
<td></td>
</tr>
<tr>
<td>69950-4</td>
<td>Hemoglobin [Mass/volume] in Pericardial fluid</td>
<td>Hemoglobin</td>
<td>MCnc</td>
<td>Pt</td>
<td>Pericard flid</td>
<td>Qn</td>
<td>Oximetry</td>
<td>g/L</td>
<td>g/L</td>
<td></td>
</tr>
<tr>
<td>718-7</td>
<td>Hemoglobin [Mass/volume] in Blood</td>
<td>Hemoglobin</td>
<td>MCnc</td>
<td>Pt</td>
<td>Blod</td>
<td>Qn</td>
<td></td>
<td>g/dL</td>
<td>g/dL</td>
<td></td>
</tr>
<tr>
<td>20509-6</td>
<td>Hemoglobin [Mass/volume] in Blood by calculation</td>
<td>Hemoglobin</td>
<td>MCnc</td>
<td>Pt</td>
<td>Blod</td>
<td>Qn</td>
<td>Calculated</td>
<td>g/dL</td>
<td>g/dL</td>
<td></td>
</tr>
<tr>
<td>42243-8</td>
<td>Deprecated Hemoglobin [Mass/volume] in Blood</td>
<td>Hemoglobin</td>
<td>MCnc</td>
<td>Pt</td>
<td>Blod</td>
<td>Qn</td>
<td>HPLC</td>
<td>g/dL</td>
<td>g/dL</td>
<td></td>
</tr>
<tr>
<td>95782-7</td>
<td>Hemoglobin [Mass/volume] in Blood by Oximetry</td>
<td>Hemoglobin</td>
<td>MCnc</td>
<td>Pt</td>
<td>Blod</td>
<td>Qn</td>
<td>Oximetry</td>
<td>g/dL</td>
<td>g/dL</td>
<td></td>
</tr>
<tr>
<td>54289-4</td>
<td>Hemoglobin [Mass/volume] in Blood from Blood product unit</td>
<td>Hemoglobin</td>
<td>MCnc</td>
<td>Pt</td>
<td>Blod*BPU</td>
<td>Qn</td>
<td></td>
<td>g/dL</td>
<td>g/dL</td>
<td></td>
</tr>
<tr>
<td>61180-6</td>
<td>Hemoglobin [Mass/volume] in Blood from Fetus</td>
<td>Hemoglobin</td>
<td>MCnc</td>
<td>Pt</td>
<td>Blod+Fetus</td>
<td>Qn</td>
<td></td>
<td>g/dL</td>
<td>g/L</td>
<td></td>
</tr>
<tr>
<td>30312-1</td>
<td>Hemoglobin [Mass/volume] in Arterial blood</td>
<td>Hemoglobin</td>
<td>MCnc</td>
<td>Pt</td>
<td>BlodA</td>
<td>Qn</td>
<td></td>
<td>g/dL</td>
<td>g/dL</td>
<td></td>
</tr>
<tr>
<td>14775-1</td>
<td>Hemoglobin [Mass/volume] in Arterial blood by Oximetry</td>
<td>Hemoglobin</td>
<td>MCnc</td>
<td>Pt</td>
<td>BlodA</td>
<td>Qn</td>
<td>Oximetry</td>
<td>g/dL</td>
<td>g/L</td>
<td></td>
</tr>
</tbody>
</table>
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 data checks - Conformance

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

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

<table>
<thead>
<tr>
<th>Type</th>
<th>Check</th>
<th>Description</th>
<th>Cycle Added</th>
</tr>
</thead>
<tbody>
<tr>
<td>Investigative</td>
<td>DC 3.01</td>
<td>Average # of diagnoses with known diagnosis type per encounter is below threshold</td>
<td>1</td>
</tr>
<tr>
<td>Investigative</td>
<td>DC 3.02</td>
<td>Average # of procedures with known procedure type per encounter is below threshold</td>
<td>1</td>
</tr>
<tr>
<td>Investigative</td>
<td>DC 3.03</td>
<td>More than 10% of records have missing/unknown values for selected fields</td>
<td>1</td>
</tr>
<tr>
<td>Required</td>
<td>DC 3.04</td>
<td>Less than 50% of patients with encounters have DIAGNOSIS records</td>
<td>2</td>
</tr>
<tr>
<td>Required</td>
<td>DC 3.05</td>
<td>Less than 50% of patients with encounters have PROCEDURES records</td>
<td>2</td>
</tr>
<tr>
<td>Investigative</td>
<td>DC 3.06</td>
<td>More than 10% of IP &amp; EI encounters with a diagnosis are missing principal diagnosis</td>
<td>2</td>
</tr>
<tr>
<td>Investigative</td>
<td>DC 3.07</td>
<td>DX, PX, &amp; encounter records in AV, ED, EI, IP setting are &lt;75% complete 3 months prior to current month</td>
<td>3</td>
</tr>
<tr>
<td>Investigative</td>
<td>DC 3.08</td>
<td>Less than 80% of prescribing orders mapped to a Tier 1 RXCUI (encodes ingredient, strength, &amp; dose form)</td>
<td>3</td>
</tr>
<tr>
<td>Investigative</td>
<td>DC 3.09</td>
<td>Less than 80% of lab results mapped to LOINC</td>
<td>3</td>
</tr>
<tr>
<td>Investigative</td>
<td>DC 3.10</td>
<td>Less than 80% of quantitative lab results specify the normal range</td>
<td>3</td>
</tr>
<tr>
<td>Investigative</td>
<td>DC 3.11</td>
<td>Vital, Rx, Lab records are &lt;75% complete 3 months prior to current month</td>
<td>4</td>
</tr>
<tr>
<td>Investigative</td>
<td>DC 3.12</td>
<td>Less than 80% of quantitative lab results mapped to LOINC specify SPECIMEN_SOURCE &amp; RESULT_UNIT</td>
<td>5</td>
</tr>
<tr>
<td>Investigative</td>
<td>DC 3.13</td>
<td>The percentage of patients with selected lab tests is below threshold</td>
<td>8</td>
</tr>
<tr>
<td>Type</td>
<td>Check</td>
<td>Description</td>
<td>Cycle Added</td>
</tr>
<tr>
<td>------------</td>
<td>-------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Investigative</td>
<td>DC 4.01</td>
<td>More than a 5% decrease in the number of patients or records in a CDM table</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>DC 4.02</td>
<td>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.</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>DC 4.03</td>
<td>More than a 5% decrease in the number of records for ICD9 or ICD10 diagnosis or procedure codes or CPT/HCPCS procedure codes.</td>
<td>6</td>
</tr>
</tbody>
</table>
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
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.

* 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. *
Results of selected completeness measures
## Data persistence

### Persistence Measures

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC 4.01</td>
<td>More than a 5% decrease in the number of patients or records in a CDM table.</td>
</tr>
<tr>
<td>DC 4.02</td>
<td>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.</td>
</tr>
<tr>
<td>DC 4.03</td>
<td>More than a 5% decrease in the number of records for ICD9 or ICD10 diagnosis or procedure codes or CPT/HCPCS procedure codes.</td>
</tr>
</tbody>
</table>

- **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
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

### ENCOUNTERR Table Implementation Guidance

- Each ENCOUNTER will generally reflect a unique combination of PATID, ADMIT_DATE, PROVIDERID and ENC_TYPE.
- Every diagnosis and procedure recorded during the encounter should be entered 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 rendered 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.
- Rebills or voided transactions and other adjustments should be processed before populating this table.
- Although “at risk” 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 in the appropriate encounter type, which is typically “OA” for outpatient visits.

### ENCOUNTERR Table Specification

<table>
<thead>
<tr>
<th>Field Name</th>
<th>RDBMS Data Type</th>
<th>SAS Data Type</th>
<th>Predefined Value Sets and Descriptive Text for Categorical Fields</th>
<th>Definition / Comments</th>
<th>Data Element Provenance</th>
<th>Field-level Implementation Guidance</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENCOUNTERID</td>
<td>RDBMS Text(3)</td>
<td>SAS Char(1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PATID</td>
<td>RDBMS</td>
<td>SAS Char(1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### DIAGNOSIS Table Specification

<table>
<thead>
<tr>
<th>Field Name</th>
<th>RDBMS Data Type</th>
<th>SAS Data Type</th>
<th>Predefined Value Sets and Descriptive Text for Categorical Fields</th>
<th>Definition / Comments</th>
<th>Data Element Provenance</th>
<th>Field-level Implementation Guidance</th>
</tr>
</thead>
<tbody>
<tr>
<td>DX_ORIGIN</td>
<td>RDBMS Text(2)</td>
<td>SAS Char(2)</td>
<td>OD=Order</td>
<td>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.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
|                |                 |               | BI=Billing                                                          |                       |                         | [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.]
|                |                 |               | CL=Claim                                                            |                       |                         |                                      |
|                |                 |               | NI=No information                                                   |                       |                         |                                      |
|                |                 |               | UN=Unknown                                                          |                       |                         |                                      |
|                |                 |               | OT=Other                                                            |                       |                         |                                      |
Impact of Studies – Prescribing

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

<table>
<thead>
<tr>
<th>IN/IN</th>
<th>Ingredient</th>
<th>Precise Ingredient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetaminophen</td>
<td>HYDROcodone Bitartrate</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BN</th>
<th>Brand Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lorcet</td>
<td></td>
</tr>
<tr>
<td>Lortab</td>
<td></td>
</tr>
<tr>
<td>Norco</td>
<td></td>
</tr>
<tr>
<td>Xadol</td>
<td></td>
</tr>
</tbody>
</table>

### Navigating RxNorm Drugs

<table>
<thead>
<tr>
<th>SCDC</th>
<th>Clinical Drug Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetaminophen 325 MG</td>
<td></td>
</tr>
<tr>
<td>HYDROcodone Bitartrate 10 MG</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SCD/SPCK</th>
<th>Clinical Drug or Pack</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetaminophen 325 MG / HYDROcodone Bitartrate 10 MG Oral Tablet</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SCDG</th>
<th>Clinical Dose Form Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetaminophen / HYDROcodone Oral Product</td>
<td></td>
</tr>
<tr>
<td>Acetaminophen / HYDROcodone Pill</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DFG</th>
<th>Dose Form Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oral Product</td>
<td></td>
</tr>
<tr>
<td>Pill</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SBD/BPCK</th>
<th>Branded Drug or Pack</th>
</tr>
</thead>
<tbody>
<tr>
<td>APAP 325 MG / HYDROcodone Bitartrate 10 MG Oral Tablet [Lorcet]</td>
<td></td>
</tr>
<tr>
<td>Norco 10/325 (HYDROcodone / APAP) Oral</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SBDG</th>
<th>Branded Dose Form Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lorcet Oral Product</td>
<td></td>
</tr>
<tr>
<td>Lorcet Pill</td>
<td></td>
</tr>
<tr>
<td>Lortab Oral Product</td>
<td></td>
</tr>
<tr>
<td>Lortab Pill</td>
<td></td>
</tr>
</tbody>
</table>

### Chart Illustration

A chart illustrating the distribution of prescriptions for the mentioned drug, showing a clear trend over time.
Impact of Studies – Prescribing (2)

Variability in prescribing data led to updates in IG

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.**

---

### Implementation Guidance Reference Table 4: Ordering of RxNorm Term Types


<table>
<thead>
<tr>
<th>RxNorm Term Type</th>
<th>Information incorporated</th>
<th>Code</th>
<th>Description</th>
<th>Ingredient(s)</th>
<th>Strength</th>
<th>Base Form</th>
<th>Brand Name</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most Preferred</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SBD, SBDG, SBDG, SBDCC, IN</td>
<td></td>
<td>X</td>
<td>Semisynthetic Drug</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCB, SCBCC</td>
<td></td>
<td>X</td>
<td>Semisynthetic Clinical Drug</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCD</td>
<td></td>
<td>X</td>
<td>Semisynthetic Drug Component</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDC</td>
<td></td>
<td>X</td>
<td>Semisynthetic Clinical Drug Component</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDF</td>
<td></td>
<td>X</td>
<td>Semisynthetic Drug Form</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDG</td>
<td></td>
<td>X</td>
<td>Semisynthetic Drug Form Group</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDCC</td>
<td></td>
<td>X</td>
<td>Semisynthetic Clinical Drug Form</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDCC</td>
<td></td>
<td>X</td>
<td>Semisynthetic Clinical Drug Component</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDF</td>
<td></td>
<td>X</td>
<td>Semisynthetic Drug Form</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDG</td>
<td></td>
<td>X</td>
<td>Semisynthetic Drug Form Group</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDCC</td>
<td></td>
<td>X</td>
<td>Semisynthetic Clinical Drug Form</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDCC</td>
<td></td>
<td>X</td>
<td>Semisynthetic Clinical Drug Component</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDF</td>
<td></td>
<td>X</td>
<td>Semisynthetic Drug Form</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDG</td>
<td></td>
<td>X</td>
<td>Semisynthetic Drug Form Group</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDCC</td>
<td></td>
<td>X</td>
<td>Semisynthetic Clinical Drug Form</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDCC</td>
<td></td>
<td>X</td>
<td>Semisynthetic Clinical Drug Component</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDF</td>
<td></td>
<td>X</td>
<td>Semisynthetic Drug Form</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDG</td>
<td></td>
<td>X</td>
<td>Semisynthetic Drug Form Group</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDCC</td>
<td></td>
<td>X</td>
<td>Semisynthetic Clinical Drug Form</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDCC</td>
<td></td>
<td>X</td>
<td>Semisynthetic Clinical Drug Component</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDF</td>
<td></td>
<td>X</td>
<td>Semisynthetic Drug Form</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDG</td>
<td></td>
<td>X</td>
<td>Semisynthetic Drug Form Group</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDCC</td>
<td></td>
<td>X</td>
<td>Semisynthetic Clinical Drug Form</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDCC</td>
<td></td>
<td>X</td>
<td>Semisynthetic Clinical Drug Component</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDF</td>
<td></td>
<td>X</td>
<td>Semisynthetic Drug Form</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDG</td>
<td></td>
<td>X</td>
<td>Semisynthetic Drug Form Group</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDCC</td>
<td></td>
<td>X</td>
<td>Semisynthetic Clinical Drug Form</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDCC</td>
<td></td>
<td>X</td>
<td>Semisynthetic Clinical Drug Component</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDF</td>
<td></td>
<td>X</td>
<td>Semisynthetic Drug Form</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDG</td>
<td></td>
<td>X</td>
<td>Semisynthetic Drug Form Group</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDCC</td>
<td></td>
<td>X</td>
<td>Semisynthetic Clinical Drug Form</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDCC</td>
<td></td>
<td>X</td>
<td>Semisynthetic Clinical Drug Component</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDF</td>
<td></td>
<td>X</td>
<td>Semisynthetic Drug Form</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDG</td>
<td></td>
<td>X</td>
<td>Semisynthetic Drug Form Group</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDCC</td>
<td></td>
<td>X</td>
<td>Semisynthetic Clinical Drug Form</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDCC</td>
<td></td>
<td>X</td>
<td>Semisynthetic Clinical Drug Component</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCDF</td>
<td></td>
<td>X</td>
<td>Semisynthetic Drug Form</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.
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 I. 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 480 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 (http://comet.merckcentral.com/pqQAAAC/jp9I).

Table II. 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 diagnoses are expected to be represented in most DataMarts.

Table III. 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.
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.
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