National Patient-Centered Clinical Research Network (PCORnet) Distributed Research Network (DRN)

Questions Answered from Preparatory-to-Research to Comparative Effectiveness S03

Jessica L. Sturtevant, MS¹, Darcy Louzao, PhD², Keith Marsolo, PhD², Kathleen McTigue, MD, MPH, MS⁴, Lesley H. Curtis, PhD²

¹Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA
²Duke Clinical Research Institute, Durham, NC
³University of Pittsburgh Medical Center, Pittsburgh, PA

#AMIA2018
Disclosure

We have no relevant relationships with commercial interests to disclose.
Learning Objectives

After participating in this session the learner should be better able to:

• Describe the PCORnet Coordinating Center (CC) role in answering questions in the Distributed Research Network (DRN)

• Describe the PCORnet Front Door, Common Data Model (CDM), and analytics role in supporting both preparatory-to-research (PTR) and observational research

• Describe how researchers can bring questions to the PCORnet Front Door leading to PTR network inquiries and more complex research
Goal of PCORnet

Data from diverse health systems

Standardize data into PCORnet Common Data Model

PCORnet CDM

Research Study

PCORnet CDM

PCORnet CDM

Link to Supplemental Data
PCORnet Terminology

- Clinical Data Research Network (CDRN) can consist of 1 or multiple DataMarts
- DataMarts are the unit of query
- DataMarts can consist of 1 or multiple sites/contributing health systems data
Panel Overview

Keith Marsolo: Common Data Model & Data Curation

Darcy Louzao: Front Door

Jessica Sturtevant: Analytics & Query Fulfillment

Kathleen McTigue: PCORnet Bariatric Study
Keith Marsolo: Common Data Model & Data Curation
Evolution of the PCORnet Common Data Model (CDM)

Have focused on domains most “common” (and available) in electronic health records & claims sources

- Initial structure based on Sentinel CDM to support repurposing of analytic tools
- Version 4.0 added more ”general purpose” observation tables – allow partners to use them for internal projects
- Version 4.0 also defined concept of “external” table definitions to help partners pilot certain activities like data linkage in a more standardized fashion
- Encourage additional pilots in other domains of interest (e.g., tumor information, devices, social determinants of health)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DEMOGRAPHIC</td>
<td>CONDITION</td>
<td>DEATH</td>
<td>PRO_CM*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIAGNOSIS</td>
<td>DISPENSING</td>
<td>DEATH_CAUSE</td>
<td>MED_ADMIN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENROLLMENT</td>
<td>LAB_RESULT_CM</td>
<td>PRESCRIBING</td>
<td>PROVIDER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENCOUNTER</td>
<td>PRO_CM</td>
<td>HARVEST</td>
<td>OBS_CLIN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PROCEDURES</td>
<td>PCORENET_TRIAL</td>
<td></td>
<td>OBS_GEN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VITAL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tables introduced into the CDM in each release
Activities related to the development of CDM v4.0

<table>
<thead>
<tr>
<th>Common Data Model</th>
<th>2016</th>
<th>2017</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>November</td>
<td>December</td>
</tr>
<tr>
<td>Release of CDM v3.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discuss strategy for CDM v4.0 with PCORnet Data Committee</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Assessing data quality - Data Curation

Purpose
• Evaluate data quality and fitness-for-use across a broad research portfolio
• Generate meaningful, actionable information for network partners, investigators and other stakeholders

Resources
• Implementation Guidance
• Data quality checks
• Data curation query packages
• Analyses and reports
• Discussion Forums
CDM Implementation Guidance – Reducing variation

Created to address 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?*

---

**ENCOUNTER Table Implementation 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.

---

**DEMOGRAPHIC Table Specification**

<table>
<thead>
<tr>
<th>Field Name</th>
<th>RDMS Data Type</th>
<th>SAS Data Type</th>
<th>Definition / Comments</th>
<th>Data Element Provenance</th>
</tr>
</thead>
<tbody>
<tr>
<td>HISPANIC</td>
<td>BDMS Test(2)</td>
<td>SAS Chart(D)</td>
<td>A person of Cuban, Mexican, Puerto Rican, South or Central American, or other Spanish culture or origin, regardless of race.</td>
<td>MSCDM v4.0 with modified field size and value set</td>
</tr>
</tbody>
</table>

Compatible with “OMB Hispanic Ethnicity” (Hispanic or Latino, Not Hispanic or Latino)
The evolution of data curation

Data Quality Checking Rules and Measures

<table>
<thead>
<tr>
<th>Cycle</th>
<th>Checks</th>
<th>Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cycle 1</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Cycle 2</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>Cycle 3</td>
<td>26</td>
<td>200</td>
</tr>
<tr>
<td>Cycle 4</td>
<td>27</td>
<td>400</td>
</tr>
<tr>
<td>Cycle 5</td>
<td>31</td>
<td>1200</td>
</tr>
</tbody>
</table>

Data Quality Checks | Data Quality Measures
## 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 ENCONTERTIDs</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>DC 1.10</td>
<td>Replication errors between ENCONTERT, 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>
</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>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 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>
</tbody>
</table>
Remediation of model conformance errors

- **Required tables not present**
  - Cycle 2: DC 1.01
  - Cycle 3: DC 1.02
  - Cycle 4: DC 1.03
  - Cycle 5: DC 1.04

- **Required fields not populated**
  - Cycle 2: DC 1.05
  - Cycle 3: DC 1.06
  - Cycle 4: DC 1.07
  - Cycle 5: DC 1.08

- **Required fields do not conform to CDM**
  - Cycle 2: DC 1.09
  - Cycle 3: DC 1.10
  - Cycle 4: DC 1.11
  - Cycle 5: DC 1.12

- **Primary key errors**
  - Cycle 2: DC 1.12
  - Cycle 3: DC 1.13
  - Cycle 4: DC 1.14
  - Cycle 5: DC 1.15

- **Required fields have values outside CDM spec.**
  - Cycle 2: DC 1.16
  - Cycle 3: DC 1.17
  - Cycle 4: DC 1.18
  - Cycle 5: DC 1.19

- **Required fields have non-permissible missing values**
  - Cycle 2: DC 1.20
  - Cycle 3: DC 1.21
  - Cycle 4: DC 1.22
  - Cycle 5: DC 1.23

- **Tables contain orphan PATIDS**
  - Cycle 2: DC 1.24
  - Cycle 3: DC 1.25
  - Cycle 4: DC 1.26
  - Cycle 5: DC 1.27

- **Tables have orphan ENOUNTERIDS**
  - Cycle 2: DC 1.28
  - Cycle 3: DC 1.29
  - Cycle 4: DC 1.30
  - Cycle 5: DC 1.31

- **Replication errors between ENOUNTER, DIAGNOSIS & PROCEDURES**
  - Cycle 2: DC 1.32
  - Cycle 3: DC 1.33
  - Cycle 4: DC 1.34
  - Cycle 5: DC 1.35

- **>5% of encounters assigned to a single PATID**
  - Cycle 2: DC 1.36
  - Cycle 3: DC 1.37
  - Cycle 4: DC 1.38
  - Cycle 5: DC 1.39

- **Tables have orphan PROVIDERIDs**
  - Cycle 2: DC 1.40
  - Cycle 3: DC 1.41
  - Cycle 4: DC 1.42
  - Cycle 5: DC 1.43
Data plausibility checks over time

- >5% of records have future dates
  - DC 2.01: 3.2%, 4.5%
- >10% of records have implausible values
  - DC 2.02: 9.5%, 7.5%
- >5% of patients with implausible date relationships
  - DC 2.03: 0.0%, 1.5%
- Average # of encounters per visit
  - DC 2.04: 7.1%, 1.5%

- >5% of lab results with inappropriate specimen source
  - DC 2.05: 2.2%, 3.9%
- Median lab result is statistical outlier
  - DC 2.06: 69.8%
- Average # of principal diagnoses above threshold
  - DC 2.07: 19.3%, 21.9%
Completeness checks over time

Note: DC 3.03 examined 10 fields at baseline & 16 in Cycle 5
Laboratory Results
Assessing lab data quality

LOINC is a very granular standard – a single lab concept like “hemoglobin” can be represented in many ways based on specimen, method, result units, etc.

The majority of these differences do not matter, for most analyses but LOINC does not provide an easy way to group related tests.
Data Curation Tests (‘‘Lab Groups’’)

Developed to allow aggregation of results of the same lab concept across DataMarts

Somewhat analogous to the now-deprecated LAB_NAME field, but with additional benefits

• Partners only need to worry about assigning a LOINC code to a given result
• Groups can be generated programmatically – no need to include an enumerated list in the CDM
Defining Tests / Labs Groups

Identified using the following LOINC attributes:

- Component (e.g. Cholesterol)
- System, a.k.a. Specimen (e.g. Blood, Serum, Plasma)
- If necessary….
  - Time (e.g. 24hr)
  - Method (e.g. Electrophoresis)
  - Class (e.g. Chem)

Cycle 3 – 20 Tests (mostly top labs by volume)

Cycle 4 (and beyond) – 490 Tests, representing labs of interest as specified by the PCORnet Collaborative Research Groups (CRGs)
### Creatinine Blood / Serum / Plasma test

<table>
<thead>
<tr>
<th>LOINC</th>
<th>COMPONENT</th>
<th>SYSTEM</th>
<th>EX_UCUM_UNITS</th>
<th>CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>14682-9</td>
<td>CREATININE</td>
<td>SER/PLAS</td>
<td>UMOL/L</td>
<td>CHEM</td>
</tr>
<tr>
<td>21232-4</td>
<td>CREATININE</td>
<td>BLDA</td>
<td>MG/DL</td>
<td>CHEM</td>
</tr>
<tr>
<td>2160-0</td>
<td>CREATININE</td>
<td>SER/PLAS</td>
<td>MG/DL</td>
<td>CHEM</td>
</tr>
<tr>
<td>35203-9</td>
<td>CREATININE</td>
<td>SER/PLAS</td>
<td>CHEM</td>
<td>CHEM</td>
</tr>
<tr>
<td>38483-4</td>
<td>CREATININE</td>
<td>BLD</td>
<td>MG/DL</td>
<td>CHEM</td>
</tr>
<tr>
<td>44784-7</td>
<td>CREATININE</td>
<td>SER/PLAS</td>
<td>MG/DL</td>
<td>CHEM</td>
</tr>
<tr>
<td>59826-8</td>
<td>CREATININE</td>
<td>BLD</td>
<td>UMOL/L</td>
<td>CHEM</td>
</tr>
<tr>
<td>77140-2</td>
<td>CREATININE</td>
<td>SER/PLAS/BLD</td>
<td>UMOL/L</td>
<td>CHEM</td>
</tr>
</tbody>
</table>

### Troponin I test

<table>
<thead>
<tr>
<th>LOINC</th>
<th>COMPONENT</th>
<th>SYSTEM</th>
<th>EX_UCUM_UNITS</th>
<th>CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>10839-9</td>
<td>TROTONIN I.CARDIAC</td>
<td>SER/PLAS</td>
<td>NG/ML</td>
<td>CHEM</td>
</tr>
<tr>
<td>16255-2</td>
<td>TROTONIN I.CARDIAC</td>
<td>SER/PLAS</td>
<td>CHEM</td>
<td>CHEM</td>
</tr>
<tr>
<td>42757-5</td>
<td>TROTONIN I.CARDIAC</td>
<td>BLD</td>
<td>NG/ML</td>
<td>CHEM</td>
</tr>
<tr>
<td>49563-0</td>
<td>TROTONIN I.CARDIAC</td>
<td>SER/PLAS</td>
<td>NG/ML</td>
<td>CHEM</td>
</tr>
</tbody>
</table>
Coverage across PCORnet

# of records with valid/non-null LOINC codes ~ 11.6 Billion

# of records that fall within a DC Test ~ 9.9 Billion

# of DC Tests present in the data – 447 (of 490)
  • Missing Tests are largely those related to the presence of an analyte, instead of the value

# of unique LOINC codes (Lab class) across PCORnet – 14,128
Improvements in lab data density

Median Lab Tests Per DataMart

<table>
<thead>
<tr>
<th>Cycle</th>
<th>DLG count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cycle 2</td>
<td>9</td>
</tr>
<tr>
<td>Cycle 3</td>
<td>74</td>
</tr>
<tr>
<td>Cycle 4</td>
<td>136</td>
</tr>
<tr>
<td>Cycle 5</td>
<td>168</td>
</tr>
</tbody>
</table>
Using Lab Groups to identify outliers

Implemented a data check to look for outliers based on median results for a given test

Initial version used inter-quartile range between 25\textsuperscript{th} & 75\textsuperscript{th} percentile values to determine if a median was in-range

- Inter-quartile range (IQR): 75\textsuperscript{th} (Q3) - 25\textsuperscript{th} (Q1)

Cycle 6 will use published reference ranges from 3 clinical labs in addition to IQR
Hematocrit Q2 Distribution (# of DM: 40)

DC Normal Reference Range ( %): lower = 34.9 , upper = 50.4
Median of Q2s = 35.4
Outlier example

Hematocrit Q2 Distribution (# of DM: 40)
DC Normal Reference Range (%): lower=34.9, upper=50.4
Median of Q2s=35.4

# of DMs mapped to the above LOINC

# of records mapped to the above LOINC
Outlier example

Hematocrit Q2 Distribution (# of DM: 40)
DC Normal Reference Range( % ): lower=34.9 ,upper=50.4
Median of Q2s=35.4

Outliers
Outlier example continued

- RESULT_NUM value notation does not match to the expected value notation from LOINC units of measure:
  - Median of Q2s=35.4%

<table>
<thead>
<tr>
<th>DC_LAB_GROUP</th>
<th>DataMart</th>
<th>LAB_LOINC</th>
<th>LOINC UOM</th>
<th>DC report RESULT_NUM Q2</th>
<th>Expected data notation to be aligned with LOINC UOM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hematocrit</td>
<td>DM_1</td>
<td>32354-3</td>
<td>%</td>
<td>0.3</td>
<td>30</td>
</tr>
<tr>
<td>Hematocrit</td>
<td>DM_1</td>
<td>41654-5</td>
<td>%</td>
<td>0.4</td>
<td>40</td>
</tr>
</tbody>
</table>

- Recommendation:
  - Review the source data and LOINC mapping along with the verification of the source Unit of Measure (UOM) in the data
  - Consider mapping to other LOINC codes not requiring % as UOM such as 71829-6
Darcy Louzao: Front Door
PCORnet Front Door

Intended for potential investigators, patient groups, healthcare organizations, clinicians and clinician groups, government, industry scientists, and sponsors to leverage PCORnet’s unique infrastructure and collaborate on patient-centered clinical research. Potential investigators can be within or outside of PCORnet.

Website/Online submission form: http://pcornet.org/frontdoor/

Questions: FrontDoor@PCORnet.org
Key Services available:

- General Consultation
- Network Collaborator Request
- Study feasibility review
- Data Network Request
Key Services available:

- General Consultation
  - Goal of this offering is to describe PCORnet, its structure, members, data, etc. to individuals who are interested in learning more but do not yet know how they would interact with PCORnet.
- Network Collaborator Request
- Study feasibility review
- Data Network Request
Behind the Door

Key Services available:

- General Consultation
- **Network Collaborator Request**
  - Avenue to connect investigators with members across PCORnet who may be interested in collaborating on research projects that may, or may not, utilize PCORnet
  - Study feasibility review
  - Data Network Request

AMIA 2018 | amia.org
Behind the Door

Key Services available:

• General Consultation
• Network Collaborator Request
• **Study feasibility review**
  • Determining if a proposed study is a good-fit for PCORnet and its infrastructure. Opportunity to work with the PCORnet Coordinating Center to optimize the proposal design for execution in PCORnet
• Data Network Request
Behind the Door

Key Services available:

• General Consultation
• Network Collaborator Request
• Study feasibility review

• **Data Network Request**
  • In support of a proposed study for funding, a prep-to-research query can be distributed across PCORnet to obtain preliminary counts of cohorts as well as identify sites that may be ideal collaborators.
Data Network Request

The Requestor sends a question to the PCORnet Coordinating Center through the Front Door.

The Coordinating Center converts the question into a query with an underlying executable code, and sends it to PCORnet partners.

PCORnet partners review the query and provide a response, which is sent back through the Front Door to the Requestor.
Step 1: Complete a RedCap Survey

PCORnet Front Door

The survey is expected to take about 15 minutes. If you would like, you may save your progress and return to the survey later. Feel free to address any questions to frontdoor@pcornet.org.

Contact Info

Request Date
PI's Name
PI's Title
Requesting Institution/Organization
PI's Email Address
PI's Phone Number
Secondary Contact Name
Secondary Contact Email
Secondary Contact Phone Number

Does this request originate from a PCORnet Clinical Data Research Network (CDRN), Patient Powered Research Network (PPRN), Collaborative Research Group (CRG), or Health Plan?

☐ No
☐ Yes, CDRN
☐ Yes, PPRN
☐ Yes, CRG
☐ Yes, Health Plan

Please select the CDRN

☐ ADVANCE
☐ ARCH

Data Network Request
The PCORnet Distributed Research Network Operations Center (DRN OC) can obtain counts from all of the PCORnet Network Partners and answer other data queries by using data transformed to the PCORnet Common Data Model (CDM) and querying via the PCORnet Query Tool. Complete this section if interested in optimizing site identification for recruitment/enrollment and to submit other data requests to the PCORnet DRN OC.

If you are planning on using the PCORnet Common Data Model, please refer to:
http://www.pcornet.org/?s=Common+data+model

Please describe the information you are seeking.

If possible, please address:
• What condition/characteristics would you like to identify? How is it defined?
• Are there other descriptors to assist with identifying the condition/characteristics?
• Is there a specific age group you are interested in?
• Are you interested in a specific diagnosis, procedure, medication, vital, demographic, lab result, encounter (inpatient, outpatient), etc.?
• Do you have classification codes (ICD 9, ICD 10) or procedure codes (CPT) available to assist with this request?
• Is there a specific time period for which you want your request limited to (e.g., 1 year period or a 5 year period)?
Step 2: Initial Assessment

Front Door team reviews request

Generally sets-up a brief call with requester

Review study goals and what type of data would be needed in support of proposal

Modify, as needed, to align with analytic tools
Multi-Pronged Approach

1. Data Curation Results
   - Allow for quick, univariate counts
   - How many patients does PCORnet have with a diagnosis of X?

2. Distributed Query
   - Allows for more complex, multivariate counts, inclusions, exclusions and descriptive statistics
   - In the year 2017, how many patients, have a diagnosis of X with no prior procedure Y in an inpatient setting? What is the distribution of these patients lab Z value? How many prescriptions A did they have in 2017?
Example 1

Request: examine the frequency of nursing-associated diagnoses and procedures, as well as nurse-sensitive patient problems and outcomes. A list of the relevant SNOMED code was provided.

Decision: PCORnet is not well-suited for this question because SNOMED codes are only populated for a few DataMarts and represent <0.1% of all diagnosis and procedure codes.
Example 2

Request: develop a study in PCORnet to examine the length of hospital stay for patients with myasthenia gravis (MG) hospitalized for an exacerbation or crisis and treated with intravenous immunoglobulin (IVIG)

Code lists for a diagnosis of MG and medication administration codes for IVIG were developed by the Coordinating Center

Examine Data Curation

- >50,000 unique patients in PCORnet with an ICD-10 diagnosis code for MG
- >35,000 unique patients with a record of IVIG administration

Decision: sufficient sample size for more detailed investigation – begin query specifications and programming
Example 3

Request: Among patients older than 40 years of age with a diagnosis of diabetes but no diagnosis of cardiovascular disease (CVD) in 2017, how many have:

1. A statin prescription
2. A peripheral arterial disease diagnosis
3. EGRF lab record and distribution of lab result
4. LDL lab record and distribution of lab result

Previous work in PCORnet has shown sufficient sample sizes of the cohort of interest; move to query development to examine key variables needed for the study.
Distributed Query

The Requestor sends a question to the PCORnet Coordinating Center through the Front Door.

The Coordinating Center converts the question into a query with an underlying executable code, and sends it to PCORnet partners.

PCORnet partners review the query and provide a response, which is sent back through the Front Door to the Requestor.
Jessica Sturtevant: Analytics and Query Fulfillment
PCORnet Analytics

I. Analytics & Reusable Tool Development

II. Query Fulfillment

III. Cohort Quality Assessment

IV. Example: Diabetes without Cardiovascular Disease
Why Architect Reusable Tools?

The **same** analytic program is distributed to **all** DataMarts

- Leverages data standardized into a Common Data Model (CDM)
- Enables one program to answer many questions rapidly and efficiently
  - Prevalence, incidence
  - Baseline cohort characteristics
  - Exposure, outcome
- Follows standard format to reduce DataMart burden in review
- Extensively tested to ensure efficient execution across a diverse network ecosystem of 80+ data partners
Application of a Reusable Tool

Patients 18-65 years old with bariatric surgery and no GI cancer in the 365 days prior to bariatric surgery, stratified by age, sex, race

Patients AGE RANGE with PROCEDURE and no DIAGNOSIS in the DAYS prior, stratified by AGE, SEX, RACE
Operationalizing a Distributed Query

One Program

DataMart 1

DataMart 2

DataMart 3

DataMart 4

DataMart 5

DataMart 6

DataMart N

Aggregate Counts (not patient level)

Final Report / Result

AMIA 2018 | amia.org
PCORnet Reusable Tools

Menu-Driven Query (MDQ)
• Simple, point & click cohort definition (e.g. Multiple Sclerosis diagnoses in 2017)
• Time to report: 2-3 weeks

PCORnet Modular Program (PMP)
• More complex cohort logic (e.g. Statin prescription in 2017 with no CVD or Stroke in the year prior)
• Time to report: 3-5 weeks

Cohort Quality Assessment (CQA) Module
• In-depth QA for a specific cohort

Baseline Table Module
• Descriptive baseline cohort covariates, characteristics
What does a Complex, Urgent Query look like in PCORnet?

- Heart Failure with Preserved Ejection Fraction, Comorbidities, Medications, Setting
- Initiated 10/25
- 2 days to operationalize & distribute query
- 5 days for DataMarts to respond
- 1 day to create first report
Heart Failure with Preserved Ejection Fraction

### By Medications (N, % of patients)

<table>
<thead>
<tr>
<th>Medication</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE inhibitor prescription within the day before or after HFrEF</td>
<td>46,674</td>
<td>13%</td>
</tr>
<tr>
<td>ARB prescription within the day before or after HFrEF</td>
<td>31,146</td>
<td>9%</td>
</tr>
<tr>
<td>Sacubitril-valsartan prescription within the day before or after HFrEF</td>
<td>29</td>
<td>0%</td>
</tr>
<tr>
<td>Spironolactone prescription within the day before or after HFrEF</td>
<td>15,043</td>
<td>4%</td>
</tr>
<tr>
<td>Beta Blocker prescription within the day before or after HFrEF</td>
<td>111,620</td>
<td>31%</td>
</tr>
<tr>
<td>Snap Diuretic prescription within the day before or after HFrEF</td>
<td>112,108</td>
<td>31%</td>
</tr>
<tr>
<td>Thiazide Diuretic prescription within the day before or after HFrEF</td>
<td>12,735</td>
<td>3%</td>
</tr>
<tr>
<td>Digoxin prescription within the day before or after HFrEF</td>
<td>11,473</td>
<td>3%</td>
</tr>
<tr>
<td>Insulin prescription within the day before or after HFrEF</td>
<td>10,790</td>
<td>3%</td>
</tr>
<tr>
<td>SGLT2 Inhibitor prescription within the day before or after HFrEF</td>
<td>13</td>
<td>0%</td>
</tr>
</tbody>
</table>

### By Hospitalizations (N, % of patients)

<table>
<thead>
<tr>
<th>Event</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inpatient (IP) encounter in the 30 days prior to HFrEF</td>
<td>95,001</td>
<td>20%</td>
</tr>
<tr>
<td>Emergency Department (ED) encounter in the 30 days prior to HFrEF</td>
<td>40,790</td>
<td>9%</td>
</tr>
<tr>
<td>Emergency Department to Inpatient Hospital Stay (DS) encounter in the 30 days prior to HFrEF</td>
<td>31,560</td>
<td>7%</td>
</tr>
<tr>
<td>Observation Stay (OS) encounter in the 30 days prior to HFrEF</td>
<td>3,933</td>
<td>1%</td>
</tr>
<tr>
<td>Inpatient (IP) encounter in the 365 days prior to HFrEF</td>
<td>111,435</td>
<td>30%</td>
</tr>
<tr>
<td>Emergency Department (ED) encounter in the 365 days prior to HFrEF</td>
<td>111,435</td>
<td>30%</td>
</tr>
<tr>
<td>Emergency Department to Inpatient Hospital Stay (DS) encounter in the 365 days prior to HFrEF</td>
<td>46,665</td>
<td>13%</td>
</tr>
<tr>
<td>Observation Stay (OS) encounter in the 365 days prior to HFrEF</td>
<td>111,435</td>
<td>30%</td>
</tr>
<tr>
<td>Inpatient (IP) encounter in the 1825 days prior to HFrEF</td>
<td>129,853</td>
<td>35%</td>
</tr>
<tr>
<td>Emergency Department (ED) encounter in the 1825 days prior to HFrEF</td>
<td>129,853</td>
<td>35%</td>
</tr>
<tr>
<td>Emergency Department to Inpatient Hospital Stay (DS) encounter in the 1825 days prior to HFrEF</td>
<td>46,263</td>
<td>13%</td>
</tr>
<tr>
<td>Observation Stay (OS) encounter in the 1825 days prior to HFrEF</td>
<td>129,853</td>
<td>35%</td>
</tr>
<tr>
<td>Inpatient (IP) encounter in the 365 days after HFrEF</td>
<td>125,310</td>
<td>32%</td>
</tr>
<tr>
<td>Emergency Department (ED) encounter in the 365 days after HFrEF</td>
<td>125,310</td>
<td>32%</td>
</tr>
<tr>
<td>Emergency Department to Inpatient Hospital Stay (DS) encounter in the 365 days after HFrEF</td>
<td>41,430</td>
<td>12%</td>
</tr>
<tr>
<td>Observation Stay (OS) encounter in the 365 days after HFrEF</td>
<td>125,310</td>
<td>32%</td>
</tr>
</tbody>
</table>

---

AMIA 2018 | amia.org
Analytics Complexity vs. Time to Complete

- **Menu-Driven Query**: Days to Weeks
- **PCORnet Modular Program**: Weeks to Months
- **Custom Analytic Program**: Months to Years
I. Analytics & Reusable Tool Development

II. **Query Fulfillment**

III. Cohort Quality Assessment

IV. Example: Diabetes without Cardiovascular Disease
What is Involved in Implementing Routine Querying?

• Standard process to ensure reliability, efficiency and accuracy

• Process management and tracking to coordinate distribution and response from 80+ DataMarts

• Secure and audited distribution of programs and returned output to/from DataMarts
Standardized Process Ensures Efficiency and Reliability

PROCESS

1. **Front Door / Investigator Consultation**
2. **Define Specifications**
3. **Create & Test Query Package**
4. **DataMart Executes Query Package**
5. **Deliver Report**

Roles:
- **Front Door/Requester**
- **QF Analyst**
- **DataMart**
Process Complexity in a Distributed Research Network

Front Door/Requester

QF Analyst 1

QF Analyst 2

DataMart

Submits Request

Initial Assessment

Prepares Specifications

Assembles Package

Internal Testing

Create Work Plan, Metadata

Distribute Program Package

Review Log & Output

Review & Receive Report

QC Internal Testing

QC Program Package

QC Report

Create Work Plan, Metadata

Distribute Program Package

Review Log & Output

Aggregate Output & Generate Report

Execute Program Package

Review & Return Output

First Wave Roll-Out

Aggregate Output & Generate Report

QC Report

Review & Return Output

60
PCORnet Analytics

I. Analytics & Reusable Tool Development

II. Query Fulfillment

III. Cohort Quality Assessment

IV. Example: Diabetes without Cardiovascular Disease
Cohort Specific Quality Assessment

- In-depth QA for a specific cohort
- Parameterized to return minimum necessary
- Variable & Cross-Variable frequencies, distributions

Defined Cohort

Prescribing Table Variables
Dispensing Table Variables
Lab Table Variables
Targeted Cohort Specific QA is Critical to Understanding Data Usability for Research

Examining a variable in the general population may lead to wrong conclusion, for example:

- Diabetics likely to have well-populated HbA1c lab data
- Inpatient population may have more reliably populated height, weight, race variables
- Medication supply variable population may vary between acute and chronic medications

Understanding data heterogeneity has a large impact on usability in research
PCORnet Analytics

I. Analytics & Reusable Tool Development
II. Query Fulfillment
III. Cohort Quality Assessment
IV. Example: Diabetes without Cardiovascular Disease
Timeline: Diabetes without CVD

- Initiated 10/4
- 4 days to operationalize & distribute query
- 8 days for DataMarts to respond
- 5 day to create first report

Question → Report in 17 Business Days
# Example: Diabetes without CVD

<table>
<thead>
<tr>
<th></th>
<th>Query Period 1/1/2017 - 12/31/2017</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Diabetes, No CVD</td>
</tr>
<tr>
<td><strong>Total N</strong></td>
<td>1,440,996</td>
</tr>
<tr>
<td><strong>By Sex (N, % of patients)</strong></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>774,107</td>
</tr>
<tr>
<td>Male</td>
<td>666,801</td>
</tr>
<tr>
<td>Other*</td>
<td>52</td>
</tr>
<tr>
<td><strong>By Age (N, % of patients)</strong></td>
<td></td>
</tr>
<tr>
<td>40-64</td>
<td>777,080</td>
</tr>
<tr>
<td>65+</td>
<td>653,908</td>
</tr>
</tbody>
</table>
### Cohort Quality Assessment of Labs & Medications

**Examine Lab Value Distribution**

**Examine Prescription Day Supply and Quantity Distributions**

---

#### Query Period 1/1/2017 - 12/31/2017

<table>
<thead>
<tr>
<th></th>
<th>Diabetes, No CVD</th>
<th>Diabetes No CVD and EGFR Lab</th>
<th>Diabetes No CVD and LDL Lab</th>
<th>Diabetes no CVD and PAD</th>
<th>Diabetes, No CVD and Statins Prescription</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total N</strong></td>
<td>1,440,956</td>
<td>654,236</td>
<td>520,812</td>
<td>161,430</td>
<td>428,412</td>
</tr>
<tr>
<td><strong>By Sex (N, % of patients)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>668,801</td>
<td>306,703</td>
<td>248,109</td>
<td>82,431</td>
<td>209,360</td>
</tr>
<tr>
<td>Female</td>
<td>774,107</td>
<td>347,514</td>
<td>272,694</td>
<td>78,987</td>
<td>209,036</td>
</tr>
<tr>
<td>Other*</td>
<td>52</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>By Age (N, % of patients)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40-64</td>
<td>777,080</td>
<td>347,600</td>
<td>284,431</td>
<td>67,047</td>
<td>216,157</td>
</tr>
<tr>
<td>65+</td>
<td>663,908</td>
<td>306,632</td>
<td>234,378</td>
<td>94,383</td>
<td>202,255</td>
</tr>
</tbody>
</table>

---

AMIA 2018 | amia.org
Kathleen McTigue: PCORnet Bariatric Study
Use Case: The PCORnet Bariatric Study
PCORnet provided a unique opportunity to use real-world health data from 41 health systems to compare bariatric procedures

- Severe obesity is a serious health concern affecting 7.7% of Americans

- Use of bariatric surgery has expanded considerably, with the most common US procedures being:
  - Adjustable gastric band
  - Roux-en-Y Gastric Bypass (RYGB)
  - Sleeve Gastrectomy (SG)

- Sleeve gastrectomy procedure has been used increasingly over past decade - despite a lack of data comparing its effectiveness to other procedures
Overview of PBS Quantitative Aims

To estimate the 1-, 3-, & 5-year benefits & risks of the three main surgical treatments for severe obesity

• Roux-en-y gastric bypass (RYGB)
• Adjustable gastric banding (AGB)
• Sleeve gastrectomy (SG)

Focus on outcomes that matter to adults & adolescents with severe obesity:

• Weight loss
• Improvement in diabetes
• Risk of adverse events
PBS Researchers and Stakeholders

**CORE STUDY TEAM**
- David Arterburn
- Kathleen McTigue
- Neely Williams
- Karen Coleman
- Cheri Janning
- Anita Courcoulas
- Darren Toh
- Jane Anau
- Roy Pardee
- Robert Wellman
- Yates Coley
- Andrea Cook
- Casie Horgan
- Jessica Sturtevant

**OTHER KEY INDIVIDUALS**
- Caroline Apovian, Lydia Bazzano,
- Cynthia Blalock, Jeff Brown, Jeanne
- Clark, Nirav Desai, Elizabeth Doane-Cirelli, Emily Eckert, Ana Emiliano, John
- Holmes, Thomas Inge, Elisha Malanga,
- Corrigan L. McBride, James McClay,
- Marc Michalsky, Sameer Murali, Joe
- Nadglowski, Beth Nauman, Rabih
- Nemr, Andrew Odegaard, Alberto
- Odor, Gabrielle Purcell, Laura
- Rasmussen-Torvik, Roz Saizan, David
- Schlundt, Steven R. Smith, Ali
- Tavakkoli, Tammy St. Clair, Julie Tice,
- Joseph Vitello, Stavra Xanathos, and
- Roni Zeiger
CDRNs & Health Systems Contributing Data to PBS
PBS worked with the Coordinating Center team to execute a series of CDM queries

Study-specific data characterization for not-yet-characterized data elements

Aim 1 queries, individual-level and distributed (weight loss)
- 42 data-contributing sites from 11 CDRNs

Aim 2 query (diabetes risk)
- 33 data-contributing sites from 11 CDRNs

Aim 3 query (adverse events)
- 10 data contributing sites from 4 CDRNs
Preparatory-to-research queries are critical at the proposal stage

- Provide insight into sample size
  - Useful for power calculations & identifying the most appropriate sites
- Can identify potential data limitations
  - E.g., DMs had >40% of bariatric patients missing BMI in the year before or the year after surgery
- Not all CDM fields are “required” and there is heterogeneity among DataMarts in terms of which fields are filled
Query development is a collaborative & thoughtful process

- Helpful for researchers to start with clearly defined research questions & an understanding about basics of CDM data elements
- Study timeline should plan on accommodating query development
  - Complex queries may require several months to develop
  - Modular tools considerably reduce this effort
- If conducting a distributed-analytic query, budget time for developing the cohort & regression in at least one site that provides individual-level data prior to developing a distributed analytic query
Code lists can benefit from stakeholder input … and may be re-used

• For query identification, we developed lists of procedure codes & diabetes medications
  • Surgeons provided input on relevant procedure codes
  • Patients, primary care providers & endocrinologists helped ensure that the medication list was complete
• Coordinating Center team has archived lists that can be adapted for future projects
Even with CDM standardization, researchers should plan on investing in data cleaning with real-world data

Some data appeared to have incorrect units
  - DataMarts worked to troubleshoot problems
  - Refreshed CDM & re-ran study query

Height, weight & BMI cleaned to remove implausible values

Calculating BMIs
  - Adults: Used any adult HT
  - Adolescents: Used nearest HT

Archiving/re-using data-cleaning protocols may save time/energy and promote use of comparable data across PCORnet studies
The number of vital signs records from which BMI could be calculated or observed was substantial – but varied by site

<table>
<thead>
<tr>
<th>Description</th>
<th>Adult</th>
<th>Adolescent</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>HT, WT &amp; BMI available from CDM</td>
<td>1,581,786</td>
<td>13,257</td>
<td>1,595,043</td>
</tr>
<tr>
<td>HT &amp; WT available; BMI calculated</td>
<td>698,610</td>
<td>13,565</td>
<td>712,175</td>
</tr>
<tr>
<td>HT &amp; BMI available, WT estimated</td>
<td>446,733</td>
<td>2,598</td>
<td>449,331</td>
</tr>
<tr>
<td>Only BMI available</td>
<td>1,240</td>
<td>2,194</td>
<td>3,434</td>
</tr>
<tr>
<td>WT &amp; BMI available; HT estimated</td>
<td>1,470</td>
<td>64</td>
<td>1,534</td>
</tr>
<tr>
<td>Total</td>
<td>2,729,839</td>
<td>31,678</td>
<td>2,761,517</td>
</tr>
</tbody>
</table>
Stakeholder input was invaluable in implementing aims

- At study kick-off meeting, requested three pair-wise comparisons of bariatric procedures versus two pair-wise comparisons
- After study-specific data characterization, exclusion of outpatient procedures was aborted on surgeons’ advice
- Patients & clinicians prioritized analyses of patient sub-groups
- With so many datamarts, we saw numerous instances of data outliers (e.g., fatty liver, depression prevalence)
  - Local clinicians were able to quickly confirm patterns and provide context; such findings typically reflected patient population characteristics, local coding practices, or patterns of care
Data collected for clinical purposes has more missingness than traditional research data.

<table>
<thead>
<tr>
<th>% of patients eligible for f/u who have a BMI at the time point</th>
<th>1 year (6-18 mo)</th>
<th>3 years (30-42 mo)</th>
<th>5 years (54-66 mo)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overall</strong></td>
<td>84%</td>
<td>68%</td>
<td>69%</td>
</tr>
<tr>
<td><strong>AGB</strong></td>
<td>76%</td>
<td>60%</td>
<td>55%</td>
</tr>
<tr>
<td><strong>RYGB</strong></td>
<td>86%</td>
<td>67%</td>
<td>67%</td>
</tr>
<tr>
<td><strong>SG</strong></td>
<td>84%</td>
<td>73%</td>
<td>86%</td>
</tr>
</tbody>
</table>

*Sensitivity analysis assessed how much our model estimates were likely to be impacted by lack of follow-up*
We focused attention on academic recognition for data-contributing sites & engaged stakeholders

Stakeholders expressed strong preference for authorship opportunities for each data-contributing site & goal of including stakeholder perspectives in publications

- Commitment to use of journals with group authorship indexed by author in PubMed

Logistical challenges of large writing groups

- PCORnet publication resources helpful
- Developed tips for helping stakeholders engage in the scientific writing process
- “Online survey” approach for author feedback
Among 46,510 patients from 22 states whose data contributed to weight loss analyses...

Most procedures were bypass or sleeve
- 24,982 RYGB (53%)
- 18,961 SG (41%)
- 2,567 AGB (6%)

The sample was, on average, middle-aged, mostly female & fairly racially/ethnically diverse
- Mean age 46; 80% female; 21% Hispanic; 21% African American

Patients were severely obese
- Mean BMI: 49 kg/m² with 38% BMI 50+ kg/m²

Comorbidities were common
- 60% HTN; 49% Dyslipidemia; 49% OSA; 40% GERD; 37% T2DM
We saw a dramatic shift in practice patterns
Comparative effectiveness could be estimated over relatively long-term follow-up

Weight Regain

- 3.6 kg ↑
- 8.2 kg ↑
- 7.6 kg ↑
The large sample enabled analysis of clinically important sub-groups such as 544 adolescents.
Key lessons

PCORnet CDM can provide important contributions to the medical literature

- Very large sample size from heterogeneous health systems
- Some research processes need adjustment in the distributed research setting
- Lessons learned & reusable tools/resources can facilitate efficient use of the CDM resource

Stakeholders added value throughout the scientific process

- Developing research question
- Implementing analyses
- Differentiating inter-site heterogeneity from data problems
AMIA is the professional home for more than 5,400 informatics professionals, representing frontline clinicians, researchers, public health experts and educators who bring meaning to data, manage information and generate new knowledge across the research and healthcare enterprise.
Thank you!