Table of Contents

(Click section name or page number to link to page)

Table of Contents 1

Partnership Packet Cover Letter 3

N3C Brochure 6

Contacts and Workstreams N3C 8

Governance (slides) Site Survey 11

IRB 25

Johns Hopkins University Central IRB Letter 33

Understanding JHM IRB pSite Requirements Example- 36

JHM IRB for N3C Local Context Questionnaire 43

Protocol 52

OMOP dictionary 68
Partnership Packet Cover Letter
N3C Letter to Partners

Dear Colleagues,

The COVID-19 global emergency raises many difficult care and healthcare management questions. Which drugs are the most viable candidates for a given patient? How can we efficiently and effectively assemble the right cohort for a trial? What social determinants impact course and outcome?

The National COVID Cohort Collaborative (N3C) is a collaboration among the NCATS-supported Clinical and Translational Science Awards (CTSA) Program hubs, the National Center for Data to Health (CD2H), distributed clinical data networks (PCORnet, OHDSI, ACT/i2b2, TriNetX), and other partner organizations, with overall stewardship by NCATS. The N3C aims to improve the efficiency and accessibility of analyses with COVID-19 clinical data, expand our ability to analyze and understand COVID, and demonstrate a novel approach for collaborative pandemic data sharing. For more detailed information, please view the slide deck and AMIA webinar recorded 4.13.2020: Building a Nationwide COVID-19 Cohort Through Informatics: A new initiative being coordinated by CD2H & NCATS AMIA Presentation

Joining the Effort
N3C is providing this Partnership Packet to provide information and instructions needed to begin the Partnering process. Due to the urgency of this mission, it is critical to convene key members of your institution who can implement this effort at a swift pace. The first steps will be to identify the following N3C Site Team Members within your organization: 1) Champion Leader, 2) member of your Contracts and Legal group, 3) an IRB administrator who is familiar with central IRB activities and 4) CRIO and/or Clinical Data Warehouse Director. Please relate this information by having a single member of your organizations fill out the site survey (PDF version attached; live survey here.)

The Data Transfer Agreements and Data Use Agreements will be executed with NIH. The Data Transfer Agreement (DTA) is a legal document signed between the institution and NCATS. This agreement covers the disclosure of a limited dataset from the contributing site to NCATS. Execution of this agreement is a prerequisite for transmitting data to NCATS for N3C. John Hopkins University (JHU) will serve as the central IRB for N3C.

Data Acquisition Process
Once IRB approval is in place, and the Data Transfer Agreements have been executed, N3C will provide each site with “White Glove” service to assist with the data exchange process. Sites will be provided with a 1) COVID Phenotype, scripts for the data models (PCORnet, OMOP, ACT/i2b2 or TriNet), 2) a process document for data exchange and 3) assistance with data quality checks.

Website: COVID.CD2H.org
For questions or comments, please contact the National Center for Data to Health (CD2H), the partner coordinating center for N3C, at data2health@gmail.com
Questions/Comments/Suggestions
Over the next several weeks, N3C will conduct webinars, and Question and Answer sessions to answer DTA/DUA, IRB, and technical questions. If you have any questions or would like to participate in this collaborative effort, please contact us (see Contacts document).

Even though this is a challenging time for our communities, this is an opportunity to achieve something together that will bring healing!

Website: COVID.CD2H.org
For questions or comments, please contact the National Center for Data to Health (CD2H), the partner coordinating center for N3C, at data2health@gmail.com
N3C Brochure
WHY JOIN N3C

There is no shortage of clinical data within institutions; however, in the United States these data are not structured the same way nor are they accessible for shared analytics by our nation’s scientists. It is imperative that we overcome these technical and regulatory barriers to address the COVID-19 pandemic.

The N3C aims to unite COVID-19 data, enabling innovative machine learning and statistical analyses that require a large amount of data—more than is available in any given institution. The goal is to enable rapid collaboration among clinicians, researchers, and data scientists to identify treatments, specialize care, and to reduce the overall severity of COVID-19. Visit covid.cd2h.org/join to learn more.

- Designate a central IRB & Execute a Data Transfer Agreement with NCATS
- Support a User Data Use Agreement
- Establish a Data Access Committee
- Help partners identify, extract, and contribute COVID-19 data
- Harmonize EHR data via OMOP, ACT/i2b2, PCORnet, or TriNetX into a single model
- Create the 1st nationwide patient-level dataset in a certified secure environment

1. DATA PARTNERSHIP & GOVERNANCE
2. PHENOTYPE & DATA ACQUISITION
3. DATA INGESTION & HARMONIZATION

4. COLLABORATIVE ANALYTICS
5. REVEAL ASSOCIATIONS & INSIGHTS
6. RAPID RESULTS & CREDIT

CLINICAL SCENARIOS & DATA ANALYTICS
- Define driving clinical questions
- Create exemplar workflows
- Identify methodologies for answering questions

TOOLS & RESOURCES
- Deploy user interfaces for data inquiry
- Register tools & algorithms for use in the portal

PORTALS, DASHBOARDS, & ATTRIBUTION
- Matchmake clinical problems with data science solutions
- Share resources, track analyses, and credit collaborators

SYNTHETIC DATA DERIVATIVE
- Derive fully de-identified synthetic data
- Make accessible to registered members of the public

JOIN US:
covid.cd2h.org/join | data2health@gmail.com | @data2health #N3C
Contacts and Workstreams
Contacts and Workstream Information

Please see below for important contact information and associated Documents. Each site should fill out the site survey here; a copy of this survey is included as a pdf for your reference. At the end of this survey, there is a place for you to relate the name and email addresses for your:

- Contracts/Legal lead for the CTSA
- IRB Director or Central IRB Administrator
- CRIO and/or Clinical Data Warehouse Director

Important Contact Information

<table>
<thead>
<tr>
<th>Topic</th>
<th>Contact Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCATS Data Transfer Agreement</td>
<td><a href="mailto:ncatspartnerships@mail.nih.gov">ncatspartnerships@mail.nih.gov</a></td>
</tr>
<tr>
<td>sIRB</td>
<td>Tricia Francis <a href="mailto:pfranci4@jhu.edu">pfranci4@jhu.edu</a></td>
</tr>
<tr>
<td>NCATS Web Pages and FAQ</td>
<td><a href="https://ncats.nih.gov/n3c">https://ncats.nih.gov/n3c</a></td>
</tr>
<tr>
<td></td>
<td><a href="https://ncats.nih.gov/n3c/about/program-faq">https://ncats.nih.gov/n3c/about/program-faq</a></td>
</tr>
<tr>
<td>General N3C Questions/Comments/Suggestions</td>
<td><a href="mailto:data2health@gmail.com">data2health@gmail.com</a></td>
</tr>
</tbody>
</table>
| Technical or Partnership Questions   | Ken Gersing [Kenneth.gersing@nih.gov](mailto:Kenneth.gersing@nih.gov)  
Sam Michael [Michaelsg@mail.nih.gov](mailto:Michaelsg@mail.nih.gov)  
Melissa Haendel [melissa@tislab.org](mailto:melissa@tislab.org)  
Anita Walden [Anita.Walden@sagebase.org](mailto:Anita.Walden@sagebase.org) |

NIH Funding Announcement for the N3C effort: Information on funding for CTSA hubs is available here; the NOSI is here; the N3C/CD2H is happy to provide scientific feedback on goals or letters of support for proposals, but is not involved in review or funding decisions. Please contact your NCATS program officer with funding questions.

N3C Workstreams

The N3C partnership welcomes participation from the scientific community through five workstreams that are open to anyone. Everyone can help reduce the impact of COVID-19.

The following links will allow you to access documents and communication tools. Please complete the onboarding form to gain access: [bit.ly/cd2-onboarding-form](https://bit.ly/cd2-onboarding-form)
N3C Phenotype & Data Acquisition Workstream
Lead: Emily Pfaff (UNC/PCORNet)
Latest COVID Phenotype: https://github.com/National-COVID-Cohort-Collaborative/Phenotype_Data_Acquisition/wiki/Latest-Phenotype
Slack: https://cd2h.slack.com/archives/C011W93LF4Y
Shared Drive: https://bit.ly/n3c-phenotype-drive
Notes: https://bit.ly/n3c-phenotype-notes

N3C Data Ingestion & Harmonization Workstream
Lead: Chris Chute (Johns Hopkins University)
Slack: https://cd2h.slack.com/archives/C011MFU4E92
Shared Drive: https://bit.ly/n3c-harmonization-drive
Notes: https://bit.ly/n3c-harmonization-notes
GitHub Repository: here

N3C Collaborative Analytics Workstream
Leads: Justin Guinney (Sage Bionetworks) and Joel Saltz (Stony Brook)
Slack: https://cd2h.slack.com/archives/C011W941TFA
Shared Drive: https://bit.ly/n2c-analytics-drive
Notes: https://bit.ly/n3c-analytics-notes

**Collaborative Analytics Substreams**
**Portal & Dashboards**
Leads: Dave Eichmann (UIowa), Warren Kibbe (Duke)

**Tools & Resources**
Leads: Andrew Williams (Tufts), Chunlei Wu (Scripps)

**Clinical Scenarios & Data Analytics**
Leads: Peter Robinson (JAX), Heidi Spratt (UTMB), Tell Bennett (CU)

N3C Synthetic data Workstream
Leads: Philip Payne (WashU) and Atul Butte (UCSF)
Slack: https://cd2h.slack.com/archives/C013D941H7Y
Shared Drive: Synthetic Drive Folder
Notes: Synthetic Notes

Website: COVID.CD2H.org
For questions or comments, please contact the National Center for Data to Health (CD2H), the partner coordinating center for N3C
data2health@gmail.com
N3C Governance
N3C Data Governance & Partnership
Data Access
Christine Suver and Kellie Walters
08 May 2020

@data2health
@ncats.nih.gov
https://covid.cd2h.org/
Access Principles: “Share widely and wisely”

The **end-goal** is broad access, including:

- Academic and Commercial
- Credentialed researchers* (LDS) and Individual (“citizen scientists”) (Synthetic Data)
- Domestic
- Directed to COVID-Related research
- Activities on N3C Enclave are recorded and can be audited
- Disclosure of research results to N3C Enclave for the public good
- Contributor Attribution
- No download of Limited Dataset
- Access authorization must be renewed annually

*Credentialed researcher are researchers from academic or commercial institutions who have completed Human Subject Research training*
N3C Data Disclosures

Contributing Site A

Contributing Site B

Contributing Site C

NCATS Environment

Staging DB

Prod DB

LDS

Palantir Data Enclave

Synthetic Derivative

LDS, no viewing of patient level data

LDS, view patient level data

NCATS Staff administering database

Data Phenotype and Acquisition WS Subcontractor (e.g., UNC, Pitt…)

Data Ingestion and Harmonization WS Subcontractor (e.g., Hopkins, UNC…)

Collaborative Analytics (e.g., ???)

Public Data Scientist (at university, NIH, etc.)

Qualified Research (at university, NIH, etc.)

Qualified Research (at university, NIH, etc.)

v. 0.2 DRAFT | Not meant to describe all data flows in technical detail

### In Discussion

Data transfer

Data access
N3C Data Disclosures: Planned Coverage Method

Green Boxes: Covered by Data Transfer Agreement with NCATS

Blue Boxes: Not HIPAA data

Yellow Boxes: Data Access Terms and Conditions, DAC Approval

Orange Boxes: Data Access Terms and Conditions; Study-specific IRB; DAC Approval

Gray Boxes: Unclear which of these groups will need access; coverage for access TBD

NCATS Staff administering database

Data Phenotype and Acquisition WS Subcontractor (e.g., UNC, Pitt…)

Data Ingestion and Harmonization WS Subcontractor (e.g., Hopkins, UNC…)

Collaborative Analytics (e.g., JAX, Stony Brook)

Palantir Data Enclave

Synthetic Derivative

LDS, no viewing of patient level data

LDS, view patient level data

Public Data Scientist X (at university, NIH, etc.)

Qualified Research Y (at university, NIH, etc.)

Qualified Research Z (at university, NIH, etc.)

v. 0.2 DRAFT | Not meant to describe all data flows in technical detail

### In Discussion

Data transfer

Data access
Data Contribution Requested

- Contributors are asked to submit data in one of four common data models—i2b2/ACT, PCORnet, OMOP, TriNetX. N3C will provide queries and support to facilitate this process and reduce burden on sites.
- Data to include structured data elements (demographics, visit, diagnosis, procedures, labs, vitals, meds, etc.) for all patients meeting the N3C COVID-19 phenotype.
- Data will be a limited dataset, containing the HIPAA identifiers birth month and year; dates of service; and zip code.
- Data will be transferred as flat files from contributor via secure FTP to a FedRAMP certified staging area on an NCATS server.
- Refresh rate of 2-3/week desired.

Regulatory Requirements

- Data Transfer Agreement to cover the disclosure of a limited dataset will be executed between Contributor and NCATS
- Project approved by Johns Hopkins IRB; Sites requested to cede review to Hopkins via SMART IRB
Tiered Data Use

**Synthetic Data**
- Data derivatives
- Low risk of re-identification

**Analytics output**
- Query results = parameters, no data output

**Limited data set**
- Includes dates and geolocation
- Data could be identifiable to the patient and institution

IDU: Intended Data Use will be posted on the N3C site with the requester’s name and institution
# N3C Data Access Proposal

<table>
<thead>
<tr>
<th>Access Level</th>
<th>Data Type</th>
<th>Eligible Users</th>
<th>Request Process</th>
<th>Approval Process</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Registered</td>
<td>Synthetic data</td>
<td>Anyone (researchers, independent contractors, data analysts)</td>
<td>Register on N3C Enclave Post Intended Data Use Statement</td>
<td>Automatic (DAC and DAC Advisory Board as needed)</td>
<td>N3C profile- agreement to N3C terms-</td>
</tr>
<tr>
<td>Controlled</td>
<td>Analytics</td>
<td>Bonafide academic &amp; commercial researchers in the US</td>
<td>Register on N3C Enclave Post Intended Data Use Statement Execute DUA</td>
<td>DAC (DAC Advisory Board as needed)</td>
<td>N3C profile- agreement to N3C terms- Affiliation to an academic or commercial institution in the US Institutional email (no gmail), Public ORCID COVID-related research DUA signed by data requester and signing official</td>
</tr>
<tr>
<td>Controlled+</td>
<td>Limited Dataset</td>
<td>Bonafide academic &amp; commercial researchers in the US</td>
<td>Register on N3C Enclave Post Intended Data Use Statement Execute DUA Post Rationale for needing access to LDS Post proof of IRB approval/Exemption</td>
<td>DAC (DAC Advisory Board as needed)</td>
<td>N3C profile- agreement to N3C terms- Affiliation to an academic or commercial institution in the US Institutional email (no gmail), Public ORCID COVID-related research DUA signed by data requester and signing official Use of Synthetic data is insufficient to address research question IRB approval/exemption document lists the study name and names the PI in research personnel</td>
</tr>
</tbody>
</table>
Registration: Create an account on N3C Enclave.

- Name:
- Phone/Email:
- Address (must be in US)
- Agreement to N3C terms, including but not limited to:
  - use LDS data for COVID-related research
  - disclose research results to N3C for public good
  - not re-identify data subjects or institutions
  - contributor attribution
  - recognizes DAC’s authority to arbitrate disputes
- Institution Name*: (Academic or Commercial, must be in the US)
- ORCID*: (Must be Public)

* Optional if accessing Synthetic Data
Data Partnership & Governance Workstream

Intended Data Use statement **(in English)**

- Project Title,
- Requester’s contact information and ORCID*
- Objective or purpose of the project
- Main testable hypothesis
- A general description of the actions to be performed by the scientist using the Data and possibly the anticipated results
- Use of LDS data must be directed to COVID-Related research

* Optional if accessing Synthetic Data
** IDUs will be posted on the N3C enclave-
Data Access Committee

The DAC oversees how N3C LDS data is accessed and used.

The DAC's responsibilities include:
- Reviewing requests to access N3C data and verifying that conditions for accessing Data are met.
- Authorizing/revoking data access
- Addressing questions and complaints
- Auditing access and use of data
- Work with the N3C leadership to resolve disputes

DAC Membership: 2 co-chairs + 3 members + 3 delegates + 1 administrator

DAC members to include representatives of: CTSA hubs, NCATS, Data Contributors, and Federal agencies
Criteria for Approval - Focus on objective criteria for granting access

N3C DAC will not evaluate the scientific merit of the intended data use. N3C DAC will not evaluate the expertise of the data requester.

- Data requester’s profile is complete with affiliation to an academic or commercial institution in the US and a public ORCID.
- The intended data use statement includes sufficient details to determine that the proposed use of data is for credible COVID-Related research.
- DUA is signed by the requester and a Signing official. By executing the DUA, the data requester and Signing officials confirm that:
  - the data requester is a credentialed investigator at the institution,
  - the data requester has completed human subject research training
  - The signing official has authority to oversee the data requester
  - Both agree to all the DUA stipulations

If requesting access to row-level Data:
- The research question cannot be addressed with the Synthetic data
- The proof of IRB approval/exemption for the proposed research lists the study name and the data requester in research personnel
Criteria for Rejection-

- Incomplete request
- Intended data use isn’t consistent with the requirement of COVID-Related research
- The Synthetic data could be used to answer the research question
- Requester has been disqualified from accessing N3C
Data Use Agreement

https://docs.google.com/document/d/1j7ByvjiMnTXItVylPL7bItMs2yyeAmSOYv2ZLFUUZi1o/edit?ts=5ea17be#

By signing the DUA, the Institutional Signing official confirms
- The requester has received human subject research training. The DAC will not need to collect proof of training
- The requester and all collaborators listed on the DUA are credentialed researcher from the institution/organization.

* The Signing Official is responsible for ensuring appropriate and ethical use of/access to the Data by the data requester.
Site Survey
N3C site survey

This is a PDF example of the site survey. Please fill the real survey online at: https://forms.gle/PkGSD9eDrXM8etux7

Dear colleagues, thank you for your interest and partnership in creating the National Covid Cohort Initiative (N3C). If you are interested in closer tracking the project, please onboard here (<2 min): http://bit.ly/cd2h-onboarding-form

The following questions will help CD2H coordinate experts to help your site most easily ingest your data into the N3C. We thank you for taking the time to help us help you.

1. *Required*
   1.1 Email address *

1.2 Your institution *

1.3 Your name *

1.4 Which of the following common data models (CDMs) are in use (and regularly populated) at your institution? (Check all that apply) *

Check all that apply.

- [ ] OMOP/OHDSI
- [ ] i2b2 (ACT ontology)
- [ ] PCORnet
- [ ] TriNetX
- [ ] Other:  


5. If your institution was to submit a set of covid-specific patient data to a central repository using one of these data models, which data model would be optimal? (Check multiple only if there is a "tie" for first place.) *

Check all that apply.

- [ ] i2b2 (ACT ontology)
- [ ] PCORnet
- [ ] OMOP/OHDSI
- [ ] TriNetX

Other: [ ]

6. Is your site already populating a covid-specific datamart using one of the common data models listed above? *

Mark only one oval.

- [ ] Yes
- [ ] No
- [ ] We have a covid datamart, but it's not in a CDM
- [ ] We are in the process of building a covid-specific CDM, but it is not yet complete

7. What numbered version of your chosen data model is your institution currently using? (Write “Not sure” if not sure.) *

__________________________________________________________

8. How often is the data in your chosen data model currently refreshed at your institution? *

__________________________________________________________
9. If not doing so already, would your institution be able to refresh the data in your chosen data model at a greater frequency (e.g., 2-3 times per week), for at least the subset of the patient population indicated in the covid phenotype definition (including some negative controls)? *

Mark only one oval.

- [ ] We are already refreshing at that rate
- [ ] Yes, we could support
- [ ] Yes, we could support with additional funding
- [ ] No, we cannot support

10. What database management system does your site use to host your chosen data model? *

Mark only one oval.

- [ ] Oracle
- [ ] MS SQL Server
- [ ] PostgreSQL
- [ ] MySQL
- [ ] Other: __________________________

11. Which EHR is in use at your institution? (Check all that apply) *

Check all that apply.

- [ ] Epic
- [ ] Cerner
- [ ] Allscripts
- [ ] Other: __________________________
12. Does your institution load any patient location information (i.e., home location) into your chosen data model, at any level (zip, census tract, etc.)? *

*Mark only one oval.*

- Yes
- No

13. Does your site load lab orders/results into your chosen data model if those labs are NOT LOINC-coded? *

*Mark only one oval.*

- Yes
- No
- Not sure

14. Would your institution be able to load non-LOINC-coded labs into your chosen data model, for use in covid-19 phenotyping? *

*Mark only one oval.*

- We are already loading non-LOINC-coded labs
- Yes, we could support
- Yes, we could support with additional funding
- No, we cannot support
15. Does your site load lab orders/results into your chosen data model if the result type for those labs is qualitative (i.e., text results rather than numeric, like "POSITIVE" or "Not detected")? *

*Mark only one oval.*

- Yes
- No
- Some, but not all
- Not sure

16. Would your institution be able to load qualitative lab data (at least a subset) into your chosen data model, for use in covid-19 phenotyping? *

*Mark only one oval.*

- We are already loading some or all qualitative labs
- Yes, we could support
- Yes, we could support with additional funding
- No, we cannot support

17. Please provide the name and email of the contact at your institution who handles execution of Data Use Agreements; e.g. the contracts/Legal lead for your CTSA/site.


18. Please provide the name and email of the contact at your institution who oversees your common data model(s); e.g. your CRIO and/or Clinical Data Warehouse Director.


19. Please provide the name and email of the IRB Director or Central IRB Administrator at your institution.
IRB
Johns Hopkins University
Central IRB Letter
APPLICATION APPROVAL

Review Type: Convened
Principal Investigator: Christopher Chute
Number: IRB00249128
Title: National COVID Cohort Collaborative (N3C): A national resource for shared analytics
Committee Chair: Richard Moore
IRB Committee: IRB-3

Date of Approval: May 1, 2020
Date of Expiration: May 1, 2021

The JHM IRB approved the above-referenced Application.

IRB review included the following:

45 CFR 46.116: A waiver of consent was granted based on the following criteria: 1) the research involves no more than minimal risk to subjects; 2) the waiver will not adversely affect the rights and welfare of the subjects; 3) the research could not be practicably carried out without the waiver; and 4) the IRB will advise you if it is appropriate for participants to be provided with additional pertinent information after participation.

Progress Report Required:

The Board determined that this research meets the criteria for submission of a Progress Report as an alternative to a Continuing Review Application. The Progress Report must be submitted using a Further Study Action and selecting progress report at least 6 weeks prior to the expiration date. Please note, the Progress Report must be submitted prior to the expiration date shown on this notice. If the Progress Report is not submitted prior to the expiration date all activity must stop. Before any research activity can resume, you must submit the progress report.

45 CFR 46.404 and/or 21 CFR 50.51: This study has been approved for the inclusion of children as 'research not involving greater than minimal risk'. The permission of parents/guardians is waived.

Assent is waived for all children.

Changes in Research: All proposed changes to the research must be submitted using a Change in Research application. The changes must be approved by the JHM IRB prior to implementation, with the following exception: changes made to eliminate apparent immediate hazards to participants may be made immediately,
and promptly reported to the JHM IRB.

**Unanticipated Problems:** All unanticipated problems must be submitted using a Protocol Event Report.

If this research has a commercial sponsor, the research may not start until the sponsor and JHU have signed a contract.

The JHMIRB is constituted to meet the requirements of the Privacy Rule at section 45 CFR 164.512(i)(1)(i)(B) and is authorized and qualified to serve as the Privacy Board for human subjects research applications conducted by Hopkins’ faculty members. The JHM IRB reviewed your request to waive or alter authorization for the above-referenced project. The IRB determined that all specific criteria for a waiver or alteration of authorization were met, as follows:

(A) The use or disclosure of protected health information involves no more than minimal risk to the privacy of individuals, based on, at least, the presence of the following elements;
   (1) An adequate plan to protect the identifiers from improper use and disclosure;
   (2) An adequate plan to destroy the identifiers at the earliest opportunity consistent with conduct of the research, unless there is a health or research justification for retaining the identifiers or such retention is otherwise required by law; and
   (3) Adequate written assurances that the protected health information will not be reused or disclosed to any other person or entity, except as required by law, for authorized oversight of the research study, or for other research for which the use or disclosure of protected health information would be permitted;
(B) The research could not practicably be conducted without the waiver or alteration; and
(C) the research could not practicably be conducted without access to and use of the protected health information.

**Study documents:**

**HIPAA Form 4:**
FINAL_Chute_IRB00249128_HIPAA Form 4_05012020.pdf

**Additional Supplemental Study Documents:**
DRAFT-N3C-Data Use Agreement.docx
Coordinating Center Description.docx
Final version - NCATS Data Transfer agreement 04-23-2020.docx
OMOP CDM v5 COVID Data Dictionary.docx

**Protocol:**
National COVID Cohort Collaborative V4.docx

**Johns Hopkins Study Team Members:**
Richard Zhu, Dazhi Jiao, Xiaohan Zhang, Stephanie Hong, Harold Solbrig, Harold Lehmann, Davera Gabriel

The Johns Hopkins Institutions operate under multiple Federal-Wide Assurances: The Johns Hopkins University School of Medicine - FWA00005752, Johns Hopkins Health System and Johns Hopkins Hospital - FWA00006087
Understanding JHM IRB pSite Requirements
SMART IRB Reliance Agreement

• JHM IRB requires use of the SMART IRB Reliance Agreement as the basis of reliance.

• JHM IRB also requires execution of a “Letter of Indemnification” [LOI] as indemnification is not specified in the Agreement.

• With assistance from the JHM IRB reliance team, the lead study coordinator will send introduction emails to sites based on whether they have already executed the SMART Agreement and/or LOI with JHM IRB.

Benefits:
• Eliminates the need for study-specific reliance agreement negotiation.
• Once you are a signatory to SMART, you may use SMART as your reliance agreement for any specific study that also involves institutions that are SMART signatories.
• JHM IRB has executed over 150 LOIs with sites.
Step 1: JHM IRB will assist the lead study team with on-boarding sites to the required agreements [SMART and LOI]

- Sites that have already executed the required agreements will receive an email request to confirm willingness to rely.
- Willingness to rely must be provided via email or letterhead from the site IRB/HRPP point of contact.
- JHM IRB will execute a “cede letter”, documenting the reliance relationship.
  - JHM IRB uses the SMART IRB Letter of Acknowledgment to document reliance.
How does this all work?

**Step 2:** Study approval documents are released to relying sites to perform a required local context review

- Sites will complete a JHM Local Context Questionnaire [LCQ] which must be signed by the site PI and an IRB/HRPP signatory
- The LCQ collects information such as local investigator qualifications/training, local ancillary reviews, and identification of any specific local issues

**Step 3:** Sites return completed LCQ to lead study team and sites are reviewed/approved.
Step 3: Site Review Process

- JHM Expedited Review Team includes three key players:
  - **IRB Reliance Team Member**: Ensures local context questionnaire is complete
  - **Consent Form Specialist**: [not applicable for this project] Ensures site-specific language has been provided for “editable” sections of the consent and builds the consent document
  - **Compliance Team Member**: Performs a “regulatory review” to identify whether there are areas of concern
- If issues are raised that impact the criteria for approval, the site application is assigned to the convened JHM IRB for review and site will be contacted.
Once a participating site ("pSite") is approved, a site-specific approval letter will be issued.

Note: The participating IRB/institution may require receipt of the JHM IRB site-specific approval letter BEFORE the pSite can be activated locally.

pSites should check with their IRB/organization about the types of items that will require "local" review during the life of the study—

Examples:
- Changes that affect local drug dispensation/dosing
- Changes that may trigger a local ancillary review
- Personnel changes/new conflicts
How will sites talk to the JHM IRB?

- Site documents are communicated through the lead PI/Point of Contact
- Sites are added as pSites
- Study-wide amendments controlled by overall PI
- Site-specific amendments/problem events/change in research can be submitted simultaneously
- pSites will submit annual enrollment data to the overall PI for inclusion in the continuing review application

  - Note: As this study is being conducted under a waiver of consent, the data JHM IRB would collect would differ than what would be sought if the study included a written consent process [e.g., how many participants were enrolled? Failed screening? Adverse events?]
Example- JHM IRB for N3C
Local Context Questionnaire
JOHNS HOPKINS MEDICINE (JHM) IRB LOCAL CONTEXT QUESTIONNAIRE

Your site is participating in a study where JHM IRB will be the IRB of record. When relying on the JHM IRB, relying institutions must agree to provide the following important information to help the JHM IRB conduct its review:

- The requirements of any applicable state or local laws, regulations, institutional policies, standards, or other local factors, including local site ancillary reviews, relevant to the research that would affect the conduct or approval of the research at your institution.

**Please seek guidance from your HRPP/Research Office/IRB regarding how to complete the local context review process at your institution. Most IRBs require a local submission in order to initiate the local context review process.**

The local context questionnaire contains three important sections:

- Section 1: Relying Site Study Team Information
- Section 2: Applicable Local Requirements
- Section 3: The Conduct of this Study at the Relying Site

Please carefully review the approved protocol and complete the local context questionnaire below. We strongly recommend that the local context questionnaire be completed as a collaborative effort. Often, to ensure all necessary information is captured, information from the local site PI, in addition to the IRB/institutional contact is required. **Please Note:** Signatures by both the local site PI and the Institutional Contact are required. Please be as careful as possible in completing this questionnaire so that the document does not need to be re-signed.

Study team members with questions about completing this local context questionnaire can contact [INSERT CONTACT NAME] at [INSERT EMAIL ADDRESS]. If your organization’s IRB has questions about this form, please contact Janelle Maddox-Regis at jmaddox3@jhmi.edu.

Your contributions to our IRB review process are important and we appreciate your assistance in providing your local context questionnaire.
**Study Title:** National COVID Cohort Collaborative (N3C): A national resource for shared analytics  

**Overall Study PI:** Christopher Chute  

**JHM IRB Protocol:** IRB00249128

### Section 1: Relying Site Study Team Information

<table>
<thead>
<tr>
<th>Relying Institution Information</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1a. Legal Name of Relying Institution:</strong></td>
<td>Click or tap here to enter text.</td>
</tr>
<tr>
<td><strong>1b. Name of Relying Site PI:</strong></td>
<td>Click or tap here to enter text.</td>
</tr>
<tr>
<td><strong>1c. Relying Site PI Phone #:</strong></td>
<td>Click or tap here to enter text.</td>
</tr>
<tr>
<td><strong>1d. Relying Site PI Email:</strong></td>
<td>Click or tap here to enter text.</td>
</tr>
<tr>
<td><strong>1e. Name of Relying Site Lead Study Contact:</strong></td>
<td>Click or tap here to enter text.</td>
</tr>
<tr>
<td><strong>1f. Relying Site Lead Study Contact Phone #:</strong></td>
<td>Click or tap here to enter text.</td>
</tr>
<tr>
<td><strong>1g. Relying Site Lead Study Contact Email:</strong></td>
<td>Click or tap here to enter text.</td>
</tr>
<tr>
<td><strong>1h. FWA #:</strong></td>
<td>Click or tap here to enter text.</td>
</tr>
<tr>
<td><strong>1i. FWA Expiration Date:</strong></td>
<td>Click or tap here to enter text.</td>
</tr>
<tr>
<td><strong>1j. Does your FWA require you apply 45 CFR 46 to all studies regardless of funding source (e.g., &quot;check the box&quot;)?</strong></td>
<td>YES ☐ NO ☐</td>
</tr>
<tr>
<td><strong>1k. List all institutions that are considered components under your FWA:</strong></td>
<td>Click or tap here to enter text.</td>
</tr>
</tbody>
</table>
| **1l. Does your site have an IRB?** | YES ☐ NO ☐  
*If YES, provide the IRB contact information:*  
**URL for the IRB/HRPP (if applicable):** | Click or tap here to enter text. |
| **1m. Is your site AAHRPP accredited?** | YES ☐ NO ☐  

**1n. Please review the planned list of personnel who will be engaged in human subjects research at your institution and verify that all of your institutionally-required training for the conduct of the research [including human subjects protections training, GCP training, and HIPAA training, as applicable] has been completed for each individual.**  

**Training Completed ☐**  

*Note: This form should not be submitted for central review if training has not been completed.*  

<p>| <strong>1o. Are all involved individuals from your institution credentialed and/or appropriately qualified and meet the institution’s standards for eligibility to</strong> | I confirm that all involved individuals are credentialed and/or appropriately qualified ☐ |</p>
<table>
<thead>
<tr>
<th>conduct the research as described in the approved protocol?</th>
<th>Note: This form should not be submitted for central review unless all individuals are credentialed and/or appropriately qualified.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1p. Did the institution determine there is a relevant individual or institutional financial COI for this protocol?</strong></td>
<td>YES ☐ NO ☐</td>
</tr>
<tr>
<td><strong>If yes:</strong></td>
<td></td>
</tr>
<tr>
<td>(1) Provide a summary of the conflict and management plan or attach documentation: Click or tap here to enter text.</td>
<td></td>
</tr>
<tr>
<td>(2) Provide an institutional Point Of Contact for questions related to the local management plan [This person should be someone in the office/entity who prepared the management plan]: Click or tap here to enter text.</td>
<td></td>
</tr>
</tbody>
</table>
## Section 2: Applicable Local Requirements

### 2a. Please review the protocol and identify areas where there are unique state, local or federal regulatory requirements that apply to the conduct of this study at your site (e.g., legally authorized representatives, state laws regarding confidentiality of specific types of health information, emancipated minors) and describe any steps that must be taken to adhere to these requirements.

*Note: Only include what’s relevant to the conduct of this study at your site. Please outline any specific changes needed to ensure adherence with the requirements you have identified. This information needs to be considered as part of the JHM IRB review.*

### 2b. Please review the protocol and identify any institutional requirements (e.g., policy or procedural requirements such as recruitment, data security, remuneration) that apply to this study and describe any steps that must be taken to adhere to these requirements.

*Note: Only include what’s relevant to the conduct of this study at your site. Please outline any specific changes needed to ensure adherence with the requirements you have identified. This information needs to be considered as part of the JHM IRB review.*

### 2c. Does your organization require that the IRB grant a waiver of privacy authorization under HIPAA for any of the following recruitment activities?

Check all that apply:

- Medical record review or other access to PHI (of potential subjects who are patients of the research team)
- Medical record review or other access to PHI (of potential subjects who are not patients of the research team)
<table>
<thead>
<tr>
<th><strong>2d. Please identify the ancillary reviews</strong> [e.g., radiation safety review, review for research with bio-specimens, drug/device safety review, etc.] that are <strong>applicable to this study</strong> and are required before the study may be initiated <strong>at your site</strong>.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Please confirm that these ancillary reviews have been completed and provide the outcome of those reviews (including any changes required to the conduct of the study).</td>
</tr>
<tr>
<td>□ N/A to the conduct of this study at this site</td>
</tr>
<tr>
<td>Click or tap here to enter text.</td>
</tr>
<tr>
<td>N/A – no ancillary reviews</td>
</tr>
<tr>
<td>Ancillary Reviews Completed</td>
</tr>
<tr>
<td>Provide the ancillary review outcome(s) and attach any relevant documentation.</td>
</tr>
<tr>
<td>Click or tap here to enter text.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>2e. Are there sources of support that are unique to your site?</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>YES ☐  NO ☐</td>
</tr>
<tr>
<td>If yes, Check all sources of support (pending or awarded) and indicate the source name:</td>
</tr>
<tr>
<td>☐ Monetary</td>
</tr>
<tr>
<td>☐ Material or Equipment (e.g., drugs or devices)</td>
</tr>
<tr>
<td>☐ None of the above</td>
</tr>
<tr>
<td>Click or tap here to enter text.</td>
</tr>
</tbody>
</table>
### Section 3: The Conduct of This Study at the Relying Site

<table>
<thead>
<tr>
<th>Question</th>
<th>YES</th>
<th>NO</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>3a. Are there any differences to the initial contact and/or recruitment plan at your site from that described in the protocol or associated documents based on local requirements or state law?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>If yes, please describe the differences and specify whether you have attached any site-specific recruitment materials for IRB review:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Click or tap here to enter text.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3b. Please review the protocol and verify that there are sufficient resources available at your site to carry out the research as planned, including study team members with prior clinical trial experience.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>(Please note: the answer to this must be YES prior to submission)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Click or tap here to enter text.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3c. Are there any different requirements for how data will be accessed and/or stored at your site from those described in the protocol or associated documents based on local requirements or state laws?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>N/A to this study’s conduct at this site</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3d. Are there any other different requirements for how the protocol will be implemented and/or conducted at your site based on local requirements or state laws?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>If yes, explain:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Click or tap here to enter text.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3e. Does your institution have any policies related to data security?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>If yes, please describe:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Click or tap here to enter text.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3f. Please confirm that the plans for data sharing as outlined in the protocol comply with your institutional requirements. If additional requirements [e.g. agreements, data security provisions, etc.] are required for your site please provide a summary of these requirements with your response.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Click or tap here to enter text.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question</td>
<td>Response Options</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
<td>------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3g. Does your site have a data security/governance committee that is required to review and approve this protocol prior to implementation?</td>
<td>YES ☐ NO ☐</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>If yes</strong>, please indicate their review outcome and/or attach their approval letter:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Click or tap here to enter text.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3h. Please review the protocol and identify whether there are any special characteristics/concerns of your community of which the reviewing IRB should be aware for this specific study. Please also outline any steps that must be taken to address these concerns.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Click or tap here to enter text.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>☐ None</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>☐ Characteristics/concerns have been identified.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Explain: Click or tap here to enter text.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3i. It is possible that the JHM SIRB may have additional questions about your local community. Please include the best contact below for additional questions about local site information.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Local Site Contact Name: Click or tap here to enter text.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Email Address: Click or tap here to enter text.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phone #: Click or tap here to enter text.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Signatures/Attestations**

By signing below, the signatories affirm that they have reviewed the SMART IRB Agreement, Letter of Indemnification and the responsibilities of relying institutions and attest that the information fulfills the relying institutions responsibilities for the provision of local context information.

As specified in the Agreement, Relying Institution is solely responsible for consulting with its own legal counsel to determine whether research reviewed by Reviewing IRB (including but not limited to any consent process or documentation and any HIPAA documentation), meets all other applicable federal, state, and local legal and policy requirements, including but not limited to HIPAA compliance. Relying Institution is solely responsible for identifying all ancillary reviews required by applicable regulation or policy in the Reliance Application and must notify Reviewing IRB of the outcome of such reviews prior to final protocol approval.

<table>
<thead>
<tr>
<th>Local Site Investigator Signature:</th>
<th>Institutional Contact [e.g. HRPP Lead] Signature:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Role/Title: Click or tap here to enter text.</td>
</tr>
<tr>
<td>Print Full Name: Click or tap here to enter text.</td>
<td>Print Full Name: Click or tap here to enter text.</td>
</tr>
<tr>
<td>Contact Phone Number/Email: Click or tap here to enter text.</td>
<td>Contact Phone Number/Email: Click or tap here to enter text.</td>
</tr>
<tr>
<td>Date of Signature: Click or tap here to enter text.</td>
<td>Date of Signature: Click or tap here to enter text.</td>
</tr>
</tbody>
</table>
Protocol
Study Overview and Goals

Background and Significance

Significance

Research Design and Methods

Data Partnership and Governance

Data Use Agreement

Single IRB

Data Access Committee

Phenotype & Data Acquisition

Clinical Phenotype (Query)

CDM Translations

Algorithm Execution at Contributing Sites

Data Workflow

Authentication

Data Ingestion

Collaborative Analytics

Data Ingestion & Harmonization

Data Staging Quality Control

Data Harmonization into Common Data Model

Collaborative Analytics

Data Enclave Environment

Modes of Access

Synthetic Derivative

Human Subjects

Protection of Human Subjects

Data Access Committee

Synthetic Derivative

Results Sharing

Inclusion of Women

Inclusion of Minorities

Inclusion of Children

References
Study Overview and Goals

The National COVID Cohort Collaborative (N3C) proposes to establish a central registry of patients who have been tested for COVID or have a clinical diagnosis of COVID. This will be derived by harmonizing COVID clinical data extracted from the federated clinical repositories associated with the Common Data Model (CDM) programs, enumerated below. These data ultimately are extracted from the electronic health records (EHRs) of the medical centers who will contribute data. The requested data constitutes a HIPAA limited data set\(^1\) in that it will contain dates and zip codes; both are necessary for the epidemiologic tracking of the COVID pandemic over time and space. The data will be updated from the contributing sites as frequently as practical, ideally twice a week.

The data extracted at each contributing site will be transferred by secure FTP to a FedRAMP\(^2\) certified staging area on the NIH National Center for Advancing Translational Science (NCATS) server. There it will be harmonized into a single data model, OMOP 5.3.1\(^3\) after multiple data quality checks. The merged dataset will be transferred to the Palantir analytic platform\(^4\), which can allow sophisticated data science analytics while preventing data download or human browsing of individual-level data. Access to this limited dataset will be determined by the N3C Data Access Committee.

Additionally, a synthetic dataset will be derived from the limited dataset, created by sampling from the statistical distribution of the underlying data. The synthetic dataset, or "synthetic derivative" does not contain data of real patients and cannot be traced to real patients; therefore, it can be used free of HIPAA regulations. This dataset is proposed to enable more open data science investigation, and would be accessible through a simple registration and attestation process.

Background and Significance

The Center for Data to Health (CD2H)\(^5\) is the coordinating center for informatics activities throughout the CTSA network of academic medical centers\(^6\). Leadership at NIH/NCATS requested CD2H to coordinate the contributions by CTSA hub sites to a central commons of clinical data about COVID tested and diagnosed persons. Subsequently, this proposed commons was adopted by NIH overall, with many NIH centers, institutes, and offices partnering with NCATS in its design and approval. Additionally, many federal agencies including CDC, FDA, and HHS leadership contributed to enabling this initiative. The Department of Veterans Affairs agreed to contribute relevant cases to this national COVID registry after a preliminary examination of the platform infrastructure and upon learning of its approved FedRAMP\(^2\) security status. Many CTSA hub sites have expressed enormous interest in contributing to this effort.

Most CTSA sites participate in one or more federated data cooperatives. All of these efforts involve academic medical centers creating a local repository of clinical data extracted from their
EHRs, that conforms to the collaborative data model (called a Common Data Model (CDM)). Queries are sent to the collaborating sites, where an analysis is conducted locally to address the query. Answers are returned to the coordinating center for that CDM network, where they are aggregated into a sophisticated meta-analyses.

The four major CDM collaboratives are:

Accelerating Clinical Trials (ACT): A CTSA supported collaborative of most CTSA sites operating with an i2b2 data model conformant to the ACT ontology structure. Federated queries occur through the SHRINE system.

PCORNet: The official federated network of the Patient-Centered Outcomes Research Institute is a US-based network of networks focusing on patient centered outcomes.

Observational Health Data Sciences and Informatics (OHDSI): is an international federation of clinical data sites, originating from the pharmacoepidemiology community.

TriNetX: ia an international network of clinical sites coordinated by a commercial entity established to facilitate case finding for pharmaceutical funded clinical trials.

These collaboratives excel in providing case counts, cohorts, and in many cases, answers to specific questions. The nature of federated analytics typically limits such questions to hypotheses testing: for example, whether a specific drug helps or hurts COVID patient courses. While it is possible to conduct federated logistic regression across sites, this is rarely done in practice. It is impractical to conduct many machine learning algorithms across federated datasets, as many algorithms require access to all analyzed data in the same computing environment.

Significance

Creating the N3C registry of individual-level (containing information specific to individual patients, sometimes called row-level) data as a limited--albeit protected--dataset of EHR data at a national level will be unprecedented in US clinical research. It will support novel machine learning analytics and discovery of important predictors associated with emergency visits, hospitalizations, ICU transfer, ventilator dependency, and death, amongst a myriad of related outcomes. It will have the scale, statistical power, and computing platform to address most questions the clinical and research communities seek to answer. Ideally it will be used by multidisciplinary teams engaging clinicians, statisticians, epidemiologists, biologists, and data scientists, who together can leverage this unique resource at this critical time in the COVID pandemic. Furthermore, these activities can be conducted with minimal risk of data breach or inadvertent disclosure, as any effort to view or download individual level data is generally disabled.
Research Design and Methods

CD2H established four Workstreams\textsuperscript{11} to undertake this work:

- Data Partnership & Governance
- Phenotype & Data Acquisition
- Data Ingestion & Harmonization
- Collaborative Analytics

Data Partnership and Governance

CD2H, in partnership with NCATS, has convened a community to leverage mechanisms from recent dataset integration efforts, particularly the All of Us research initiative\textsuperscript{12} at NIH. As will all workstreams, public meetings, list serves, Slack channels, and document repositories comprise efforts to engage as many CTSA hubs, stakeholders, and representatives from all four of the CDM communities.

A table of definitions of terms, shared between this protocol and the Data Use Agreement, is attached to this application.

The project is currently planned to exist for two years. Community, public benefit, and knowledge emergent throughout this time will dictate whether an extension request is warranted.

Data Use Agreement

Sage Bionetworks drafted the proposed Data Use Agreement (DUA), which is attached to this application. It describes data contribution, processing, access terms and conditions. It clarifies confidentiality agreements, intellectual property, warranties and liabilities, conflict resolution, and expiration or termination. This portion of the DUA would be signed by the clinical data contributor (such as a CTSA hub) and NCATS. The document also includes a proposed data access application for the limited dataset signed by the investigator team requesting access and NCATS. The access request stipulates the data is to be used only for research, cannot be downloaded, and all results must be shared with the community. Investigators must attest that they will not attempt to reidentify data in any way, and acknowledge the contributions of the communities that made the N3C resource possible.

Single IRB

To lower any burden associated with contributing data to the N3C, this protocol is submitted to establish a central IRB. N3C and NCATS leadership accept the offer of Johns Hopkins Medicine to operate in this capacity. Sites that contribute data to N3C are not obligated to use the central IRB, and may opt to operate through their own institutional IRB.
It is the preference of Johns Hopkins Medicine IRB to use the SMART IRB\textsuperscript{13} reliance agreement as the basis of reliance. The SMART IRB master reliance agreement was created in 2016 to harmonize and streamline the IRB review process for multisite studies. It enables reliance on a study-by-study basis, clearly defines roles and responsibilities of relying institutions and reviewing IRBs, and eliminates the need to sign reliance agreements for each study [e.g., a non-SMART IRB agreement]. 700+ institutions have already signed onto this agreement and are actively using it as the basis of reliance for multisite projects.

Data Access Committee

Requests to analyze the limited dataset will be reviewed by the N3C Data Use Committee (DAC). In general, permission and credentials to analyze will be permitted under these circumstances:

- Investigators are credentialed by a recognized academic, community, or commercial organization
- They attest to all terms and conditions in the data access DUA
  - Including returning all results and analytic code
  - Asserting they will not attempt to reidentify any data by any means
  - Acknowledge contributors to the resource
- They provide a compelling and feasible research plan appropriate for the resource

Members of the DAC will include representation from the collaboration federal agencies and contributing organizations. They will be appointed by CD2H in consultation with NIH and other federal agencies to represent the involved communities and the spectrum of skills among potential investigators.

Phenotype & Data Acquisition

Clinical Phenotype (Query)

CD2H and the Phenotype team have drafted a prototype query for clinical data extraction from the CDM platforms at the contributing sites; this document is attached to this proposal. It has been widely vetted among clinical experts, CDM representatives, and CTSA sites. While it serves as an initial consensus document we recognize that modification and updates, released as N3C phenotyping versions, will need to be made as tests, data, and clinical practice evolves through this pandemic. For example, the recently released serology tests to establish whether an individual has previously been infected with COVID may impact phenotyping strategies.

At present, there are four phenotype categories defined by the document.

Inclusion criteria:
- Patients of all ages
- Encounter on or after 1/1/2020
- Meeting any of the following criteria:
Lab Confirmed Negative
  ○ LOINC codes Negative result
  ○ Note: this may eventually comprise a large fraction of all patients at a medical center, thus we are likely to quickly implement sampling strategies for these persons. But not in version 1.

Lab Confirmed Positive
  ○ LOINC codes Positive result (based on enumerated LOINC codes)

Likely Positive
  ○ COVID Dx Code (from short list of ICD codes) only

Possible Positive
  ○ Two or more suggestive ICD codes only

Data transmitted:
The CDMs targeted contain information on the following categories:
  ● Demographics, including zip code
  ● A hashed identifier to support future linkage with imaging, viral RNA, or other major national data resources with a shared hash identifier (below).
  ● Laboratory tests and results, including dates
  ● Medications, including start and where available stop dates
  ● Vital signs, with dates
  ● Diagnoses, with dates
  ● Procedures, with dates
  ● Admission, Discharge, Transfer information with dates
    ○ Including death with dates

The information will be requested retrospectively for one year or longer, at the contributing sites discretion, prior to the query start date (1/1/2020), and through the present with serial data updates for a given patient.

Hashed identifier:
To support future linkage with other datasets, the N3C team will develop a strategy for sites to develop a unique, encrypted hashed identifier. The goal of a hashed identifier is to allow data to be linked using identifiers, such as name, date of birth, without actually disclosing those identifiers. The hashed identifier will be constructed using industry standards14.

CDM Translations
To minimize burden on contributing sites, the CD2H team will consult with representatives from each CDM community to translate the proto-query into reproducibly executable code to operate against the sites CDM repository. Four queries will be generated, one each for each CDM in the format native to each CDM. The queries will be validated on test data, and reviewed with CDM subject matter experts. These queries will be openly published on the N3C web site.
Algorithm Execution at Contributing Sites

Each site will select one and only one CDM at a time as the basis for its data contribution to minimize record duplication in the N3C repository. They may opt to change their sourcing CDM during participation. Sites may access the CDM specific phenotyping query, and execute at their site. Local modifications may be required, if for example not all COVID tests are not yet LOINC coded at their site. The CD2H will maintain a “white glove” helpdesk for each of the four CDM involved in the program. This helpdesk will be the primary support resource of the contributing sites. Each helpdesk in turn will have access to CDM subject matter experts supported on the N3C COVID supplement to CD2H. The helpdesk will help each site adopt the CDM query code, and facilitate the transfer of extracted data to the NCATS staging area via sFTP.

Data Workflow

The Data Workflow will consist of three key components: Authentication, Data Ingestion, and Collaborative Analytics.

Authentication

The Unified NCATS Authentication (UNA) system will be used to authenticate Data Platform users. The UNA system is an identity broker that facilitates collaboration by enabling networked applications to easily support multiple identity providers (IdPs) and authentication protocols. UNA is dynamically configurable, multi-tenant authorization and federation service for applications. It allows applications to authenticate users via multiple identity providers using standard protocols. The protocols are implemented as Passport.js\(^{15}\) strategies and include Open ID Connect\(^{16}\), SAML\(^{17}\), and WSFederation\(^{18}\). The authentication protocol allows users to access the platform. By utilizing a Federated solution, and if the organization is Federated with the NIH, users of various organizations can access the platform by using their own organization credentials. This authentication solution is dynamically configurable, which allows users to be authenticated using alternative identity providers (such as login.gov) in case an organization is not Federated with the NIH. This not only brings more security to the application, but provides flexibility in administration for system administrators and users alike. For additional security, this authentication solution utilizes a layered approach, where users and their credentials are verified by multiple authentication protocols and methods. Utilizing the layered structure with multiple identity providers allows the authentication solution to coordinate a seamless and undisturbed login for users. This authentication solution also serves as an Identity Broker, which allows client applications to support one or more identity providers.

Data Ingestion

The NCATS ingestion platform will be a tool called Adeptia\(^{19}\), which is a cloud based Platform as a Service. Adeptia will be highly restricted and not directly accessible by sites or
investigators. Users are added through invitation only and are monitored by system administrators. Strict Password policies are enforced, which include limited entry retries and password expiration options are available for administrators to enforce. Adeptia uses certificates in order to ensure security while data is transmitted across the platform. Adeptia is protected by protocols such as FTPS, SFTP, HTTPS, and many others that require the use of certificates in order to encrypt the data and to verify the digital signature of the application and the user sending the data.

Collaborative Analytics

The data analyses platform will be accessed through a portal, which invokes Palantir\(^4\). Palantir provides a variety of tools, which support common data analytics resources such as R, Python, and Jupyter. These tools are configured to disallow any downloading, and to prohibit individual-level data viewing by default. Palantir Federal Cloud Service has FedRAMP\(^2\) authorization at the Moderate level which is deployed on Amazon Web Services (AWS) GovCloud\(^20\) that will be leveraged for this Data Platform. Palantir utilizes a multi-tiered approach to security and compliance. Palantir utilizes automated audit and logging tools, and provides automatic alerting for high priority events. Palantir also utilizes continuous deployment methods to ensure patches and updates are delivered seamlessly without user or system-wide downside or effects. Palantir contains HIPAA compliance protocols, and enforces global standards such as Findability, Accessibility, Interoperability, and Reusability (FAIR\(^21\)) and General Data Protection Regulation (EU GDPR\(^22\)).

Data Ingestion & Harmonization

Data Staging Quality Control

Data entering the NCATS staging area by sFTP from a site will be managed in a separate partition space for each contributing site. The data will be unpacked from the transmitted tables, and reincarnated as a database under the CDM model and version forwarded by the contributing site. There will be three stages of data quality checking and validation.

First-phase: The live CDM replication of the data will undergo the suite of data quality metrics and dashboards made available by their corresponding CDM; all CDM communities support suit suites. Additionally, value set verification will be undertaken to ensure that coded data does not include obsolete or deprecated codes (widely present in laboratory and medication coding systems). Any discrepancies will be iterated with the contributing site at this phase.

Second-phase: After conversion of the CDM data into OMOP 5.3.1 (below), the transformed data will be examined for systematic transform errors. We will also run test queries on the native CDM version of the data and the transformed OMOP 5.3.1 dataset, and compare answers. Discrepancies will be addressed by the data transform team.
Third-phase: After OMOP transformation, the standard suite of data quality and dashboarding tooling from the OHDSI consortium will be run. Discrepancies and findings at this level will be shared with contributing sites, but not addressed until the next data refresh cycle.

Data Harmonization into Common Data Model

There are two phases of data harmonization work: creating the transform mappings, and executing them on contributed mappings.

The CD2H and NCATS have been working for years on methods, tooling, and strategies for model to model transformation. Several maps have been created to go from CDMs to the CDISC BRIDG\textsuperscript{23} metamodel; this allows transforms between and among each model. We are building on these foundations to create a series of pairwise data transforms, from the major versions of each CDM to OMOP 5.3.1. These transform mappings include not only structure to structure mappings, but also the value sets that are bound to those structures. For example some models use SNOMED codes for diagnoses, while others use differing versions of ICD codes. We will conform to the OMOP 5.3.1 data target with one exception, we will use ICD-10-CM codes for diagnoses to avoid any intellectual property concerns and obviate information loss engendered through data value transforms from the native ICD-10-CM sources in all US EHR data today.

The data transform mappings will be reviewed and validated by subject matter experts from each CDM. The mappings will be uploaded to a commercial extraction/transform/load (ETL) utility procured and used by NCATS, Adeptia.\textsuperscript{19}

We will use the data transform mappings on the Adeptia platform to conduct the ETL from the source CDM into OMOP 5.3. The platform also has intrinsic data quality metrics to supplement the public tools that will form the core elements of our data quality efforts.

Collaborative Analytics

The main goal of this project is to enable investigators to conduct data discovery on the COVID cohort data. Thus, the analytics environment is key.

Data Enclave Environment

We will establish a secure data enclave, which is defined as:

\begin{quote}
A data enclave is a secure network through which confidential data, such as identifiable information from census data, can be stored and disseminated. In a virtual data enclave a researcher can access the data from their own computer but cannot download or remove it from the remote server.\textsuperscript{24}
\end{quote}
The Palantir product can enforce “enclave” conditions. The platform will include the standard suite of R, Python, and Jupyter analytic resources, and can support the uploading of custom libraries.

Modes of Access

We anticipate three modes of access to N3C data:

Public data scientists: This category will have a low bar for entry, requiring only investigator registration. However, access will only be granted to the synthetic derivative (see below) of the data. We expect most teams will initially use this access path. All teams doing software development that requires viewing of individual-level data must use this mode. Should compelling findings be identified, investigators may apply to validate their findings on the limited data set enclave. Even though this data is synthetic, it will remain on the Palantir data enclave, and cannot be downloaded.

Limited data set access: This category permits investigators to run software in the data enclave against the limited dataset, though not to see the individual-level data. Software development requiring any viewing of individual-level data must be done on the synthetic derivative. Access via this mechanism will require DAC review and approval. All the conditions and stipulations of the data access DUA would pertain at this level.

Direct Data Access: In rare cases investigators may petition for direct access to the limited dataset within the enclave. They will still be prohibited from downloading data. Access at this level would require IRB approval from their home institution and approval by the DAC.

Synthetic Derivative

Synthetic data is a term of art in clinical and translational research referencing clinical data that has been “synthesized” from real clinical data in a manner that obviates any reidentification. The explicit goal is to create a non-human subjects research dataset that statistically resembles the original data, but cannot be used to reidentify any subjects. The technique analyses the true source data to describe its contents, for example laboratory data, as a mathematical distribution of those results. Data is “synthesized” by randomly sampling from that mathematical distribution. The same is done for date and demographic information. Relationships between elements, for example drug use and diagnoses, is preserved by calculating a series of correlations and associations; these are then used to inform the statistical sampling in a manner that preserves these statistical associations.

Ideally, synthetic data are nearly identical to original PHI data, and can be analyzed as if they were original data but without any privacy concerns. Not only is there significant potential to protect patient privacy through analysis of data as a synthetic derivative, but synthetic derivatives of data can enable data sharing and accelerate discovery. Once real patient data are...
synthesized, the resulting data set no longer contains data on individual patients, but rather is a collection of observations which maintain the statistical properties of the original data set. Since the data set no longer contains data on real patients, synthetic derivatives can be shared between researchers at different institutions.\(^{25}\)

The MDClone\(^ {26}\) synthetic data engine will be used to generate the synthetic data derivative in this project. MDClone’s synthetic data engine is able to take any given data set, analyze its statistical properties, and create a brand-new set of data for research and clinical decision-making. This new data is synthetic, and since it does not contain data of real patients and cannot be traced to real patients, it can be used free of HIPAA regulations.

Human Subjects

Protection of Human Subjects

We understand the principle of minimum necessary data. The core challenge with the COVID pandemic is that society does not have the luxury of leisurely proposing hypotheses that can be addressed through the more conventional federated data networks. N3C intends to allow the research community the opportunity to leverage a harmonized clinical dataset in support of a variety of secondary data analyses, including machine learning algorithms. Among the characteristics of machine learning algorithms is their capacity to discover features and predictors that might otherwise be unanticipated. We are not proposing “black box” prediction algorithms, but rather an emphasis on looking at all the data to discover interpretable elements that can help medicine fight this condition, armed with information.

These aspirations require the creation of a central registry, and of making it a limited dataset. How the pandemic moves and behaves through time and over geographic areas requires dates and zip codes ideally at the five digit level.

Given these realities, it is our obligation to protect this large, limited data set to the best of our ability, minimizing risk of reidentification, breach, or inadvertent disclosure. Moreover, use of these data is limited to secondary data analyses; there will be no participant contact.

Enclave Computing Platform

We believe the most secure action we can take to balance the need for legitimate access with patient privacy and confidentiality is to restrict access to a data enclave environment. We are fortunate that NCATS has been establishing such an environment with the Palantir resource. The enclave will disallow any data downloading. In addition, except for extremely limited use cases (above in Mode of Access), it will prohibit the individual-level viewing of the limited dataset.
Data Access Committee

All access to the limited data set must be approved by the DAC, described above. Their role is to identify bone fide researchers with compelling questions. Investigators are expected to have explored the synthetic derivative prior to requesting access to the limited dataset. Investigators requesting such access must also agree to all conditions and stipulations in the data access DUA.

Synthetic Derivative

To maximize the public benefit of this resource, a synthetic derivative of the data, which by its very nature will not contain any identifiers, will be generated, as described above. The synthetic derivative does not contain data of real patients and cannot be traced to real patients; therefore, it can be used free of HIPAA regulations. While barriers to this data will be lower, investigators are still expected to attest that they will use this data only for COVID research and disclose all results for the public good. They will not be permitted to download the data, and must do analyses and software development on the Palantir platform.

Results Sharing

Any risk associated with a large, aggregated dataset must be offset by benefit. To ensure timely benefit in the face of this urgent pandemic, all investigators will be required to share intermediate and final results, as well as the computer code to generate these. This is easily enforced on the Palantir platform, which can record all analyses made. This would include software generated on the synthetic copy, and of course all software run on the limited dataset.

Inclusion of Women

The distribution of women in the N3C repository will correspond to the distribution of women in the clinical contributing sites fulfilling the phenotyping query. No effort will be made to exclude or otherwise select for gender.

Inclusion of Minorities

The distribution of minorities in the N3C repository will correspond to the distribution of minorities in the clinical contributing sites fulfilling the phenotyping query. No effort will be made to exclude or otherwise select for race or ethnicity.

Inclusion of Children

The distribution of children in the N3C repository will correspond to the distribution of children in the clinical contributing sites fulfilling the phenotyping query. No effort will be made to exclude or otherwise select for patients under 21.
References


OMOP dictionary
COVID-19 Clinical Data Warehouse Data Dictionary

Based on OMOP Common Data Model
Specifications Version 5.3.1

Contents

1 PERSON ............................................................................................................... 2 2
OBSERVATION_PERIOD ................................................................. 4 3
SPECIMEN ................................................................................................. 4 4
DEATH ........................................................................................................ 6 5
VISIT_OCCURRENCE ........................................................................ 6 6
PROCEDURE_OCCURRENCE ..................................................... 8 7
DRUG_EXPOSURE ........................................................................... 9 8
DEVICE_EXPOSURE ........................................................................ 11 9
CONDITION_OCCURRENCE ...................................................... 12 10
MEASUREMENT ............................................................................. 14 11
OBSERVATION .................................................................................. 16

Standardized Health System Data Tables ...........................................
17 12 LOCATION ...............................................................................
19 13 CARE_SITE ..............................................................................
19 14 PROVIDER ................................................................................
20 OHDSI Release notes v5.3 .......................................................... 21
This document is a modification of an existing OHDSI document.

The NIH COVID-19 Data Warehouse is an NIH data sharing resource, operated under a contract containing clinical and imaging data from individuals who have received a Coronavirus Disease 2019 (“COVID-19”) tested or whose symptoms are consistent with COVID-19. Data will also be collected from individuals infected with pathogens such as SARS 1, MERS, and H1N1 to support comparative studies.

Note the tables and fields have contain both “required” and “non-required data”. The required information will be expected as part of the data pull on each patient where non-required fields are optional.

The field person_id is a unique identifier of an individual patient however it is NOT the institutions individual medical record number, (MRN). The person_id is generated at the time of de-identification and will be used to concatenate individual’s data in the longitudinal data warehouse.

1 PERSON The person domain contains records that uniquely identify each patient in the source data who is time at- risk to have clinical observations recorded within the source systems. Each person record has associated demographic attributes which are assumed to be constant for the patient throughout the course of their periods of observation. All other patient-related data domains have a foreign-key reference to the person domain.

Field Required Type Description
person_id Yes integer A unique identifier for each person.

gender_concept_id Yes integer A foreign key that refers to a standard concept identifier in the Standardized Vocabularies for the gender of the person.

year_of_birth Yes integer The year of birth of the person. For data sources with date of birth, the year is extracted. For data sources where the year of birth is not available, the approximate year of birth is derived based on any age group categorization available.

month_of_birth No integer The month of birth of the person. For data sources that provide the precise date of birth, the month is extracted and stored in this field.

race_concept_id Yes integer A foreign key that refers to a standard concept identifier in the Standardized Vocabularies for the race of the person.

ethnicity_concept_id Yes integer A foreign key that refers to the standard concept identifier in the Standardized Vocabularies for the ethnicity of the person.

location_id No integer A foreign key to the place of residency for the person in the location table, where the detailed address information is stored.

provider_id No integer A foreign key to the primary care provider the person is seeing in the provider table.

NCATS Pre-decisional Document: Tables provided for informational purposes only Version: 04.27.20

Field Required Type Description

care_site_id No integer A foreign key to the site of primary care in the care_site table, where the details of the care site are stored.

person_source_value No varchar(50) An (encrypted) key derived from the person identifier in the source data. This is necessary when a use case requires a link back to the person data at the source dataset.

gender_source_value No varchar(50) The source code for the gender of the person as it appears in the source data. The person’s gender is mapped to a standard gender concept in the
Standardized Vocabularies; the original value is stored here for reference.

gender_source_concept_id No Integer A foreign key to the gender concept that refers to the code used in the source.

race_source_value No varchar(50) The source code for the race of the person as it appears in the source data. The person race is mapped to a standard race concept in the Standardized Vocabularies and the original value is stored here for reference.

race_source_concept_id No Integer A foreign key to the race concept that refers to the code used in the source.

ethnicity_source_value No varchar(50) The source code for the ethnicity of the person as it appears in the source data. The person ethnicity is mapped to a standard ethnicity concept in the Standardized Vocabularies and the original code is stored here for reference.

ethnicity_source_concept_id No Integer A foreign key to the ethnicity concept that refers to the code used in the source.

CONVENTIONS
• Valid Gender, Race and Ethnicity Concepts belong to the "Demographic" domain.
• Person source data attributes are race, gender, and ethnicity.
• Ethnicity in the OMOP CDM follows the OMB Standards for Data on Race and Ethnicity: Only distinctions between Hispanics and Non-Hispanics are made.
• Additional information is stored through references to other tables about the home address (location_id) and the primary care provider.
• The provider refers to the primary care provider (General Practitioner).
• The care site refers to where the provider typically provides the primary care.
• All persons are required to have a valid gender and year of birth.
presents only Version: 04.27.20

- The person table requires only one value for each attribute. While it is possible for a person to change genders, locations, and providers over time, it is the responsibility of the data holder to select the one value to use in the CDM.

2 OBSERVATION_PERIOD The observation period domain contains records which uniquely define the spans of time for which a person is at-risk to have clinical events recorded within the source systems. One person may have one or more disjoint observation periods, during which times analyses may assume that clinical events would be captured if observed, and outside of which no clinical events may be recorded.

Field Required Type Description

observation_period_id Yes integer A unique identifier for each observation period.

person_id Yes integer A foreign key identifier to the person for whom the observation period is defined. The demographic details of that person are stored in the person table.

observation_period_start_date Yes date The start date of the observation period for which data are available from the data source.

observation_period_end_date Yes date The end date of the observation period for which data are available from the data source.

period_type_concept_id Yes Integer A foreign key identifier to the predefined concept in the Standardized Vocabularies reflecting the source of the observation period information

CONVENTIONS

- Each Person can have more than one valid OBSERVATION_PERIOD record, but no two observation periods can overlap in time for a given person.

- During an Observation Period, any clinical event that happens to the patient is expected to be recorded. Conversely, the absence of data indicates that no clinical events occurred to the patient.

- No clinical data are valid outside an active Observation Period. Clinical data that refer to a time outside (diagnoses of previous conditions such as "Old MI" or medical history) of an active Observation Period are recorded as Observations. The date of the Observation is the first day of the first Observation Period of a patient.

- For claims data, observation periods are inferred from the enrollment periods to a health benefit plan.

3 SPECIMEN The specimen domain contains the records identifying each biological
sample from a person.

**Field Required Type Description**

specimen_id Yes integer A unique identifier for each specimen.

person_id Yes integer A foreign key identifier to the person for whom the specimen is recorded.

NCATS Pre-decisional Document: Tables provided for informational purposes only Version: 04.27.20

**Field Required Type Description**

specimen_concept_id Yes integer A foreign key referring to a standard concept identifier in the Standardized Vocabularies for the specimen.

specimen_type_concept_id Yes integer A foreign key referring to the predefined concept identifier in the Standardized Vocabularies reflecting the system of record from which the specimen was represented in the source data. specimen_date Yes date The date the specimen was obtained from the person.

specimen_time No time The time on the date when the specimen was obtained from the person.

quantity No float The amount of specimen collection from the person during the sampling procedure.

unit_concept_id No integer A foreign key to a standard concept identifier for the unit associated with the numeric quantity of the specimen collection.

anatomic_site_concept_id No integer A foreign key to a standard concept identifier for the anatomic location of specimen collection.

disease_status_concept_id No integer A foreign key to a standard concept identifier for the disease status of specimen
CONVENTIONS

• Anatomic site is coded at the most specific level of granularity possible, such that higher level classifications can be derived using the Standardized Vocabularies

4 DEATH The death domain contains the clinical event for how and when a person dies. A person can have up to one record if the source systems contain evidence that s/he is deceased, such as:

i. Condition Code in the Header or Detail information of claims ii. Status of enrollment into a health plan iii. Explicit record in EHR data

Living patients should not contain any information in the death table.

Field Required Type Description
person_id Yes integer A foreign key identifier to the deceased person.
The demographic details of that person are stored in the person table. death_date Yes date The date the person was deceased. If the precise date including day or month is not known or not allowed, December is used as the default month, and the last day of the month the default day. death_type_concept_id Yes integer A foreign key referring to the predefined concept identifier in the Standardized Vocabularies reflecting how the death was represented in the source data. cause_concept_id No integer A foreign key referring to a standard concept identifier in the Standardized Vocabularies for conditions. cause_source_value No varchar(50) The source code for the cause of death as it appears in the source data. This code is mapped to a standard concept in the Standardized Vocabularies and the original code is stored here for reference. cause_source_concept_id No integer A foreign key to the concept that refers to the code used in the source. Note, this variable name is abbreviated to ensure it will be allowable across database platforms.

CONVENTIONS

• Each Person may have more than one record of death in the source data. It is the task of the Extract Transform and Load (ETL) to pick the most plausible or most accurate records to be aggregated and stored as a single record in the Death table.

• If the Death Date cannot be precisely determined from the data, the best approximation should be used.

5 VISIT_OCCURRENCE The visit domain contains the spans of time a person continuously receives medical services from one or more providers at a care site in a given setting within the health care system. Visits are classified into 4 settings: outpatient care, inpatient confinement, emergency room, and long-term care. Persons may transition between these settings over the course of an episode of care. If applicable, relationships between visits within an episode of care may be represented in the FACT_RELATIONSHIP table.

Visits are recorded in various data sources in different forms with varying levels of standardization. For example:

i. Medical Claims include Inpatient Admissions, Outpatient Services, and Emergency Room visits.

ii. Electronic Health Records may capture Person visits as part of the activities recorded.
<table>
<thead>
<tr>
<th>Field Required Type Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>visit_occurrence_id Yes integer A unique identifier for each person's visit or encounter at a healthcare provider.</td>
</tr>
<tr>
<td>person_id Yes integer A foreign key identifier to the person for whom the visit is recorded. The demographic details of that person are stored in the person table.</td>
</tr>
<tr>
<td>visit_concept_id Yes integer A foreign key that refers to a visit concept identifier in the Standardized Vocabularies.</td>
</tr>
<tr>
<td>visit_start_date Yes date The start date of the visit.</td>
</tr>
<tr>
<td>visit_start_time No time The time the visit started.</td>
</tr>
<tr>
<td>visit_end_date Yes date The end date of the visit. If this is a one-day visit the end date should match the start date.</td>
</tr>
<tr>
<td>visit_end_time No time The time the visit ended.</td>
</tr>
<tr>
<td>visit_type_concept_id Yes Integer A foreign key to the predefined concept identifier in the Standardized Vocabularies reflecting the type of source data from which the visit record is derived.</td>
</tr>
<tr>
<td>provider_id No integer A foreign key to the provider in the provider table who was associated with the visit.</td>
</tr>
<tr>
<td>care_site_id No integer A foreign key to the care site in the care site table that was visited.</td>
</tr>
<tr>
<td>visit_source_value No Varchar(50) The source code for the visit as it appears in the source data.</td>
</tr>
<tr>
<td>visit_source_concept_id No Integer A foreign key to a concept that refers to the code used in the source.</td>
</tr>
</tbody>
</table>

**CONVENTIONS**

- A Visit Occurrence is recorded for each visit to a healthcare facility.
- Valid Visit Concepts belong to the "Visit" domain.
- Standard Visit Concepts are defined as Inpatient Visit, Outpatient Visit, Emergency Room Visit and Long-Term Care Visit. Source concepts from place of service vocabularies are mapped into these standard visit concepts in the Standardized Vocabularies.
Each Visit is standardized by assigning a corresponding Concept Identifier based on the type of facility visited and the type of services rendered.

- At any one day, there could be more than one visit.

One visit may involve multiple providers, in which case the ETL must specify how a single provider id is selected or leave the provider_id field null.

One visit may involve multiple care sites, in which case the ETL must specify how a single care_site id is selected or leave the care_site_id field null.

### 6 PROCEDURE_OCCURRENCE

The procedure domain contains records of activities or processes ordered by and/or carried out by a healthcare provider on the patient to have a diagnostic and/or therapeutic purpose. Procedures are present in various data sources in different forms with varying levels of standardization. For example:

iii. Medical Claims include CPT-4, ICD-9-CM (Procedures), and HCPCS procedure codes that are submitted as part of a claim for health services rendered, including procedures performed.

iv. Electronic Health Records that capture CPT-4, ICD-9-CM (Procedures), HCPCS or OPCS-4 procedures as orders.

#### Field Required Type Description

- **procedure_occurrence_id**: Yes integer A system-generated unique identifier for each procedure occurrence.
- **person_id**: Yes integer A foreign key identifier to the person who is subjected to the procedure. The demographic details of that person are stored in the person table.
- **procedure_concept_id**: Yes integer A foreign key that refers to a standard procedure concept identifier in the Standardized Vocabularies.
- **procedure_date**: Yes date The date on which the procedure was performed.
- **procedure_type_concept_id**: Yes integer A foreign key to the predefined concept identifier in the Standardized Vocabularies reflecting the type of source data from which the procedure record is derived.
- **modifier_concept_id**: No integer A foreign key to a standard concept identifier for a modifier to the procedure (e.g. bilateral). Quantity No integer The quantity of procedures ordered or administered.
- **provider_id**: No integer A foreign key to the provider in the provider table who was responsible for carrying out the procedure.
- **visit_occurrence_id**: No integer A foreign key to the visit in the visit table during
which the procedure was carried out.

procedure_source_value No varchar(50) The source code for the procedure as it appears in
the source data. This code is mapped to a standard procedure concept in the Standardized Vocabularies and the
original code is, stored here for reference. Procedure source codes are typically ICD-9-Proc, CPT-4, HCPCS or
OPCS-4 codes. procedure_source_concept_id No integer A foreign key to a procedure concept that
refers to the code used in the
source.

NCATS Pre-decisional Document: Tables provided for informational
purposes only Version: 04.27.20

Field Required Type Description qualifier_source_value No varchar(50) The source code for the qualifier as
it appears in the source
data.

CONVENTIONS

• Valid Procedure Concepts belong to the "Procedure" domain. Procedure Concepts are based on a
variety of vocabularies: SNOMED-CT, ICD-9-Proc, CPT-4, HCPCS and OPCS-4.

• Procedures are expected to be carried out within one day.

• Procedures could involve the application of a drug, in which case the procedural component is
recorded in the procedure table and simultaneously the administered drug in the drug exposure
table when both the procedural component and drug are identifiable.

• If the quantity value is omitted, a single procedure is assumed.

• The Procedure Type defines from where the Procedure Occurrence is drawn or inferred; for
administrative claims records, the type indicates whether a Procedure was primary or secondary
and their relative positioning within a claim.

• The Visit during which the procedure was performed is recorded through a reference to the
VISIT_OCCURRENCE table. This information is not always available.

• The Provider carrying out the procedure is recorded through a reference to the PROVIDER table.
This information is not always available.

7 DRUG_EXPOSURE The drug exposure domain captures records about the inferred
utilization of a biochemical substance with a physiological effect when ingested or otherwise
introduced into the body. Drugs include prescription and over-the-counter medicines, vaccines, and
large-molecule biologic therapies. Drug exposure is inferred from clinical events associated with
orders, prescriptions written, pharmacy dispensing, procedural administrations, and other
patient-reported information.

Drug Exposure records are recorded from a variety of source
information:
i. The “Prescription” section of an EHR captures prescriptions written by physicians or from electronic ordering systems

ii. The ”Medication list” section of an EHR for both non-prescription products and medications prescribed by other providers

iii. Prescriptions filled at dispensing providers such as pharmacies, and then captured in reimbursement claim systems

iv. Drugs administered as part of a Procedure, such as chemotherapy or vaccines

Only drugs with active pharmaceutical ingredients are recorded. Radiological devices ingested or applied locally do not count as drugs.

Field Required Type Description

<table>
<thead>
<tr>
<th>Field</th>
<th>Required</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>drug_exposure_id</td>
<td>Yes</td>
<td>Integer</td>
<td>A system-generated unique identifier for each drug utilization event.</td>
</tr>
<tr>
<td>person_id</td>
<td>Yes</td>
<td>Integer</td>
<td>A foreign key identifier to the person who is subjected to the drug.</td>
</tr>
<tr>
<td>drug_concept_id</td>
<td>Yes</td>
<td>Integer</td>
<td>A foreign key that refers to a standard concept identifier in the Standardized Vocabularies for the drug concept.</td>
</tr>
<tr>
<td>drug_exposure_start_date</td>
<td>Yes</td>
<td>date</td>
<td>The start date for the current instance of drug utilization. Valid entries include a start date of a prescription, the date a prescription was filled, or the date on which a drug administration procedure was recorded.</td>
</tr>
<tr>
<td>drug_exposure_end_date</td>
<td>No</td>
<td>date</td>
<td>The end date for the current instance of drug utilization. It is not available from all sources.</td>
</tr>
</tbody>
</table>
| drug_type_concept_id       | Yes      | Integer   | A foreign key to the predefined concept identifier in the Standardized Vocabularies reflecting the type of drug exposure recorded. It indicates how the drug exposure was represented in the source.
stop_reason No varchar(20) The reason the medication was stopped, where available. Reasons include regimen completed, changed, removed, etc.

Refills No integer The number of refills after the initial prescription. The initial prescription is not counted, values start with 0.

Quantity No float The quantity of drug as recorded in the original prescription or dispensing record.

days_supply No integer The number of days of supply of the medication as recorded in the original prescription or dispensing record.

Sig No CLOB The directions ("signetur") on the drug prescription as recorded in the original prescription (and printed on the container) or dispensing record.

route_concept_id No integer A foreign key to a predefined concept in the Standardized Vocabularies reflecting the route of administration.

effective_drug_dose No float Numerical value of drug dose for this drug_exposure record.

dose_unit_concept_id No integer A foreign key to a predefined concept in the Standardized Vocabularies reflecting the unit the effective_drug_dose value is expressed.

lot_number No varchar(50) An identifier to determine where the product originate d

provider_id No integer A foreign key to the provider in the provider table who initiated (prescribed) the drug exposure.
**Field Required Type Description**

- **visit_occurrence_id** No integer A foreign key to the visit in the visit table during which the drug exposure initiated.

- **drug_source_value** No varchar(50) The source code for the drug as it appears in the source data. This code is mapped to a standard drug concept in the Standardized Vocabularies.

  - **drug_source_concept_id** No Integer A foreign key to a drug concept that refers to the code used in the source.

- **route_source_value** No varchar(50) The information about the route of administration as detailed in the source.

- **dose_unit_source_value** No varchar(50) The information about the dose unit as detailed in the source.

**CONVENTIONS**

- Valid Drug Concepts belong to the "Drug" domain. Most Concepts in the Drug domain are based on RxNorm, but some may come from other sources. Concepts are members of the Clinical Drug or Pack, Branded Drug or Pack, Drug Component or Ingredient classes.

- Source drug identifiers, including NDC codes, Generic Product Identifiers, etc. are mapped to standard drug Concepts in the Standardized Vocabularies (e.g., based on RxNorm). When the Drug Source Value of the code cannot be translated into standard Drug Concept IDs, a Drug exposure entry is stored with only the corresponding source_concept_id and drug_source_value and a drug_concept_id of 0.

- The Drug Concept with the highest content of information is preferred during the mapping process: Concept Classes Branded Drug or Pack, followed by Clinical Drug, followed by Drug Component, and only if no other information is available the Ingredient. If only the drug class is known, no drug record should be written.

- A Drug Type is assigned to each Drug Exposure to track from what source the data were drawn or inferred.

- The Effective Drug Dose and the Dose Unit Concepts are provided in cases when the dose is explicitly provided, as it is typically for pediatric and chemotherapeutic treatments, and can only refer to a single active ingredient. Combination products which have doses for each ingredient need to be recorded as separate records.

- If possible, the visit in which the drug was prescribed or delivered is recorded through a reference to the visit table.
The device exposure domain captures records about a person's inferred exposure to a foreign physical object or instrument that which is used for diagnostic or therapeutic purposes through a mechanism beyond chemical action. Devices include implantable objects (e.g. pacemakers, stents, artificial joints), durable medical equipment and supplies (e.g. bandages, crutches, syringes), and other instruments used in medical procedures (e.g. sutures, defibrillators).

**Field Required Type Description**

- **device_exposure_id** Yes integer A system-generated unique identifier for each device.
- **person_id** Yes integer A foreign key identifier to the person who is subjected to the procedure. The demographic details of that person are stored in the person table.
- **device_concept_id** Yes integer Only the DI portion of the UDI would be captured as a Concept in the Standardized Vocabularies.
- **device_exposure_start_date** Yes date The date the device or supply was applied or used.
- **device_exposure_end_date** No date The date the device or supply was removed from use.
- **device_type_concept_id** Yes integer Provenance for the data, e.g. procedure device, from registry, etc.
- **unique_device_id** No varchar(50) The entire UDI or equivalent.
- **quantity** No integer The number of individual devices used for the exposure.
- **provider_id** No integer A foreign key to the provider in the provider table who was responsible for using the device.
- **visit_occurrence_id** No integer A foreign key to the visit in the visit table during which the device was used.
- **device_source_value** No varchar(50) The source code for the device as it appears in the
source data. This code is mapped to a standard device concept in the Standardized Vocabularies and the original code is stored here for reference.

device_source_concept_id No integer A foreign key to a device concept that refers to the code used in the source.

CONVENTIONS

• Valid Device Concepts belong to the "Device" domain.

• The distinction between devices or supplies and procedures are sometimes blurry, but the former are physical objects while the latter are actions, often to apply a device or supply.

• For medical devices that are regulated by the FDA, a Unique Device Identification (UDI) is required if available in the data source and is recorded in the unique_device_id field.

• The DI portion of that UDI is used to define concepts in the CONCEPT table. However, devices are also defined based on other source vocabularies, like HCPCS.

• The Visit during which the device was first used is recorded through a reference to the VISIT_OCCURRENCE table. This information is not always available.

• The Provider exposing the patient to the device is recorded through a reference to the PROVIDER table. This information is not always available.

9 CONDITION_OCCURRENCE The condition occurrence domain captures records of clinical observations of a person suggestive of the existence of disease or a medical condition based on diagnoses, signs and/or symptoms observed by a provider or reported by a patient.

Conditions are recorded in different sources and levels of standardization. For example:

NCATS Pre-decisional Document: Tables provided for informational purposes only Version: 04.27.20

i. Medical claims data include ICD-9-CM diagnosis codes that are submitted as part of a claim for health services and procedures.

ii. EHRs may capture Person conditions in the form of diagnosis codes and symptoms as ICD-9-CM codes but may not have a way to capture out-of-system conditions.

Field Required Type Description

condition_occurrence_id Yes integer A unique identifier for each condition occurrence
person_id Yes integer A foreign key identifier to the person who is experiencing the condition. The demographic details of that person are stored in the person table.

condition_concept_id Yes integer A foreign key that refers to a standard condition concept identifier in the Standardized Vocabularies.

condition_start_date Yes date The date when the instance of the condition is recorded.

condition_end_date No date The date when the instance of the condition is considered to have ended. If this information is not available, set to NULL.

condition_type_concept_id Yes integer A foreign key to the predefined concept identifier in the Standardized Vocabularies reflecting the source data from which the condition was recorded, the level of standardization, and the type of occurrence. For example, conditions may be defined as primary or secondary diagnoses, problem lists and person statuses.

stop_reason No varchar(20) The reason, if available, that the condition was no longer recorded, as indicated in the source data. Valid values include discharged, resolved, etc. Note that a stop_reason does not necessarily imply that the condition is no longer occurring.

provider_id No integer A foreign key to the provider in the provider table who was responsible for determining (diagnosing) the condition.

visit_occurrence_id No integer A foreign key to the visit in the visit table during which the condition was determined (diagnosed).

condition_source_value No varchar(50) The source code for the condition as it appears in the source data. This code is mapped to a standard condition concept in the Standardized Vocabularies and the original code is, stored here for reference. Condition source codes are typically ICD-9-CM diagnosis codes from medical claims or discharge status/visit diagnosis codes from EHRs.

condition_source_concept_id No integer A foreign key to a condition concept that refers to the code used in the source.
CONVENTIONS

• Valid Condition Concepts belong to the "Condition" domain. Standard Condition Concepts are based on SNOMED-CT.

NCATS Pre-decisional Document: Tables provided for informational purposes only Version: 04.27.20

• Condition records are typically inferred from diagnostic codes recorded in the source data. Such code system, like ICD-9-CM, ICD-10-CM, read etc., provide a comprehensive coverage of conditions. However, if the code does not define a condition, but rather an observation or a procedure, then such information is not stored in the CONDITION_OCCURRENCE table, but in the respective tables instead.

• Source Condition identifiers are mapped to Standard Concepts for Conditions in the Standardized Vocabularies. When the source code cannot be translated into a Standard Concept, a CONDITION_OCCURRENCE entry is stored with only the corresponding source_concept_id (if available) and source_value and a condition_concept_id of 0.

• Family history and past diagnoses ("history of") are not recorded in the CONDITION_OCCURRENCE table. Instead, they are listed in the OBSERVATION table.

• Codes written in the process of establishing the diagnosis, such as "question of" and "rule out", are not represented here. Instead, they are listed in the OBSERVATION table if they are used for analyses.

• A Condition Occurrence Type is assigned based on the data source and type of condition attribute, including:
  o ICD-9-CM Primary Diagnosis from Inpatient and Outpatient Claims
  o ICD-9-CM Secondary Diagnoses from Inpatient and Outpatient Claims
  o Clinician diagnoses or problem Concepts from EHRs

10 MEASUREMENT A measurement is the capture of a structured value (numerical or categorical) obtained through systematic examination of a person or sample. The Measurement domain captures measurement orders and measurement results. The measurement domain can contain laboratory results, vital signs, quantitative findings from pathology reports, etc.

Field Required Type Description

measurement_id Yes integer A unique identifier for each measurement.

person_id Yes integer A foreign key identifier to the person about whom the measurement was recorded. The demographic details of that person are
measurement_concept_id Yes integer A foreign key to the standard measurement concept identifier in the Standardized Vocabularies.

measurement_date Yes date The date of the Measurement.

measurement_time No time The time of the Measurement

measurement_type_concept_id Yes integer A foreign key to the predefined concept identifier in the Standardized Vocabularies reflecting the type of data on which the measurement record is based.

operator_concept_id No integer A foreign key identifier to the mathematical operator that is applied to the value_as_number. Operators are <, ≤, =, ≥, >.

NCATS Pre-decisional Document: Tables provided for informational purposes only Version: 04.27.20

Field Required Type Description

value_as_number No float A measurement stored as a number. This is applicable to measurement where the result is expressed as a numeric value.

value_as_concept_id No integer A foreign key to a measurement stored as a concept identifier. This is applicable to measurements where the result can be expressed as a standard concept from the Standardized Vocabularies (e.g., positive/negative, present/absent, low/high, etc.).

unit_concept_id No integer A foreign key to a standard concept identifier of measurement units in the Standardized Vocabularies.

range_low No float The lower limit of the normal range of the measurement. The lower range is assumed to be in the same units of measure as the measurement value.

range_high No float The upper limit of the normal range of the measurement. The lower range is assumed to be in the same units of measure as the measurement value.
provider_id No integer A foreign key to the provider in the provider table who was responsible for making the measurement.

visit_occurrence_id No integer A foreign key to the visit in the visit table during which the measurement was recorded.

measurement_source_value No varchar(50) The measurement name as it appears in the source data. This code is mapped to a standard concept in the Standardized Vocabularies and the original code is stored here for reference.

measurement_source_concept_id No integer A foreign key to a concept that refers to the code used in the source.

unit_source_value No varchar(50) The source code for the unit as it appears in the source data. This code is mapped to a standard unit concept in the Standardized Vocabularies and the original code is stored here for reference.

value_source_value No varchar(50) The source value associated with the structured value stored as numeric or concept. This field can be used in instances where the source data are transformed to produce the structured value.

CONVENTIONS
- Valid Measurement Concepts for both the measure (measurement_concept_id) and the measure result (value_as_concept) belong to the "Observation" domain. Measurement Concepts are based mostly on the LOINC vocabulary, with some additions from SNOMED-CT.
- Measurements are stored as attribute value pairs, where the attribute is the measure and the value represent the result. The value can be a concept (stored in value_as_concept), or a numerical value (value_as_number). The availability of a result is not mandatory.
- If reference ranges for upper and lower limit of normal as provided (typically by a laboratory) are stored in the range_high and range_low fields. Ranges have the same unit as the value_as_number.
- The Visit during which the observation was made is recorded through a reference to the VISIT_OCCURRENCE table. This information is not always available.
• The Provider making the observation is recorded through a reference to the PROVIDER table. This information is not always available.

11 OBSERVATION The observation domain captures any clinical facts about a patient obtained in the context of examination, questioning or a procedure. The observation domain supports capture of data not represented by other domains, including unstructured measurements, medical history and family history.

Field Required Type Description

<table>
<thead>
<tr>
<th>Field</th>
<th>Required</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>observation_id</td>
<td>Yes</td>
<td>integer</td>
<td>A unique identifier for each observation.</td>
</tr>
<tr>
<td>person_id</td>
<td>Yes</td>
<td>integer</td>
<td>A foreign key identifier to the person about whom the observation was recorded. The demographic details of that person are stored in the person table.</td>
</tr>
<tr>
<td>observation_concept_id</td>
<td>Yes</td>
<td>integer</td>
<td>A foreign key to the standard observation concept identifier in the Standardized Vocabularies.</td>
</tr>
<tr>
<td>observation_date</td>
<td>Yes</td>
<td>date</td>
<td>The date of the observation.</td>
</tr>
<tr>
<td>observation_time</td>
<td>No</td>
<td>time</td>
<td>The time of the observation.</td>
</tr>
<tr>
<td>observation_type_concept_id</td>
<td>Yes</td>
<td>integer</td>
<td>A foreign key to the predefined concept identifier in the Standardized Vocabularies reflecting the type of the observation.</td>
</tr>
<tr>
<td>value_as_number</td>
<td>No</td>
<td>float</td>
<td>The observation result stored as a number. This is applicable to observations where the result is expressed as a numeric value.</td>
</tr>
<tr>
<td>value_as_string</td>
<td>No</td>
<td>varchar(60)</td>
<td>The observation result stored as a string. This is applicable to observations where the result is expressed as verbatim text.</td>
</tr>
<tr>
<td>value_as_concept_id</td>
<td>No</td>
<td>Integer</td>
<td>A foreign key to an observation result stored as a concept identifier. This is applicable to observations where the result can be expressed as a standard concept from the Standardized Vocabularies (e.g., positive/negative, present/absent, low/high, etc.).</td>
</tr>
<tr>
<td>qualifier_concept_id</td>
<td>No</td>
<td>integer</td>
<td>A foreign key to a standard concept identifier for a qualifier (e.g., severity of drug-drug interaction alert)</td>
</tr>
<tr>
<td>unit_concept_id</td>
<td>No</td>
<td>integer</td>
<td>A foreign key to a standard concept identifier of measurement units in the Standardized Vocabularies.</td>
</tr>
</tbody>
</table>
Field Required Type Description

provider_id No integer A foreign key to the provider in the provider table who was responsible for making the observation.

visit_occurrence_id No integer A foreign key to the visit in the visit table during which the observation was recorded.

observation_source_value No varchar(50) The observation code as it appears in the source data. This code is mapped to a standard concept in the Standardized Vocabularies and the original code is stored here for reference.

observation_source_concept_id No integer A foreign key to a concept that refers to the code used in the source.

unit_source_value No varchar(50) The source code for the unit as it appears in the source data. This code is mapped to a standard unit concept in the Standardized Vocabularies and the original code is stored here for reference.

qualifier_source_value No varchar(50) The source value associated with a qualifier to characterize the observation

CONVENTIONS

- Valid Observation Concepts for the object (observation_concept_id) belong to the "Observation" domain. Observation Concepts are based mostly on the LOINC vocabulary, with some additions from SNOMED-CT.

- Valid Observation Concepts and the finding (value_as_concept_id) are not enforced by a domain but should be Standard Concepts.

- Observations must have an object represented as a concept, and a finding, represented as a concept, a numerical value or a verbatim string. There should be no observations records without an associated value. Observations which appear to be suggestive statements of positive assertion
should have a recorded value as concept of ‘Yes’.

• Observations obtained using standardized methods (e.g. laboratory assays) that produce discrete results are recorded by preference in the MEASUREMENT table.

• The Visit during which the observation was made is recorded through a reference to the VISIT_OCCURRENCE table. This information is not always available.

• The Provider making the observation is recorded through a reference to the PROVIDER table. This information is not always available.

**Standardized Health System Data**

**Tables**

These tables describe the healthcare provider system responsible for administering the healthcare of the patient, rather than the demographic or clinical events the patient is involved in.

Below provides an entity-relationship diagram highlighting the tables within the Standardized Health System portion of the OMOP Common Data Model:
12 LOCATION The Location table represents a generic way to capture physical location or address information. Locations are used to define the addresses for Persons and Care Sites.

Field Required Type Description

- location_id Yes integer A unique identifier for each geographic location.
- zip No varchar(9) The zip or postal code. For US addresses, valid zip codes can be 3, 5 or 9 digits long, depending on the source data.
- location_source_value No varchar(50) The verbatim information that is used to uniquely identify the location as it appears in the source data.

CONVENTIONS

- Each address or Location is unique and is present only once in the table.
- Locations do not contain names. In order to construct a full address that can be used on the Postal Service, the address information from the Location needs to be combined with information from the Care Site. The Person table does not contain name information.
- All fields in the Location tables contain the verbatim data in the Source. None of them are mandatory, but a valid Location record should at least contain either a Location Name or Location Zip.
- Zip codes are handled as strings of up to 9 characters length. For US addresses, these represent either a 3-digit abbreviated Zip code as provided by many Sources for Patient protection reasons, or the full 5-digit Zip code or the 9-digit (ZIP + 4) codes are recorded. Unless for specific reasons, analytical methods should expect and utilize only the first 3 digits. For international addresses, different rules apply.

13 CARE_SITE The Care Site table contains a list of uniquely identified physical or organizational units where healthcare delivery is practiced (offices, wards, hospitals, clinics, etc.).

Field Required Type Description

- care_site_id Yes Integer A unique identifier for each organization. Here, an organization is defined as a collection of one or more care sites that share a single EHR database.
care_site_name No varchar(255) The description of the care site

place_of_service_concept_id No Integer A foreign key that refers to a place of service concept identifier in the Standardized Vocabularies.

location_id No Integer A foreign key to the geographic location of the administrative offices of the organization in the location table, where the detailed address information is stored.

care_site_source_value No varchar(50) The identifier for the organization in the source data, stored here for reference.

place_of_service_source_value No varchar(50) The source code for the place of service as it appears in the source data, stored here for reference.

CONVENTIONS

• There can be hierarchical and business relationships between Care Sites (e.g., wards can belong to clinics, which in turn can belong to hospitals, which in turn can belong to hospital systems, which in turn can belong to HMOs). These relationships should be defined in the FACT_RELATIONSHIP table.

• The Care Site Source Value typically contains the name of the Care Site.

• The Place of Service Concepts belongs to the Domain "Provider". These Concepts are based on a catalog maintained by the CMS

14 PROVIDER The Provider table contains a list of uniquely identified health care providers. These are typically physicians, nurses, etc.

Field Required Type Description

provider_id Yes Integer A unique identifier for each provider.

npi No varchar(20) The National Provider Identifier (NPI) of the provider.

specialty_concept_id No Integer A foreign key to a standard provider’s specialty concept identifier in the Standardized Vocabularies.

care_site_id No Integer A foreign key to the main care site where the provider
CONVENTIONS

• Providers are not duplicated in the table.

• Valid Specialty Concepts for both the test (measurement_concept_id) belong to the "Provider" domain. The Specialty Concepts are based on the CDC specialty classification.

• This table is used to represent fixed relationship between Providers and Care Sites. Providers are also linked to Care Sites through Condition, Procedure and Visit records.
• #79 Adds the METADATA table
• #92 Fixes modifier typo in PROCEDURE_OCCURRENCE
• #120 Adds the following fields to PAYER_PLAN_PERIOD:
  o PAYER_CONCEPT_ID
  PAYER_SOURCE_CONCEPT_ID
  PLAN_CONCEPT_ID
  PLAN_SOURCE_CONCEPT_ID
  SPONSOR_CONCEPT_ID
  SPONSOR_SOURCE_CONCEPT_ID
  STOP_REASON_CONCEPT_ID
  STOP_REASON_SOURCE_VALUE
  STOP_REASON_SOURCE_CONCEPT_ID
Additional Resources

- [https://ncats.nih.gov/n3c](https://ncats.nih.gov/n3c)
- [https://ncats.nih.gov/n3c/about](https://ncats.nih.gov/n3c/about)
- [https://ncats.nih.gov/n3c/about/program-faq](https://ncats.nih.gov/n3c/about/program-faq)
- [https://ncats.nih.gov/n3c/about/resources](https://ncats.nih.gov/n3c/about/resources)
- [https://ncats.nih.gov/n3c/funding](https://ncats.nih.gov/n3c/funding)