Guide to N3C

3/14/23

Table of contents

Welcome

The Guide to N3C is designed to guide research with the National COVID Cohort Collaborative (N3C).

Contributing Content

The N3C and its domain teams are healthiest when assimilating contributions from reseachers of different skills (*e.g.*, clinicians & informaticians), specialties (*e.g.*, endocrinology & gerontology), programming languages (*e.g.*, Python, R, & SQL) and experiences (*e.g.*, students & PIs).

Accordingly, the guide-to-n3c-v1 repository welcomes input from you.

If you see a small mistake or unclear language, we invite you to inform us in a new issue or to edit the source and submitting a pull request. When an editor reviews and accepts your change, the website will be updated within minutes.

If you have an idea for something more substantial such as a chapter or section, please start a new issue and the N3C Educational Committee will coordinate with you where it fits best.

Platform

As the chapters are being written, talk to the chapter's Lead Author to learn if it's being written in GitHub or Google Docs. Eventually all Google Docs chapters will be translated to Markdown. Once a chapter has been finally converted to Markdown…

To make small changes like spelling corrections, we recommend editing the source directly in GitHub. It handles the details without your knowledge (like starting a fork and prompting your pull request). From the appropriate page of the book, click on the "Edit this page " button and type your change in the GitHub editor.

Substantial edits and writing are better accommodated by a text editor on your local machine that can preview the rendered content as you type. We suggest Visual Studio Code or RStudio.

You don't have to understand the rest to contribute, but for those interested:

- The majority of this book is written in a collection of Markdown documents and assembled by the Quarto.
- After your change is pushed to GitHub, a GitHub Action spawns a small VM that (a) collects all the Markdown documents, (b) calls Quarto/Pandoc to convert them to html, and (c) moves the rendered html files to the "gh-pages" branch.
- GitHub Pages serves the contents of the `gh-pages` branch to anyone with a browser.

Funding and Licensing

The N3C is supported by NIH National Center for Advancing Translational Sciences (NCATS).

Part I

Getting Started

1 Introduction

> **Note**
>
> This chapter is being drafted in Google Docs at https://drive.google.com/drive/u/0/folders/16QkU2vonX5iZzjsLCxIEREPEdZ9S6wza
> See a draft of the chapter outline at https://docs.google.com/document/d/1ttUKgwVcIZHM87elrlUNV6Qi9thzOwKBg8GegKObEtg/

> **Warning**
>
> At this point, any edits to this chapter should be made in Google Docs. The current Markdown is for testing only. It is NOT the source of truth (yet).

Additional Contributors:

1.1 Mission

The National COVID Cohort Collaborative, or N3C, is an open-science community stewarded by the National Center for Data To Health (CD2H) and the NIH National Center for Advancing Translational Sciences (NCATS), with significant contributions from many partners including the Clinical and Translational Science Awards (CTSA) program, Centers for Translational Research (CTRs), and thousands of researchers from hundreds of participating institutions in the US and abroad.

Faced with the COVID-19 pandemic, the issue addressed by N3C is clear and direct: the US has no centralized data repository for health records and related information, hindering the response of the scientific community.[1] Although healthcare providers are mandated by law to utilize electronic health records (EHR), little guidance coordinates how or exactly what information to collect and store. Commercial EHR suites (e.g. Epic) are widely used, and controlled vocabularies (e.g. ICD10 and SNOMED) provide standards for representing medical information, but there are many such standards in use and EHR software is highly configurable to the needs of individual organizations. As a result, databases of EHR information across the

[1] This article in MIT Technology Review provides a good overview of N3C and the surrounding landscape: It took a pandemic, but the US finally has (some) centralized medical data.

US are largely non-interoperable, presenting challenges to researchers hoping to use this vast national store of information in practice.

1.2 Common Data Models and N3C

In recent years, the common solution to these issues have been the creation of *Common Data Models* (CDMs). A Common Data Model is an agreed-upon structure and format for databases containing clinical information, into which diverse organizational EHR databases can be standardized for research purposes. Even so, healthcare organizations are reluctant to share their data directly, because these risks associated with data breaches for protected health information (PHI) defined by HIPAA are high. As a result, several groups of healthcare and research organizations have formed _federated_research networks, where each organization in a

Figure 1.1: A visual representation of disparate, non-compatible EHR databases across the United States.

network translates some subset of their EHR data into an agreed-upon common data model, and affiliated researchers can then write queries intended for data in that format. Examples of such *federated* networks include PCORNet and i2b2, each of which utilize their own common

data model. In a federated model, data queries are generated by researchers, but executed locally by the data owners and only summarized or specifically requested results are sent back to researchers. This ensures protection for the raw data (which never leaves the boundaries of any individual organization), but prevents exploratory data investigation and other techniques (like many machine-learning methods) which require direct access to the totality of the data.

Driven by the imperative of addressing COVID-19, in concert with the use of a FedRAMP-certified cloud-based analysis ecosystem we call the N3C Data Enclave,[2] N3C is partnering with EHR data providers across the nation to collect billions of EHR data points for millions of patients with and without COVID-19 in a single, secure, accessible database for research use. N3C simultaneously moderates controlled access to these data by research teams from across the country (and beyond), including from private companies, community colleges, universities, medical schools, and government entities.

Like federated research networks, N3C also uses a common data model, known as OMOP, chosen for its strong community support and open nature, support of scientific use cases, and availability of tools for translating and working with data (we'll discuss the OMOP data format in more detail in later chapters).[3] To rapidly collect data from around the country, N3C leverages the existing work data owners have already done to convert their organization-unique data to one of a handful of N3C-supported "source" common data models: PCORNet, i2b2, TriNetX, ACT, and OMOP. A potential data partner with data in PCORNet format, for example, will locally run a set of N3C-generated "PCORNet to OMOP" translation scripts prior to transferring the result to N3C via a secure channel. The process of coalescing multiple such data payloads into a unified whole is known as *harmonization*, and is a complex task even after everything has been mapped to OMOP initially. Two overlapping teams of EHR data experts participate in this process: one works closely with data partners to make it as easy as possible to contribute data to N3C, and another handles the post-ingestion harmonization and comprehensive quality checks of the incoming data.

1.3 The N3C "Enclave" and Data Access

Once harmonized and stored in the secure Enclave, the data are made available via a web-based interface to research teams using common tools such as SQL, Python, and R, as well as a number of code-light graphical user interfaces.

[2]The Federal Risk and Authorization Management Program (FedRAMP) is a rigorous, standardized certification program with an emphasis on security and information protection. The Enclave is an installation of Palantir Technologies' Foundry platform, a FedRAMP-certified data analytics suite.

[3]OMOP was originally developed by its namesake, the Observational Medical Outcomes Partnership, but is now stewarded by the Observational Health Data Sciences and Informatics (OHDSI, pronounced "odyssey"), an international group of researchers and clinicians. For complete information about OHDSI and OMOP, see the Book of OHDSI.

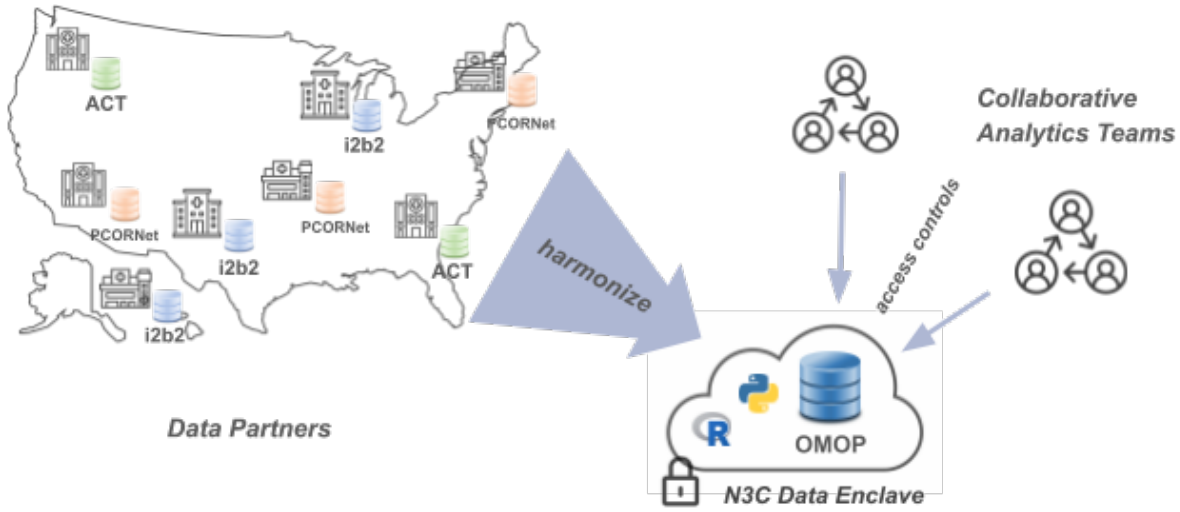

Figure 1.2: A visual representation of N3C's data harmonization from source-CDM model, and team-based access to these data in a secure cloud-based enclave.

Mere access to the Enclave, however, doesn't automatically provide access to any of the protected data itself (although we do make other, publicly-available datasets available with minimal restriction for practice and learning purposes). Multiple "levels" of the data are available with different anonymization techniques applied, facilitating "just enough" access to research teams depending on their needs and ability to access protected health information. Accessing the most secure level, for example, requires obtaining approval by an Institutional Review Board (IRB) who validates the appropriateness of human subjects research, while the lowest level is heavily anonymized and accessible by private individuals (citizen scientists) with only certain legal and training requirements.

Because effective analysis of EHR data requires a diverse set of skills–especially clinical and data science/statistical expertise–N3C provides organizational structures and resources to rapidly create and support multidisciplinary research teams, many of which are geographically diverse as well. As of December 2021, dozens of these "Domain Teams" support nearly 300 ongoing research projects, contributed to by over 2500 researchers hailing from 250+ different institutions and organizations. Sixty-nine data partners provide EHR data for 10 million patients (of whom have had COVID-19), representing 5.5 billion lab results, 1.6 billion medication records, 1 billion clinical observations, and 500 million clinical visits. For up-to-date information on these numbers and more, visit our dashboard at https://covid.cd2h.org/dashboard.

Figure 1.3: Summary statistics for N3C patients as of Aug, 2022. Confirmed COVID-19 patients are those with a known positive PCR or Antigen lab test, possible patients are those with likely symptomatology.

1.4 Benefits of Participation

For researchers, N3C provides a unique opportunity to participate in next-generation, distributed team science. Investigators with expertise in multiple domains come together across organizational boundaries to answer questions critical to the understanding and management of COVID-19 and its impact on the health of individuals and communities across the United States. Clinical, health informatics, data science, epidemiology, biostatistics, and public health experts join forces in true team science manner to …. and form Domain Teams to support and encourage….

For researchers:

- Participate in next-generation distributed team science
- Build connections and collaborations
- Support from fellow researchers and domain experts (computational, clinical, etc)
- See chapter "Onboarding, Enclave Access, N3C Team Science"
- "N3C Clinical Domain Teams enable researchers with shared interests to analyze data within the N3C Data Enclave more efficiently and to collaborate on multi-site research. The Clinical Domain Teams, developed within the Clinical Scenarios subgroup, focus on specific clinical questions surrounding COVID-19's impact on health conditions. Clinical Domain Teams are enabled by Slack channels for discussion, meetings, and document management, and are supported by N3C workstreams and administration. N3C encourages researchers of all levels to join a Domain Team that represents their interests, or to suggest new clinical areas to explore."
  - "N3C Domain Teams enable researchers with shared interests analyze data within the N3C Data Enclave and collaborate more efficiently in a team science environment. These teams provide an opportunity to collect pilot data for grant submissions, train algorithms on larger datasets, inform clinical trial design, learn how to use tools for large scale COVID-19 data, and validate results. Domain Teams

are enabled by Slack channels for discussion, meetings, and document management and are supported by N3C workstreams. N3C encourages researchers of all levels to join a Domain Team that represents their interests, or to suggest new clinical areas to explore. A Domain Team can submit one or more research projects, but collaboration is encouraged for similar concepts."
  - "Multi-discipline Clinical Domain Teams focus on clinical questions surrounding COVID-19's impact on health conditions and consist of clinical and subject matter experts, statisticians, informaticists, and machine learning specialists. Cross-Cutting Domain Teams have a varied focus that applies to multiple domains."

- Largest database of de-identified EHR data in the US
  - "one of the largest, most secure clinical data resources for accelerating and collaborating on COVID-19 research" from N3C website
- Big-data capabilities and powerful tools (R, python, packages, sql)
- Accessible - no compute infrastructure needed other than a browser
- Learn and gain experience working with clinical data in a common data model
- Work with EHR data that has been checked for quality, completeness, and consistency

For institutions w/ researchers (why sign a DUA):

- Increase research productivity and profile
- Encourage cross-institutional/geographically-diverse connections
- Provide clinical translational research opportunities to faculty and students
- Enable real-world learning opportunities for students

For data partners (why contribute data):

- Raise institutional research profile
- Get feedback on local data quality and completeness in comparison to peers
  - Compare your local research results to results on data from partners across the US to assess generalizability
- Place your data in a secure enclave that supports broad access and reproducibility
- Grow the userbase of researchers utilizing your data, including at your own institution

1.5 News articles about N3C

VOX one

MIT press article

1.6 Exemplars of great projects

Query the community to see who wants to be featured :)

Projects that have been in the news (press/news articles) and/or have been published, potentially highly cited

Student papers/projects

Maybe restrict to those that have a peer-reviewed publication?

BARDA challenge

2 A Research Story

40,000-foot view of a research project, from onboarding to publishing

Additional Contributors: Sharon Patrick, Shawn O'Neil

Now that we have introduced N3C and described its motivation and importance, we'll walk through the lifecycle of an example project from onboarding to publishing. This path typically takes at least 6 months and 6 collaborators. It is difficult to do by yourself, but fortunately the N3C has attracted a large and diverse set of researchers. Coupled with a large and diverse set of patients, it is possible to complete a research project within a year.

If you are starting a project by yourself, you'll likely be able to recruit collaborators with a complementary set of skills (in addition to the resources such as instructional material and office hours. If you would like to join an existing project, there are domain teams and ongoing projects that likely will fit your interests and benefit from your abilities.

Voice of Morgan Freeman

Our story begins in your office. Your own piece of heaven. As a researcher of scurvy, you have wondered, "do patients receiving the newest medications have more favorable covid outcomes than patients receiving the previous generation?". Based on the meds' relationships with other diseases, you expect there is a modest improvement. But in order to detect a modest effect size, many patients are required, and your local institution's EMR doesn't have enough meeting the inclusion criteria. In fact, no single institution's EMR has enough.
Yesterday you attended a local N3C presentation and became interested because it likely has enough qualifying scurvy patients to detect even small signals. Your mind wanders as you get a little greedy; additional hypotheses enter your daydream. *Does the relationship attenuate as you move inland?* You realize that a massive national dataset can not only better address your existing question, it could also allow you to ask newer and more nuanced questions. *Is the relationship more pronounced in other racial/ethnic groups?* The stream persists throughout the night.

{Should we use 2nd person or 3rd person?}

2.1 Onboarding

The startup costs are more expensive for an N3C investigation compared to most, but the incremental costs are cheaper. Even with strong institutional support, the university's agreement with the NIH takes several months in legal and administrative channels. Yet after clearing that first (tall) hurdle for your site, each specific project takes only a week or two to be processed by the N3C staff. That's a remarkably short time considering the scale of available data. It's likely quicker than initiating a project based on a single EMR from your site –much quicker than EMRs from 70+ sites.

Voice of someone like Bill Kurtis, but without the connotation of murder

The next afternoon you are chatting with your institution's Navigator.[a] She organized the local N3C presentation and invited any interested attendees to contact her.

[a]This role may be called something differently at your Institution; the roles are defined below in Section **??**.

Hover over a footnote to see the popup, without jumping to the bottom of the page.

- **Navigator**: I'm glad you think the N3C might help your research. As I wrote in this morning's email, the agreement between the university and the NIH was established last year, so don't worry about that.[1] There are two remaining steps. First, complete your personal paperwork.[2] Second, submit a DUR tailored to your hypotheses.[3]
- **Investigator**: Remind me what a DUR is?
- **N**: A *d*ata *u*se *r*equest describes your upcoming project. Once a committee approves your proposal, your project's code and data are protected in this workspace allotted on the NIH cloud.[4] Everyone on your project uses this dedicated workspace too. But they don't have to submit additional DURs –your grant them permission to join yours.[5]
- **I**: Umm, I think I got it.
- **N**: It will make sense once you get it into it. Skim the example DUR proposals I'm sending now. Then start filling out this online form. Get as far as you can, and then I'll help with the rest. If there's something I don't know, I'll ask a friend. The DUR application process will take about an hour. Then the proposal will likely be approved within a week or two. In the meantime, we can talk about potential collaborators.

[1]Read about the institutional-level DUA in Chapter **??**.
[2]See Chapter **??**.
[3]Project-level paperwork are discussed in Chapter **??**.
[4]The NIH "Enclave" is detailed in Chapter **??**.
[5]DURs are the topic of Chapter **??**.

2.2 Team building & collaborating

The next step is to build a team to leverage retrospective medical records. Like most contemporary research teams, heterogenous skills are important. Ideally a team has at least:

1. a **navigator** who has learned the administrative and IRB requirements and is able to facilitate the investigation,
2. a **data engineer** who understands the challenges of EMRs and is able to extract and transform information,
3. a **statistician** who understands the limitations of observational collection and is able to model retrospective data,
4. a **subject matter expert** (SME) who has clinical experience with the disease of interest and is able to inform decisions with EMR variables, and
5. a **principal investigator** who knows the literature and is able form testable hypotheses and write the manuscript.

N3C teams have some differences from conventional research teams at single sites. Some trends we have noticed are:

1. Most N3C teams have researchers from at least three institutions. In the experience of the authors and editors, this encourages more diverse opinions and more willingness to express constructive criticism. Researchers from a single institution/lab are sometimes are more reluctant to generate contrary views.

2. The role of navigator is even more important. Your local EMR investigations are likely guided by someone with years of experience with the institutional safeguards and the personnel who can help when something stalls. N3C is bigger and younger than your site's EMR research team, so an N3C project will benefit when guided by a bright, patient, and persistent navigator.

If your team needs someone, consider asking a relevant domain team for helping identifying and approaching a potential collaborator.

> Voice of Jamie Foxx
>
> Recruiting your crew…

2.3 Research Team's First Meeting

Voice of ...

Three weeks later...
Once the team is assembled, the first discussion is usually a variation of this exchange:

- **Investigator**: Welcome everyone. We'd like to know if Drug A or Drug B is associated with better outcomes.
- **Statistician**: No problem. I can longitudinally model the type and amount of each medication received by each patient, relative to their intake date.
- **Data Engineer**: Hmmm. I'm happy to produce a dataset with the `dose` and `frequency` columns[6], but you may not find it useful. Those two columns are sparsely populated and they look inconsistent across sites.[7]
- **I**: Bummer. Then what's realistic or feasible?
- **Subject Matter Expert**: Maybe this simplifies the picture... In my clinical experience, a patient rarely switches between Drugs A & B. Based on the initial presentation, their provider will pick A *or* B, and complete the regimen unless there's an adverse event.
- **S**: In that case, should my initial model have three levels for treatment: A, B, and A+B?
- **I**: Probably. In the N3C database, can someone tell me how many patients get both during the same visit?
- **DE**: I'm already logged into the Enclave[8]. Give me 2 minutes to whip up something in SQL.[9]
- **I**: Oh my goodness, is that your cat? What a cutie! [10]
- **DE** *after a few minutes*: Ok, I got it. [Unmutes himself.] Ok, I got it. 40% of patients are Drug A only, 52% are Drug B only, while 8% have at least one administration of both Drug A & B in the same visit.
- **SME**: Weird. 8% is a lot more than I expected. I was thinking around 1%.
- **DE**: Hmm, let me check. Give me another minute.[11]
- **DE** *after a few minutes*: I see what you mean. It looks like the bulk of the combo patients were admitted in the spring of 2020. After Jan 2021, only 3% of patients have both Drug A & B.
- **S**: I was already planning to model the phase of the pandemic. I'll test if there's a significant interaction between time and treatment.

[6] Read about the OMOP Standard Tables in Chapter **??**, specifically the medications are in the `drug_exposure` table.
[7] Conformance is a topic in Chapter **??**.
[8] See Chapter **??** for accessing the N3C Enclave.
[9] Read about SQL, Python, and R transforms in Code Workbooks in Chapter **??**.
[10] There is a brief discussion of SME's cat.
[11] There is a brief discussion of S's daughter strutting in the background wearing a cowboy hat and waiving a fairy wand.

- **I**: I like that as a starting point. Regarding the question about dose and frequency… For now let's assume the providers were following the current dosing guidelines. Therefore the `dose` and `frequency` variables can be dropped from the analyses.
- **S**: Phew. I didn't want to admit this. But I skimmed the dosing guidelines you emailed yesterday. It looked complicated. I wasn't sure if I could appropriately incorporate those variables in the model.
- **I**: Well, that's everything I wanted to cover today. See you in two weeks. Wait. I can't believe I forgot. Sorry -our Navigator is sick this week and I'm almost worthless in her absence. Is everyone still on the call? For our secondary hypothesis, we want everything to connect to a patient's diagnoses. …before, during, and after their covid hospitalization.
- **DE**: Bad news. This is kinda like the `dose` and `frequency` situation a few minutes ago. The structure of the OMOP diagnosis table theoretically can connect a patient's diagnoses across different locations. But the quality of the historical records really depends on the site. Some places like Delaware leverage their state's HIE[12] to populate their N3C dataset. However other places are not as well connected. If a patient doesn't have diagnosis records, it's tough to determine if they are healthy, or if their primary care provider uses a siloed EMR.[13]
- **I**: Ugh. Good point.
- **DE**: But I've got good news. All the N3C contributors comprehensively capture all conditions diagnosed *during* the visit. Furthermore the diagnosis codes are standardized really well across sites. That's because all the providers enter ICD codes into the EMR, which eventually can be cleanly mapped to OMOP's standard concepts.[14]
- **I**: Well, that's fine for this paper. Maybe our next manuscript will follow up with N3C's death records.[15]
- **SME**: Sorry everybody, I have clinic this week, and they're calling me. I need to drop.[16]
- **S**: Can I go back and ask a question about medications? I see that Drug A has 15 different brand names. I don't recognize half of them. How should I classify them?
- **DE**: It's actually worse than that. Sorry I'm a downer today. Can you see my screen? Drug A has 15 brand names and 200 different RxNorm codes; each package is uniquely identified by the NIH's NLM. SME and I started on a concept set Thursday. We're operationalizing the drug classes by their RxNorm ingredient. There are five ingredients that are conceptualized as Drug A. A friend showed me how she used the OMOP tables in a different project.[17] I'll roll up the meds into the patient-level dataset. It will have one integer for the number of medication records tied to a Drug A ingredient and another integer for Drug B records. You'll probably want to transform the two counts into two

[12] An HIE is a health information exchange.

[13] The benefits and caveats of real-world data are a theme throughout the book, particularly in the best practices discussed in Chapter **??**.

[14] Authoring and using concept sets are described in Chapter **??**. Mapping an ICD to SNOMED diagnosis code is an example of mapping a "non-standard" to a "standard" concept, discussed in Chapter **??**.

[15] TODO: is the book planning to have a section on the CMS & death records?

[16] Everyone says goodbye to the cat.

[17] The `concept_relationship` table is discussed with the OMOP concept hierarchy in Chapter **??**.

booleans.

- **S**: And if I change my mind and decide to use the counts, then at least I'll know.
- **Shoreleave**: and knowing is half the battle.

2.4 Protocol, variables, & definitions

This aspect of the scientific process is probably both the most familiar and most vague. Most researchers have several years of graduate-level courses and real-world experience.

1. Tradeoffs are inevitable when selecting variables. Rarely will an investigator's first choice be available.
2. Retrospective medical records are extracted from a larger dataset. An investigation can use only a fraction of the terabytes in an EMR. Many decisions are involve to include only the relevant variables among the qualifying patients.

{Mention CD2H's Informatics Playbook}

(Wu and C2DH 2022, chap. 1)

2.5 Creating an analysis-ready dataset

{Conventional data engineer role. Dataset is created with input from the analyst.}

2.6 Learning and using OMOP (e.g. concept sets)

OMOP originated in 2014 to facilitate the detection of small but significant side effects from new pharmaceuticals. Detecting a small signal requires a large datasets –larger than any single health care database (Sciences and Informatics 2019, chap. 1). Since then, the foundation has supported many other research goals. It is well-suited for N3C because:

- It has evolved from 10? years and accommodates a wide range of data sources
- It has an established community and documentation to help institutions convert their EMR to OMOP and to help researchers analyze their hypotheses.

{3-4 sentence description of the original OMOP motivation. It standardizes (a) tables & columns and (b) vocabulary. Spend 1-2 paragraphs on concept set, focusing more on motivations than the mechanics.}

2.7 Analyses

- Developing the Analyses
- finalizing analysis
  - Pinning to a release
  - DRR
  - figures

2.8 Draft paper, pub committee

> Voice of Sam Elliot
>
> Nearing the trail head...

3 Data Lifecycle - From Patients to N3C Researchers

> **Note**
>
> This chapter is being drafted in Google Docs at https://drive.google.com/drive/u/0/folders/1ZjnFezddZk0YllqIDZN1mgexn1yM51tH
> See a draft of the chapter outline at https://docs.google.com/document/d/1ttUKgwVcIZHM87elrlUNV6Qi9thzOwKBg8GegKObEtg/

> **Warning**
>
> At this point, any edits to this chapter should be made in Google Docs. The current Markdown is for testing only. It is NOT the source of truth (yet).

4 Governance, Leadership, and Operations Structures

> **Note**
>
> This chapter is being drafted in Google Docs at https://drive.google.com/drive/u/0/folders/19lryu26FaMAVrDHHARYcFJFpb18vOxcT
> See a draft of the chapter outline at https://docs.google.com/document/d/1ttUKgwVcIZHM87elrlUNV6Qi9thzOwKBg8GegKObEtg/

> **Warning**
>
> At this point, any edits to this chapter should be made in Google Docs. The current Markdown is for testing only. It is NOT the source of truth (yet).

5 Onboarding, Enclave Access, N3C Team Science

Additional Contributors:

> **Note**
>
> This chapter is being drafted in Google Docs at https://drive.google.com/drive/u/0/folders/1zOGR2rGGgr1lxP8mV5XmtkuxEZI_R7II
> See a draft of the chapter outline at https://docs.google.com/document/d/1ttUKgwVcIZHM87elrlUNV6Qi9thzOwKBg8GegKObEtg/

> **Warning**
>
> At this point, any edits to this chapter should be made in Google Docs. The current Markdown is for testing only. It is NOT the source of truth (yet).

The N3C has an extensive secure onboarding process due to the sensitivity of the data within the enclave. There are several steps that need to be completed in order for a researcher or N3C user to gain access to the enclave.

5.1 Researcher Eligibility

Citizen scientists, researchers from foreign institutions and researchers from U.S.-based institutions are all eligible to have access to the N3C Data Enclave. Everyone with an N3C Data Enclave account has access to the tools and public datasets that are available in the Enclave.

There are several levels of Electronic Health Record (EHR) data that are available within the N3C Data Enclave. (For more information about the levels of data, see the section 'Description of Levels 1, 2, 3' in the 'Getting & Managing Data Access' chapter. LINK NEEDS TO BE ADDED HERE)

Citizen scientists are only eligible to access synthetic data (Level 1). This data is artificial but statistically-comparable to, and computationally derived from, the original EHR data.

Researchers from foreign institutions are eligible to access synthetic data (Level 1) and patient data that has been deidentified by removal of protected health information (PHI) (Level 2). (PHI includes 18 elements defined by the Health Insurance Portability and Accountability Act (HIPAA).)

Researchers from U.S.-based institutions are eligible to access synthetic data (Level 1), deidentified patient data (Level 2) and patient data that includes dates of service and patient zip code (Level 3). (The latter data set is referred to as a limited dataset because it contains only 2 or the 18 PHI elements.)

5.2 Registration

5.2.1 ORCiD; InCommon vs Login.gov

5.2.2 NIH IT security training

5.2.3 Human Subjects Training

Due to the secure nature of the data that is available in the N3C Enclave registration of users is key. There are two options in which users can log in and create an account, InCOmmon or Login.gov. The InCommon pathway is available to select institutions that participate in that identity management service. You can clock of the link to confirm if your organization participates. If your institution does not participate with InCOmmon, you will need to create a login.gov account. Use the link to Login.gov and complete the required fields to create an account. Once you know which pathway you will use to create an enclave account there are other security measures that are put in place, you will need to have a ORCiD, complete NIH security Trainingn, and also human subjects tratining.

ORCiD, which stands for Open Researcher and Contributor ID, is a unique identifier free of charge to researchers.

The N3C Data enclave is hosted by National Center for Advancing Translational Sciences and all researchers must complete the "Informational Securty, Counterintelligence, Privacy Awareness, Records Management Refresher, Emergency Prepardeness Refresher" course. The course can be accessed at https://irtsectraining.nih.gov/public.aspx. The course take approximately 60-90 minutes to complete and you should print your certificate of completion. Users need to complete Human Subjects training that aligns with their institution's guidelines. You will need to provide the date of completion as part of enclave creation.

Overall, users will need to confirm if they use the InCOmmon or Login.gov pathway, register for an ORCiD, have completed NIH Security Training, and completed institution human subjects training.

5.3 Data Use Agreements

The data use agreement (DUA) establishes the permitted uses of the data in the N3C Data Enclave. By signing the agreement, an institutional official is assuring that users from their institution will abide by the terms defined in the agreement.

A DUA must be executed by National Center for Advancing Translational Science (NCATS) and a research institution. The DUA must be signed by authorized institutional officials who have the authority to bind all users at their institution to the terms of the DUA. (A citizen scientist who is not affiliated with an institution must execute a data use agreement with NCATS.) A DUA will be in effect for five years from the DUA Effective Date.

Every individual who has access to the N3C Data Enclave must be covered by a DUA. This DUA must be in place before an account for the N3C Enclave is requested. If your institution has an active DUA, there is no additional action required with regards to the DUA. A list of institutions with DUAs in place can be found at List of DUA Signatories: (https://covid.cd2h.org/duas).

The Institutional Data Use Agreement form is available at:

https://ncats.nih.gov/files/NCATS_N3C_Data_Use_Agreement.pdf

For more information see:

https://ncats.nih.gov/n3c/resources/data-access

5.4 Enclave Access

5.5 Research Project Teams

5.5.1 Project Lead vs COllaborations

5.5.2 Common roles and expectations (PIs, PMs, SMEs, Analysts, ...)

5.5.2.1 Expertise needed

5.6 Domain Teams

The N3C Data enclave is built for multi-site collaboration, and aims to bring together researchers of different backgrounds with similar questions using domain teams. Because N3C is multi-site, it can be difficult to collaborate with researchers of different backgrounds from different sites. Domain Teams exist to alleviate this difficulty. Some collaboration examples

could be collecting pilot data for grant submission, sharing methodology and cohort logic, or learning how to use tools for large-scale data like machine learning.

For example, let's say your institution just signed the DUA and you have some questions about the relationship between rurality and COVID treatments. You can look at the list of domain teams to see rural health. Then you can get in contact and go to the next upcoming meeting. At the meeting you can find out whether your questions are already part of an existing project within the domain team, or if a new project should be created.

If you don't see your type of questions belonging to any existing domain teams you can create a new one here:

https://n3c-help.atlassian.net/servicedesk/customer/portal/2/group/3/create/58

See here for a list of existing domain teams:

https://covid.cd2h.org/domain-teams

5.7 Browsing Researchers/Projects/Institutions

5.7.1 Object Explorer, Public Dashboard

Once you have an enclave account you can log in and use the object explorer to browse researchers and research projects. The object explorer can be found on the left hand side of your view on the enclave homepage. Click on Object Explorer and there are several object-type groups, to search researchers and projects, click on N3C Admin, from there are you can search Data Use Requests, N3C Researchers, and Research Projects. If you are looking for a particular N3C researcher, you can click on that box and in the search bar type in their name and hit enter. A new box will be displayed and you can clock on that researcher's name and from there you can see the research projects that are lead of or a collabotaro on. You can go back to the Object Explorer, object type groups, click N3C Admin again, and search research projects by clicking that box. Using the search bar at the top of the page you can search by key word. Type in the key word and click enter and a results box will be displayed. You can view all results to find the project in which you are interested in joining. From that screen, you can select the title of the project that you are interested in joining or reading about. There is a public-facing version of searching projects in addition to using the enclave as a search method. Users can search https://covid.cd2h.org/projects or https://covid.cd2h.org/dashboard/exploration#projects and search for title, lead investigator name and also the institution. There are many features that are available to search using the public-facing dashboard. There are four categories of.

Part II

Investigating

6 Getting & Managing Data Access

> **Note**
>
> This chapter is being drafted in Google Docs at https://drive.google.com/drive/u/0/folders/1rrYYjSm5cWni1wyvs___-ByPl9Q6uRy10
> See a draft of the chapter outline at https://docs.google.com/document/d/1ttUKgwVcIZHM87elrlUNV6Qi9thzOwKBg8GegKObEtg/

> **Warning**
>
> At this point, any edits to this chapter should be made in Google Docs. The current Markdown is for testing only. It is NOT the source of truth (yet).

7 Understanding the Data

> **Note**
>
> This chapter is being drafted in Google Docs at https://drive.google.com/drive/u/0/folders/1OHv1P2DKGucKBpSNEiQp8lGfR2xYHPAW
> See a draft of the chapter outline at https://docs.google.com/document/d/1ttUKgwVcIZHM87elrlUNV6Qi9thzOwKBg8GegKObEtg/

> **Warning**
>
> At this point, any edits to this chapter should be made in Google Docs. The current Markdown is for testing only. It is NOT the source of truth (yet).

8 Introducing Enclave Analysis Tools

Additional Contributors: Johanna Loomba, Evan French, etc...

> **Note**
>
> This chapter is being drafted in Google Docs at https://drive.google.com/drive/u/0/folders/1RefwFn6mIitASq9IfPccN-T33iMohUEb
> See a draft of the chapter outline at https://docs.google.com/document/d/1ttUKgwVcIZHM87elrlUNV6Qi9thzOwKBg8GegKObEtg/

> **Warning**
>
> At this point, any edits to this chapter should be made in Google Docs. The current Markdown is for testing only. It is NOT the source of truth (yet).

Be sure to read the Notes for Contributors!

8.1 7 - Introducing Enclave Analysis Tools

Chapter Leads: Amy Olex, Andrea Zhou

8.2 Introduction

This chapter introduces tools in the N3C Enclave used to analyze data, view results, track project progress, and obtain shared data and code developed in the N3C community. The focus is on accessing and using each tool, the skill level needed, as well as what types of analyses each tool is geared toward. It is expected that you know how data are organized in the N3C Enclave, including the OMOP data model, vocabulary, and concept sets (see Chapter X for details).

Figure 1: High-level overview of an N3C project.

Due to the complexities of analyzing large clinical datasets, such as that compiled in the N3C Enclave, it is common, and many times necessary, to work in multidisciplinary collaborative

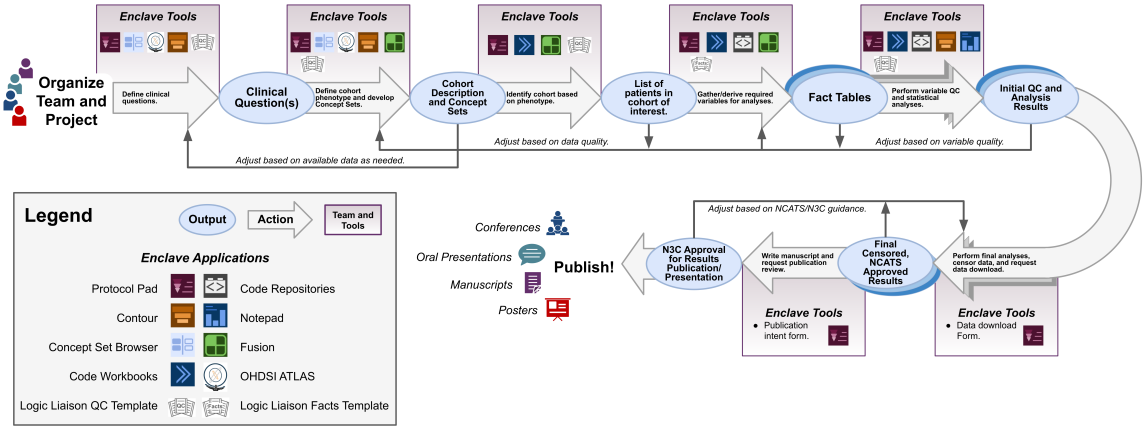

Figure 8.1: alt_text

teams to answer a research question. Figure 1 provides a high-level overview of the process behind forming a team and performing research using the N3C Enclave, along with the recommended field expertise needed during each phase. It is important to note that while certain team members may take the lead at various stages, a project benefits if all team members are engaged to some degree at all phases. Managing these collaborative and multi-faceted projects requires good record keeping. The N3C Protocol Pad (see Section X in Chapter X) is designed specifically for N3C research and to aid teams in designing, implementing, reporting, and publishing their work in a Findable, Accessible, Interoperable, and Reusable (FAIR) manner [1][2]. Thus it is recommended that you utilize this tool throughout the implementation of your project.

A research project in N3C starts with organizing a team with the required expertise (clinical, informatics, statistical, etc), followed by defining clinical questions around COVID-19 and characterizing the cohorts needed to answer each, i.e. clinical phenotyping. Because the N3C contains real-world EHR data that is harmonized from multiple data models and dozens of institutions, some information needed to identify an ideal clinical phenotype may be missing or incomplete. Thus, it is important to assess what information is needed to create an N3C computational phenotype for your cohorts. This could include using conditions, labs, or medications as proxies to identify a cohort if some information is not available. Generally, clinicians or other subject matter experts are leading this process with informaticians/data scientists providing guidance on what information is included in the N3C Enclave, what is missing or sparse, and overall data quality (see Chapter X). Some data quality aspects can be easily obtained through the use of Logic Liaison Templates (see Section X) accessible through the N3C Knowledge Store. The N3C Enclave application Contour (see Section X) can be utilized at this stage, along with Code Workbooks (see Section X) for quick querying and visualizing of the data. Additionally, Fusion (see Section X) can be utilized to keep track of

developed concept sets and utilized to easily input them into Logic Liaison Templates.

The generation of a computational phenotype overlaps with the generation of concept sets (see Section X in Chapter X for details.), and is often a cyclical process. Well-vetted concept sets are key to obtaining robust cohorts, thus, having a team member familiar with the organization of data and the OMOP vocabulary, such as a data liaison, who can work closely with a clinician is beneficial. Concept set generation can be done using the N3C Enclave Concept Set Browser, or externally through OHDSI ATLAS.

Informaticians and data scientists then utilize the computational phenotype and vetted concept sets to generate fact tables (i.e. datasets containing information about each patient like demographics, comorbidities, lab results, etc) for the cohorts of interest using the raw OMOP tables, which requires specific knowledge of how to work with large datasets in a Spark environment. Fact tables include all the information needed to characterize a cohort and perform downstream analyses to answer your research questions. Facts can include patient demographics, socioeconomic status, COVID status/severity, medications, comorbidities, etc. Logic Liaison Fact Table Templates can provide you a boost by allowing fast and robust generation of commonly used facts using N3C vetted concept sets and peer-reviewed code as a starter table. You can then append this base fact table to include project-specific facts needed for analyses. Figures 6 and 7 in the N3C Knowledge Store section of this chapter provide a more detailed view of how Logic Liaison Templates can be integrated into a project to expedite fact table generation. The generation of the original fact tables from raw OMOP tables can be done using Code Workbooks (see Section X) or Code Repositories (see Section X).

Data scientists and statisticians can then analyze the extracted and formatted fact tables. This includes statistical tests, summary tables, visualizations, and reports for the team to discuss. Data analysis is also a cyclical process with all team members engaged in assessing results and circling back to further refine the computational phenotype and concept sets if needed. Depending on the type of analysis needed, Code Workbooks (see Section X) or Contour (see Section X) can be utilized at this step, followed by Foundry's Notepad (see Section X) for reporting out results for secure team dissemination within the Enclave environment.

Once you obtain results that you wish to share with others, all tables, figures, and other data needed for reporting in publications, conference submissions, presentations, or any other activity outside the N3C Enclave environment must be submitted as a Data Download Request for a download review by NCATS (see Chapter X). The download request is meant to ensure no prohibited data is being downloaded as per the N3C Data Download Policy summarized in the Publishing and Sharing Your Work chapter. After approval, your results can be included in research outputs, such as publications, and then submitted to the Publication Review Committee (see Chapter X). This step is necessary to ensure data are being reported properly in the context of the research project and that proper attribution is being given to all those who contributed to the success of the research, either directly or indirectly. Upon approval, you are free to submit to the venue of choice and freely present the approved data to anyone at any time. Data download requests are performed within the Enclave environment, followed by submitting a Google Form to the Publication Review Committee.