Guide to N3C

2022-11-08T07:18:17+00:00

Table of contents

II Investigating   28

6 Getting & Managing Data Access   29

7 Understanding the Data   30

8 Analyzing the Data   31

9 Best Practices and Important Data Considerations   42

10 Publishing and Sharing Your Work   43

III Special Topics   44

11 Help and Support   45

12 Machine Learning   46

13 Advanced Enclave Coding Techniques   47

14 Start to finish examples or worked examples   48

Welcome

The Book of 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 book-of-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

Part I

Getting Started

1 Introduction

Karen Crowley ([Brown University](#))
Shawn O'Neil ([University of Colorado, Anschutz](#))

> **ℹ 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).

Chapter Contributors:

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

1.0.2 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

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,

---

[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.

[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.

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

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.

1.0.3 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.
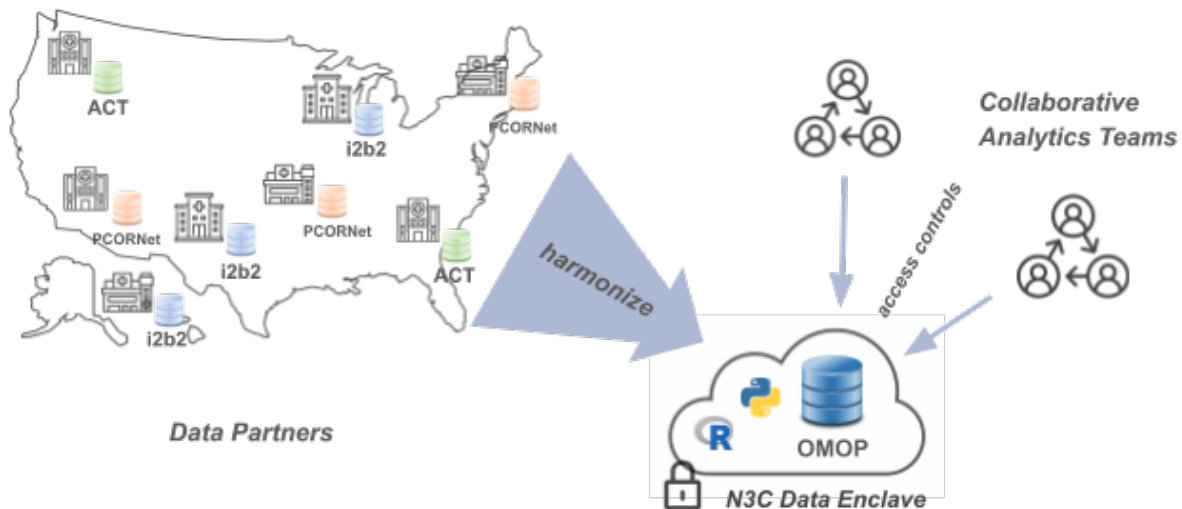
Figure 1.2: A visual representation of N3C's data harmonization from source-CDM model, and team-based access to these data in a secure cloud-based enclave.

---

[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.

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.

15,035,327
TOTAL N3C PATIENTS

5,729,896
CONFIRMED COVID-19 (+)

147,254
POSSIBLE COVID-19 (+)

73.0
SITES

17.3b
TOTAL ROWS

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.

1.0.4 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

News articles about N3C

VOX one

MIT press article

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

William H Beasley (University of Oklahoma Health Sciences Center)
Alfred H Anzalone (University of Nebraska Medical Center)

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

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.

2.1 Onboarding

Before the first data point is viewed, three things should be established by you and your institution:

- {A person's paperwork. Include ORCiD}
- {DUA coverage & An institution's paperwork.}
- {DUR & a project's paperwork.}

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 DUA will take several months in legal and administrative channels. Yet after clearing that first (tall) hurdle for your site, each specific project takes only a few days to be processed by the N3C staff. That's a remarkably short time for scale of available data. It's likely quicker than initiating a project based on a single EMR from your site –much less EMRs from 70+ sites.

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:

- 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.

- 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.

2.3 Common Discussion at the Team's First Meeting

Once the team is assembled, the first discussion is usually a variation of this exchange:

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

- **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[1], but you may not find it useful. Those two columns are sparsely populated and they look inconsistent across sites.[2]
- **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[3]. Give me 2 minutes to whip up something in SQL.[4]
- **I**: Oh my goodness, is that your cat? What a cutie! [5]
- **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.[6]
- **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.
- **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.

---

[1]Read about the OMOP Standard Tables in Chapter 7, specifically the medications are in the `drug_exposure` table.

[2]Conformance is a topic in Chapter 3.

[3]See Chapter 6 for accessing the N3C Enclave.

[4]Read about SQL, Python, and R transforms in Code Workbooks in Chapter 8.

[5]There is a brief discussion of SME's cat.

[6]There is a brief discussion of S's daughter strutting in the background wearing a cowboy hat and waiving a fairy wand.

- **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[7] 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.[8]
- **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.[9]
- **I**: Well, that's fine for this paper. Maybe our next manuscript will follow up with N3C's death records.[10]
- **SME**: Sorry everybody, I have clinic this week, and they're calling me. I need to drop.[11]
- **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.[12] 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 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.

---

[7]An HIE is a health information exchange.

[8]The benefits and caveats of real-world data are a theme throughout the book, particularly in the best practices discussed in Chapter 9.

[9]Authoring and using concept sets are described in Chapter 8. Mapping an ICD to SNOMED diagnosis code is an example of mapping a "non-standard" to a "standard" concept, discussed in Chapter 7.

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

[11]Everyone says goodbye to the cat.

[12]The `concept_relationship` table is discussed with the OMOP concept hierarchy in Chapter 7.

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 (Observational Health Data 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

> **ℹ Note**
>
> Note that there are five types of callouts, including: `note`, `warning`, `important`, `tip`, and `caution`.

> **💡 Tip With Caption**
>
> This is an example of a callout with a caption.

> **🔥 Expand To Learn About Collapse**
>
> This is an example of a 'folded' caution callout that can be expanded by the user. You can use `collapse="true"` to collapse it by default or `collapse="false"` to make a collapsible callout that is expanded by default.

3 Data Lifecycle - From Patients to N3C Researchers

> **ℹ 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

> **ℹ 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

> **ℹ 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).

Chapter Leads: Sharon Patrick, Jonathan Emery, Suzanne McCahan, Mary Helen Mays

Chapter Contributors:

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

> **ℹ 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

> ℹ **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 Analyzing the Data

Chapter Leads: Amy Olex, Andrea Zhou

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

8.1 Identifying Concept Sets

As data is ingested from the ever-growing list of data partners, the electronic health information coded in the various vocabularies used across the country are mapped to a single vocabulary of SNOMED CT. This is the core terminology for the OMOP common data model used within the enclave and a designated standard used to represent clinical meanings in a hierarchical structure. By leveraging the hierarchical structure, parent codes and descendants can be captured with ease to create intentional concept sets for use in analysis. For more details around the process of concept set creation, read the [authoring concept sets]#Authoring-Concept-Sets section.

The Concept Set Browser is an N3C specific tool that allows researchers to explore and modify existing concept sets as well as create new concept sets to fit their exact study needs. New researchers can start their search for a concept set with the list of N3C Recommended concept sets. These concept sets have been frozen in their validated state by the Logic Liaisons and the Data Liaisons after obtaining clinician and informatic reviews. These concept sets are the ones used to identify common comorbidities and other facts on the Phenotype Explorer and in the Logic Liaison Fact Table templates. The other method of finding commonly used concept

sets is by exploring bundles within the Concept Set Browser. These are sets of concept sets that are often used together in a group. These are created by *** and used when ***.

8.2 Using Concept Sets

Once concept sets have been identified for use in an analysis through the concept set browser, the concept set members table becomes the link between concepts and the encompassing concept set. Researchers can use concept sets by referencing the concept set name or the concept set version id also known as the codeset id. When searching for a concept set by name, it is recommended to also use most_recent_version = true to obtain the most up to date list of concept ids at any point in the analysis. This method is highly favored if the concept set is marked as N3C Recommended since these concept sets can only be updated by a small number of approved users after they have gone through re-validation, if necessary, following updates such as vocabulary changes. Researchers may choose to look up concept sets in the concept set members table by their codeset id if they wish to use the current most recent version of a concept set that is not N3C Recommended as any user can make edits to these concept sets. This will allow researchers to perform their analysis from start to finish without worry that someone has modified their concept set midstream without their knowledge. The other instance in which a researcher may choose to reference a concept set by codeset id is when choosing to use a specific prior version of an existing concept set.

8.3 Using the Knowledge Store

The N3C Knowledge Store is an application where Enclave users can discover shared Code Templates, External Datasets, Reports, Cohorts, and Python Libraries (all also known as Knowledge Objects) and share similarly re-usable Knowledge Objects of their own with other Enclave users, regardless of the specific DUR from which the resource originated. Majority of Knowledge Store objects come about as core contributors and researchers alike develop resources they believe may be useful for others either within or outside of their research project team and wish to share them with the broader community. If you find yourself in this situation, you can easily create a KS resource that can be shared using Palantir's documentation around templatization and submitting to the Knowledge Store. Otherwise, more specifics on how to navigate the Knowledge Store can be found in this [guide]#Knowledge-Store-Guide within the Enclave. Of the many types of Knowledge Objects, the most common are code templates and datasets.

Datasets in the Knowledge Store can be internal or external datasets. Internal datasets are generated from data inside the enclave, typically by researchers as part of their project, and are often of patient level granularity. External datasets found in the Knowledge Store provide a wealth of information from public datasets that have been brought into the Enclave along with

the crosswalks necessary for joining these aggregate data to person level data at various levels of granularity. Either type of dataset can be imported into a workbook or code repository to be used as a starting point for further transformation or analysis.

Depending on the author's intended use, some code templates can be applied to a researcher's custom input dataset while other code templates produce a dataset that can be joined with a researcher's study cohort. The code templates can also be imported and customized to produce a dataset for study analysis or simply used as example logic for Enclave users who are newer to coding. A few helpful starter templates are those produced by the [Logic Liaisons]#N3C-Logic-Liaisons. Through surveying the Domain Team Leads to establish a list of commonly derived variables and continuous feedback from the N3C Community to refine and update this list, the Logic Liaisons have developed, disseminated, and continue to maintain two master fact templates. These two master fact templates each produce visit-level and person-level data frames of commonly used derived variables for [all N3C patients]#All-Patients-Template as well as a subset who have an index date for their acute COVID-19 infection, the [confirmed COVID-19 Positive patients]#Confirmed-Covid-Pts-Template (PCR/AG positive or U07.1 COVID-19 diagnosed).

The Logic Liaisons have also developed, disseminated, and continue to maintain a handful of sister fact templates and data quality templates in the Knowledge Store. The sister templates utilize the visit-level and person-level datasets of the master fact template to efficiently generate additional derived variables based on broadly requested and applicable logic such as study-specific fact indexing, co-occurrence, and CCI score calculations. The data quality templates provide a variety of visualizations for looking at [data density of the OMOP domain tables]#Data-Density. When looking to assess study specific variables, the [Systematic Missingness template]#Systematic-Missingness provides an all or nothing fact density indicator while the [Fact Density by Site template]#Fact-Density provides relative densities of a fact across sites. [Training materials]#Logic-Liaison-Tutorials for getting started with Logic Liaison templates are available within the Enclave.

While it is not necessary to utilize Knowledge Store resources when conducting your research project, it does allow you to get a jumpstart on gathering and understanding the data by avoiding effort duplication and providing a general starting point. The Logic Liaison fact templates are specifically designed to provide a validated and community agreed-upon method for calculating particular facts at the encounter level and/or the patient level. The Logic Liaison data quality templates provide the same structure for analyzing data missingness, density, and contribution quality by site.

8.4 Enclave Applications

This section will cover the usage of various applications made available in the N3C Enclave, including Contour, Code Workbooks, Reports, and more (a complete list of Foundry applications can be found here). However, before designing and running an analysis utilizing data in

the Enclave, it helps to understand the concept of a "data transformation" and how the data is stored and accessed via SPARK on a distributed file system.

8.4.1 10,000 foot view of distributed file systems, SPARK, and data transformations

Data in the N3C Enclave is so large it cannot be stored all in one giant table. Instead it is stored on multiple servers in what is called a "distributed file system" where sets of rows in a table may be stored across multiple physical servers (Figure 1). Accessing these multiple servers to get data from one table requires a lot of coordination behind the scenes. SPARK is the program coordinating the fetching, writing, and analysis of data. You, as an Enclave user, can interface with SPARK through familiar Python, R, and SQL commands in order to retrieve data and run your analysis. You don't necessarily need to know all of the details in how SPARK operates to run basic code; however, when working with data in the N3C Enclave it is highly recommended to get more acquainted with how SPARK operates in order to optimize your code to make it run faster (like WAY faster). You can reference Chapter X of this book, or Palatir's SPARK Fundamentals and SPARK Optimization modules for more details on SPARK and how to optimize your code. For the rest of this chapter we are just going to focus on how SPARK uses "data transformations'".

The basics of a data transformation are illustrated in Figure 2A. *One or more tables are specified as the input* and are transformed into a single output table. Multiple transformations can be strung together (Figure 2B) to create an analysis pipeline (see Palatir's Anatomy of a Data Pipeline module for more detailed information, as well as Best Practices in Chapter X). *The primary and required output is always a single table*; however, visualizations such as graphs and charts can also be generated and saved along the way. You, the user, define the transformations using Enclave applications like Contour and Code Workbooks. Fusion sheets can be utilized to manually create tables for input into an analysis (e.g. to hold configuration options or frequently used codesets), and Reports are used to compile and display analysis results such as summary tables, graphs, and charts. In the following sections, each of these applications is discussed in more detail.

8.4.2 Contour

Foundry's Contour application is a programming-free interface to the N3C Enclave that allows those with limited knowledge of Python, R, and SQL to create top-down analysis pipelines in a point-and-click fashion, as well as interactive dashboards for sharing results. It is best used to quickly summarize and visualize data, and usage of Contour's expression language allows for more complex querying and data aggregation. One of the advantages of Contour is the ability to quickly and easily create summary figures with various complexity from source tables without having to code. Contour has a variety of figure options, including bar charts,

histograms and heatmaps. A detailed orientation to Contour can be found in the Foundry Documentation here.

8.4.3 Fusion

Palantir Documentation: Fusion Sheet Overview

- Useful for writing back datasets for use within the Enclave
- Leverage cell references and spreadsheet functions
- Index datasets
- Sync tables to a dataset to use in other Foundry applications
- Import xls data
- Create charts
- Allows customization and flexibility
-

Fusion is a spreadsheet application within the Enclave analogous to Excel or Google Sheets. Palantir has curated extensive documentation describing its core functionality as well as providing select how-to tutorials. The primary utility of Fusion is to allow users to sync specific cell ranges of values within a spreadsheet to datasets, which can subsequently be [imported]#Importing-and-Viewing-Data into any other Enclave application. This tool is an excellent option for use cases which require manual data entry, such as curating lists of [concept sets]#Concept-Set-Section to configure the [Logic Liaison Fact Templates]#LL-Fact-Templates. Unlike many other Enclave applications, Fusion is not suitable for large datasets; it has a maximum size restriction of 50 MB per document. Similar to Google Sheets, Fusion allows users to simultaneously edit the same document and view other users' changes in real time.

As any Excel super user knows, a spreadsheet is not merely a mechanism for storing data, but also a powerful tool for analyzing, visualizing, and applying complex logical operations on data in its own right. Fusion provides many features familiar to other spreadsheet applications such as cell-referencing formulas, formatting, and a charting library to name a few. External .xls/.xlsx formatted files (up to 4 MB) can be directly imported into the Enclave as Fusion sheets with all cell references, formulas, and formatting preserved. [caveat].

In addition to standard spreadsheet functionality, Fusion has additional features which allow it to bi-directionally integrate with the rest of your Enclave environment. The first such feature is the ability to import other Enclave datasets into Fusion. Prior to using any dataset in Fusion (including datasets synced from Fusion sheets), it must first be indexed into a non-distributed format using the _Data _tab in the top menu. You can view available indexed datasets using the *Data>Datasets available* menu option. Objects created within Fusion, such as formatted tables can be embedded in reports#Reports. Finally, Fusion sheets can be templatized to facilitate replication of similar functionality.

8.4.4 Code Workbooks

Tutorial: Intro to Code Workbook

Palantir Documentation: Code Workbook Overview

- graphical organization of logic
- Simplification of code
- easy reuse of pre-authored logic
- Add visualizations from here to reports
- Supports Python, R, and SQL
- Branching facilitates collaboration
- Workspace to reuse templatized logic

Code Workbook is a GUI-based application for users to apply code-based transformations to datasets for the purpose of creating new datasets and visualizations. The explicit goals of the application are to facilitate a collaborative environment in which users can quickly iterate over logic to produce artifacts interoperable with the suite of Enclave applications. The default Code Workbook interface is structured as a directed graph in which nodes represent either datasets or transformations which can output datasets. Edges represent the flow of data through the graph such that upstream datasets are inputs for logical operations performed by downstream code transforms (see Figure). Any dataset which a user has access to within the workspace where the Code Workbook is located can be imported as an input to the various types of transform.

8.4.4.1 Types of Transform

- **Manual Entry** transforms allow users to manually curate non-persisted datasets directly in the Code Workbook as a "quick and dirty" alternative to manually curating persisted datasets with Fusion#Fusion.
- **Python Code** transforms let users write a Python function with input datasets as input parameters.
- **R Code** transforms let users write an R function with input datasets as input parameters.
- **SQL Code** transforms let users write a SparkSQL query to create a new dataset from the available inputs
- **Template** transforms are parameterized blocks of reusable code which users can configure from a point-and-click interface. Many code templates are available in the [Knowledge Store]#Knowledge-Store, but users can also create their own templates scoped to their DUR workspace. Multiple single-node templates can be combined to create a multi-node template to allow reuse of entire configurable pipelines.

- **Visualize** transforms offer a point-and-click interface for users to create charts, distributions, histograms, and pivot tables. Note this transform is only available for saved datasets.

Both Python and R transforms can optionally return a single dataset and produce visualizations using Python libraries/R packages. Any visualization produced in a Code Workbook and subsequently be embedded in a Reports#Reports document. Datasets returned by a transform are ephemeral by default, that is, the transform must be recomputed each time the dataset is used as a downstream input, but options exist to conserve compute power by either caching or saving the output. Caching stores the output temporarily, while saving the dataset stores it permanently in the Enclave. It is recommended that transforms in a pipeline requiring significant compute be saved as datasets to reduce iteration time during development. Once a dataset is saved, users can set triggers and schedules for it to be automatically updated when certain input datasets are updated or at regular temporal intervals.

Tooling is available within Code Workbooks to facilitate the application objective of quick iteration. The console feature, located in the top right, allows users to interactively explore logic and syntax in Python, R, or SQL outside of a transform node. Also in the top right, is the Global Code section where users can globally import Python libraries and R packages or define custom functions to be used in any transform in the Code Workbook. Code transforms include a Logs feature for users to view console output generated by code and to view detailed stack traces when transforms fail due to an error.

Many Python and R transforms rely on external libraries and packages which can be made available via the Environment Configuration. The Enclave provides a number of default configurations tailored to common use cases, for instance, the *default* profile, which includes common packages like pandas and tidyverse is suitable for routine analysis, whereas the *profile-high-memory* profile includes packages like founder_ml to facilitate machine learning. Users can create their own Environment Configurations which include packages meeting their specific needs. Not all libraries and packages are included in the list of options, but users can [submit a ticket]#Tickets to request additional packages be made available. The Enclave maintains instances of the non-custom configurations on warm standby, allowing them to be quickly initialized when a user requests a new environment. Custom configurations require more time for initialization as instances of these must be started from scratch rather than merely assigned. For this reason, it is recommended to use non-custom configurations when possible.

Following best practices for collaborative software development, Code Workbook allows for branching of the logic within a workbook. As with other popular source control technologies (i.e. git), branching allows users to make copies of a workbook which they can develop independently of the source workbook. Once the development in a particular branch is deemed complete, it can be merged back into the originating branch. Prior to the merge, users can preview both line-level differences within each node, as well as node-level differences of nodes that have been added/removed. Good practice dictates that individual users perform all development on individual branches, which are then merged back into a common *master* branch.

Because the *master* branch can change in the interval between a user cutting a branch and merging it back in, previewing merge changes is an important step for ensuring that the individual's changes are both correct and compatible with the current state of the *master* branch. Another prime use case for code branching is to ensure the reproducibility of a given dataset used in a research project. Because the OMOP and N3C-curated datasets are also versioned, teams can create a code branch in which all input datasets are set to the same version release to effectively freeze a dataset used in a specific analysis for later reproducibility while still allowing the possibility of adding additional features. User generated datasets are set to the same branch as the Code Workbook in which they were created.

Palantir has created extensive documentation of the Code Workbook application including tutorials. N3C has also published training materials.

8.4.5 Reports

Palantir Documentation: Reports Overview

- display charts and visualizations in combination with text descriptions about the analysis
- Collaborate with others through comments on the reports
- Add context by attaching images and other files
- Create a story around analysis results
- Useful when sharing results or even just putting together preliminary look at the progress

Many research projects in the Enclave are complex, involving multiple summary datasets, statistical analyses, and visualizations scattered across multiple applications and documents. Reports is a tool for consolidating various research artifacts from multiple sources within the Enclave into a single coherent document. Formatted Fusion#Fusion tables, Contour#Contour charts, Python/R-generated images from Code Workbooks#Code-Workbooks, and more are all embeddable in a Reports document, with the option to add a title and/or caption for each embedded artifact. Users can also create sections and provide narrative structure to their documents using MarkDown blocks. All components of a Reports document can be arranged and configured using a point-and-click interface.

All embedded objects can be configured to refresh automatically when the underlying data sources update, allowing Reports to function effectively as dashboards. Reports are also useful for presenting an executive summary of results for internal stakeholders as well as external presentations. [Logic Liaison Templates]#LL-Templates in the [Knowledge Store]#Knowledge-Store generally includes a README which is created using Reports. Reports can be requested for download using the Download Request Form. Palantir has curated documentation for creating and editing Reports.

8.4.6 Object Explorer

Palantir Documentation: Object Explorer

Tutorial: Exploring N3C Projects and Collaborators with Object Explorer

- Easily find objects of interest
- Compare and contrast items

A step upstream of the Knowledge Store is Palantir's Object Explorer which allows users to explore and analyze objects of interest through a point and click interface. Detailed information around how to use this feature can be found in [Palantir's documentation here]#Object-Explorer. Researchers can find unpublished Knowledge Objects, [explore N3C Projects and Collaborators]#Object-Explorer-Tutorial, view object usage statistics, etc. using this application.

8.4.7 Data Lineage (aka Monocle)

Whether you're creating a pipeline for data analysis specific to your research project or investigating one you came across in the Knowledge Store to determine if it will be useful to your study, you'll likely want the capability to holistically assess how that dataset came about. The Data Lineage tool within the Foundry platform allows users to do just that - easily explore data pipelines from start to finish. With upstream on the left and downstream on the right, this tool makes for an intuitive way to visualize the relationships between datasets and their ancestors or descendants with enhanced views made possible through color-coding and grouping based on a dataset's details. This application is particularly useful if a researcher needs to schedule a build for a dataset to update with the weekly refresh of data in the enclave. Palantir Documentation provides a short tutorial and additional descriptions of this tool's functionality.

8.4.8 Code Repositories

8.4.9 Protocol Pad

Quick Guide

Detailed Guide

Research studies can span many months and pass through the hands of many team members before reaching a stage where the researcher may want to share the results through publication or other approved means. The Protocol Pad serves as an electronic lab notebook to help organize tasks, track progress, and document results in a cohesive format throughout the process of reaching a study's final state.

8.5 DRAFT Language

To demonstrate the use of Contour we will work through a simple use-case using the SynPuf Synthetic dataset to summarize patient demographics as a bar chart, filter a table, join in additional information from another table, and calculate a patient's age.

8.5.0.1 Interface Overview

Before we get started, let's get oriented to the interface for the Contour application. To create a new Contour analysis click on the "+ New" button (Figure <new_button>) and select the "Analysis" option (Figure <analysis_option>). A new Contour analysis will open, and you will see a window similar to that in Figure <contour_interface>. To change the name of this analysis, click on the last entry in the path (Figure <contour_interface>, A). This path can also be used to navigate back to previous parent directories in the Enclave. To the right of the analysis name is a drop-down indicating you are in "Editing mode" (Figure <contour_interface>, B). Note that if more than one person has the same workbook open only one person will be given Editing privileges while the other will be in Viewing Mode.

The center of the screen is your working space. Currently it is empty because we have not added any datasets for analysis. To add a new data set for analysis, click on the "+ Create new path" button in the center of the screen (Figure <contour_interface>, C), or the plus "+" button towards the top of the window ((Figure <contour_interface>, D). The term "path" in Contour is used to indicate the location of a data set. Each time you "Create a Path" you are telling the application which data set to start with, whether it be a source OMOP table or the output from another Contour analysis.

8.5.0.2 Importing and Viewing Data

As discussed in the introduction to this chapter, all analyses work through transformations, and Contour is no exception. Analyses in Contour are built as a pipeline working from top to bottom. To start an analysis we need to import a starting dataset (i.e. create a new path). Click on the "+ Create new path" button, then navigate to the OMOP "person" table in the Synpuf Dataset by going to "All" > "Data Catalog" > SynPuf Synthetic Data" > "person" (Figure <create_path>). Your window should now look similar to Figure <analysis_start>.

In this view there are 3 sections to be aware of:

1. Starting dataset
2. Transformations (currently there are none)
3. Resulting dataset (can be saved as a table or imported into another analysis as the starting path)

To add a transformation (a.k.a. board), just click on one of the suggestions (Figure <analysis_start>, B), or you can hover over one of the arrows before or after the suggestions and select "Insert Board" (Figure <insert_board>). These arrows will also appear before and after each transformation so you insert a transformation anywhere in the pipeline. Just be aware it will change all results downstream!

Notice at the top where the "+" button is (Figure <analysis_Start>, red box) there is a new tab named "person". You can create multiple analysis workflows within the same Contour analysis and navigate through them using the tabs.

Now we have imported our first data set, however notice it is not displaying the table for a preview. Within Contour, to see the data you specifically have to tell it to create a table as one of the transformations. To preview the data in the "person" table, click on the "Visualize" button and scroll down to "Table" (Figure <table_viz>). Your screen should look like Figure <table_preview>.

8.5.0.3 Creating Summary Figures

One of the advantageous things about Contour is that it is easy to quickly create summary figures of various complexity from source tables without having to code. Contour has a variety of figure options, including bar charts, histograms and heatmaps.

To create a bar chart of our example data, look through the table preview to identify the column name that has the variable you want to plot. In our instance it will be "race_concept_name". Create a new transformation below the table preview for a bar chart (named "Chart" in the list, Figure <bar_chart_options>). Initially the chart is empty because we have to tell it which columns to use. In the "Data" tab on the right, set the X-Axis to be "race_concept_name", and the Y-Axis to be "person_id" and "Unique count" (Figure <bar_char_options_set>), then click on
Compute & save" to view the chart (Figure <bar_chart_figure>). You can utilize the "segment by" option to segment your primary column by a secondary column (Figure <bar_chart_segmented>), as well as explore some of the other options available. Note that each visualization option will have its own parameters. You can review Foundry's documentation for more details on each of these.

9 Best Practices and Important Data Considerations

> **ℹ Note**
>
> This chapter is being drafted in Google Docs at https://drive.google.com/drive/u/0/folders/1ExkYChsnO3hYZk6HCI5cEfQdQJ9F-ynw
> 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).

10 Publishing and Sharing Your Work

> ℹ **Note**
>
> This chapter is being drafted in Google Docs at https://drive.google.com/drive/u/0/folders/1kmrjxsdrwbspPucPTU3hiOHMXJRQQpPp
>
> 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).

Part III

Special Topics

11 Help and Support

> **ℹ Note**
>
> This chapter is being drafted in Google Docs at https://drive.google.com/drive/u/0/folders/1SnhEKmA-A1GJHN5kBYRnX_za5dX_n1EC
> 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).

12 Machine Learning

> ℹ **Note**
>
> This chapter is being drafted in Google Docs at https://drive.google.com/drive/u/0/folders/1HZ3IGv17zUl9t8RxZSl4uOq_FRzrgTp_
>
> 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).

13 Advanced Enclave Coding Techniques

> ℹ **Note**
>
> This chapter is being drafted in Google Docs at https://drive.google.com/drive/u/0/folders/1K660Qn7m1z4TswwepM06CKgAPTojjt7q
> 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).

14 Start to finish examples or worked examples

> **ℹ Note**
>
> This chapter is being drafted in Google Docs at https://drive.google.com/drive/u/0/folders/1rWxFtzk1kyUSRJPDgPWwlmVCjnWEGf6i
>
> 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).

Part IV

Back Matter

References

Observational Health Data Sciences and Informatics. 2019. *The Book of OHDSI: Observational Health Data Sciences and Informatics.* United States: Independent. https://ohdsi.github.io/TheBookOfOhdsi/.

Wu, Chunlei, and C2DH. 2022. "Informatics Playbook." https://playbook.cd2h.org/.