Introduction to Analyzing Real-World Data Using the National COVID Cohort Collaborative (N3C)

Spring 2024 - N3C Education and Training Domain Team

Organizers/Instructors

Jerrod Anzalone, MS
University of Nebraska Medical Center
Will Beasley, PhD
University of Oklahoma Health Sciences Center
Stephanie Hong, FAMIA
Johns Hopkins University
Johanna Loomba, ME
University of Virginia
Shawn O'Neil, PhD, MS
University of Colorado Anschutz
Sharon Patrick, MS, MPA
West Virginia University
Ken Wilkins, PhD
National Institute of Diabetes and Digestive and Kidney Diseases
Andrea Zhou, ME
University of Virginia

Description

The National COVID Cohort Collaborative (N3C) is an open science community focused on analyzing patient-level data from many clinical centers to reveal patterns in COVID-19 patients. Over 10 billion rows of data are stored in the secure N3C Data Enclave in the OMOP common data model for Electronic Health Records. The Enclave (Palantir Foundry) supports analysis with SQL, R, and Python for conducting statistical and machine learning analyses, backed by the distributed-computing framework Apache Spark. This virtual short course introduces researchers (students, staff, and faculty) to N3C and will include hands-on experiences in computing, exploring and analyzing EHR data, and good research practices in the context of N3C. The knowledge and skills obtained through this course will prepare participants for developing their own N3C studies and understanding the secondary use of de-identified EHR data in general.

Overview

Six Thursday sessions: (2-4 pm CST)

- January 18: **Introduction to N3C**
- January 25: **Observational Research, OHDSI Tools, and N3C**
- February 1: **Analysis Using Synthetic Data**
- February 8: **Building an Analytic Dataset Using Shared N3C Resources**
- February 15: **Analyzing Data Curated from N3C: A Replication Study**
- February 22: **N3C Policies and Procedures: From Access to Publication**

Textbooks (both freely available online):

- ***The Researcher's Guide to N3C (G2N3C)***
  O'Neil ST, Beasley W, Loomba J, Patrick S, Wilkins KJ, Crowley KM., Anzalone, AJ (Eds.) (2023). *The Researcher's Guide to N3C: A National Resource for Analyzing Real-World Health Data.* DOI: 10.5281/zenodo.7749367
- ***The Book of OHDSI (BoO)***

Observational Health Data Sciences and Informatics (OHDSI). (2021). *The Book of OHDSI: Observational Health Data Sciences and Informatics.* https://ohdsi.github.io/TheBookOfOhdsi/

Point of Contact: Sharon Patrick: sharon.patrick@hsc.wvu.edu

Location: Virtual via Zoom

Office Hours (optional):
Students are encouraged to attend the general N3C office hours Tuesdays and Thursdays at 10a PST/1p EST. N3C support staff and course instructors will be available.

Read more here: https://covid.cd2h.org/support
Zoom link

Prerequisites:

- SQL and/or Python and/or R

Prerequisite Assignments for class:	Reading(s):	Due Date
• Access to Enclave	• *G2N3C*: Chapter 1, 2, 3, 4, 5	• January 18, 2024

Session 1. Introduction to N3C
January 18, 2024

Objectives: Session 1 introduces students to the N3C and the real-world data research ecosystem in 2024. Students will learn about the history and future of N3C, the platform, and the course structure.

Assignments:	Reading(s):	Due Date
• Request access to L2 DUR, Practice Enclave Navigation	• *G2N3C*: Chapter 6, 7 • *BoO (optional)*: Chapters 4, 5	• January 24, 2024, 11:59 PM

Session 2. Observational Research, OHDSI Tool, and N3C
January 25, 2024

Objectives: Session 2 delves into using electronic health record data for research and its application in N3C. Students will learn the building blocks of research studies in N3C using the platform and OHDSI tools, including how to use Athena, Atlas, and the N3C Concept Set Browser. Student will leave with a better understanding of the OMOP common data model, standard medical terminologies used in OHDSI, and their translation into N3C.

Assignments:	Reading(s):	Due Date

• Concept set creation and basic use	• *G2N3C*: Chapter 8, 9 • *BoO*: Chapter 8	• January 31, 2024, 11:59 PM

Session 3. Analysis Using Synthetic Data
February 1, 2024

Objectives: Session 3 will introduce students to data analysis in the N3C platform using entirely synthetic data. This session will cover using code workbooks in N3C to analyze data in SQL, R, and Python.

Assignments:	Reading(s):	Due Date
• TBD	• *G2N3C*: Chapter 10, 11	• February 7, 2024, 11:59 PM

Session 4. Building an Analytic Dataset Using Shared N3C Resources
February 8, 2024

Objectives: Session 4 will introduce the N3C Logic Liaison (LL) templates and their use in research studies in the platform. The LL templates are part of the N3C Knowledge Store, which includes many community-derived computable artifacts developed by the research community to support reproducible workflows and better science. Students will leave with an understanding of how to use the LL template(s), make modifications for study-specific purposes, and use it for downstream analyses.

Assignments:	Reading(s):	Due Date
• LL Template Documentation • Assignment TBD	• *Ponce J, Anzalone AJ, Bailey K, et al. Impact of malnutrition on clinical outcomes in patients diagnosed with COVID-19. JPEN J Parenter Enteral Nutr. 2022;46(8):1797-1807. doi:10.1002/jpen.2418* • *BoO (optional)*: Chapter 9, 10	• February 14, 2024, 11:59 PM

Session 5. Analyzing Data Curated from N3C: A Replication Study
February 15, 2024

Objectives: Session 5 will build upon the materials learned in prior lessons and put it all together by replicating an previously published study from N3C. Students will use the concept sets from

Session 4 to define a final study cohort and run statistical analyses on these mirroring those in the published study.

Assignments:	Reading(s):	Due Date
● Assignment TBD	● Attribution and Publication Principles for N3C ● Publication Review Checklist	• February 21, 2024, 11:59 PM

Session 6. N3C Policies and Procedures: From Access to Publication
February 22, 2024

Objectives: Session 6 will introduce the governance structures in N3C to ensure data privacy and alignment with the DUA and DTAs in place. Students will learn about procedural requirements to download aggregated results and model output in N3C, submit work for publication to the N3C Publication Committee, and best practices when preparing and sharing data from N3C studies.

Assignments:	Reading(s):	Due Date
● Course Feedback Survey	● None!	• February 28, 2024, 11:59 PM

!!! Important Reminders !!!

- N3C Data Enclave Institutional Data Use Agreement (Confidentiality and Data Security)
    - **Users agree to not attempt to re-identify or contact any individuals** who are the subjects of the Data or any known living relatives unless required by law to maintain public health and safety.
    - **Users agree not to attempt to use the Data to identify or contact any Data Contributors or healthcare providers** unless such identification is needed for Data preparation and management purposes and only at the request of NCATS or required by law to maintain public health and safety.
    - **Users agree to not photograph, create screenshots, nor download data** viewed on the NCATS N3C Data Enclave. Furthermore, Users' access to the NCATS N3C Data Enclave will be terminated after 1 year. Users' access to their analyses and findings on the NCATS N3C Data Enclave may continue with DAC approval of DUR renewal.

- **N3C Code of Conduct**
    - **Make no attempt to identify** institutions, communities or identifiable populations associated with particular N3C data; to re-identify or contact individuals, their relatives or relevant groups from whom data represented

within the N3C Data Enclave were collected; or to generate information that could allow individual identities to be readily ascertained.
    - **Maintain the confidentiality of the data** accessible within the enclave and not distribute, provide access to, attempt to download or otherwise capture views of N3C data for use or disclosure to any entity or individual beyond those specified in the approved DUR.

- **No** recording, hand copying and pasting, photographs/screenshots (or emailing of screenshots), verbal description of data, or sharing of access to data

- Everything should stay within the N3C Data Enclave (we can look at any work in there)
    - Only share screen with those with same level access DUR (e.g., Level 2 de-identified) for in person or virtual (e.g., Zoom) meetings
    - Be sure to lock your computer whenever you step away, even for a short period of time.

- **If you are not sure about something, ask the course organizers/instructors first!**

Materials and Communication

- Materials will be distributed through the course GitHub repo: https://github.com/National-COVID-Cohort-Collaborative/short-course-2024-january

- Communication will be done using the #n3c-2024-short-course Slack channel in the CD2H workspace:
    - Use for troubleshooting and questions during class (rather than Zoom chat)
    - Use for questions outside of class (avoid direct messages)