IMPACT OF COVID-19 INFECTION IN PATIENTS WITH PULMONARY NON-TUBERCULOUS MYCOBACTERIUM (NTM) INFECTION: A NATIONAL COVID COLLABORATIVE COHORT (N3C) STUDY

By

Carlos E. Figueroa Castro, MD

A CAPSTONE

Presented to the Department of Medical Informatics and Clinical Epidemiology
and the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of

Master of Science

December 2020

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Master's Capstone Project of

Carlos E. Figueroa Castro

"Impact of COVID-19 infection in patients with pulmonary non-tuberculous mycobacterium (NTM) infection: a National COVID Collaborative Cohort (N3C) study"

Has been approved

William Hersh, MD

Table of Contents

Acknowledgments

The analyses described in this report were conducted with data or tools accessed through the National Center for Advancing Translational Sciences (NCATS) National COVID Collaborative Cohort (N3C) Data Enclave (https://ncats.nih.gov/n3c/about) and supported by NCATS U24 TR002306. This research was possible because of the patients whose information is included within the data and the organizations and scientists who have contributed to the ongoing development of this community resource. Support from Bradley W. Taylor, Chief Research Informatics Officer at the Medical College of Wisconsin (MCW) Clinical and Translational Science Institute (CTSI), April Haverty, MCW Director Grants and Contracts Research Office, and Anita Walden, MS, Assistant Director, N3C Administration

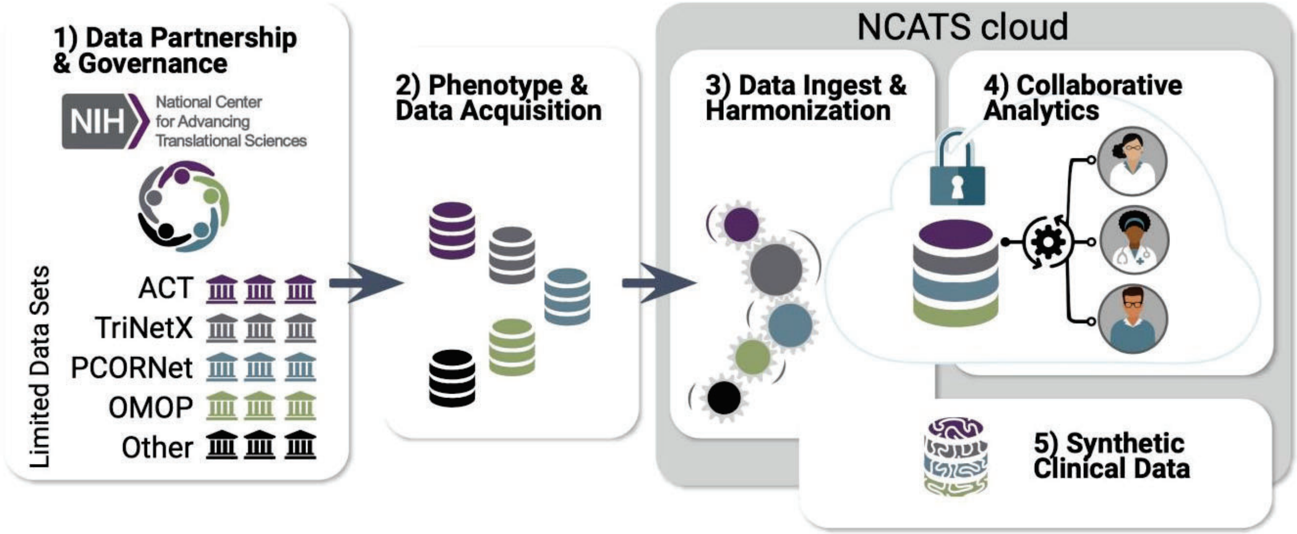
COVID-19 is an acute respiratory infection caused by a novel coronavirus, with certain risk factors associated to poor outcomes. It is not known if patients with pulmonary non-tuberculous *Mycobacterium* (PNTM) infections are at a higher risk of infection, or its severe manifestations. Patient data from the National COVID Cohort Collaborative (N3C) was used to study this population. We described the N3C structure and steps to gain access to patient data, the development of a PNTM infection phenotype, and the use of tools within the N3C analytical platform, the N3C Data Enclave, to answer this question. We identified 277 PNTM patients, who were mostly female (64.3%), of advanced age (mean age: 65.16 years), and not of Hispanic ethnicity (92.8%). The most common NTM was *Mycobacterium avium intracellulare*. Eighteen patients had COVID-19 infection based on PNTM infection phenotype, and they were older than COVID-19 patients without PNTM infection (mean age 60.27 vs 43.21, $p = 0.0004$). Analytics used a combination of point-and-click and command line tools available in the N3C Data Enclave. Additional work to refine the PNTM phenotype definition; and to evaluate differences in hospital course, and the effect of confounders commonly seen in PNTM patients, are required. N3C is a promising tool to study the effects of COVID-19 on special populations, and become a model for large-scale observational studies of novel infectious diseases.

Introduction

COVID-19 is an infectious disease caused by SARS-CoV-2, a novel betacoronavirus first linked to a seafood wholesale market in Wuhan, China [1]. Clinical manifestations varies from asymptomatic infection to severe acute respiratory infection and death. Based on the COVID-19 tracking project by the Center for Systems Science and Engineering at Johns Hopkins University (https://coronavirus.jhu.edu/map.html) [2], this virus has affected at least 66 million people, and caused more than a million and a half deaths worldwide as of December 6, 2020. In the United States, the virus has infected more than fourteen million individuals, and caused more than 280,000 deaths. Risk factors associated to poor outcomes have been identified on multiple retrospective cohort studies, including older age, cardiovascular disease, chronic lung disease, diabetes, and obesity [3–8]. However, it is difficult to establish whether patients with a specific clinical condition with low prevalence in the general population is at a higher risk of COVID-19 infection, or serious adverse outcomes. In this scenario, cohort studies are limited by the number of patients that could be described, and in most situations, the descriptions are based on single-center studies, which limits generalization of results. Early descriptive studies of special populations during the COVID-19 pandemic, like people with HIV infection [9,10], and transplant recipients [11], are some examples of this problem. This sort of study design is usually a compromise related to lack of access to patient's data, which is paradoxical, given the growing abundance of patient's clinical data. Unfortunately, accessing, organizing, and querying large clinical data sources is out of reach for a single institution or researcher. To solve this situation, established network of research consortia like the National Patient-Centered Clinical Research Network (PCORnet) [12], Accrual to Clinical Trials Network (ACT) [13], and TriNetX [14], and newly formed entities like the 4CE Consortium [15], have been used to facilitate analysis of the impact of COVID-19 infection in special populations, like HIV [16], solid organ transplant [17], and discovery of novel risk factors [18].

Non-tuberculous *Mycobacterium* (NTM) infections are caused by a diverse number of *Mycobacterium* species (> 190) [19]. Even though these bacteria can affect any organ system, the primary manifestation is pulmonary, which largely affect people with certain conditions like chronic obstructive pulmonary disease (COPD), cystic fibrosis (CF), bronchiectasis, and people with some immunodeficiency conditions. These infections have a worldwide distribution. The most common cause of pulmonary NTM (PNTM) is *M.avium intracellulare* complex (MAC). The incidence and prevalence of these infections seem to be increasing in the United States, especially among women and elder populations [20]. The mainstay of therapy is antibiotic combination (usually a macrolide, ethambutol, and rifampin for MAC infection) for 18-24 months, and reinfection is common [21]. It is not known whether these patients are at an intrinsic higher risk for COVID-19 infection, or its severe manifestations, or because of the presence of known risk factors for both PNTM infection and severe COVID-19 infection. To evaluate this hypothesis, access to patient data contained in the newly formed National COVID Cohort Collaborative (N3C) was obtained [22]. N3C is a partnership among the National Center for Advancing Translational Sciences (NCATS) and Clinical and Translational Science Awards (CTSA) Program hubs, and the National Center for Data to Health (CD2H), N3C provides researchers with clinical data derived from electronic health records from ACT, TriNetX, PCORnet, and others using the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) [23] to answer critical research questions by developing a collaborative analytics platform, the N3C Data Enclave [24], which uses harmonized data following the OMOP CDM [25], and the development of synthetic clinical data, which is an artificial, statistically-comparable, computational derivative of the original data without protected health information (PHI) as defined by the Health Insurance Portability and Accountability Act (HIPAA) (Figure 1).

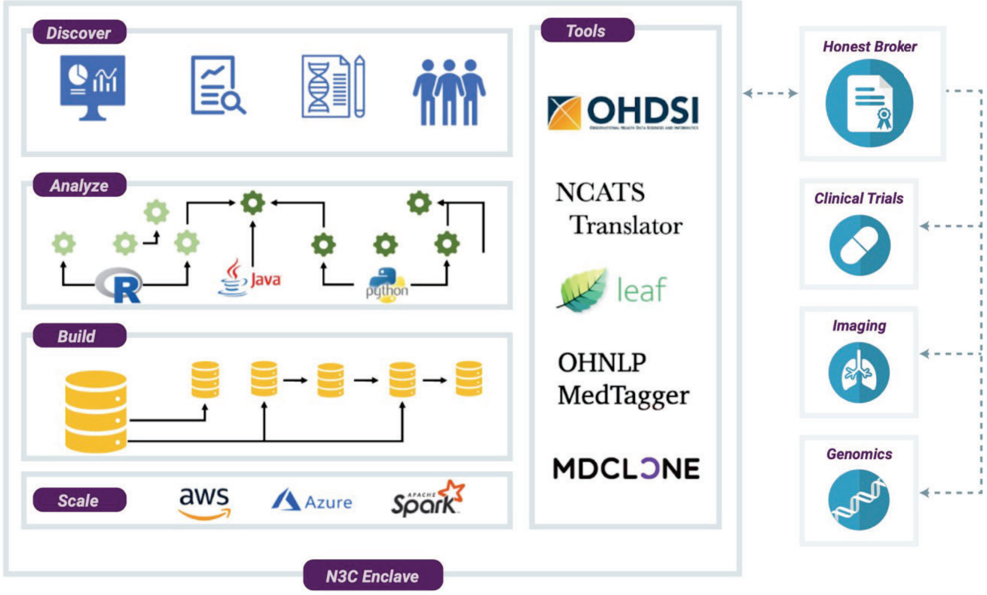
Figure 1. N3C activities and community workstreams (from Reference [22])



1) Data Partnership & Governance
NIH National Center for Advancing Translational Sciences
Limited Data Sets
ACT
TriNetX
PCORNet
OMOP
Other
2) Phenotype & Data Acquisition
NCATS cloud
3) Data Ingest & Harmonization
4) Collaborative Analytics
5) Synthetic Clinical Data

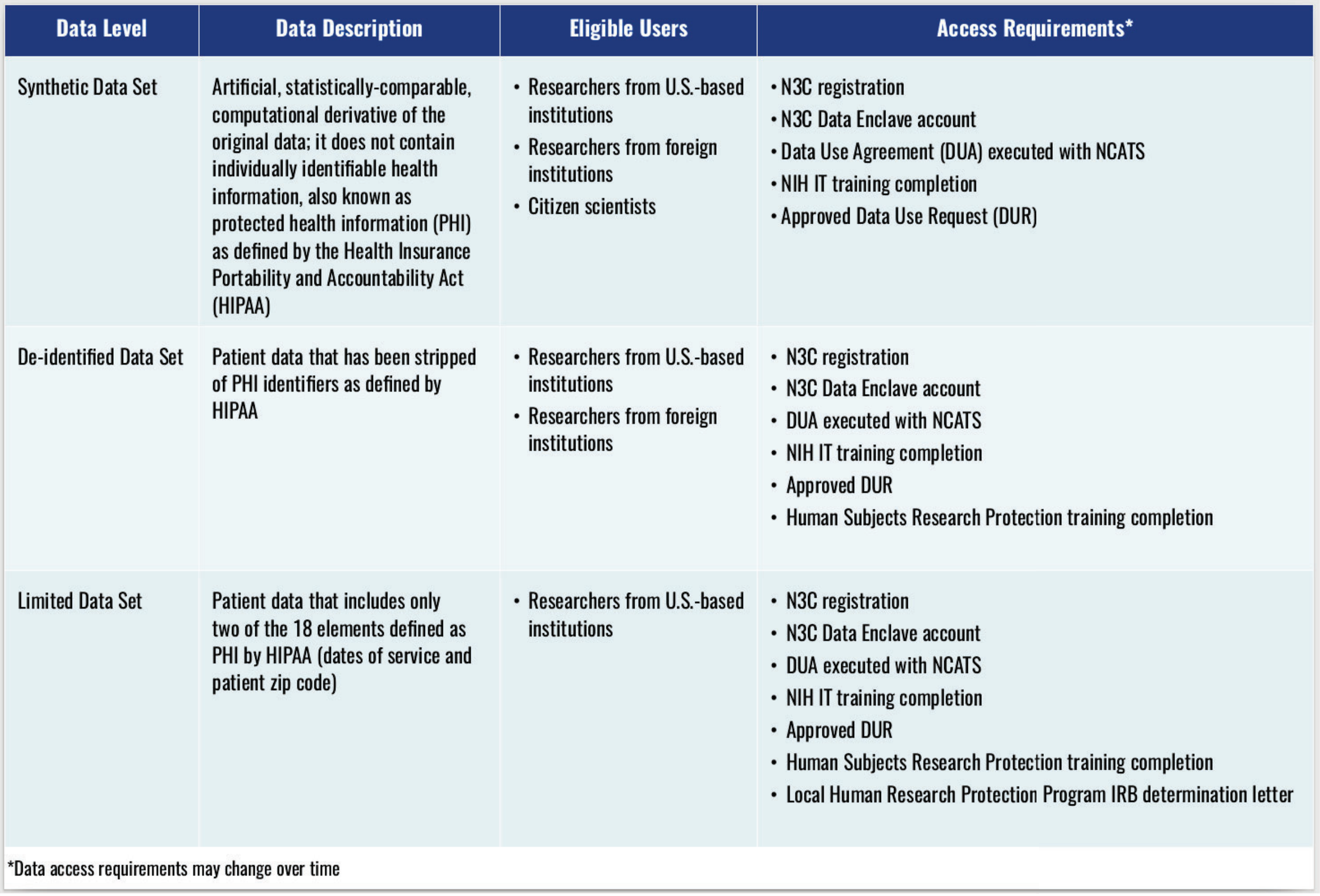
N3C is designed to turn data into knowledge during the COVID-19 pandemic, especially data analytics and statistics that require a large amount of data which could be adapted to other medical conditions [22]. N3C Patient data is kept in the Palantir platform (Palantir Foundry), which resides in Amazon Web Services (AWS) GovCloud, a service provided by Amazon Web Services designed to host sensitive data and regulated workloads, with multiple safeguards including role-based access controls, full system log entries, granular host and network level logging, robust end-to-end encryption (SSL/TLS, authentication, white-listing mechanisms), and comprehensive auditing of all data processing and access within the cloud platform. Users can only analyze data within the platform (Figure 2), with multiple tools for statistical, analytical, and machine learning.

Figure 2. N3C Data Enclave (from Reference [22])



N3C data will only be used for clinical and translational research and public health surveillance of COVID-19. In order to gain access to the N3C Data Enclave, a data use agreement (DUA) between NCATS and the institution has to be signed. Data provision to the N3C Data Enclave is performed under a Data Transfer Agreement (DTA), but it is not required to have A DTA to establish a DUA. As of December 12th, 2020, seventy four institutions had executed a DTA (https://ncats.nih.gov/n3c/resources/data-contribution/data-transfer-agreement-signatories), and 134 were DUA signatories (https://covid.cd2h.org/duas). Access to the data is free to DUA signatories, but data access within the enclave requires an approved data use request (DUR) by the N3C Data Access Committee, effective for one year after NCATS grants access. Three tiers of data are available (Figure 3).

Figure 3. NCATS access requirements by data level (from N3C Data Overview, https://ncats.nih.gov/n3c/about/data-overview#access-requirements)

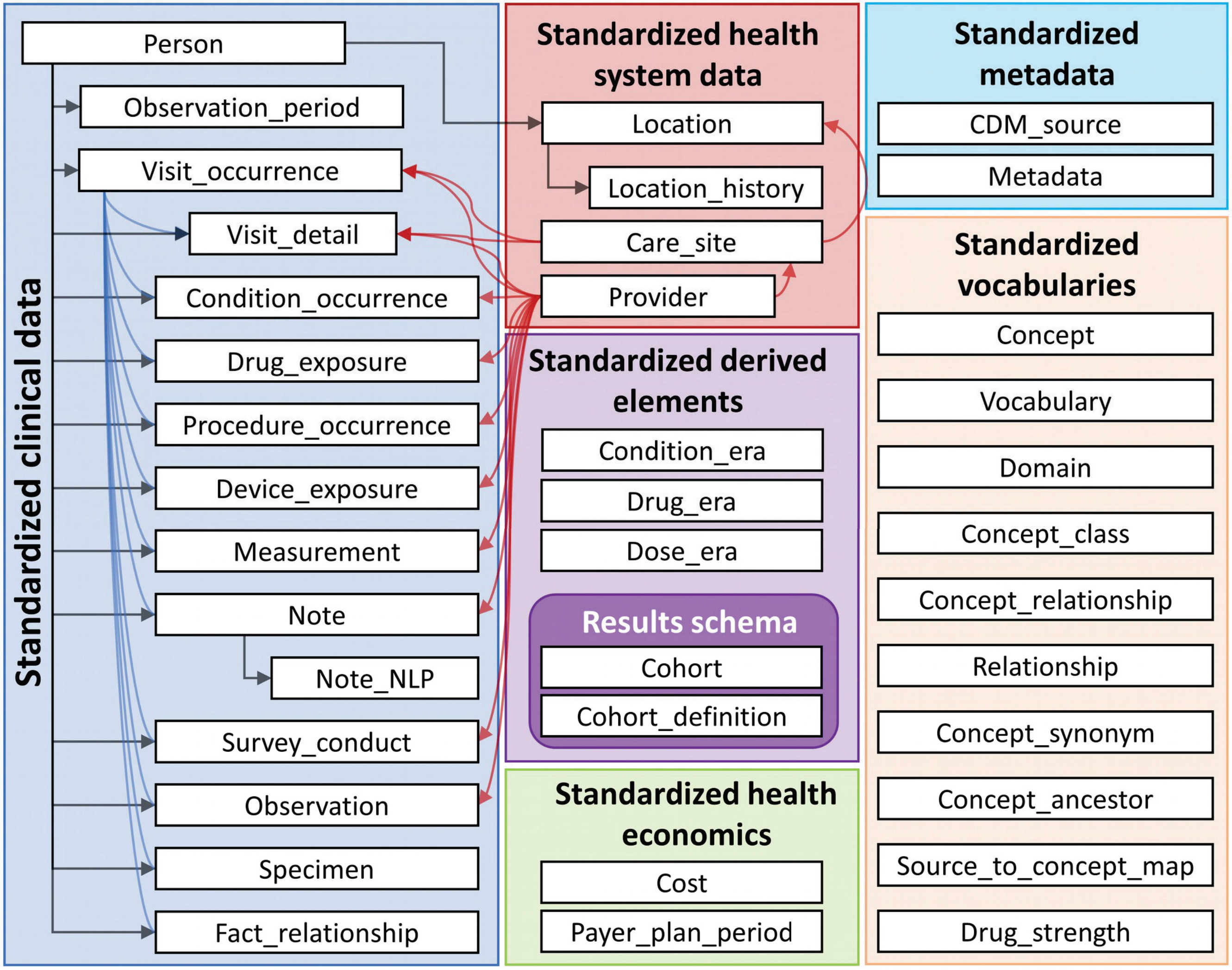


Data Level	Data Description	Eligible Users	Access Requirements*
Synthetic Data Set	Artificial, statistically-comparable, computational derivative of the original data; it does not contain individually identifiable health information, also known as protected health information (PHI) as defined by the Health Insurance Portability and Accountability Act (HIPAA)	• Researchers from U.S.-based institutions • Researchers from foreign institutions • Citizen scientists	• N3C registration • N3C Data Enclave account • Data Use Agreement (DUA) executed with NCATS • NIH IT training completion • Approved Data Use Request (DUR)
De-identified Data Set	Patient data that has been stripped of PHI identifiers as defined by HIPAA	• Researchers from U.S.-based institutions • Researchers from foreign institutions	• N3C registration • N3C Data Enclave account • DUA executed with NCATS • NIH IT training completion • Approved DUR • Human Subjects Research Protection training completion
Limited Data Set	Patient data that includes only two of the 18 elements defined as PHI by HIPAA (dates of service and patient zip code)	• Researchers from U.S.-based institutions	• N3C registration • N3C Data Enclave account • DUA executed with NCATS • NIH IT training completion • Approved DUR • Human Subjects Research Protection training completion • Local Human Research Protection Program IRB determination letter

*Data access requirements may change over time

For the purposes of extraction, transformation, and loading of data from the research networks, N3C uses the Observational Health Data Sciences and Informatics (OHDSI) CDM [26] , which allows N3C to harmonize patient data from diverse sources into a common data standard, and develop a standardized analytic to be executed on the data (Figure 4). The common data model is well documented, and improves reproducible analysis. N3C provides documentation about the CDM used in the enclave that is freely available [27].

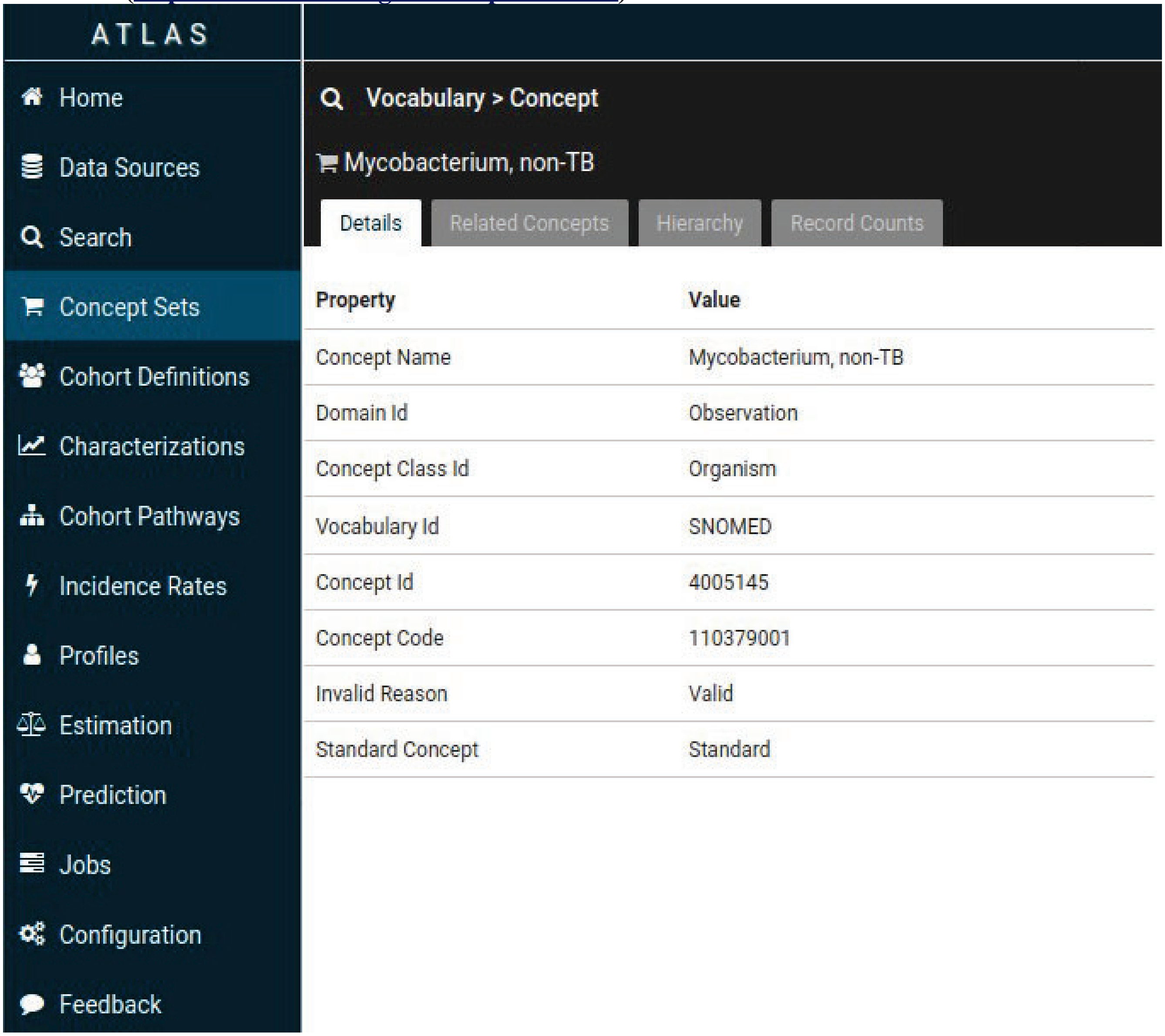
Figure 4. Overview of all tables in the Observational Health Data Sciences and Informatics (OHDSI) common data model version 6.0 (from Reference [26])



Methods

In order to study our hypothesis, access to the N3C Data Enclave was gained after establishing a DUA and DTA between N3C and the Medical College of Wisconsin (MCW) Clinical and Translational Science Institute (CTSI, https://ctsi.mcw.edu) on 10/1/2020, which required submitting a brief outline of the intended use of data within the N3C Data Enclave, mirroring the information provided in the N3C Data Enclave DUR submission site. After registration to N3C (https://labs.cd2h.org/registration/), completion of NIH Information Security and Information Management Training (https://irtsectraining.nih.gov/public.aspx), and uploading proof of completion of human subjects research training requirements (CITI program, https://www.citiprogram.org), we submitted a DUR to access de-identified patient data, as defined by HIPAA and outlined in Figure 3. The DUR was approved by the N3C Data Access Committee under the project identification RP-1C6E5B (Impact of COVID-19 infection in patients with pulmonary non-tuberculous Mycobacterium (NTM) infection: A cohort study) on 10/27/2020. We planned to identify patients with PNTM and COVID-19 infection within the N3C Data Enclave based on computational phenotypes created by using the OHDSI collaborative Vocabulary Repository (ATHENA) [28], and ATLAS, an open source application developed as a part of OHDSI intended to provide a unified interface to patient level data and analytics [29]. N3C Data Enclave includes clinical data from patients who meet criteria in the N3C COVID-19 phenotype [30] from sites across the United States dating back to January 2018 [22]. Using ATLAS (Figure 5), a case of PNTM infection was defined as a patient in the `condition_occcurrence` table with the concept name "Mycobacterium, non-TB", concept Id 4005145 (Figure 5).

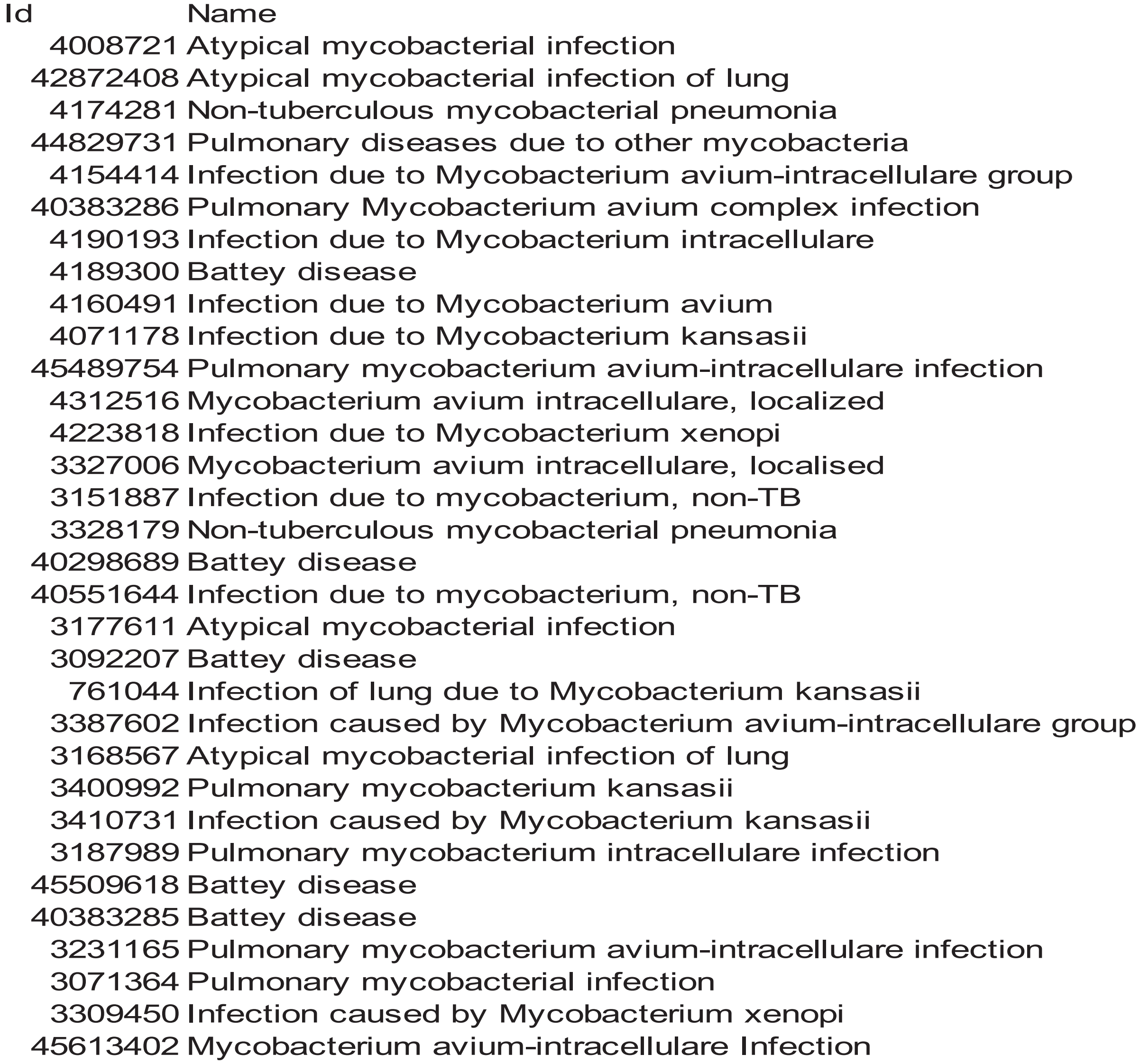
Figure 5. Concept set for "Mycobacterium, non TB", as extracted from vocabulary search in ATLAS [29] (https://atlas.ohdsi.org/#/concept/4005145)



Property	Value
Concept Name	Mycobacterium, non-TB
Domain Id	Observation
Concept Class Id	Organism
Vocabulary Id	SNOMED
Concept Id	4005145
Concept Code	110379001
Invalid Reason	Valid
Standard Concept	Standard

Manual exploration of its related concepts to create a list of search terms relevant to PNTM infection was made (Figure 6), as defined by the N3C COVID-19 Clinical Data Warehouse Data Dictionary, which is based on OMOP CDM specifications [27].

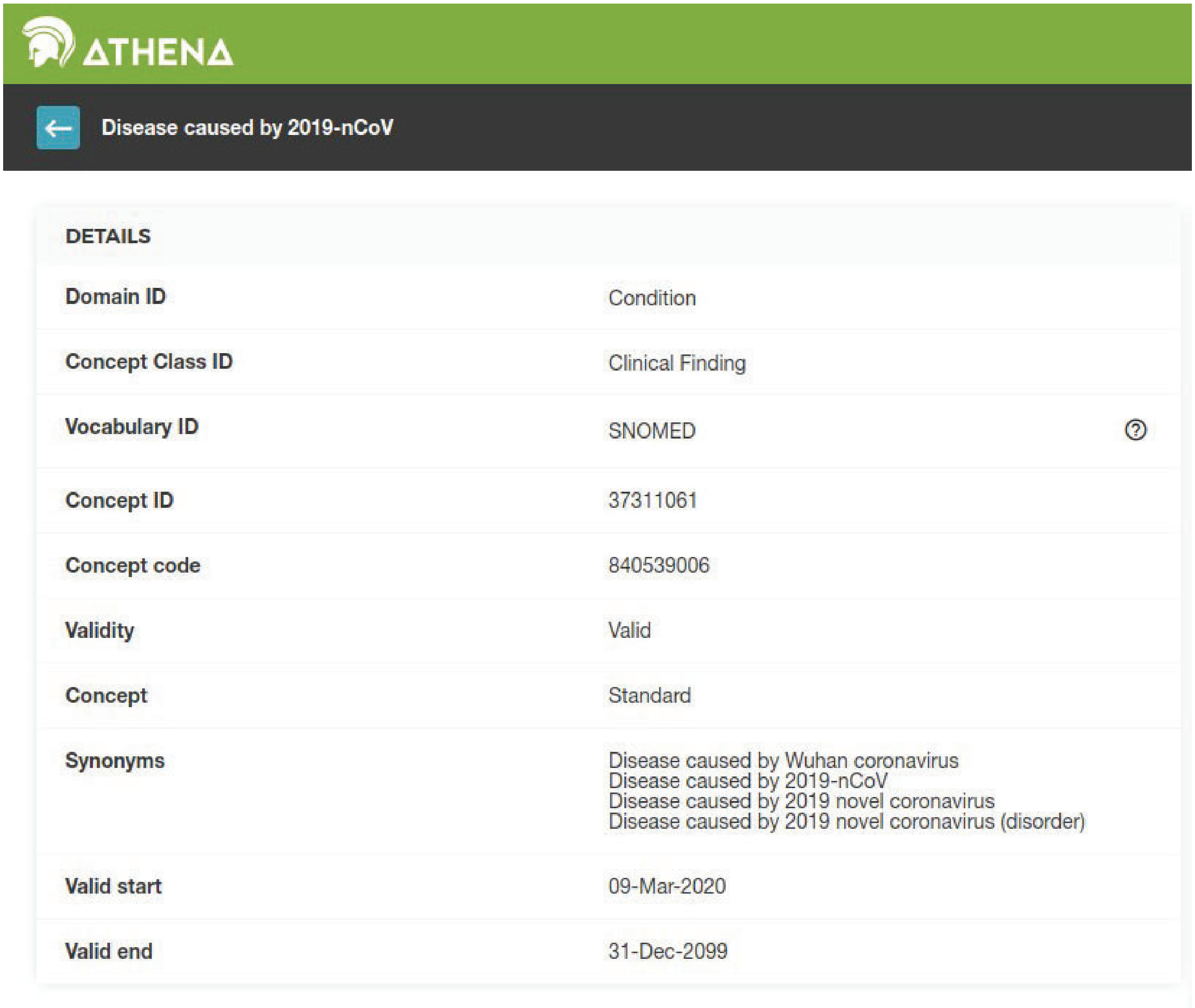
Figure 6. List of concepts used to define the pulmonary non-tuberculous *Mycobacterium* (PNTM) infection phenotype.



Id	Name
4008721	Atypical mycobacterial infection
42872408	Atypical mycobacterial infection of lung
4174281	Non-tuberculous mycobacterial pneumonia
44829731	Pulmonary diseases due to other mycobacteria
4154414	Infection due to Mycobacterium avium-intracellulare group
40383286	Pulmonary Mycobacterium avium complex infection
4190193	Infection due to Mycobacterium intracellulare
4189300	Battey disease
4160491	Infection due to Mycobacterium avium
4071178	Infection due to Mycobacterium kansasii
45489754	Pulmonary mycobacterium avium-intracellulare infection
4312516	Mycobacterium avium intracellulare, localized
4223818	Infection due to Mycobacterium xenopi
3327006	Mycobacterium avium intracellulare, localised
3151887	Infection due to mycobacterium, non-TB
3328179	Non-tuberculous mycobacterial pneumonia
40298689	Battey disease
40551644	Infection due to mycobacterium, non-TB
3177611	Atypical mycobacterial infection
3092207	Battey disease
761044	Infection of lung due to Mycobacterium kansasii
3387602	Infection caused by Mycobacterium avium-intracellulare group
3168567	Atypical mycobacterial infection of lung
3400992	Pulmonary mycobacterium kansasii
3410731	Infection caused by Mycobacterium kansasii
3187989	Pulmonary mycobacterium intracellulare infection
45509618	Battey disease
40383285	Battey disease
3231165	Pulmonary mycobacterium avium-intracellulare infection
3071364	Pulmonary mycobacterial infection
3309450	Infection caused by Mycobacterium xenopi
45613402	Mycobacterium avium-intracellulare Infection

Patients with COVID-19 infection were identified using a narrower COVID-19 computational phenotype which differs from the N3C COVID-19 phenotype definition [30]. This decision was based in the fact that the N3C phenotype definition captures laboratory-confirmed, suspected, and possible cases of COVID-19 [22]. Our definition focused in identify COVID-19-related visits, a more stringent phenotype, in order to simplify analysis. Based on ATHENA, COVID-19 patients were identified as entries in the `condition_occcurrence` table with the concept name “Disease caused by 2019-nCoV”, concept Id 37311061 (Figure 7).

Figure 7. Concept set for "Disease caused by 2019-nCoV", as extracted from vocabulary search in ATHENA [28] (https://athena.ohdsi.org/search-terms/terms/37311061)



Disease caused by 2019-nCoV

DETAILS	
Domain ID	Condition
Concept Class ID	Clinical Finding
Vocabulary ID	SNOMED ⓘ
Concept ID	37311061
Concept code	840539006
Validity	Valid
Concept	Standard
Synonyms	Disease caused by Wuhan coronavirus Disease caused by 2019-nCoV Disease caused by 2019 novel coronavirus Disease caused by 2019 novel coronavirus (disorder)
Valid start	09-Mar-2020
Valid end	31-Dec-2099

After identifying these two cohort, we performed multiple data transformations to describe their demographic characteristics; and study whether there are statistically significant differences between patients afflicted with COVID-19 infection in regards to the presence or absence of PNTM infection, by using multiple analytical tools available in the N3C Data Enclave analytical platform, Palantir Foundry. Methods in this study protocol and reporting follow recommendations by the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement [31].

Results

As of December 9, 2020, the N3C Data Enclave contained data of 2,635,160 patients, with 372,716 patients having COVID-19 infection, based on N3C phenotype, from 36 participating sites. Code Workbook, a tool to create, edit, and extend Palantir Foundry pipelines with code or point-and-click transforms in a graph environment, was created to write an SQL command to identify PNTM patients based on our phenotype definition, by creating the table `pntm_phenotype`, followed by creating the table `pntm_patients` with unique patient identifier (`person_id`), a foreign key from the `person` table, as defined by the OMOP CDM (Table 1). The SQL variant supported in Code Workbook is Spark SQL. De-identified patient data is located in the N3C Data Enclave at the OMOP Safe Harbor dataset folder.

Table 1. Structured query language commands to identify unique patients with pulmonary non-tuberculous *Mycobacterium* (PNTM) infection phenotype

```
pntm_phenotype
```

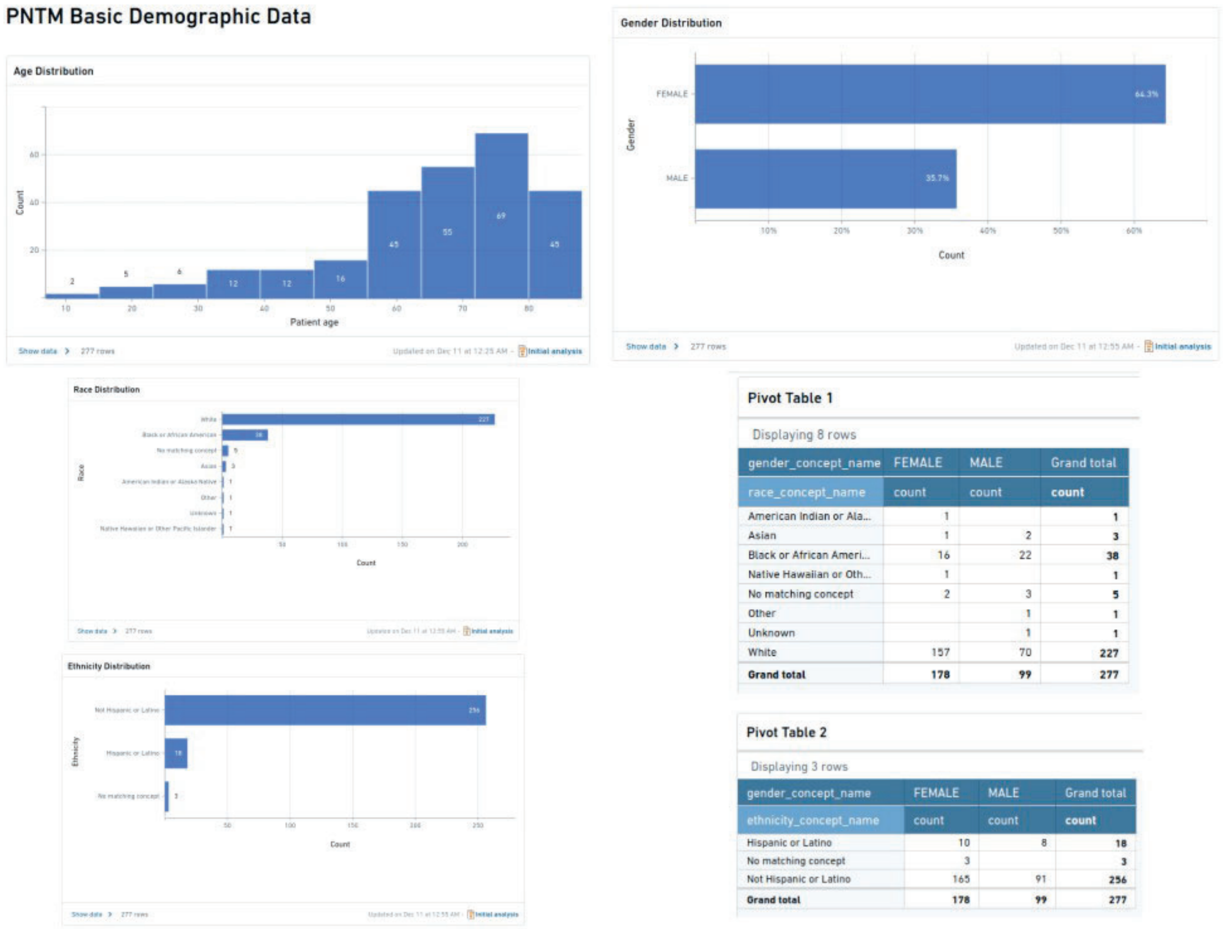
```
SELECT *
FROM condition_occurrence
WHERE condition_occurrence.condition_concept_id IN ('4008721','42872408',
'4174281', '44829731', '4154414', '40383286', '4190193', '4189300', '4160491',
'4005145', '4071178', '45489754', '4312516', '4223818', '3327006', '3151887',
'3328179', '40298689', '40551644', '3177611', '3092207', '761044', '3387602',
'3168567', '3400992', '3410731', '3187989', '45509618', '40383285', '3231165',
'3071364', '3309450', '45613402')
```

```
pntm_patients
```

```
SELECT DISTINCT person_id
FROM pntm_phenotype
```

Using this computational phenotype definition, we identified 277 patients, with PNTM infection. After removing rows without data, the cohort has a female predominance (178/277, 64.3%), and an age distribution skewed to advanced age (mean: 65.16, standard deviation: 16.15). The majority of patients were white (227/277, 81.9%), and not of Hispanic ethnicity (256/277, 92.8%). For comparison, basic demographic data from all patients in the N3C Data Enclave (2,635,160 patients) were mostly female (1,465,805, 55.6%), white (67.4%), and not of Hispanic ethnicity (2,013,532, 76.4%). This demographic data, and associated dashboard, were built within Contour, a tool available in N3C Data Enclave to analyze large data sets with filters, joins, and visualizations (Figure 8).

Figure 8. Basic demographic dashboard for patients with pulmonary non-tuberculous *Mycobacterium* (PNTM) infection, based on report exported from Contour.



Pivot Table 1

Displaying 8 rows

gender_concept_name	FEMALE	MALE	Grand total
race_concept_name	count	count	count
American Indian or Ala...	1		1
Asian	1	2	3
Black or African Ameri...	16	22	38
Native Hawaiian or Oth...	1		1
No matching concept	2	3	5
Other		1	1
Unknown		1	1
White	157	70	227
Grand total	**178**	**99**	**277**

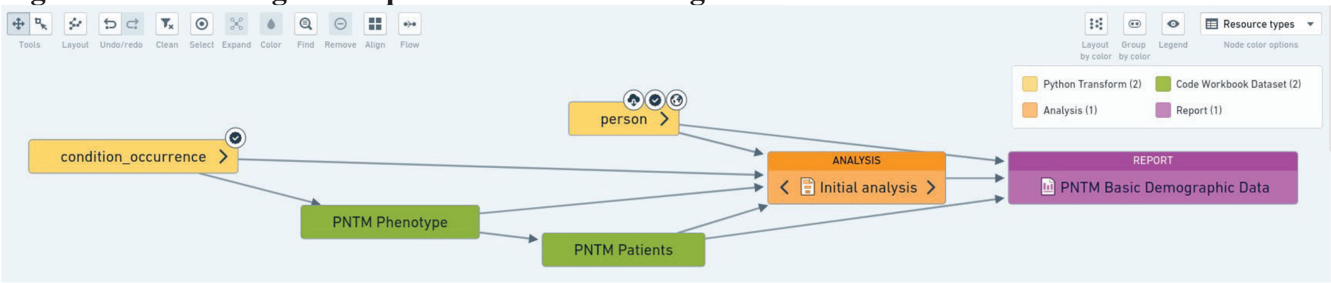
Pivot Table 2

Displaying 3 rows

gender_concept_name	FEMALE	MALE	Grand total
ethnicity_concept_name	count	count	count
Hispanic or Latino	10	8	18
No matching concept	3		3
Not Hispanic or Latino	165	91	256
Grand total	**178**	**99**	**277**

The results can be exported as new data sets, and interactive reports can be created and shared with other team participants within the N3C Data Enclave. A data lineage graph showing data sets pipelines used in this report build was created in Data Lineage, a visual tool in N3C Data Enclave to explore data pipelines (Figure 9).

Figure 9. Data lineage for report build shown in Figure 2.



Analysis from `pntm_phenotype` revealed 1320 entries from 10 distinct data partners, and 312 patients. Most common descriptors found included Non-tuberculous mycobacterial pneumonia (619/1320, 46.9%), Infection due to *Mycobacterium avium-intracellulare* group (345/1320, 26.1%), pulmonary *Mycobacterium avium* complex infection (167/1320, 12.7%), atypical mycobacterial infection (110/1320, 8.3%), infection due to *Mycobacterium avium* (58/1320, 4.4%), atypical mycobacterial infection of lung (8/1320, 0.6%), infection due to *Mycobacterium kansasii* (7/1320, 0.5%), and Battey disease (6/1320, 0.5%). One data partner (`data_partner_id` number 181) provided more than half of all entries (701/1320, 53.1%), a significant amount of unique patients (98/312, 31.4%), and all the entries identifying *M.kansasii*, which were traced back to two patients. Further analysis revealed one of them had both "infection due to *Mycobacterium kansasii*" and "Infection due to *Mycobacterium avium-intracellulare* group" reported, but it did not reveal a stop date for either condition, and *M.kansasii* was the first one reported. *Mycobacterium xenopi* infection (concept codes 4223818, 3309450) were not identified. To build our COVID-19 phenotype, we created two tables, `covid_condition` to identify COVID-19-related visits, and a unique list of patients, `covid_cases`.

Table 2. Structured query language commands to identify patients with visits related to COVID-19 infection

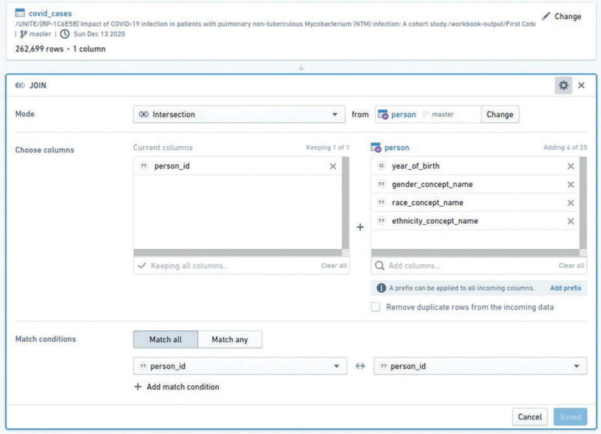
`covid_condition`

`SELECT *`
`FROM condition_occurrence`
`WHERE condition_concept_id = '37311061'`

`covid_cases`
`SELECT DISTINCT person_id FROM covid_cases`

Using Contour, a visual tool to create left, right, inner, and full joins as an alternative to create joins by writing SQL code in Code Workbook (Figure 10), we performed an inner join between `covid_cases` and person, and its result was exported as a new table, `covid_patient_demo`, in order to perform basic demographic descriptive analysis (Figure 11).

Figure 10. Join build dialog in Contour to create COVID-19 basic demographic data report.



covid_cases
/UNITE/[RP-1C6E5B] Impact of COVID-19 infection in patients with pulmonary non-tuberculous Mycobacterium (NTM) infection: A cohort study./workbook-output/First Code
| master | Sun Dec 13 2020
262,699 rows · 1 column

JOIN

Mode: Intersection from person master

Choose columns
Current columns — Keeping 1 of 1
person_id
Keeping all columns...

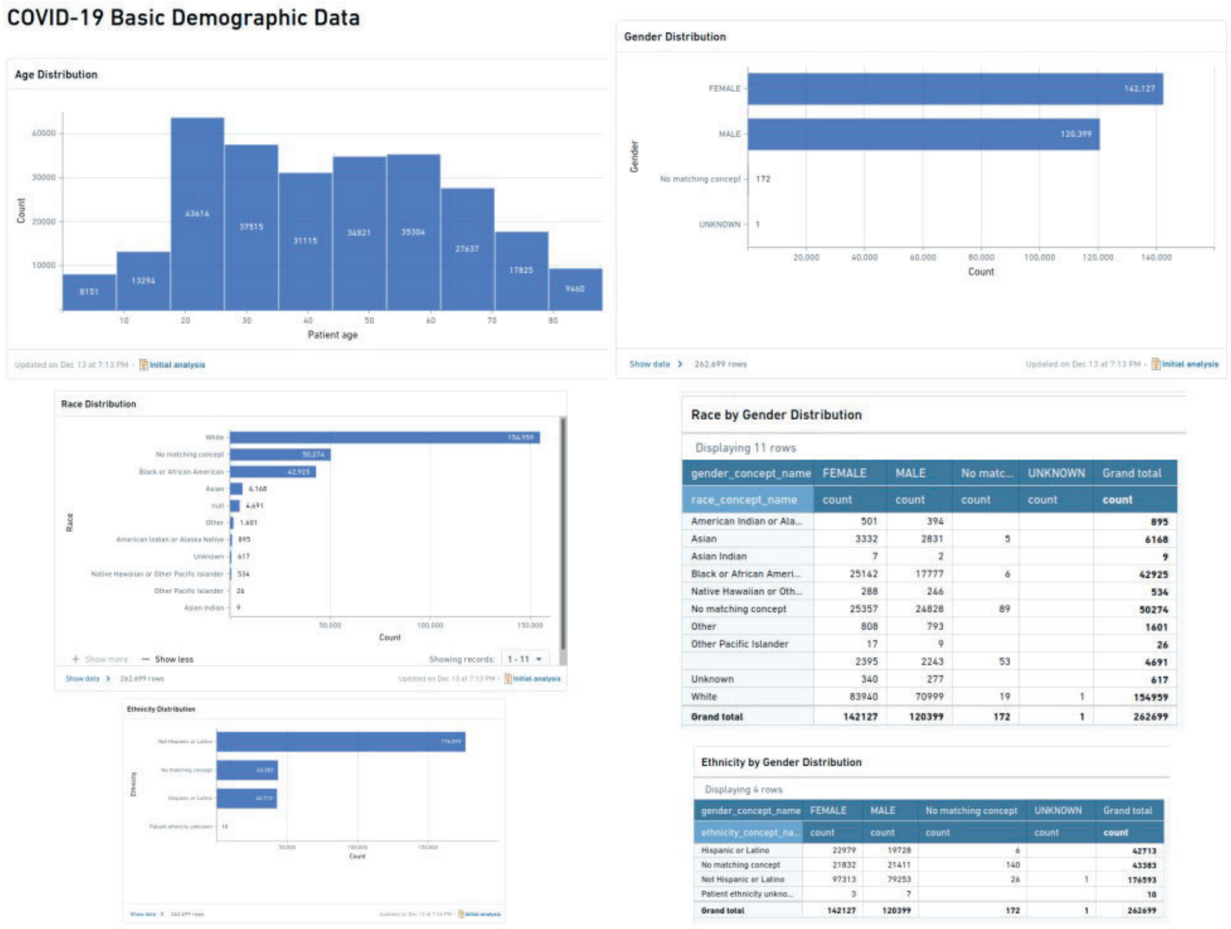
person — Adding 4 of 25
year_of_birth
gender_concept_name
race_concept_name
ethnicity_concept_name

A prefix can be applied to all incoming columns.

□ Remove duplicate rows from the incoming data

Match conditions: Match all
person_id ↔ person_id

Figure 11. Basic demographic dashboard for patients with COVID-19 infection.



Race by Gender Distribution

Displaying 11 rows

gender_concept_name	FEMALE	MALE	No matc...	UNKNOWN	Grand total
race_concept_name	count	count	count	count	count
American Indian or Ala...	501	394			895
Asian	3332	2831	5		6168
Asian Indian	7	2			9
Black or African Ameri...	25142	17777	6		42925
Native Hawaiian or Oth...	288	246			534
No matching concept	25357	24828	89		50274
Other	808	793			1601
Other Pacific Islander	17	9			26
	2395	2243	53		4691
Unknown	340	277			617
White	83940	70999	19	1	154959
Grand total	**142127**	**120399**	**172**	**1**	**262699**

Ethnicity by Gender Distribution

Displaying 4 rows

gender_concept_name	FEMALE	MALE	No matching concept	UNKNOWN	Grand total
ethnicity_concept_na...	count	count	count	count	count
Hispanic or Latino	22979	19728	6		42713
No matching concept	21832	21411	140		43383
Not Hispanic or Latino	97313	79253	26	1	176593
Patient ethnicity unkno...	3	7			10
Grand total	**142127**	**120399**	**172**	**1**	**262699**

With this tool, `covid_patient_demo` analysis revealed 262,699 patients, (mean age: 43.21, SD: 20.25), with female (54.1%), and white (58.9%) predominance. Hispanic ethnicity constituted 16.3% of this cohort. We identified 18 patients who have had COVID-19 with known PNTM infection by performing an inner join between `covid_patient_demo` and `ptnm_patients` in Contour. They were older (mean age: 60.27, SD: 17), and were mostly female (12/18, 66.7%), white (12/18, 66.7%), and not of Hispanic ethnicity (15/18, 83.3%). Two out of eighteen were found in the `death` table, which is available at OMOP Safe Harbor in the N3C Data Enclave. Further evaluation of the resulting join did not reveal a cause of death, and one entry had a date of birth two months after this analysis was performed, so these entries were not attributed to COVID-19 infection. To build the two groups for comparison (COVID-19 cases based on the presence or absence of PNTM infection), Contour was used to remove the 18 rows by using its Filter tool, and a new table, `covid_no_ntm_patients` was built. Table 3 illustrates the demographic differences in patients with COVID-19 based on PNTM status. Statistical significance was set at α = 0.05 for 2-tailed tests by performing unpaired t-test for comparison of means, and 2-sample z-test to compare sample proportions.

Table 3. Demographic characteristics in patients with COVID-19 based on pulmonary non-tuberculous *Mycobacterium* (PNTM) infection status.

	COVID-19		C+/NTM+		C+/NTM-		
Demographics	**Total**	**Percentage**		**Percentage**		**Percentage**	***p* value**
Number	262699		18		262681		
Age (mean, SD)	43.21, 20.25		60.27, 17		43.21, 20.25		0.0004
Gender							
Female	142127	0.54	12	0.67	142115	0.54	0.951
Male	120399	0.46	6	0.33	120393	0.46	
No matching concept	172	0.00	0	0.00	172	0.00	
Unknown	1	0.00	0	0.00	1	0.00	
Race							
White	154959	0.59	12	0.67	154947	0.59	0.9711
No matching concept	50274	0.19	0	0.00	50274	0.19	
Black or African American	42925	0.16	5	0.28	42920	0.16	0.9465
Asian	6168	0.02	0	0.00	6168	0.02	
null	4691	0.02	0	0.00	4691	0.02	
Other	1601	0.01	0	0.00	1601	0.01	
American Indian or Alaska Native	895	0.00	1	0.06	894	0.00	0.9323
Unknown	617	0.00	0	0.00	617	0.00	
Native Hawaiian or Other Pacific Islander	534	0.00	0	0.00	534	0.00	
Other Pacific Islander	26	0.00	0	0.00	26	0.00	
Ethnicity							
Not Hispanic or Latino	176593	0.67	15	0.83	176578	0.67	0.9395
No matching concept	43383	0.17	0	0.00	43383	0.17	
Hispanic or Latino	42713	0.16	3	0.17	42710	0.16	
Patient ethnicity unknown	10	0.00	0	0.00	10	0.00	

To identify whether these patients required inpatient management associated to COVID-19, a data set to capture inpatient visits, `inpatient_and_er_visit`, was created.

Table 4. Structured query language commands to identify patients with inpatient and emergency room visits.

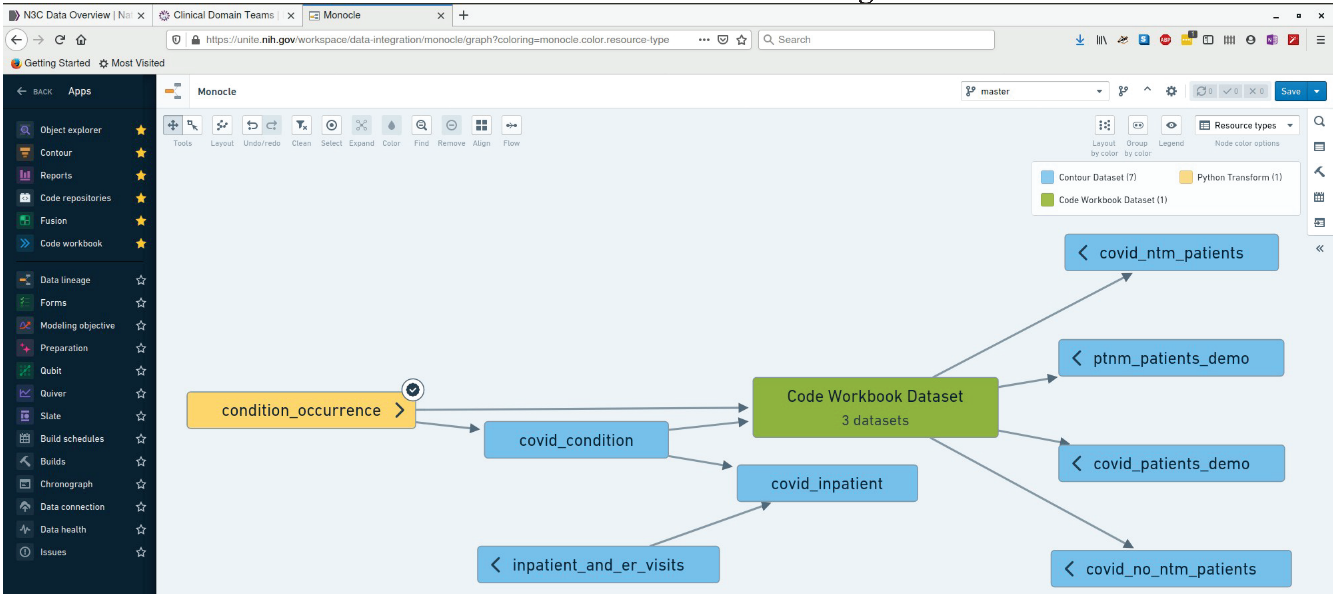
```
inpatient_and_er_visit
```

```
SELECT *
FROM visit_occurrence v
WHERE v.visit_concept_id IN ('262','8717','9201', '9203', '32037', '581379')

-- 262 Emergency Room and Inpatient Visit
-- 8717 Inpatient Hospital
-- 9201 Inpatient Visit
-- 9203 Emergency Room Visit
-- 32037 Intensive Care
-- 581379 Inpatient Critical Care Facility
```

To identify inpatient visits related to COVID-19 infection, a join between `inpatient_and_er_visit` and `covid_condition`, with a visit start date on or after January 1st, 2020, was created in Contour, and saved in a new table, `covid_inpatient`. When performing an inner join with our PNTM infection cohort data set, we identified three patients with PNTM infection requiring inpatient admission related to COVID-19 infection (mean age: 38, SD: 15.12). Two of them were female, and all were white. No deaths were noted in this group, after performing an inner joint with the data set `death`. A visual analysis of the data sets used in this analysis using Monocle, another analytical tool available in Palantir Enclave, is shown (Figure 12). A quick lookup allowed detection of an outdated data set, which was then rebuilt in Contour.

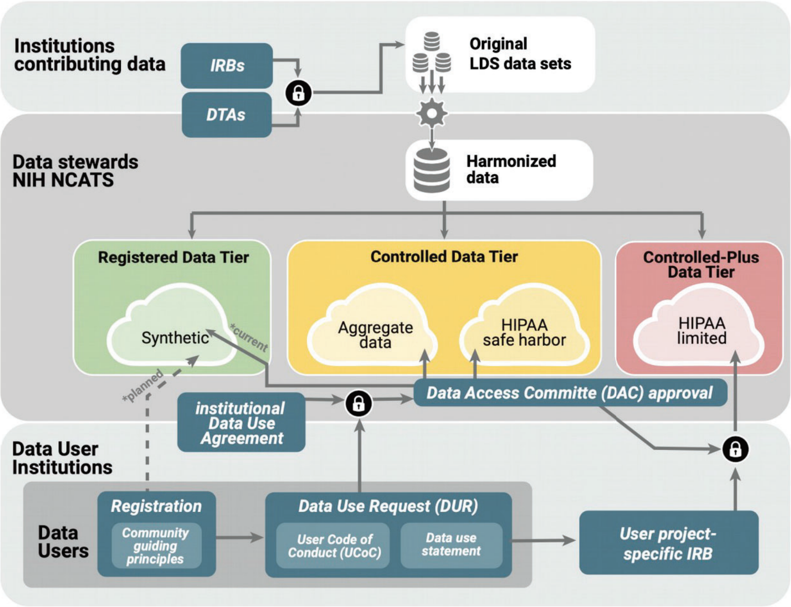
Figure 12. Data pipeline built for this study, using Monocle, a tool within Palantir Enclave. A list of tools available in the enclave are seen at the left side of the figure.



Discussion

From an informatics standpoint, the goal of gaining access to the N3C Data Enclave was completely fulfilled within the amount of time stipulated to complete this capstone project. In contrast to the straightforward process to gain access to the enclave, analyzing the data in the enclave came with a steep curve (Figure 13).

Figure 13. N3C Data Enclave regulatory steps and user access (from Reference [22])



As of this writing, documentation about Palantir Foundry, including implementation of popular tools in data science like R, Python, and SparkSQL, are out of reach for people without a DUA and a DUR. However, Contour proved to be a very flexible tool within the enclave, especially with its embedded data visualization and reporting capabilities. Even though other tools were used for this analysis (Data Lineage, Code Workbook), most if not all data sets created with the SQL commands in Code Workbook could have been built within Contour. For simple data set queries and new data set builds, as described in some of our builds and queries, this tool will become a very popular choice within the enclave, especially if the user is not comfortable with available data science tools. The analysis of large data to solve complex questions is one of the main reasons N3C was built, and it is difficult to conceive Contour being enough for more complex analyses. One crucial aspect of the project is the absolute need to familiarize with the CDM, but one of the main strengths of N3C is the use of the OHDSI CDM, which has excellent documentation, and an active community of developers. N3C Data Enclave can integrate other tools to analyze OMOP data, including OHDS ATLAS, LOINC2HPO, NCATS Biomedical Data Translator, and Leaf [22].

Multiple limitations are present in our study. Our definition of COVID-19 differs from the phenotype described by the enclave, but we chose a more stringent definition of a positive COVID-19 case to simplify the analysis. Even though the use of de-identified data obviated the need for local institutional review board (IRB) oversight, the loss of PHI makes date and time analysis difficult, and geographical information is lost, which has been shown to be of importance in PNTM infection [32]. The computational definition for PNTM infection was limited to the more relevant NTM species causing pulmonary infection, and it is likely other species will be reported with more frequency in the near future [19]. No deaths were reported in patients with PNTM infection and COVID-19 infection, but the small number of cases makes extrapolation to the overall PNTM infection population problematic, especially when the source of patient data is confined to the United States healthcare system, and furthermore, to the relatively reduced number of institutions that have signed a DTA with NCATS. Additional statistical analysis of associated clinical conditions like the impact of body mass index, race and ethnicity, and the presence of comorbidities and/or confounders like COPD, bronchiectasis, and cystic fibrosis, would be the logical step to follow. The project is expected to continue for the remainder of the DUR agreement, and some of our goals is to continue the proposed analyses, followed by publication of our findings in a peer-reviewed journal. N3C has published guidelines for publication of documents using N3C data, including analysis reports, data, resources, abstracts, presentations, preprints, and publications [33].

We would like to consider refinements to our PNTM infection computational phenotype, including detection of patients by querying drug exposure, and observation tables, as defined by the CDM, and to share this information with the N3C community for additional feedback. N3C has created clinical domains teams to enable researchers with shared interests to analyze data within the enclave and collaborate more efficiently in a team science environment (https://covid.cd2h.org/domain-teams), by weekly discussions via video conference, Slack, and Google Groups. However, finding a "home" for this project within current clinical domain teams was not feasible, with the consequent loss of the social and collaborative component embedded in the N3C Data Enclave design. Interestingly enough, an infectious diseases research team at OHSU reached out to join efforts in analyzing this group of patients, after DUR data that is available to the public in the N3C website. We would like to develop a JavaScript Object Notation (JSON, a lightweight data-interchange format) PNTM infection cohort definition with ATLAS, which can be used with a Python script available in N3C Knowledge Base, in order to increase standardization and reproducibility of data analysis.

Summary and Conclusions

In conclusion, we were able to use patient data from N3C Data Enclave data sets to perform a proof-of-concept study to determine the impact of COVID-19 infection in patients with PNTM infection. We developed computational phenotypes for both conditions by using freely available tools before we gained access to the data. Multiple tools within the analytical platform in N3C were used to identify patients with PNTM infection, who tended to be female, older, and mostly of white race when compared to the overall N3C cohort. The majority of these infections were associated to *Mycobacterium-avium intracellulare* complex. There was a statistically significant difference in regards to age between patients with PNTM and COVID-19 infection when compared to COVID-19 patients without PNTM infection. Three patients with PNTM infection required inpatient admission, and no fatalities related to COVID-19 were present. Additional research work to evaluate whether there are differences in hospital course, and to quantify the effect of other medical conditions that are commonly present in patients with PNTM infection, and are known risk factors for severe COVID-19 infection, are required. N3C provides multiple options to researchers to study patients with COVID-19 infection, including open and propriertary solutions, which will be beneficial when studying the effects of COVID-19 on special populations. Given the use of OHDSI CDM as the data model for loading and transforming data, it is expected N3C will continue to provide information about COVID-19 infection, and become a model for organizing data to facilitate large-scale observational studies of novel infectious diseases.

Funding

Source of funding: None

Bibliography

1. Zhu N, Zhang D, Wang W, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med*. 2020;382(8):727-733. doi:10.1056/NEJMoa2001017
2. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis*. 2020;20(5):533-534. doi:10.1016/S1473-3099(20)30120-1
3. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet (London, England)*. 2020;395(10223):497-506. doi:10.1016/S0140-6736(20)30183-5

4. Wang D, Hu B, Hu C, et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA*. 2020;323(11):1061-1069. doi:10.1001/jama.2020.1585

5. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA*. 2020;323(13):1239-1242. doi:10.1001/jama.2020.2648

6. Wu C, Chen X, Cai Y, et al. Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China. *JAMA Intern Med*. 2020;180(7):934-943. doi:10.1001/jamainternmed.2020.0994

7. Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet (London, England)*. 2020;395(10229):1054-1062. doi:10.1016/S0140-6736(20)30566-3

8. Gandhi RT, Lynch JB, Del Rio C. Mild or Moderate Covid-19. *N Engl J Med*. 2020;383(18):1757-1766. doi:10.1056/NEJMcp2009249

9. Alberici F, Delbarba E, Manenti C, et al. A single center observational study of the clinical characteristics and short-term outcome of 20 kidney transplant patients admitted for SARS-CoV2 pneumonia. *Kidney Int*. 2020;97(6):1083-1088. doi:10.1016/j.kint.2020.04.002

10. Vizcarra P, Pérez-Elías MJ, Quereda C, et al. Description of COVID-19 in HIV-infected individuals: a single-centre, prospective cohort. *lancet HIV*. 2020;7(8):e554-e564. doi:10.1016/S2352-3018(20)30164-8

11. Montagud-Marrahi E, Cofan F, Torregrosa J-V, et al. Preliminary data on outcomes of SARS-CoV-2 infection in a Spanish single center cohort of kidney recipients. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2020;20(10):2958-2959. doi:10.1111/ajt.15970

12. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby J V, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc*. 2014;21(4):578-582. doi:10.1136/amiajnl-2014-002747

13. Visweswaran S, Becich MJ, D'Itri VS, et al. Accrual to Clinical Trials (ACT): A Clinical and Translational Science Award Consortium Network. *JAMIA open*. 2018;1(2):147-152. doi:10.1093/jamiaopen/ooy033

14. Topaloglu U, Palchuk MB. Using a Federated Network of Real-World Data to Optimize Clinical Trials Operations. *JCO Clin cancer informatics*. 2018;2:1-10. doi:10.1200/CCI.17.00067

15. Brat GA, Weber GM, Gehlenborg N, et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. *NPJ Digit Med*. 2020;3:109. doi:10.1038/s41746-020-00308-0

16. Hadi YB, Naqvi SFZ, Kupec JT, Sarwari AR. Characteristics and outcomes of COVID-19 in patients with HIV: a multicentre research network study. *AIDS*. 2020;34(13):F3-F8. doi:10.1097/QAD.0000000000002666

17. Ranabothu S, Kanduri SR, Nalleballe K, Cheungpasitporn W, Onteddu S, Kovvuru K. Outcomes of COVID-19 in Solid Organ Transplants. *Cureus*. 2020;12(11):e11344. doi:10.7759/cureus.11344

18. Harrison SL, Fazio-Eynullayeva E, Lane DA, Underhill P, Lip GYH. Comorbidities associated with mortality in 31,461 adults with COVID-19 in the United States: A federated electronic medical record analysis. *PLoS Med*. 2020;17(9):e1003321. doi:10.1371/journal.pmed.1003321

19. Tortoli E, Fedrizzi T, Meehan CJ, et al. The new phylogeny of the genus Mycobacterium: The old and the news. *Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis*. 2017;56:19-25. doi:10.1016/j.meegid.2017.10.013

20. Winthrop KL, Marras TK, Adjemian J, Zhang H, Wang P, Zhang Q. Incidence and Prevalence of Nontuberculous Mycobacterial Lung Disease in a Large U.S. Managed Care Health Plan, 2008-2015. *Ann Am Thorac Soc*. 2020;17(2):178-185. doi:10.1513/AnnalsATS.201804-236OC
21. Daley CL, Iaccarino JM, Lange C, et al. Treatment of Nontuberculous Mycobacterial Pulmonary Disease: An Official ATS/ERS/ESCMID/IDSA Clinical Practice Guideline. *Clin Infect Dis an Off Publ Infect Dis Soc Am*. 2020;71(4):905-913. doi:10.1093/cid/ciaa1125
22. Melissa H, Christopher C, Kenneth G. The National COVID Cohort Collaborative (N3C): Rationale, Design, Infrastructure, and Deployment. *J Am Med Inform Assoc*. August 2020. doi:10.1093/jamia/ocaa196
23. Stang PE, Ryan PB, Racoosin JA, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann Intern Med*. 2010;153(9):600-606. doi:10.7326/0003-4819-153-9-201011020-00010
24. National Institutes of Health (NIH). National Center for Advancing Translational Sciences (NCATS). National COVID Cohort Collaborative Data Enclave Repository. https://covid.cd2h.org/enclave. Published 2020. Accessed August 12, 2020.
25. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform*. 2015;216:574-578.
26. Observational Health Data Sciences and Informatics. *The Book of OHDSI*. Monee, IL; 2019. https://ohdsi.github.io/TheBookOfOhdsi/.
27. National Institutes of Health (NIH). National Center for Advancing Translational Sciences (NCATS). OMOP Common Data Specifications. https://ncats.nih.gov/files/OMOP_CDM_COVID.pdf. Published 2020. Accessed August 12, 2020.
28. Odysseus Data Services Inc. Athena – OHDSI Vocabularies Repository. https://athena.ohdsi.org/search-terms/start. Published 2020. Accessed August 12, 2020.
29. Observational Health Data Sciences and Informatics (OHDSI). ATLAS. https://atlas.ohdsi.org/#/home. Published 2020.
30. National COVID Cohort Collaborative. Latest Phenotype. https://github.com/National-COVID-Cohort-Collaborative/Phenotype_Data_Acquisition/wiki/Latest-Phenotype. Published 2020. Accessed August 12, 2020.
31. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med*. 2007;147(8):573-577. doi:10.7326/0003-4819-147-8-200710160-00010
32. Spaulding AB, Lai YL, Zelazny AM, et al. Geographic Distribution of Nontuberculous Mycobacterial Species Identified among Clinical Isolates in the United States, 2009-2013. *Ann Am Thorac Soc*. 2017;14(11):1655-1661. doi:10.1513/AnnalsATS.201611-860OC
33. N3C Consortium. Attribution and Publication Principles for N3C (National Covid Cohort Collaborative). doi:10.5281/zenodo.4000322