The National COVID Cohort Collaborative (N3C): Rationale, Design, Infrastructure, and Deployment

*Melissa A. Haendel; Oregon Health & Sciences University; Oregon State University, OR, USA
*Christopher G. Chute; Johns Hopkins University Baltimore, MD, USA
Please see attached supplemental files for masthead and contributing authors.

(Version 2020-07-14; IN PRESS IN JAMIA OPEN)

ABSTRACT

Objective

COVID-19 poses societal challenges that require expeditious data and knowledge sharing. Though organizational clinical data are abundant, these are largely inaccessible to outside researchers. Statistical, machine learning, and causal analyses are most successful with large-scale data beyond what is available in any given organization. Here, we introduce the National COVID Cohort Collaborative (N3C), an open science community focused on analyzing patient-level data from many centers.

Methods

The Clinical and Translational Science Award (CTSA) Program and scientific community created N3C to overcome technical, regulatory, policy, and governance barriers to sharing and harmonizing individual-level clinical data. We developed solutions to extract, aggregate, and harmonize data across organizations and data models, and created a secure data enclave to enable efficient, transparent, and reproducible collaborative analytics.

Organized in inclusive workstreams, in two months we created: legal agreements and governance for organizations and researchers; data extraction scripts to identify and ingest positive, negative, and possible COVID-19 cases; a data quality assurance and harmonization pipeline to create a single harmonized dataset; population of the secure data enclave with data, machine learning, and statistical analytics tools; dissemination mechanisms; and a synthetic data pilot to democratize data access.

Discussion

The N3C has demonstrated that a multi-site collaborative learning health network can overcome barriers to rapidly build a scalable infrastructure incorporating multi-organizational clinical data for COVID-19 analytics. We expect this effort to save lives by enabling rapid collaboration among clinicians, researchers, and data scientists to identify treatments and specialized care and thereby reduce the immediate and long-term impacts of COVID-19.

LAY SUMMARY

COVID-19 poses societal challenges that require expeditious data and knowledge sharing. Though medical records are abundant, they are largely inaccessible to outside researchers. Statistical, machine learning, and causal research are most successful with large datasets beyond what is available in any given organization. Here, we introduce the National COVID Cohort Collaborative (N3C), an open science community focused on analyzing patient-level data from many clinical centers to reveal patterns in COVID-19 patients. To create N3C, the community had to overcome technical, regulatory, policy, and governance barriers to sharing patient-level clinical data. In less than 2 months, we developed solutions to acquire and harmonize data across organizations and created a secure data environment to enable transparent and reproducible collaborative research. We expect the N3C to help save lives by enabling collaboration among clinicians, researchers, and data scientists to identify treatments and specialized care needs and thereby reduce the immediate and long-term impacts of COVID-19.

INTRODUCTION

Rationale

The severe acute respiratory syndrome-coronavirus 2 (SARS-CoV-2) has infected 12.6 million people—and the 2019 Novel Coronavirus Disease (COVID-19) has caused 562,000 deaths—worldwide as of July 11, 2020, according to the Johns Hopkins[1]. Scientists warn that recurrences are likely after the current initial pandemic, particularly if SARS-CoV-2 immunity wanes over time[2]. To curb this trajectory, in addition to public health measures to contain the virus as much as possible, it is crucial to gather large amounts of data in a comprehensive and unbiased fashion[3]. These data enable the global community to understand the natural history and complications of the disease, ultimately guiding approaches to effectively prevent infection and manage care for individuals with COVID-19.

Key challenges of a new pandemic disease include understanding pathophysiology and symptom progression over time; addressing biological, environmental, and socioeconomic risk and protective factors; identifying treatments; and rapidly building clinical decision support (CDS) and practice guidelines. The pandemic raises many difficult questions: Which drugs are most likely to benefit a given patient? What treatments, risk factors, and social determinants of health (SDoH) impact disease course and outcome? How do we develop, adapt, and deploy CDS to keep up with a dynamic pandemic? To address these questions, it is critical to analyze a high volume of reliable patient-level, accurately-attributed, nationally-representative data.

Currently, the research community's access to electronic health record (EHR) data is limited within given organizations or consortia of local and regional organizations. Research consortia such as Accrual to Clinical Trials Network (ACT)[4], National Patient-Centered Clinical Research Network (PCORnet)[5], Observational Health Data Sciences and Informatics (OHDSI)[6], FDA's Sentinel Initiative[7], TriNetX[8], and the recently established international Consortium for Characterization of COVID-19 by EHR (4CE)[9] support querying structured data across participating organizations using a common data model (CDM). These networks are a vital resource for responding to the COVID-19 crisis, revealing key patterns in the disease[9,10]. However, their distributed nature would greatly complicate certain types of analyses that require a centralized approach to enable timely analyses. Study questions and data queries that can be pre-specified, such as testing for associations between one or a group of comorbidities and laboratory results, are often answerable using federated networks. In contrast, centralized resources can greatly simplify implementation of iterative processes such as training deep learning algorithms and carrying out clustering for phenotype development. [11–14] A centralized resource also enables rapid integration with knowledge graphs and other translational knowledge and data sources to aid discovery, prioritization, and weighting of results. Federated machine learning algorithms will likely ultimately play important roles in allowing model training on distributed datasets [15–19]. While these methods show great promise, we have chosen not to pursue this approach at this time to avoid adding complexity to an already ambitious project. Creating a massive corpus of harmonized EHR data for analytics would support rapid collaboration and discovery, and also build upon the substantial resources (e.g. CDM-specific data quality tools, etc.) developed within the federated consortia.

The recent retractions in the Lancet[20] and the New England Journal of Medicine[21] have underscored the need for fully provenanced and reproducible EHR analyses as major policy decisions that can hinge on EHR results. Moreover, the pathway for obtaining permissions to reuse data must be clear and well-documented. The ideal data resources are FAIR (findable, accessible, interoperable, reusable), particularly in a pandemic where analyses must be fast, verifiable, and based on the latest data[22].

N3C Overview

The National COVID Cohort Collaborative (N3C; covid.cd2h.org) aims to aggregate and harmonize EHR data across clinical organizations in the United States (US), especially the Clinical and Translational Science Awards (CTSA) Program hubs that encompass more than 60 organizations and their partners[23]. In just two months, the N3C was built on a foundation of established, productive research communities and their existing resources. It comprises a collaborative network of more than 600 individuals and 100 organizations and is growing. N3C enables broad access and analytics of harmonized EHR data, demonstrating a novel approach for collaborative data sharing that could transcend current and future health emergencies. The primary features of N3C are: national collaboration and governance, regulatory strategies, COVID-19 cohort definitions via community-developed phenotypes, data harmonization across four CDMs, and development of a collaborative analytics platform to support deployment of novel algorithms of data aggregated from the US. The N3C supports community-driven, reproducible, and transparent analyses with COVID-19 data, promoting rapid dissemination of results and atomic attribution and demonstrating that open science can be effectively implemented on EHR data at scale.

N3C is built upon principles of partnership, inclusivity, transparency, reciprocity, accountability, and security:

- **Partnership**: N3C members are trusted partners committed to honoring the N3C Community Guiding Principles and User Code of Conduct.
- **Inclusivity**: N3C is open to any organization that wishes to contribute data. N3C also welcomes registered researchers who follow our governance processes, including citizen and community scientists, to access the data.
- **Transparency**: Open and reproducible research is the hallmark of N3C. Access to data is project-based. Descriptions of projects are posted and searchable to promote collaborations.
- **Reciprocity**: Contributions are acknowledged and results from analyses, including provenance and attribution, are expected to be shared with the N3C community.
- **Accountability**: N3C members take responsibility for their activity and hold each other accountable for achieving N3C objectives.
- **Security**: Activities are conducted in a secure, controlled-access, cloud-based environment, and are recorded for auditing and attribution purposes.

The analytics platform or N3C Enclave, hosted by a secure National Center for Advancing Translational Science (NCATS)-controlled cloud environment, includes clinical data from patients who meet criteria in the N3C COVID-19 phenotype from sites across the US dating back to January 2018[24]. Privacy-preserving record linkage will be developed to allow association with additional regulatory approvals to other datasets, such as imaging, genomic, or clinical trial data. Additionally, N3C will pilot the creation of algorithmically-derived synthetic data sets. The N3C data is available to researchers to conduct a broad range of COVID-19-related analyses. N3C activities are divided into five workstreams as shown in Figure 1.

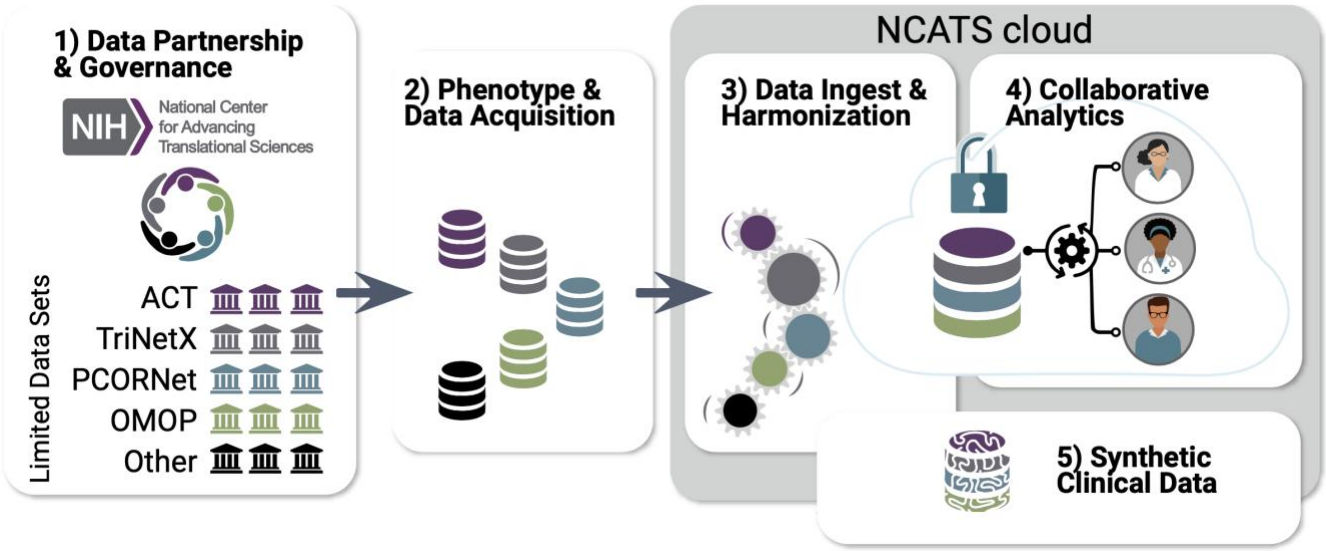

Figure 1. Establishing N3C socio-technical processes and infrastructure via community workstreams. *Each workstream includes representatives from National Center for Advancing Translational Sciences (NCATS)[25], the CTSA hubs[26], the Center for Data to Health (CD2H)[27], sites contributing data, and other members of the research community. 1)* ***Data Partnership & Governance:*** *This workstream designs governance and makes regulatory recommendations to NIH for their execution. Organizations sign a Data Transfer Agreement (DTA) with NCATS and may use the central IRB. 2)* ***Phenotype and Data Acquisition****: The community defines inclusion criteria for the N3C COVID-19 cohort and supports organizations in customized data export. 3)* ***Data Ingest & Harmonization****: Data resides within different organizations in different CDMs. This workstream quality-assures and harmonizes data from different sources and CDMs into a unified dataset. 4)* ***Collaborative Analytics workstream****: Data are made accessible for collaborative use by the N3C community. A secure data enclave (N3C Enclave), from which data cannot be removed, houses analytical tools and supports reproducible and transparent workflows. Formulation of clinical research questions and development of prototype machine learning and statistical workflows is collaboratively coordinated; Portals and dashboards support resource, data, expertise, and results navigation and reuse; 5)* ***Synthetic Clinical Data****: A pilot to determine the degree to which synthetic derivatives of the limited dataset (LDS) are able to approximate analyses derived from original data, while enhancing shareable data outside the N3C Enclave.*

DATA PARTNERSHIP & GOVERNANCE

The Data Partnership and Governance Workstream focuses on collaboratively developing a governance framework to support open science, while preserving patient privacy and promoting ethical research. With this goal in mind we borrowed best practices from prior work including centralized data sharing models (All of Us Research Program researcher hub[28], Human Tumor Atlas Network[29], the Synapse platform[30–35]) and consulted governance frameworks of other networks (Global Alliance for Genomics and Health[36], International Cancer Genome Consortium[37], ACT Network[38]). The N3C governance framework was drafted and refined iteratively with feedback from partners, especially from sites contributing data. This framework is composed of interlocking elements: (1) a secure analytic environment; (2) governing documents; (3) data transfer and access request processes and the Data Access Committee; (4) community guiding principles; and (5) an attribution and publication policy. The regulatory steps for organizations and users are shown in Figure 2, which provides details on the many layers of security, approvals, and policy-meeting required to ensure the dual goals of the highest security for and broad usage of the data. N3C supports four tiers of data: HIPAA limited data, HIPAA Safe-harbor data, aggregate data, and synthetic data[39,40] (see Table 1).

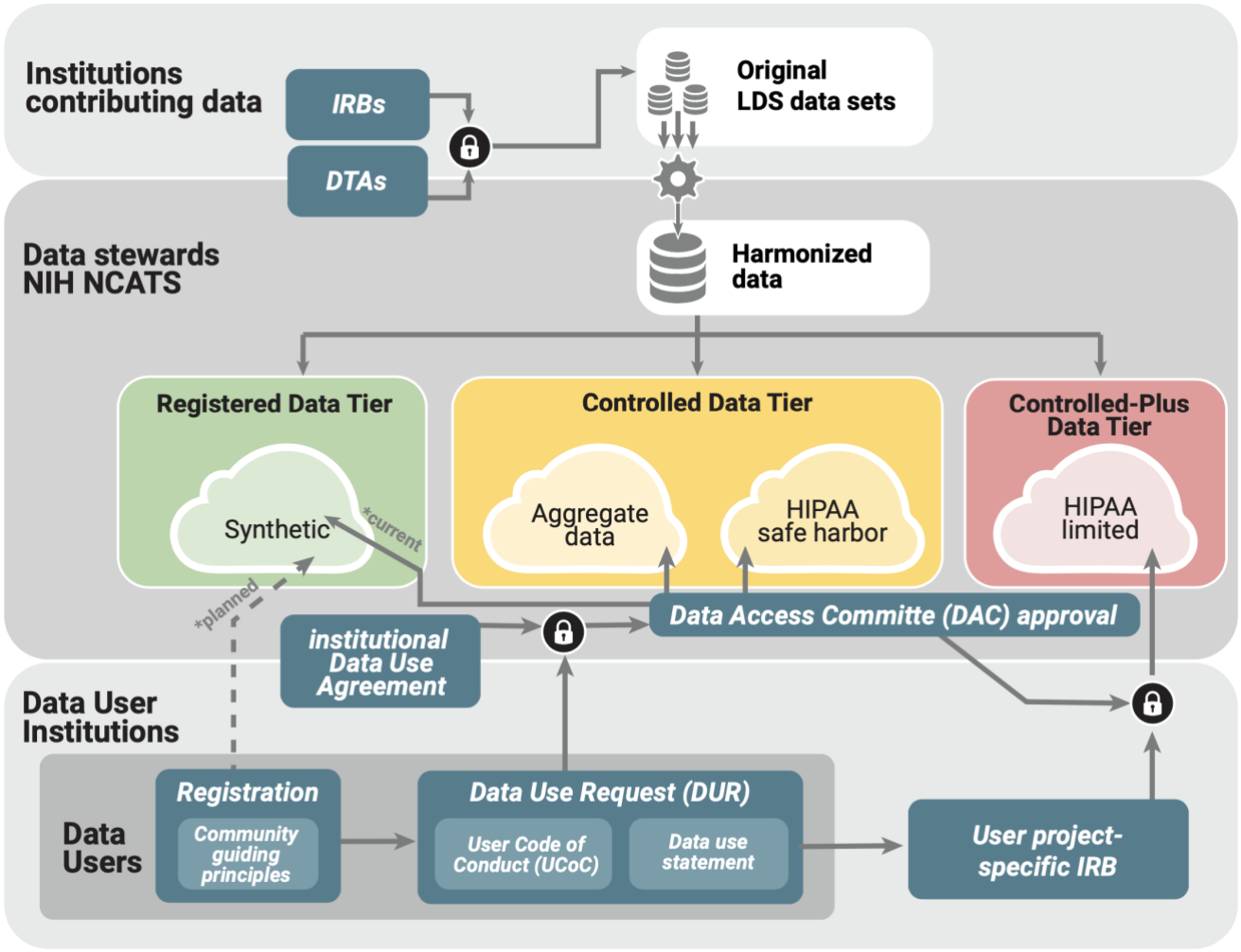

Figure 2. Regulatory steps and user access. *Organizations can operate as data contributors or data users or both; contribution is not required for use. For contributing organizations, the first step is a Data Transfer Agreement (DTA) which is executed between NCATS and the contributing organization (and its affiliates where applicable). For organizations using data, a separate, umbrella/institute-wide Data Use Agreement (DUA) is executed between organizations and NCATS. Interested investigators submit a Data Use Request (DUR) for each project proposal, which is reviewed by a Data Access Committee (DAC). The DUR includes a brief description of how the data will be used, a signed User Code of Conduct (UCoC) that articulates fundamental actions and prohibitions on data user activities, and if requesting access to patient-level data a proof of additional IRB approval. The DAC reviews the DUR and upon approval, grants access to the appropriate data tier within the N3C Enclave. Synthetic data currently follow the same procedure, but if the pilot is successful, we aim to make access available by simple registration if provisioned by the organizations. The lock symbol references steps where multiple conditions must be met.*

Security, privacy, and ethics

N3C has designed and tested processes and protocols to protect sensitive data and provide ethical and regulatory oversight. The N3C Enclave, which provides the only external access to the combined data set, is protected by a Certificate of Confidentiality[41]. This allows for granular access controls, activity auditing, and prevention of data downloads. NCATS acts as the data steward on behalf of contributing organizations.

Community Guiding Principles

Shared expectations and trust are essential for the success of the N3C community. Our goal is to ensure that N3C provides the ability to easily engage and onboard to a collaborative environment, for the broadest possible community. To this end, the workstream developed Community Guiding Principles

(CGP), which describe behavioral and ethical expectations, our diversity statement, and a conflict resolution process.

Data Transfer and Data Use Agreements

The Data Partnership and Governance Workstream worked closely with NCATS to develop two governing agreements: the Data Transfer Agreement (DTA), which is signed by contributing organizations and NCATS, and the Data Use Agreement (DUA), which is signed by accessing organizations and NCATS. Under the HIPAA Privacy Rule[40], a limited dataset may be shared if an agreement exists between the disclosing and the receiving parties. The NCATS DTA and DUA meet these HIPAA requirements and include provisions prohibiting any attempts to re-identify the data or use it beyond COVID-19 research. The decision to cover data transfer and data use as separate agreements was intentional, as it allows organizations to access data even if they do not contribute data.

IRB Oversight

Submission of data to N3C must be approved by an Institutional Review Board (IRB). To lower the burden associated with individual IRB submissions, and in accordance with the revised Common Rule[42] we established a central IRB at Johns Hopkins University School of Medicine (JHM) via the SMART IRB[43] Master Common Reciprocal reliance agreement. Contributing sites are encouraged to rely on the central IRB, but may choose to undergo review through their local IRB. This initial IRB approval is intended to cover only contribution of data to N3C and does not cover research using N3C data.

Data Use Request and Approvals

The Data Partnership and Governance Workstream and NCATS collaboratively developed a Data Use Request (DUR) framework, with the dual aims of protecting patient data and ensuring a transparent process for data access. Our tiered approach to data access allows us to reduce regulatory burden on investigators, while ensuring appropriate regulatory approvals are in place. There are three tiers of access: Registered, Controlled, and Controlled-Plus as described in Table 1.

Investigators wishing to access the data must have an N3C user profile linked to a public Open Researcher and Contributor Identifier (ORCID)[44]. Access requirements and approval processes vary depending on the level of access requested. For each project for which a user wishes to access data, he or she must submit a DUR with their intended data use statement and include a non-confidential abstract of the research project that will be publicly posted within N3C for transparency and to encourage collaborations. Data requesters must also sign a User Code of Conduct to affirm their agreement to the N3C terms and conditions. The N3C Data Access Committee (DAC), composed of representatives from NIH with occasional consulting from N3C community members, will review the DUR and verify that the conditions for access (see Table 1) are met. The DAC's role is to evaluate DURs; it does not exist to evaluate the scientific merit of the project.

Table 1. N3C data access tiers and conditions for COVID-19 related research.

Access Level	**Registered**	**Controlled**		**Controlled-Plus**
Data Type	**Synthetic Data** [45,46] **(pending pilot)**	**Aggregate Data (i.e., counts)**	**HIPAA Safe Harbor** [40,47,48]	**HIPAA Limited** [49]
Description	Computational data derivative that statistically resembles the original data	Counts and summary statistics representing 10 or more individuals	Data stripped of 18 direct identifiers per HIPAA rules	Data that may contain 3 direct identifiers per HIPAA rules (dates, full zip code, and any age)
Capabilities				
Downloadable data	Planned: pending validation & organizational agreement	Downloadable query results	No	No
Custom software	Yes	Yes - on downloaded query results	Yes with DAC approval	Yes - with independent IRB and DAC approval
Access Prerequisites				
Affiliation Requirement	Planned: no affiliation required (pending pilot validation)	Academic or commercial research organizations	Academic or commercial research organizations	Academic or commercial research organizations
Data Use Agreement Signed by Home Organization	Required at present (planned to be not required pending pilot)	Required	Required	Required
Human Subjects Training	Required at present	Required	Required	Required
NIH Security Training	Required at present	Required	Required	Required
Request Submission and Approval Steps				
Data Use Request	Required at present	Required	Required	Required
Rationale for Accessing Identified Data	N/A	N/A	N/A	Required
General description of research project (objectives, testable hypothesis, planned analysis)	Yes	Yes	Yes	Yes
Abstract of research project (posted publicly)	Yes	Yes	Yes	Yes
Approval Process	DAC	DAC	DAC	DAC + IRB

Attribution and Publication Policy

N3C participants share a commitment to the dissemination of scientific knowledge for the public good. The Attribution and Publication Policy extends FAIR[22,50] to encompass all contributions to the N3C. Analyses posted within the N3C Enclave leverages the Contributor Attribution Model[51] to track the transitive credit[52] of all upstream contributors. A publication committee assists in tracking N3C outcomes. This first N3C manuscript was developed through an open call for contributions from the entire N3C and is an exemplar of the Attribution Policy.

N3C Data Linkage

Clinical data have high utility for COVID-19-related research; however, N3C recognizes the need to analyze clinical data along with data from other sources. Therefore, we will establish a privacy-preserving strategy to enable linkages within and external to the N3C data set. In this way, genomic, radiology, pathology imaging, and other data can be analyzed in conjunction with the N3C clinical data. It will also lay the groundwork for future studies to deduplicate patients.

PHENOTYPE AND DATA ACQUISITION

The purpose of this workstream is threefold: (1) to determine the data inclusion/exclusion criteria for import to N3C (computable phenotype); (2) to create and maintain a set of scripts to execute the computable phenotype in each of four CDMs—ACT, OMOP, PCORnet, and TriNetX—and extract relevant data for that cohort; and (3) to provide direct support to sites throughout the data acquisition process.

Computable Phenotype Definition

Given our evolving understanding of COVID-19 signs and symptoms, it is challenging to define stable computable phenotypes that can accurately identify COVID-19 patients from their EHR data. To ensure that the data in N3C encompass these varied and fluctuating perspectives, we chose to bring together existing inclusion criteria and codesets from a number of organizations (e.g., CDC coding guidance[53,54], PCORnet[55], OHDSI[56], LOINC[57], etc.) into a “best-of-breed” phenotype. The draft phenotype was iterated within the N3C community and remains open to public comment. The N3C phenotype[58] is designed to be inclusive of any diagnosis codes, procedure codes, lab tests, or combination thereof that may be indicative of COVID-19, while still limiting the number of extracted records to meaningful and manageable levels (see Table 2). Notably, the N3C COVID-19 phenotype purposely includes patients who tested negative for COVID-19; thus inclusion in the N3C cohort is not equivalent to “positive for COVID-19,” but rather “relevant for COVID-19-related analysis” as defined by their categorization as “lab-confirmed negative,” “lab-confirmed positive,” “suspected positive,” or “possible positive” (see the N3C phenotype documentation[59] for detailed definitions of these categories).

Table 2. Scale comparison of three sites’ positive COVID-19 cases, their N3C-relevant cohort, and their denominator (number of patients seen in a one-year period). All numbers rounded to nearest 10.

	Site 1	Site 2	Site 3
COVID-positive patients as publicly reported by site[a]	2,550	5,540	390
N3C-relevant cohort[b]	67,350	46,500	12,000
Denominator[c]	1,271,510	1,259,330	172,000

[a] The number of COVID-positive patients publicly reported by this site as of the week of 6/8/2020
[b] The number of patients qualifying for the N3C COVID-relevant phenotype at this site as of the week of 6/8/2020

[c] The number of unique patients seen in a one-year period at this site

To encourage maximal community input into the phenotype definition, we chose to use GitHub[60] to host all versions of the phenotype definition in both machine-readable (SQL) format and human-readable descriptions[61]. The phenotype is updated approximately every two weeks, reflecting (for example) when the emergence of new variants of COVID-19 lab tests necessitate adding new LOINC codes, or to incorporate suggestions from the community.

Data Extraction Scripts

Once the N3C community agreed on the initial phenotype logic, the initial phenotype logic was translated into SQL to run against each of four common data models at participating sites: ACT, OMOP, PCORnet, and TriNetX. Multiple SQL dialects support the different relational database management systems in use.

The use of existing CDMs allows for rapid startup and minimizes the burden of participation by contributing sites. Most CTSA sites and many other medical centers host one or more CDMs. In particular, the following four CDMs are frequent in these communities, and form the basis for data submission to N3C:

- **ACT Network**: A federated network, data model, and ontology for CTSA sites that consists of i2b2 data repositories that are integrated by the SHRINE (Shared Health Research Information Network)[62] platform, focused on real time querying across sites[4].
- **PCORnet**: The official federated network and data model for the Patient-Centered Outcomes Research Institute (PCORI)[63] is a US-based network of networks focusing on patient-centered outcomes.
- **OHDSI**: A multi-stakeholder, open science collaborative focused on large-scale analytics in an international network of researchers and observational health databases maintaining and using the OMOP CDM[64].
- **TriNetX**: An international network of clinical sites coordinated by a commercial entity (TriNetX, Inc.) providing clinical data for cohort identification, site selection, and research to investigators in healthcare and life sciences[8,65].

Contributing organizations are expected to submit data using one of these models.

N3C's SQL scripts serve two functions for participating sites: (1) to identify the qualifying patient cohort in a site's CDM of choice and store that cohort in a table, and (2) to extract longitudinal data for the stored cohort into flat files, ready for transmission to the central N3C data repository. The scripts extract the majority of the tables and fields in each of the CDMs, with the exception of tables and fields that are unique to a single model and cannot be successfully harmonized. At a high level, data domains extracted by N3C include: demographics, encounter details, medications, diagnoses, procedures, vital signs, laboratory results, procedures, and social history; specific variables included in these domains for each of the data models can be found in each model's documentation[66–68]. Like the phenotype definition, all scripts are publicly posted on GitHub[69] for community comment and peer-review.

Data Transfer Process

The guiding principle for these scripts is to minimize customization at the local site level. The workstream devised four different methods of data extraction and transfer (see Table 3), allowing sites to use the technology stack with which they are most comfortable, while complying with our guiding principles.

Once a site joins N3C and is ready to contribute data, members of the Phenotype and Data Acquisition workstream make themselves available via web-based "office hours" to onboard the new site and explain the process for script execution and data transmission.

Table 3. Data extraction and transfer methods that sites may use to submit data to N3C.

	Human (Manual) Steps	Automated Steps
R Package	1. Download the R and SQL code. 2. Configure local variables (DB connection, schema names, etc.)	3. Run phenotype and extract scripts. 4. Extract results to individual files, following N3C naming and structure conventions. 5. SFTP extract to N3C.
Python Package	1. Download the Python and SQL code. 2. Configure local variables (DB connection, schema names, etc.)	3. Run phenotype and extract scripts. 4. Extract results to individual files, following N3C naming and structure conventions. 5. SFTP extract to N3C.
TriNetX	(Automated step first) 1. Download data from TriNetX. 2. SFTP extract to N3C.	1. TriNetX runs phenotype and extract scripts on the site's behalf.
SQL	1. Download the SQL code. 2. Configure local variables (schema names, etc.) 3. Run phenotype script. 4. Run extract scripts, one at a time. 5. Extract results to individual files using the N3C directory structure, naming conventions, file format. 6. SFTP extract to N3C.	None

DATA INGESTION AND HARMONIZATION

N3C aims to support consistency in the data acquisition process across the four CDMs. Simply aggregating those data together is insufficient. Not only does each model have different structures and values, but heterogeneity exists within models. The goal of the Data Ingestion and Harmonization workstream is to align and harmonize the syntax and semantics of data from all contributing sites into a single data model, retaining as much specificity and original clinical intent as possible as well as data quality and transparency. These steps support N3C's ultimate goal of producing comparable and consistent data to enable effective and efficient analytics[70,71].

Target Data Model Selection

A single data model enables scalable analytics. The emergent Health Level Seven International (HL7) Fast Healthcare Interoperability Resources (FHIR)[72] standard may form a pluripotent research data model in complete synchrony with EHR source data[73]. The CD2H[74] has been working through its Next Generation Data Sharing core and catalyzing the formation of the Vulcan FHIR Accelerator for Translational Research[75] to advance this strategic goal. However, FHIR is not sufficiently mature in its specification and, more pertinently, its development of "bulk" multi-patient research data transfers. The most expedient alternative was to select among the four contributing CDMs. All the CDMs enjoy large, dedicated communities continuously contributing to their development, and all are valuable to COVID-19 research. As a tactical choice, OMOP 5.3.1[76] was selected as the canonical model of N3C due to its maturity, documentation, and open source quality monitoring library, data maintenance, term mapping, and analytic tools[77,78].

Model Harmonization Mappings

With OMOP 5.3.1 selected as the target data model, it was first necessary to map tables, fields, and

value sets from ACT 2.0, PCORnet 5.1, and TriNetX to OMOP 5.3.1 to serve as a foundation for N3C's extract - transform - load (ETL) processes. Fortunately, as part of the Common Data Model Harmonization (CDMH)[79] project, CD2H and related federal projects had initiated mapping from each CDM to the BRIDG[80] and FHIR standards. N3C was able to leverage this previous work to jumpstart the required mappings between each CDM and OMOP 5.3.1.

N3C worked with contractors and colleagues from the CDMH project to build two sets of harmonization data for each source CDM:

1. Syntactic mapping for each CDM field to a corresponding table/field in OMOP with conversion logic
2. Semantic mapping of where in the OMOP vocabulary each value in each value set should be mapped.

N3C hosted numerous review and validation meetings for each set of source-to-target mappings. All meetings included subject matter experts (SMEs) from the source CDMs, and SMEs from the OHDSI community. All mappings at all stages of development are publicly available on GitHub[81].

Extract - Transform - Load

When a participating site submits a data payload to N3C, the data submission flows through an ETL pipeline that leverages the aforementioned mappings. The pipeline is powered by Adeptia[82], a cloud-based Platform as a Service on the secure NCATS Amazon Web Services (AWS) production cloud. Prior to loading a given data payload into the production N3C database, the payload must first undergo a series of data quality checks as part of the ingestion process. This process, described below, ensures that any errors can be corrected, and that site-specific idiosyncrasies can be accounted for and known to downstream users.

Data Quality Processes

In large data aggregation projects, where many sources combine to form a larger dataset, there are issues caused by the data heterogeneity which impact data quality (DQ)[83,84]. DQ measures, including consistency, correctness, concordance, currency, and plausibility, are important to support analysis and computation[85,86]. Many large-scale data aggregation projects benefit from focusing on a set of contextual use cases or a defined population research domain.[87–89] For N3C, we developed an approach to DQ that addresses the downstream application of the data for machine learning and statistical analytics .

In order to establish a starting point, the N3C Data Ingestion and Harmonization workstream became familiar with a wide array of available DQ tools and processes. They met with SMEs from each of the source CDMs, focusing on the DQ approaches and tools employed in their native implementations (see Table 4). These native approaches became a foundation on which N3C could build its own DQ processes.

Table 4. Data quality tools and methods evaluated.

Tool Type		**Tool**
Native CDM DQ Processes	*PCORnet*	Data Check Scripts (v8.0)[90]
	ACT	"Smoke" tests[91]
	TriNetX	Focused Curation Process
	Adeptia Platform Processes	Process automation support[92]
		Data & Map validation functions[93]
OHDSI Collaborative Tools	*Data Quality Dashboard*	Data quality tests of OMOP databases[89]
	Atlas	Design/execute analytics on OMOP databases[94]
	Achilles	Data characterization of OMOP databases[95]
	White Rabbit	ETL preparation and support[96]
	Custom Scripts	SQL, R

N3C Ingestion and Harmonization Data Quality Checks

The Data Ingestion & Harmonization workstream developed strategies to assess and improve DQ within the N3C ingestion pipeline. This group considered (1) what DQ requirements were appropriate for N3C, (2) which tools and methods could be used to support DQ, and (3) where in the ingestion pipeline DQ checks should be instantiated.

In these discussions, the group agreed that a "light touch" was the best approach to DQ for N3C; to pass along the data as they are, and only in some cases make "cleaning" corrections. These cleaning steps would seek to correct the data only to the extent required to support computation and OMOP data model conformance. The exception to this is data related to COVID-19 tests, as we anticipate variance in how organizations code COVID-19 tests, particularly in the very early stages of the pandemic. Due to the criticality of these data for N3C, we corrected erroneous coding using text data indicating COVID-19 status, which would otherwise be lost.[97]

To ensure that data loss was minimized in the data transformation process, we made the decision to retain the raw source data during and after the mapping and transformation process to preserve contextual details about the data for meta-analyses downstream. Additional detail about the N3C Data Quality Checks and ingestion process is provided in Figure 3.

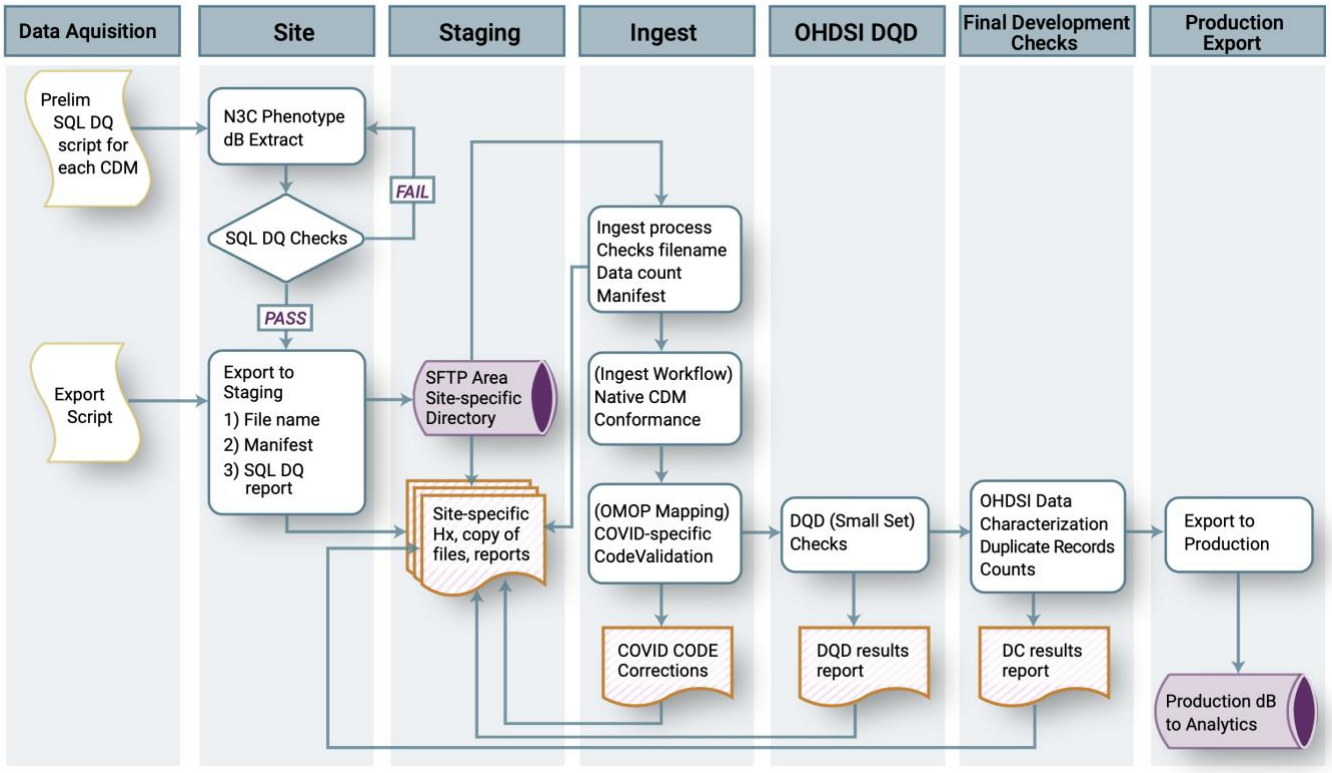

Figure 3. N3C Data Quality Checks. *At the sites, the extraction script performs a check for duplicate primary keys; if duplicate keys are found, the extraction fails until the site resolves the error. When extraction is successfully completed, a data "manifest" is created that contains metadata about the site and the payload. Site personnel then sFTP the data to N3C to be queued for ingestion. The first step in the ingestion process checks that the payload is consistent with the formatting requirements and the manifest file. Next, the payload is loaded into a database modeled after the payload's native CDM, which ensures source data model conformance. Next, a series of data quality checks including a subset of COVID-19-specific code validations are performed, and if needed, minimal corrections are made. Any corrections are recorded and added to the payload documentation. Next, the payload is transformed to OMOP 5.3.1 using the validated maps from the payload's native CDM. Once in OMOP 5.3.1, a subset of the OHDSI Data Quality Dashboard tests are run, and the results of these are added to the payload documentation. The payload is then exported to a merged database containing all the previously harmonized site data payloads, where it is then checked for conformance again before export to the analytics pipeline.*

COLLABORATIVE ANALYTICS & THE N3C ENCLAVE

The goals of the Collaborative Analytics workstream are to ensure secure stewardship of N3C data; design and disseminate analyses; integrate community tools and resources; provide tracking and attribution of users, results, and contributions; and enable novel approaches to data sharing (Figure 4).

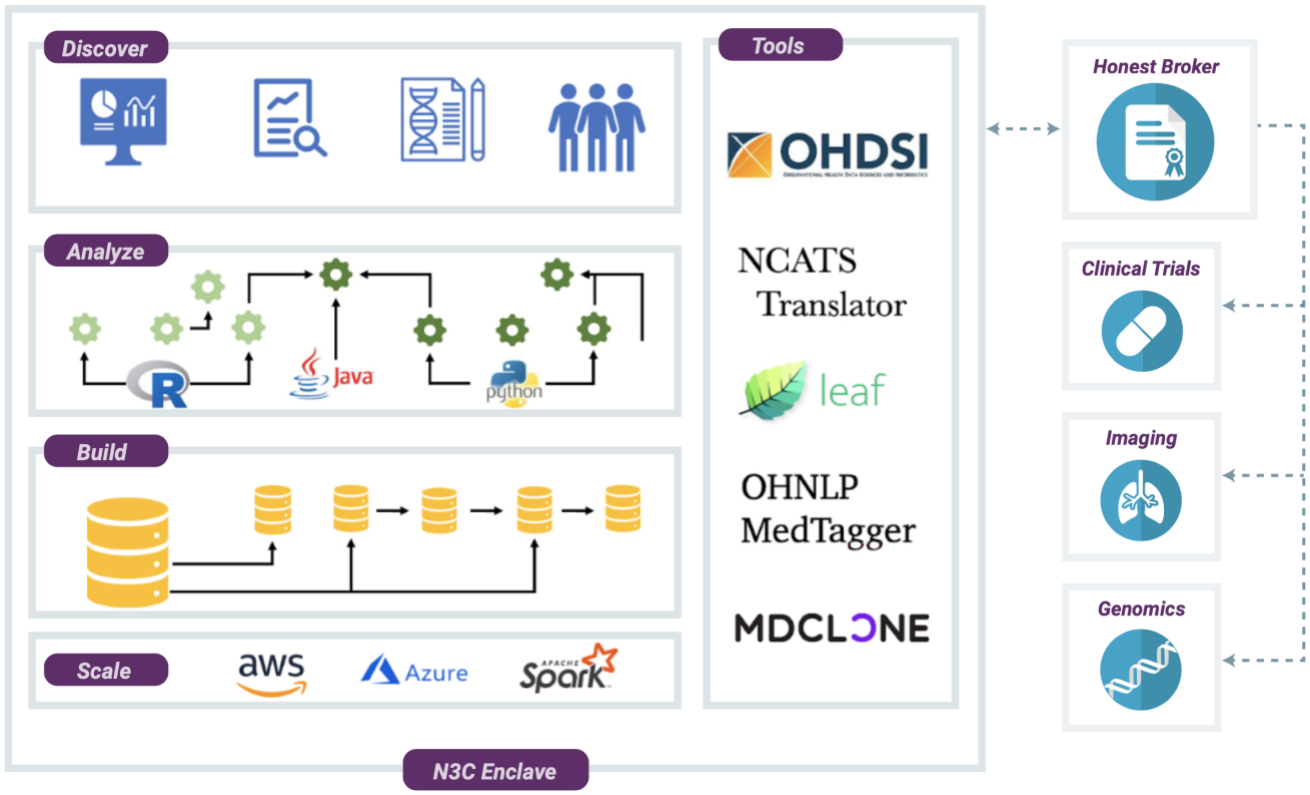

Figure 4. N3C Enclave. *The analytical environment for N3C is a secure, virtualized, cloud-based platform. Within the N3C Enclave, researchers have access to raw data, as well as transformed and harmonized datasets derived by other researchers. Analytical tools hosted within the environment support complex ETL, generation of COVID-19 specific data elements, statistical analysis, machine learning, and rich visualizations. Third-party tools contributed by the community can be integrated into the environment; current tools include OHDSI tools and the Leaf patient cohort builder. N3C is developing methods for integration of genomic, imaging, and other data modalities.*

A “data enclave” is a secure data and computing environment, designed to facilitate virtual access to hosted data with safeguards to prohibit or limit data export[98]. The N3C Enclave meets this definition as a virtual, secure, cloud-based data enclave—controlling user access with regulatory and technical protections, and prohibiting the download of patient-level data from the N3C environment—while enabling COVID-19 analysis by the research community. The N3C Enclave is managed by NCATS, which serves as the legal custodian of all data within the environment (see Governance). Hosted within the N3C Enclave is Palantir Foundry, a data science platform enabling complex and reproducible analysis using standard open-source, analytical packages in languages such as Python, R, SQL, and Java, as well as point-and-click and dashboard-style analytical tools. Standard packages for statistical analysis and machine learning, such as Tensorflow, scikit-learn, and others are available, and backed by Apache Spark allowing operations at very large data scales. Community contributed tools and resources are also being made available, the first deployments are listed in Table 5.

The platform is certified as Federal Risk and Authorization Management Program (FedRAMP) Moderate[99], a government security standard for unclassified but highly sensitive data. To enable research collaboration on sensitive EHR data, the N3C Enclave supports fine-grained access controls and auditing mechanisms, allowing multiple users to work securely in a single system. The system provides “limited realms”, where users are granted access to specifically designated data and resources such as Limited DataSet (LDS) and Safe-Harbor data. Additional security and auditing mechanisms include the ability to limit patient-level data access, read and write access to data sets, and user access, auditing, and tracing.

As outlined in Figure 2, investigators have restricted access to LDS data without project specific IRB reviews. This is mediated by the designation of a few software agents, such as cross tabulation, logistic regression, mapping and other related visualizations, as having privileged access to the LDS data in a manner that 1) prohibits users from seeing the underlying patient-level data, and 2) inhibits the display of tables or cells that comprise less than 10 patients. Through this enclave functionality, secure analyses of data containing limited PHI (LDS) can proceed without compromising privacy or confidentiality. The outputs from these specially designated software packages are regarded as results, and are not subject to human subjects data restrictions.

Table 5. Examples of community contributed tools integrated within the N3C computing environment.

Tool	Description
OHDSI Atlas	**OMOP-optimized tools for cohort querying and analysis.** Data quality; data and cohort definition; rapid & reliable phenotype development[100]; phenotype performance evaluation[101]; integration of validated phenotypes definitions into study skeletons that learn and validate predictive models[102] and execute a variety of comparative cohort study designs using empirically validated best practices[103–105].
LOINC2HPO	**Mapping of LOINC-encoded laboratory test results to Human Phenotype Ontology (HPO)**. Interoperability for lab results or radiologic findings with OMOP CDM; phenotypic summarization for use in machine learning algorithms, semantic algorithms, and knowledge graph-based applications[106].
NCATS Biomedical Data Translator	**Translational integration with basic research data and literature knowledge.** Symptom-based diagnosis of diseases linked to research-based molecular and cellular characterizations[107,108][109]; suite of resources include the Biolink Model[110], a distributed API architecture, and a variety of knowledge graphs (KGs) covering a range of biological entities such as genes, biological processes, and diseases; the KG-COVID-19[111] knowledge graph also includes literature annotation.
Leaf	**Web-based cohort builder.** Study feasibility for clinician investigators with limited informatics skills[112]; hierarchical concepts and ontologies to construct SQL query building blocks, exposed by a simple drag-and-drop user interface.

Transparency and reproducibility are fundamental to the prescribed use of the N3C Enclave[113]. The platform automatically builds a provenance graph for every dataset and analysis. Each artifact in the platform is stored as an immutable object, enabling full Git-like traceability on all changes. Each workflow includes extensive metadata describing all of the inputs, the user who triggered it, the build environment, and the required packages. Researchers can confidently share results as “reports,” which include a precise record of how they were generated, allowing other researchers to replicate and build on the analyses. Key capabilities are:

- **Raw data provenance**: Support for provenance capture of imported data, and recording of metadata for understanding the origins of each dataset.
- **Data lineages**: Data transformations recorded as a dependency graph, enabling full (re)construction of data lineage.
- **Versioning**: Data versioning, allowing full analytical reproducibility.
- **Validation and errors**: Runtime characteristics monitored and recorded.
- **Attribution**: Fine-grained attribution of individuals, groups, and organizations and a record of their contributions according to the Contributor Attribution Model (Figure 5)**.**

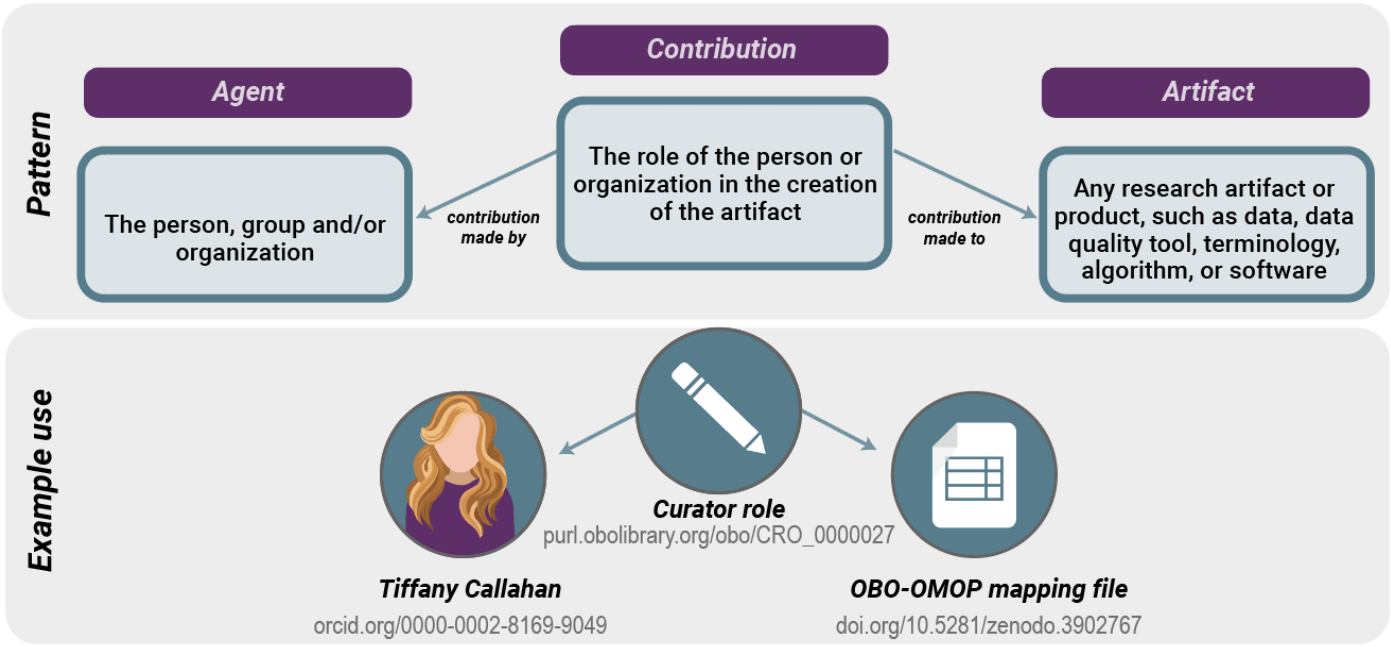

Figure 5. The Contributor Attribution Model (CAM). *In the N3C Enclave, CAM is used to aggregate all contributions to any given workflow or report generated with a specific declaration of what exactly each person contributed, supporting the notion of transitive credit[52]. ORCID identifiers are used to identify users. An example contributor to an artifact used in the N3C is shown on the lower panel.*

SYNTHETIC CLINICAL DATA PILOT

The creation of synthetic clinical data represents a unique opportunity for N3C to more widely disseminate and provide greater utility for the N3C dataset. Current state-of-the-art approaches for the generation of synthetic clinical data can be broadly classified as:

- **Statistical simulation:** Statistical models or profiles of normal human physiology and/or disease states are created based upon real-world data. The ensuing simulated patients and their data are generally consistent with population-level norms[114–116].
- **Computational derivation:** Computational models of real-world data are produced on-demand, which can be used to produce novel data in a multi-dimensional space (e.g., features) that adhere to the quantitative distributions and co-variance of the original source data. When generating these types of models, data content and statistical features are a function of the input dataset. The process can be repeated multiple times with the same source data producing multiple derivative synthetic datasets. Further, such computationally derived synthetic datasets do not share mutual information with source data, minimizing the potential for re-identification[45,46,117–119].

N3C has launched a pilot to evaluate the creation of synthetic data from the N3C LDS, and will focus on validating the synthetic data for key analyses against those performed on the LDS in areas such as identifying patients for whom COVID-19 testing can/should impact clinical management; anticipating severity of disease, risk of death, and potential response to therapies; and geospatial analytics for enhanced insights into geographic hotspots and for quantifying the contribution of zip code level SDoH in predictive analytics.

DISCUSSION

Analytical innovation and open science on sensitive data

The N3C architecture, dataset, and analytic environment is a powerful platform for developing machine learning algorithms, statistical models, and clinical decision support tools. Analytic models are able to use time series, clinical, and laboratory information to predict progression, assess need and efficacy of clinical interventions, and predict long term sequelae. Researchers are able to leverage both "raw" EHR data, and carefully curated derivatives, building upon the work of prior or parallel studies. The platform also supports translational informatics by making available basic research data and knowledge in the form of knowledge graphs and related tools, mined and annotated literature, and clinical EHR data in the same analytical space. Semantic interoperability enables questions to aid drug and mechanism discovery efforts such as: "What protein targets are activated by drugs that show effectiveness among patients with COVID-19 infection? What genetic variants are associated with recovery from COVID-19 infection? What biological pathways contribute to disease severity among patients infected with COVID-19?."

N3C offers an innovative model for deeply collaborative analytics on clinical data, promoting open and transparent research practices on sensitive EHR data at scale. Recent high-profile manuscript retractions in prominent journals underscore the imperative for transparency and reproducibility in COVID-19 research[120][21,121]. Attribution is native to the system, and supports the notion of transitive credit[52] for all contributors. Investigators are encouraged to pre-specify hypotheses or other study goals in a publicly-available and versioned study protocol and to maintain full documentation of all code and protocol revisions in order to mitigate the risk of p-hacking and promote the legibility and traceability of all major study design and analytic choices[122]. The N3C Enclave allows and, indeed, requires sharing of software, results, and methods. It is our belief that by allowing the research community to work together in this way, we aree able to rapidly increase our collective understanding of COVID-19 and identify effective approaches for prevention and treatment, ultimately curbing the effects of this pandemic on our nation and world.

Status of data availability within the N3C Enclave

At the time of publication, there have been 49 DTAs executed, 27 IRB protocols approved (23 reliance, 4 local), and 11 institutions have deposited data within the N3C (5 PCORnet, 3 OMOP, 2 TriNetX, and 1 ACT). Overall, the cohort from the 11 sites to date represents 33K positive COVID-19 cases out of ~513K total patients. We anticipate ~150K positive cases to be included in the N3C Enclave by Sept. 1st, 2020.

What kinds of analyses are enabled?

COVID-19 has proven to be a novel, heterogeneous disease, particularly in terms of range of symptoms and signs, severity and clinical course. By integrating data from multiple sites, we **enable researchers to explore questions with vastly more statistical power** than is achievable at individual sites and to use machine learning methods at scale.

N3C enables us to address several important questions related to the **diagnosis and management** of COVID-19. For example, how are different types of antigen and antibody tests for SARS-CoV-2 being used across the country? What other laboratory and imaging protocols are being used in conjunction with viral testing in ambulatory, urgent care, and emergency department environments? What place (if any) does convalescent plasma have in COVID-19 treatments? What are the markers for and best practices to prevent COVID19-related clotting disorders? What are the best practices for inflammatory monitoring prior to cytokine storm syndrome? The first three of these might be answerable in a federated network, but the last two require a centralized data resource such as N3C.

N3C is a well-suited resource to **clinically characterize and deeply phenotype** a very large cohort of patients with COVID-19. In addition to frequently reported metrics such as rates of hospitalization and ICU admission, ventilator, and renal replacement therapy utilization, these analyses can assess variation

in duration of need for intensive clinical support. Detailed temporal analyses of the progression of respiratory and other organ system dysfunctions are possible. Prevalence and predictors of complications such as cardiomyopathy, thrombosis, acute kidney injury, hypoxemia, stroke, and delirium can be evaluated. For populations with rare complications, such as the emergence of Kawasaki disease-like inflammatory symptoms, a centralized dataset provides the statistical power to characterize emerging adverse effects. Once accurate models to predict complications are available, tools can be implemented for prevention, early detection, and intervention. For prediction tasks based on longitudinal data, a variety of methods based on recurrent neural network architectures can be leveraged[123]. To characterize patient subtypes, tensor factorization approaches have been shown to be quite effective for similar tasks[124]. Accurate machine learning-based CDS tool development requires algorithm optimization, a process that is greatly facilitated by a centralized data resource.

Detailed medication and other clinical data in N3C also **enable analyses of treatment pathways and patient response**. These analyses can encompass medications received prior to and concurrent with the disease course as well as specific drug therapies, such as antivirals like remdesivir or hydroxychloroquine, tocilizumab, corticosteroids, broad-spectrum antibiotics, antifungals, and therapeutic anticoagulation. They can also provide evidence for best practices in clinical care such as supplemental oxygen, proning[125], noninvasive positive pressure ventilation, invasive ventilation, and extracorporeal membrane oxygenation. N3C will be well-positioned to generate immediately testable hypotheses about combinations and sequences of therapies, helping researchers to better design, prioritize, and analyze randomized trials. Analyses can take into account known outcome predictors including 1) medical history, comorbidities, and indicators such as hypertension, diabetes, and body mass index; 2) progression of vital signs; and 3) laboratory data such as electrolytes, markers of organ dysfunction, measures of inflammation, and indicators of possible thrombosis or approaching cytokine storm[126]. Investigators can develop tools to predict treatment response based on these analyses. Clinicians could match a patient's phenotype to one or more distinct groups of patients in the N3C dataset with known clinical outcomes. Such patient matching can be done at the point of care and provide real-time precision reference information for CDS, potentially based on patient similarity learning[127]. Furthermore, N3C facilitates the use of specific algorithms that can increase the unbiased selection of cohorts that have complete data, and which can be applied to most EHR studies[128,129].

The size and national coverage of N3C data make it a **unique source of COVID-19 data for population health segmentation and risk stratification**. Segmenting the population for the risk of various outcomes (e.g., clinical, utilization) allows more efficient and effective resource allocation and interventions[130] as well as enable healthcare providers to measure and balance the risk of COVID-19 complications versus other clinical conditions and morbidities. For example, identifying patients who will benefit the most from the anticipated COVID-19 vaccination is of utmost importance[131]. Assessing heterogeneity of treatment/vaccine effect at the scale necessary is facilitated by the centralized nature of N3C.

The pandemic has amplified and exacerbated the effects of systemic racism and long-standing social and economic disparities on health and healthcare[132–135]. N3C-based studies can support healthcare providers to **identify clinical outcome disparities and social determinants of health** (SDoH), as well as to help public health officials and policy makers to identify inequalities on a systemic level (e.g., analyzing statewide claims or EHR data using models developed based on N3C data). The N3C can expedite analytics regarding the impact of COVID-19 on different segments of the population, including racial and ethnic groups, rural population, children, pregnant women and newborns, and residents of communal living. Several sites are contributing structured data about the SDoH (e.g. race, ethnicity, zip code), and geo-derived SDoH factors or environmental pollution can also be associated based on the zip code. N3C also provides a unique opportunity to enhance the role of data science and population health informatics in bridging the gap between clinical care, public health, and social services[136]; thus, collectively aiming for predictive models promoting equity for all minorities[137] in the current and potential future COVID-19 outbreaks.

Integrating data from multiple clinical research systems has proven effective for estimating potential research cohorts, identifying eligible patients, supporting current studies, and enabling new analyses [73,138]. However, there are a number of caveats and N3C is no exception. Patient care data and the processes that generate and capture them differ from good research practices[139]. EHR data captured in real time are often wrong (e.g., incorrect diagnosis) or may have originated from a different patient. The available data may not convey the complete clinical picture due to fragmentation of patient care. For example, a patient's initial coronavirus test results may be performed by a government laboratory and not transmitted to the patient's EHR. Finally, patient care data rarely have completeness, reliability, granularity, and competent coding found in good, prospective clinical studies. This is not to say that research using the N3C Enclave will be flawed. The sheer magnitude of the dataset provides a buffer against the effects of systematic reporting bias. A number of methods can be used for considering data from multiple institutions, for example by applying methods used in meta-analysis[140].

CONCLUSION

N3C has been driven by passionate individuals through a complicated world of regulation and habituation by healthcare organizations. By opening the door to a broad analytic community, we bring to the table new skill sets, diverse viewpoints, and additional opportunities for novel approaches. N3C is driving new standards in openness for collaboration on sensitive clinical data, and builds upon the infrastructure developed nationwide over the past decades.

Specifically, the N3C model will continue to be refined and streamlined to provide a scalable approach that can be leveraged to help manage future waves of COVID-19, unforeseen novel diseases, and other major health crises, as well as long-standing challenges in healthcare. While N3C is focused on the US, this is a global pandemic and we must identify ways to collaborate with other international groups who are building similar infrastructure for a global approach; such conversations are underway[141,142].

FUNDING STATEMENT

This work was supported by the National Institutes of Health, National Center for Advancing Translational Sciences Institute grant number U24TR002306.

COMPETING INTERESTS STATEMENT

N3C includes a number of commercial partners, without whom N3C would not be possible; they are: Adeptia, Inc., TriNetX, Inc., Palantir Technologies, Microsoft Corporation, MDClone, IQVIA, and Amazon.

Author conflict of interest statements: Melissa Haendel and Julie McMurry have a founding interest in Pryzm Health; Ms. Kristin Kostka is an employee of IQVIA; Andrew Girvin, Amin Manna, Harish Ramadas, Benjamin Amor, and Nabeel Qureshi are employees of Palantir Technologies; Matvey Palchuk and Lora Lingrey are employees of TriNetX Inc; Clair Blacketer is an employee of Janssen Research & Development; Cody Rutherford is an employee of Noblis, Inc.; John Liu is an employee of Optum, Inc.; Micahel Larionov is an employee of Spok. Inc.; Ofer Mendelevitch and Michael Lesh are founders and shareholders of Syntegra USA Inc; Andrew Kanter is the CMO of Intelligent Medical Objects, Inc; Hua Xu has research-related financial interests in Melax Technologies, Inc.

ETHICS STATEMENT

While no IRB review is required for the work presented in this manuscript, we describe the creation of a central IRB at JHU for use by member organizations. The central IRB protocol has been made public[143].

CONTRIBUTIONS STATEMENTS

MA Haendel: I have led the creation and implementation of N3C, coordinating all workstreams, resource development and integration, and generally playing data and people yenta. • CG Chute: Program initiation, governance, guidance, technical strategy, and promotion. • TD Bennett: I contributed by participating in decision-making as co-lead of the Clinical Scenarios and Analytics subgroup and by offering language for and editing the manuscript. • DA Eichmann: I co-lead the Portal & Dashboards subgroup and am a member of the CD2H steering committee, and hence have been involved in decision making from the outset, as well as managing the N3C web site and COVID-19 publication search engine. • J Guinney: I have led the Collaborative Analytics workstream. • WA Kibbe: Contributing to the policy and governance components and the N3C data enclave portal and dashboards. • PRO Payne: Conceptualization of synthetic data work stream and architecture, development of content for manuscript, • ER Pfaff: I lead the Phenotype and Data Acquisition Workstream, and designed the data acquisition process in collaboration with a group of coders from across the CTSA consortium. • PN Robinson: I have led the Clinical Scenarios group and have contributed to the development of OMOP-HPO interoperability. • JH Saltz: Writing, contributed to decision making, contributed to system use case development and definition of covid 19 data elements • H Spratt: N3C Clinical Scenarios & Data Analytics subgroup co-lead • C Suver: governance WG lead - developed governance policies and documents • J Wilbanks: I have co-led the governance workstream. • AB Wilcox: Advisory role (overall goal and strategy); author role (writing review and editing); data role (data modeling, data transformation, data validation); conceptualization role (contributed the original data use principles, the data modeling approach for specific use cases, and the centralized data approach) • AE Williams: My contribution fits the writing review and editing role in the contribution role ontology • C Wu: I have co-led the Tools subgroup on collaborative analytics. • C Blacketer: I have served the N3C consortium as an OMOP subject matter expert providing guidance around the transformation of PCORNet, I2B2, and TriNetX to the OMOP model as well as offering data quality insights and solutions to be employed in the data pipeline. • RL Bradford: contributions to phenotype development from scientific and technical standpoint. Design and testing of phenotype and extraction scripts • JJ Cimino: Dr. Cimino served as a project advisor, and reviewed and provided comments on the manuscript. • M Clark: Working on the phenotype and data acquisition stream,

I developed the SQL code for the phenotype to run against PCORNet CDM, developed a python extraction script to help sites automate this process and contributed to the stream through daily meetings. • EW Colmenares: I contributed to the design and code of the phenotype as well as the scripts to extract data from the common data models. • PA Francis: Project management, regulatory, team management, project administration • D Gabriel: I provide content expertise and project management for the data ingestion and harmonization workstream • A Graves: I've contributed code and design work for web and visualization development, helped to support communication with design work, and supported project administration on a local institutional level. • R Hemadri: Design and architect DI&H system • SS Hong: I have designed and implemented the data ingestion and harmonization workflow both in SQL and as well as in Adeptia tool. • G Hripscak: I have helped with phenotyping and data harmonization related to OMOP. • D Jiao: Contributed code to the data harmonization process • JG Klann: Edited & provided comments on manuscript; participate in governance and data harmonization discussions; offer expertise on i2b2/ACT data model • K Kostka: OMOP SME / Coder for Phenotype & Data Acquisition + Data Harmonization & Ingestion WGs • AM Lee: Contribution to phenotype and data intake development, processing, and PCORNet SME. • HP Lehmann: I participate in daily Data Ingest & Harmonization meetings, have site data collection and curation responsibility for JHU, and have reviewed the MS. • L Lingrey: Developed TriNetX phenotype and extract logic. • RT Miller: Contributed code and data • M Morris: Phenotype small work group code and ACT to OMOP harmonization SME • SN Murphy: I have helped with phenotyping and data harmonization related to ACT. • K Natarajan: I have helped with phenotyping, analytics and code. I've also provided general input on OMOP. • MB Palchuk: representative of TriNetX - responsible for N3C submissions; active member of Phenotype and Data Acquisition working group • U Sheikh: My contribution has been to manage the NCATS infrastructure implementation and coordinate all of the vendors. • H Solbrig: Advisory and data transformation • S Visweswaran: Participating in decision-making in the Phenotype and Data Acquisition workstream. • A Walden: I have contributed content and edits to the manuscript. My role in N3C as Assistant Director is operationalizing the infrastructure for the N3C, project management planning around the governance activities, organizing groups and helping to design and implement user-centered design principles for website and portals. • KM Walters: Active member of Governance and Phenotype workstreams; revised manuscript text for those sections; reviewed and revised governance documentations; contributed to governance planning discussions; drafted data acquisition documentation; project management for data acquisition team. • GM Weber: Participated in decision-making in the data quality group and patient hashing meetings. • XTanner Zhang: I am an active member responsible for design and implementation in data ingestion and harmonization workstream. • RL Zhu: Data Ingestion & Harmonization Workstream • AT Girvin: I worked on the design and implementation of the Palantir platform for the N3C consortium, along with the first set of analyses done within it. • N Qureshi: I led the design and implementation of the Palantir Analytic platform for the N3C consortium, along with the first set of analyses done within it. • MG Kurilla: I have led the coordination of NIH programs • SG Michael: As the CIO of NCATS i help support the entire N3C initiative from a design, implementation, operation and standpoint. • LM Portilla: I have led the regulatory coordination with institutions and supported development with the governance workstream. • LM Portilla: I drafted the Data Transfer Agreements and Data Use Agreements for N3C • JL Rutter: I have led the regulatory implementation at NIH and provided programmatic vision and oversight. • CP Austin: Participated in decision making, scoping, and design of the project • KR Gersing: I have led the implementation of the architecture at NCATS as well as co-led the entire N3C effort across all workstreams. • S Al-Shukri: attending multiple workstream meetings and webinars, contributing data • A Alaoui: COntributing to data governance principles and N3C data access • B Amor: Engineering and workflow lead for the Foundry software backing the N3C platform • A Baghal: Contributing data from institutional EDW. • PD Banning: Focused on the laboratory data specific comments and plans for analysis; including surveillance for impending inflammatory tipping point leading to cytokine storm syndrome/hemodynamics instability and other ways in which laboratory values can help COVID-19 patient management. • EM Barbour: Contributor of data to the N3C consortium • MJ Becich: I lead (as chair) the iEC lead team and connect CTSA hubs to CD2H and N3C. I am a collaborator in all five N3C workstream actively exchanging ideas and expertise. A team I lead is also contributing the ACT COVID phenotype and data pipelines for other N3C sites. • A Beheshti: Provided edits to the manuscript. • GR Bernard: Significant data contribution. • S Bhattacharyya: Contributions in developing clinical scenarios •

L Boulware: Document review, contribution to ideas on use cases • s bozzette: Participated in governance and clinical groups • DE Brown: Provided a series of inputs to the draft DUA that were both wording suggestions and areas requiring further delineation and also provided input on how synthetic data should be considered in the DUA. • JB Buse: Reviewed and edited content of the manuscript, provided advice and assistance in setting up the collaborative and work teams • BJ Bush: contribution of data and revision of governance documents • TJ Callahan: Provide detailed editing of manuscript, converted figures into tables, and helped with generation/formatting/maintenance of all citations via PaperPile • TR Campion: Contributing data from New York City • E Casiraghi: code contribution, statistical analysis of patient data, deep learning for text context classification, deep learning for image classification, radiomics • AA Chaudhry: contributed to the review of the manuscript and contributed to the discussion section of the manuscript, • G Chen: participating in Collaborative Analytics • A Chen: In addition to participating in workstream calls and discussions, I added a "patient matching" use case under the enabled analyses section in the paper, for which I am working on connecting N3C analysis results to free real-time clinical collaboration tools for rapid and wide dissemination. • GD Clifford: contributing data • MP Coffee: I edited and added additions related to the clinical applications of this consortium. • T Conlin: port schema to open source database; minor edits • C Cook: Program Manager for CD2H, providing key operational support for the N3C initiative. • KA Crandall: Contributions include participation in both the Phenotypes & Data Acquisition and Collaborative Analytics workflows, local data governance processing and data contribution, and manuscript review and editing. • M Deacy: Project Management • RR Dietz: I am the key scheduler for all official N3C meetings and webinars. I am also operational support for N3C meetings. • NJ Dobbins: I'm the lead developer of Leaf and will be leading Leaf implementation and adoption for N3C deployment, as well as contributing to overall design and code. • PL Elkin: I have worked on the Analytic methods and for that workstream. I have worked with the data ingestion group and provided input to many of the other workstreams. I have edited the manuscript and provided content for the Codification and ReCodification Section of the manuscript. I have worked with a number of hubs to ensure that their data contributions include codified lab test names, clinical problems in SNOMED CT and HPO and drugs in RxNorm.. • PJ Embi: I have contributed to the N3C consortium by contributing to and/or lead key elements of the governance and operational data management processes, both for the limited and synthetic data workstreams, including chairing a governance sub-committee, enabling honest-broker functions, contributing to the creation of governance documents, and assisting with overall decision making. I also contributed to the critical review and editing of the manuscript. • JC Facelli: As PI of the Utah CCTS N3C project we have contributed data to N3C • K Fecho: In reviewing the guidelines for consortial authors/contributors, I realize that I am obliged to report my level of involvement. Specifically, I contributed text on the NCATS Biomedical Data Translator program to the N3C paper. I have not been an active contributor to the N3C consortium itself, but rather I have been passively observing the consortium's work and cross-referencing it in related initiatives such as a CTSA administrative supplement that is under review. • X Feng: Contributing code. • RE Foraker: Contributing data and drafting manuscript • TS Gal: Contributing data. • L Ge: I reviewed and edited the manuscript and participated work groups • G Golovko: Organized and facilitated data transfer for UTMB CTSA COVID19 Dataset, as well as prepared internal governance documents to get clearance for N3C data ingestion • R Gouripeddi: Contribute data, paper review. • CS Greene: Writing, Review and Editing • S Gupta: governance documents • A Gupta: participating in decision making, document draft • JG Hajagos: I provided sample OHDSI 5.3 files and I have been developing analytic queries in the Unite enclave for Covid patients based on analyses run at Stony Brook. • DA Hanauer: Contributing data and providing expertise regarding data acquisition for COVID-related data for PCORnet, which is being used to populate N3C. • JRichard Harper: I've been an active member of Phenotype and collaborative analytics workstreams and revised the methods manuscript text for those sections; reviewed and revised other documents and contributed to planning. etc • NL Harris: I helped edit the manuscript. • PA Harris: I met with Chris Chute about lessons learned on governance from AoU program and also participated in numerous meetings with iEC Leads discussing the N3C platform and approach. • MR Hassan: Contributing to the research of tools and how they can/will help researchers. Providing support to implement tools. • Y He: I contribute to the data modeling and standardization (esp. using ontology) and graph representation. I also contribute to the acute kidney injury (AKI) use case study. • EL Hill: Active member of the Collaborative Analytics team; Read the paper and added a few minor changes, and have

participated on Slack; hoping to be more involved going forward! • ME Hoatlin: Read and edited manuscript based on subject matter expertise in data analysis and virology • KL Holmes: Participation in decision making, contributions to documents and other materials, role in resources, infrastructure, metadata, and information. • L Hughes: writing review and editing role • RS Jawa: Edited the manuscript, our institution will be contributing data. • G Jiang: I contributed to creating the PCORNet CDM to OMOP CDM mappings and minor editing on manuscript. • X Jing: I review, edit the manuscript and contribute conceptual ideas to figures and graphs. • mP Joachimiak: Contributed the KG-COVID-19 knowledge graph with team, participated in discussions about machine learning and platform design. • SG Johnson: Participated in N3C Governance and Analytics workgroups, reviewed manuscript • R Kamaleswaran: Participated in discussions for defining analytics environment and underlying data engineering. • TGeorge Kannampallil: contributing to the data and its management • AS Kanter: contributed experience and data to assist in data normalization and discussing terminology modeling • R Kavuluru: I am part of the NLP and analytics streams of N3C to provide feedback on ML/DL/NLP aspects. I reviewed the paper and focused on changes in the "Collaborative Analytics" section. I will continue to make further changes as needed. • K Khanipov: Contributed in editing the manuscript and as a data submitter to N3C • H Kharrazi: contributed to the review of the manuscript and drafted a section on population health analytics • D Kim: Porting code for data extraction from ACT (i2b2) • BM Knosp: contributing data • A Krishnan: Design • T Kurc: contributing to data and tools from Stony Brook University • AM Lai: Contributing to the design of the synthetic data pipeline and participating in the data governance workstream. • CG Lambert: Contributing institutional data, participating in decision-making, advising on OHDSI implementation • M Larionov: I suggested edits to the paper to improve clarity. Also fixed minor grammar errors. • SB Lee: Participating in discussions and idea generation. Help with edits • MD Lesh: I will lead the effort to provide the Syntegra synthetic data engine to create and validate synthetic versions of COVID clinical data • O Lichtarge: Contributing to consortial discussions and editing • H Liu: Contributing to collaborative analytics working stream, leading NLP subgroup. • S Liu: I am the major contributor of the N3C cloud NLP engine: OHNLP MedTagger. I participated in decision making regarding its design on NLP architecture and algorithms. • J Liu: Phenotype NLM SNOMED value sets • JJ Loomba: Contributions have included conceptualization and definition of data access standards, as well as discussions are around infrastructure and methodology. • SK Mallipattu: Read the manuscript and our institution contributed to data and code for the N3C • CK Mamillapalli: Editing and review of manuscript. • A Manna: Technical lead of UNITE (N3C Enclave) environment • CE Mason: Data and code contribution and overall design • JP Mathew: Will Provide data; Will edit manuscript as needed • JC McClay: Document editing, data contribution, NIGMS CTR consortium governance • JA McMurry: Paper contributions: conceptualization, writing, editing, figures; N3C contributions: operational data integration. • PP Mehta: manuscript editing and review • O Mendelevitch: I will be responsible for designing and implementing the Syntegra synthetic data engine for COVID19 clinical data. • S Meystre: Lead the local contribution of clinical data, helped pilot the administrative enrollment process and contributed to the manuscript review • RA Moffitt: Code for Data curation, analysis, and visualization. Manuscript writing. • JH Moore: Contributed to the analytics design through weekly meetings, offered advice about how to maximize the effectiveness of the analytics, worked with the team at Penn Medicine to secure the approvals needed to share data. • H Morizono: HM contributed in a writing review and editing role, and as a data submitter role to N3C • CJ Mungall: Provision of COVID-19 Knowledge Graph • MC Munoz-Torres: Contributing to governance documents, contributing to design and deployment of N3C website, contributing to design and implementation of Project Management Pipeline, contributing with Project Management. • AJ Neumann: Contributing to design and implementation of Project Management Pipeline, contributing with Project Management of Workstreams and Sub Groups. • X Ning: I have contributed to the manuscript draft by providing comments and edits. • JE Nyland: Contributing to writing review and editing. • L O'Keefe: Project Management for Phenotyping/Data Acquisition group • A O'Malley: Engineering and configuring the Palantir platform • ST O'Neil: Infrastructure coordination (fielding suggestions and questions to/from Palantir team and others), training • JS Obeid: coordination of data provision effort and regulatory approval at our institution, manuscript contribution and edits, data team supervision at our institution. • EL Ogburn: thinking about how N3C can inform RCTs, introducing causal language where needed in the paper • J Phuong: I helped navigate local data governance at UW, get approvals for the external IRB, extract and ingest the UW COVID OMOP limited data set into N3C, and helped co-champion and drive clinical scenarios and data

analytics developments. • J Posada: Aid in the efforts to prepare and contribute Stanford data to N3C on behalf of our CTSA • P Prasanna: Contributed to manuscript, part of Stony Brook University team working on curation of certain data elements • F Prior: Participating in work groups, implementing a project to link N3C data to COVID image data in TCIA, minor contributions to the manuscript. • J Prosser: I provide technical expertise with secure architectures and infrastructure • ALienau Purnell: participating in decision making and design • A Rahnavard: Contributing in design, analyses, code, and methods development • H Ramadas: I contributed code and logic to the modeling efforts, helped set up the Unite infrastructure and trained users. • JT Reese: I contributed original software and biological expertise (software engineering role, advisory role, statistical analysis role) • JL Robinson: Commented and provided editorial suggestions in the manuscript • DL Rubin: I reviewed the paper and participated in discussions and in decision-making • CD Rutherford: Contributions of governance and design reviews and documentation for data analytics. • EM Sadhu: I have contributed to various CD2H & N3C related discussions and am preparing the data for submission from our site. • A Saha: Contributing code • MMorrison Saltz: I have been instrumental in clinical support to develop code used to classify COVID subjects, their associated co-morbidities, treatment, medications and outcomes. This code will be used on the aggregated N3C data to perform similar analyses. In addition I have been participating in the Governance process, ensuring that access to the data be fair and equitable for all, and as accessible as possible to drive the best possible science around this pandemic. • T Schaffter: Organizing the EHR COVID-19 DREAM Challenge • TKL Schleyer: Detailed edits of draft of 6/15/2020. • S Setoguchi: I have critically reviewed the manuscript and provided edits/suggestions. I am leading NJ ACTS (Rutgers CTSA's) effort to contribute our EHR data to the consortium, • NH Shah: Lead efforts to prepare and contribute Stanford data to N3C on behalf of our CTSA • N Sharafeldin: Read-proofed and provided edits to the manuscript draft • E Sholle: Contributed to data harmonization workflow by providing AMC-OMOP stakeholder perspective • JC Silverstein: Participation in governance discussion and related documents and may be responsible for data contribution, and our team engagement with i2b2 to OMOP translation. • A Solomonides: Data contribution. I am very interested in creating a lay summary of the paper. Also using the paper for Data Governance permissions at NorthShore. • J Solway: Participating in data partnership and governance committee and edited manuscript • J Su: Participating in decision-making • V Subbian: I am part of two N3C workstreams (phenotype and analytics). I reviewed and provided feedback on the entire manuscript, including critical feedback on ethical issues related data sharing, governance, and analytics. • H Tak: Expansion of research field and data deployment • BW Taylor: I reviewed and edited the manuscript, participated in discussions for defining analytics environment, participating in work groups • AE Thessen: I contributed during governance meetings and editing manuscript text. • JA Thomas: I contributed with data collection, data curation, data quality assurance, data transformation, writing review and editing. • U Topaloglu: I am the local PI for the N3C and waiting for PCORNet model to be finalized and also participate in the workstream • DR Unni: Involved in integration of NCATS Data Translator resources into N3C. CRO:0000040, CRO:0000041, CRO:0000072 • JT Vogelstein: Made a number of minor suggestions to the text, have been in conversation with many other co-authors about various concepts presented in manuscript. "writing review and editing role", "modifier role" • AM Volz: Provide communications (weekly updates, website content, edits to manuscript, SOPs, FAQs, meeting notes, etc.) for the N3C consortium. • DA Williams: data and manuscript review/edit • KM Wilson: Palantir Unite platform trainer and assistance with setting up the Palantir environment • H Xu: Editing the online document and co-leading the N3C Collaborative Analytics NLP subgroup • CB Xu: I am on the N3C Meetings for Collaborative Analytics, Data Ingestion and Harmonization, Portals and Dashboards. My team at University of Wisconsin-Madison created a COVID reporting dashboard for our partner institution UW Health (data modeling role, data transformation role) • Y Yan: I reviewed and edited the manuscript and contributed codes for COVID phenotyping. • E Zak: I contributed data and gathered research resources and datasets. • L Zhang: contributing to design or governance documents • C Zhang: suggested edits • J Zheng: participating in work groups. reviewing and editing manuscript

ACKNOWLEDGEMENTS STATEMENT

We acknowledge the Oregon Clinical and Translational Research Institute for their guidance and review of N3C plans and regulatory processes as they unfolded.

REFERENCES

1 COVID-19 Map. Johns Hopkins Coronavirus Resource Center. https://coronavirus.jhu.edu/map.html (accessed 12 Jul 2020).

2 Kissler SM, Tedijanto C, Goldstein E, *et al.* Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science* 2020;**368**:860–8.

3 Williamson EJ, Walker AJ, Bhaskaran K, *et al.* OpenSAFELY: factors associated with COVID-19 death in 17 million patients. *Nature* Published Online First: 8 July 2020. doi:10.1038/s41586-020-2521-4

4 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**:147–52.

5 Fleurence RL, Curtis LH, Califf RM, *et al.* Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc* 2014;**21**:578–82.

6 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**:574–8.

7 Findlay S. *The FDA's Sentinel Initiative*. Project HOPE 2015.

8 Topaloglu U, Palchuk MB. Using a Federated Network of Real-World Data to Optimize Clinical Trials Operations. *JCO Clin Cancer Inform* 2018;**2**:1–10.

9 Brat GA, Weber GM, Gehlenborg N, *et al.* International Electronic Health Record-Derived COVID-19 Clinical Course Profiles: The 4CE Consortium. Infectious Diseases (except HIV/AIDS). 2020. doi:10.1101/2020.04.13.20059691

10 Carton TW, Marsolo K, Block JP. *PCORnet COVID-19 Common Data Model Design and Results*. 2020. doi:10.5281/zenodo.3897398

11 Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *N Engl J Med* 2019;**380**:1347–58.

12 Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng* 2018;**2**:719–31.

13 Kramer WG, Perentesis G, Affrime MB, *et al.* Pharmacokinetics of dilevalol in normotensive and hypertensive volunteers. *Am J Cardiol* 1989;**63**:7I – 11I.

14 Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med* 2016;**375**:1216–9.

15 Wang Y, Zhao Y, Therneau TM, *et al.* Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records. *J Biomed Inform* 2020;**102**:103364.

16 Li T, Sahu AK, Talwalkar A, *et al.* Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Process Mag* 2020;**37**:50–60.

17 Zerka F, Barakat S, Walsh S, *et al.* Systematic Review of Privacy-Preserving Distributed Machine Learning From Federated Databases in Health Care. *JCO Clin Cancer Inform* 2020;**4**:184–200.

18 Yang Q, Liu Y, Chen T, *et al.* Federated Machine Learning: Concept and Applications. *ACM Trans Intell Syst Technol* 2019;**10**:1–19.

19 Brisimi TS, Chen R, Mela T, *et al.* Federated learning of predictive models from federated Electronic Health Records. *Int J Med Inform* 2018;**112**:59–67.

20 Mehra MR, Ruschitzka F, Patel AN. Retraction-Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. *Lancet* Published Online First: 5 June 2020. doi:10.1016/S0140-6736(20)31324-6

21 Mehra MR, Desai SS, Kuy S, *et al.* Retraction: Cardiovascular Disease, Drug Therapy, and

Mortality in Covid-19. N Engl J Med. DOI: 10.1056/NEJMoa2007621. *N Engl J Med* Published Online First: 4 June 2020. doi:10.1056/NEJMc2021225

22 Wilkinson MD, Dumontier M, Aalbersberg IJJ, *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;**3**:160018.

23 CTSA Program Hubs | National Center for Advancing Translational Sciences. National Center for Advancing Translational Sciences. 2015.https://ncats.nih.gov/ctsa/about/hubs (accessed 13 Jun 2020).

24 *Phenotype_Data_Acquisition*. Github https://github.com/National-COVID-Cohort-Collaborative/Phenotype_Data_Acquisition (accessed 20 Jun 2020).

25 National Center for Advancing Translational Sciences. National Center for Advancing Translational Sciences. https://ncats.nih.gov/ (accessed 7 Jun 2020).

26 CTSA Program Hubs | National Center for Advancing Translational Sciences. National Center for Advancing Translational Sciences. 2015.https://ncats.nih.gov/ctsa/about/hubs (accessed 7 Jun 2020).

27 CD2H. https://ctsa.ncats.nih.gov/cd2h/ (accessed 7 Jun 2020).

28 All of Us Research Hub. https://www.researchallofus.org/; (accessed 18 Jun 2020).

29 Human Tumor Atlas Network. Human Tumor Atlas Network. https://humantumoratlas.org/ (accessed 18 Jun 2020).

30 Grayson S, Suver C, Wilbanks J, *et al.* Open Data Sharing in the 21st Century: Sage Bionetworks' Qualified Research Program and Its Application in mHealth Data Release. Published Online First: 11 December 2019. doi:10.2139/ssrn.3502410

31 All of Us Research Hub. https://www.researchallofus.org/ (accessed 18 Jun 2020).

32 Human Tumor Atlas Network. Human Tumor Atlas Network. https://humantumoratlas.org/ (accessed 18 Jun 2020).

33 Regulatory & Ethics Toolkit. https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/ (accessed 18 Jun 2020).

34 Data Access Compliance Office. https://icgc.org/daco (accessed 18 Jun 2020).

35 i2b2: Informatics for Integrating Biology & the Bedside. https://www.i2b2.org/ (accessed 18 Jun 2020).

36 Regulatory & Ethics Toolkit. https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/ (accessed 18 Jun 2020).

37 Data Access Compliance Office. https://icgc.org/daco (accessed 18 Jun 2020).

38 i2b2: Informatics for Integrating Biology & the Bedside. https://www.i2b2.org/ (accessed 18 Jun 2020).

39 govinfo. https://www.govinfo.gov/app/details/CFR-2011-title45-vol1/CFR-2011-title45-vol1-sec164-514 (accessed 18 Jun 2020).

40 HIPAA Privacy Rule and Its Impacts on Research. https://privacyruleandresearch.nih.gov/pr_08.asp (accessed 18 Jun 2020).

41 Certificates of Confidentiality (CoC) - Human Subjects | grants.nih.gov. https://grants.nih.gov/policy/humansubjects/coc.htm (accessed 15 Jun 2020).

42 Office for Human Research Protections (OHRP). The Revised Common Rule's Cooperative Research Provision (45 CFR. Published Online First: 1 August 2019.https://www.hhs.gov/ohrp/regulations-and-policy/single-irb-requirement/index.html (accessed 20 Jun 2020).

43 SMART IRB | National IRB Reliance Initiative. https://smartirb.org/ (accessed 14 Apr 2020).

44 Sprague ER. ORCID. *J Med Libr Assoc* 2017;**105**:207.

45 Raab GM, Nowok B, Dibben C. Guidelines for Producing Useful Synthetic Data. arXiv [stat.AP]. 2017.http://arxiv.org/abs/1712.04078

46 Snoke J, Raab GM, Nowok B, *et al.* General and specific utility measures for synthetic data. *J R Stat Soc A* 2018;**181**:663–88.

47 govinfo. https://www.govinfo.gov/app/details/CFR-2011-title45-vol1/CFR-2011-title45-vol1-sec164-514 (accessed 13 Jun 2020).

48 Office for Civil Rights (OCR). Methods for De-identification of PHI. Published Online First: 6 November 2015.https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html (accessed 20 Jun 2020).

49 HIPAA Privacy Rule and Its Impacts on Research. https://privacyruleandresearch.nih.gov/pr_08.asp (accessed 13 Jun 2020).

50 Haendel M, Su A, McMurry J. *FAIR-TLC: Metrics to Assess Value of Biomedical Digital Repositories: Response to RFI NOT-OD-16-133*. 2016. doi:10.5281/zenodo.203295

51 Welcome to the Contributor Attribution Model — Contributor Attribution Model documentation. https://contributor-attribution-model.readthedocs.io/en/latest/ (accessed 20 Jun 2020).

52 Transitive Credit and JSON-LD. *Journal of Open Research Software* 2015;**3**:14.

53 Centers for Disease Control and Prevention. ICD-10-CM Official Coding Guidelines - Supplement Coding encounters related to COVID-19 Coronavirus Outbreak. 2020.https://cdc.gov/nchs/data/icd/interim-coding-advice-coronavirus-March-2020-final.pdf

54 Center for Disease Control and Prevention. ICD-10-CM Official Coding and Reporting Guidelines April 1, 2020 through September 30, 2020. 2020.https://www.cdc.gov/nchs/data/icd/COVID-19-guidelines-final.pdf

55 PCORnet® COVID-19 Common Data Model Launched, Enabling Rapid Capture of Insights on Patients Infected with the Novel Coronavirus | The National Patient-Centered Clinical Research Network. The National Patient-Centered Clinical Research Network. 2020.https://pcornet.org/news/pcornet-covid-19-common-data-model-launched-enabling-rapid-capture-of-insights/ (accessed 8 Jun 2020).

56 Burn E, You SC, Sena AG, *et al.* An international characterisation of patients hospitalised with COVID-19 and a comparison with those previously hospitalised with influenza. *medRxiv* Published Online First: 25 April 2020. doi:10.1101/2020.04.22.20074336

57 SARS-CoV-2 and COVID-19 related LOINC terms – LOINC. LOINC. https://loinc.org/sars-cov-2-and-covid-19/ (accessed 8 Jun 2020).

58 *Phenotype_Data_Acquisition*. Github https://github.com/National-COVID-Cohort-Collaborative/Phenotype_Data_Acquisition (accessed 21 Jun 2020).

59 *Phenotype_Data_Acquisition*. Github https://github.com/National-COVID-Cohort-Collaborative/Phenotype_Data_Acquisition (accessed 21 Jun 2020).

60 *National COVID Cohort Collaborative*. Github https://github.com/National-COVID-Cohort-Collaborative (accessed 14 Jun 2020).

61 *Phenotype_Data_Acquisition*. Github https://github.com/National-COVID-Cohort-Collaborative/Phenotype_Data_Acquisition (accessed 8 Jun 2020).

62 Weber GM, Murphy SN, McMurry AJ, *et al.* The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. *J Am Med Inform Assoc* 2009;**16**:624–30.

63 Patient-Centered Outcomes Research Institute (PCORI). https://www.pcori.org/ (accessed 12 Apr 2020).

64 OHDSI – Observational Health Data Sciences and Informatics. https://ohdsi.org/ (accessed 12 Apr 2020).

65 TriNetX. TriNetX. https://www.trinetx.com/ (accessed 12 Apr 2020).

66 PCORnet. PCORnet Common Data Model v5.1 Specification. 2019. https://pcornet.org/wp-content/uploads/2019/09/PCORnet-Common-Data-Model-v51-2019_09_12.pdf

67 Box. https://pitt.app.box.com/s/qoj5afssw4oz3v27ipmfidhitmgya9nt (accessed 21 Jun 2020).

68 *CommonDataModel*. Github https://github.com/OHDSI/CommonDataModel (accessed 21 Jun 2020).

69 *Phenotype_Data_Acquisition*. Github https://github.com/National-COVID-Cohort-Collaborative/Phenotype_Data_Acquisition (accessed 17 Jun 2020).

70 Chute CG. Clinical classification and terminology: some history and current observations. *J Am Med Inform Assoc* 2000;**7**:298–303.

71 Haendel MA, Chute CG, Robinson PN. Classification, Ontology, and Precision Medicine. *N Engl J Med* 2018;**379**:1452–62.

72 Health Level 7 (HL7). Fast Healthcare Interoperability Resources (FHIR). https://www.hl7.org/fhir/ (accessed 21 May 2020).

73 Chute CG, Huff SM. The Pluripotent Rendering of Clinical Data for Precision Medicine. *Stud Health Technol Inform* 2017;**245**:337–40.

74 Center for Data to Health (CD2H). https://ctsa.ncats.nih.gov/cd2h/ (accessed 12 Apr 2020).

75 Health Level 7 (HL7). Vulcan Accelerator Home - Vulcan Accelerator - Confluence. https://confluence.hl7.org/display/VA/Vulcan+Accelerator+Home (accessed 21 May 2020).

76 CDM v5.3.1. https://ohdsi.github.io/CommonDataModel/cdm531.html (accessed 21 Jun 2020).

77 Kahn MG, Batson D, Schilling LM. Data model considerations for clinical effectiveness researchers. *Med Care* 2012;**50 Suppl**:S60–7.

78 Ogunyemi OI, Meeker D, Kim H-E, *et al.* Identifying appropriate reference data models for comparative effectiveness research (CER) studies based on data from clinical information systems. *Med Care* 2013;**51**:S45–52.

79 HHS Office of the National Coordinator. Common Data Model Harmonization | HealthIT.gov. https://www.healthit.gov/topic/scientific-initiatives/pcor/common-data-model-harmonization-cdm (accessed 7 Jun 2020).

80 CDISC. BRIDG. https://www.cdisc.org/standards/domain-information-module/bridg (accessed 13 Apr 2020).

81 *Data-Ingestion-and-Harmonization*. Github https://github.com/National-COVID-Cohort-Collaborative/Data-Ingestion-and-Harmonization (accessed 14 Jun 2020).

82 Banga J, Tyagi MR, Hans S. B2B Integration Platform for next-gen business connectivity | Adeptia. https://adeptia.com/ (accessed 13 Apr 2020).

83 Kahn MG, Brown JS, Chun AT, *et al.* Transparent reporting of data quality in distributed data networks. *EGEMS (Wash DC)* 2015;**3**:1052.

84 Khare R, Utidjian L, Ruth BJ, *et al.* A longitudinal analysis of data quality in a large pediatric data research network. *J Am Med Inform Assoc* 2017;**24**:1072–9.

85 Weiskopf NG, Hripcsak G, Swaminathan S, *et al.* Defining and measuring completeness of

electronic health records for secondary use. *J Biomed Inform* 2013;**46**:830–6.

86 Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc* 2013;**20**:144–51.

87 Zozus M. *The Data Book: Collection and Management of Research Data*. CRC Press, Taylor & Francis Group 2017.

88 Kahn Mg Eliason. Quantifying clinical data quality using relative gold standards. *AMIA Annu Symp Proc* 2010; **2010**:356-360.

89 Execute and View Data Quality Checks on OMOP CDM Database. https://ohdsi.github.io/DataQualityDashboard/ (accessed 20 Jun 2020).

90 PCORnet: The National Patient-Centered Clinical Research Network. PCORnet Data Checks v8. The National Patient-Centered Clinical Research Network. 2020.https://pcornet.org/wp-content/uploads/2020/03/PCORnet-Data-Checks-v8.pdf (accessed 20 Jun 2020).

91 Wikipedia contributors. Smoke testing (software). Wikipedia, The Free Encyclopedia. 2020.https://en.wikipedia.org/w/index.php?title=Smoke_testing_(software)&oldid=962025059 (accessed 12 Jul 2020).

92 Hans S, Adeptia. Explore B2B Process Automation Solutions for Integration Needs. https://adeptia.com/solutions/b2b-process-automation (accessed 20 Jun 2020).

93 ETL Data Integration Software for Connecting Business Data. https://adeptia.com/products/etl-data-integration (accessed 20 Jun 2020).

94 ATLAS. https://atlas.ohdsi.org/#/home (accessed 20 Jun 2020).

95 Creates Descriptive Statistics Summary for an Entire OMOP CDM Instance. https://ohdsi.github.io/Achilles/ (accessed 20 Jun 2020).

96 Eagleton MJ, Kashyap VS. Introduction. *J Vasc Surg* 2020;**72**:e4–5.

97 Dong X, Li J, Soysal E, *et al.* COVID-19 TestNorm - A tool to normalize COVID-19 testing names to LOINC codes. *J Am Med Inform Assoc* Published Online First: 22 June 2020. doi:10.1093/jamia/ocaa145

98 Lane J, Schur C. Balancing access to health data and privacy: a review of the issues and approaches for the future. *Health Serv Res* 2010;**45**:1456–67.

99 FedRAMP.gov | FedRAMP.gov. https://www.fedramp.gov/ (accessed 21 Jun 2020).

100 Hripcsak G, Shang N, Peissig PL, *et al.* Facilitating phenotype transfer using a common data model. *J Biomed Inform* 2019;**96**:103253.

101 Swerdel JN, Hripcsak G, Ryan PB. PheValuator: Development and evaluation of a phenotype algorithm evaluator. *J Biomed Inform* 2019;**97**:103258.

102 Reps JM, Schuemie MJ, Suchard MA, *et al.* Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *J Am Med Inform Assoc* 2018;**25**:969–75.

103 J. Schuemie M, Soledad Cepede M, A. Suchard M, *et al.* How Confident Are We About Observational Findings in Health Care: A Benchmark Study. *Harvard Data Science Review* Published Online First: 31 January 2020. doi:10.1162/99608f92.147cc28e

104 Schuemie MJ, Ryan PB, Hripcsak G, *et al.* Improving reproducibility by using high-throughput observational studies with empirical calibration. *Philos Trans A Math Phys Eng Sci* 2018;**376**. doi:10.1098/rsta.2017.0356

105 Schuemie MJ, Hripcsak G, Ryan PB, *et al.* Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proc Natl Acad Sci U S A*

2018;**115**:2571–7.

106 Zhang XA, Yates A, Vasilevsky N, *et al.* Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery. *NPJ Digit Med* 2019;**2**. doi:10.1038/s41746-019-0110-4

107 Biomedical Data Translator Consortium. Toward A Universal Biomedical Data Translator. *Clin Transl Sci* 2019;**12**:86–90.

108 Biomedical Data Translator Consortium. The Biomedical Data Translator Program: Conception, Culture, and Community. *Clin Transl Sci* 2019;**12**:91–4.

109 Austin CP, Colvis CM, Southall NT. Deconstructing the Translational Tower of Babel. *Clin Transl Sci* 2019;**12**:85.

110 Biolink Model. https://biolink.github.io/biolink-model (accessed 21 Jun 2020).

111 *kg-covid-19*. Github https://github.com/Knowledge-Graph-Hub/kg-covid-19 (accessed 20 Jun 2020).

112 Dobbins NJ, Spital CH, Black RA, *et al.* Leaf: an open-source, model-agnostic, data-driven web application for cohort discovery and translational biomedical research. *J Am Med Inform Assoc* 2020;**27**:109–18.

113 Brito JJ, Li J, Moore JH, *et al.* Recommendations to enhance rigor and reproducibility in biomedical research. *Gigascience* 2020;**9**. doi:10.1093/gigascience/giaa056

114 Walonoski J, Kramer M, Nichols J, *et al.* Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *J Am Med Inform Assoc* 2018;**25**:230–8.

115 Baowaly MK, Lin C-C, Liu C-L, *et al.* Synthesizing electronic health records using improved generative adversarial networks. *J Am Med Inform Assoc* 2019;**26**:228–41.

116 Chen J, Chun D, Patel M, *et al.* The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. *BMC Med Inform Decis Mak* 2019;**19**:44.

117 Hayes J, Melis L, Danezis G, *et al.* LOGAN: Membership Inference Attacks Against Generative Models. arXiv [cs.CR]. 2017.http://arxiv.org/abs/1705.07663

118 Erez L. … dedicated computer clients specially programmed to generate synthetic non-reversible electronic data records based on real-time electronic querying and methods of …. US Patent 10,235,537. 2019.https://patents.google.com/patent/US10235537B2/en

119 Foraker R, Mann DL, Payne PRO. Are Synthetic Data Derivatives the Future of Translational Medicine? *JACC Basic Transl Sci* 2018;**3**:716–8.

120 Mehra MR, Desai SS, Ruschitzka F, *et al.* RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. *Lancet* Published Online First: 22 May 2020. doi:10.1016/S0140-6736(20)31180-6

121 Mehra MR, Desai SS, Kuy S, *et al.* Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19. *N Engl J Med* 2020;**382**:e102.

122 Head ML, Holman L, Lanfear R, *et al.* The extent and consequences of p-hacking in science. *PLoS Biol* 2015;**13**:e1002106.

123 Shickel B, Tighe PJ, Bihorac A, *et al.* Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE J Biomed Health Inform* 2018;**22**:1589–604.

124 Luo Y, Wang F, Szolovits P. Tensor factorization toward precision medicine. *Brief Bioinform* 2017;**18**:511–4.

125 Thompson AE, Ranard BL, Wei Y, *et al.* Prone Positioning in Awake, Nonintubated Patients With COVID-19 Hypoxemic Respiratory Failure. *JAMA Intern Med* Published Online First: 17 June 2020. doi:10.1001/jamainternmed.2020.3030

126 Mehta P, McAuley DF, Brown M, *et al.* COVID-19: consider cytokine storm syndromes and immunosuppression. *Lancet* 2020;**395**:1033–4.

127 Suo Q, Ma F, Yuan Y, *et al.* Deep Patient Similarity Learning for Personalized Healthcare. *IEEE Trans Nanobioscience* 2018;**17**:219–27.

128 Belhadjer Z, Méot M, Bajolle F, *et al.* Acute heart failure in multisystem inflammatory syndrome in children (MIS-C) in the context of global SARS-CoV-2 pandemic. *Circulation* Published Online First: 17 May 2020. doi:10.1161/CIRCULATIONAHA.120.048360

129 Lin KJ, Rosenthal GE, Murphy SN, *et al.* External Validation of an Algorithm to Identify Patients with High Data-Completeness in Electronic Health Records for Comparative Effectiveness Research. *Clin Epidemiol* 2020;**12**:133–41.

130 Kharrazi H, Lasser EC, Yasnoff WA, *et al.* A proposed national research and development agenda for population health informatics: summary recommendations from a national expert workshop. *J Am Med Inform Assoc* 2017;**24**:2–12.

131 Kharrazi H, Chi W, Chang H-Y, *et al.* Comparing Population-based Risk-stratification Model Performance Using Demographic, Diagnosis and Medication Data Extracted From Outpatient Electronic Health Records Versus Administrative Claims. *Med Care* 2017;**55**:789–96.

132 Williams DR, Cooper LA. COVID-19 and Health Equity-A New Kind of 'Herd Immunity'. *JAMA* Published Online First: 11 May 2020. doi:10.1001/jama.2020.8051

133 Glover RE, van Schalkwyk MC, Akl EA, *et al.* A framework for identifying and mitigating the equity harms of COVID-19 policy interventions. *J Clin Epidemiol* Published Online First: 8 June 2020. doi:10.1016/j.jclinepi.2020.06.004

134 Price-Haywood EG, Burton J, Fort D, *et al.* Hospitalization and Mortality among Black Patients and White Patients with Covid-19. *N Engl J Med* Published Online First: 27 May 2020. doi:10.1056/NEJMsa2011686

135 Millett GA, Jones AT, Benkeser D, *et al.* Assessing Differential Impacts of COVID-19 on Black Communities. *Ann Epidemiol* Published Online First: 14 May 2020. doi:10.1016/j.annepidem.2020.05.003

136 Gamache R, Kharrazi H, Weiner JP. Public and Population Health Informatics: The Bridging of Big Data to Benefit Communities. *Yearb Med Inform* 2018;**27**:199–206.

137 Obermeyer Z, Powers B, Vogeli C, *et al.* Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;**366**:447–53.

138 Cimino JJ, Ayres EJ, Remennik L, *et al.* The National Institutes of Health's Biomedical Translational Research Information System (BTRIS): design, contents, functionality and experience to date. *J Biomed Inform* 2014;**52**:11–27.

139 Hersh WR, Weiner MG, Embi PJ, *et al.* Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care* 2013;**51**:S30–7.

140 Hersh W, Cimino J, Payne PRO, *et al.* Recommendations for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2013;**1**:14. doi:10.13063/2327-9214.1018

141 EMBL-EBI launches COVID-19 Data Portal. https://www.ebi.ac.uk/about/news/press-releases/embl-ebi-launches-covid-19-data-portal (accessed 21 Jun 2020).

142 ELIXIR support to COVID-19 research | ELIXIR. https://elixir-europe.org/services/covid-19 (accessed 21 Jun 2020).

143 Chute CG. *National COVID Cohort Collaborative (N3C): A national resource for shared analytics*. 2020. doi:10.5281/zenodo.3902948