Journal of the American Medical Informatics Association, 28(3), 2021, 427–443
doi: 10.1093/jamia/ocaa196
Advance Access Publication Date: 17 August 2020
Research and Applications

Research and Applications

The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment

Melissa A. Haendel [1,2], Christopher G. Chute [3], Tellen D. Bennett [4], David A. Eichmann [5], Justin Guinney [6], Warren A. Kibbe [7], Philip R.O. Payne [8], Emily R. Pfaff [9], Peter N. Robinson [10], Joel H. Saltz [11], Heidi Spratt [12], Christine Suver [6], John Wilbanks [6], Adam B. Wilcox [13], Andrew E. Williams [14], Chunlei Wu [15], Clair Blacketer [16], Robert L. Bradford [9], James J. Cimino [17], Marshall Clark [9], Evan W. Colmenares [18], Patricia A. Francis [19], Davera Gabriel [19], Alexis Graves [20], Raju Hemadri,[21] Stephanie S. Hong,[19] George Hripscak,[22] Dazhi Jiao [19], Jeffrey G. Klann,[23] Kristin Kostka,[24] Adam M. Lee,[25] Harold P. Lehmann [19], Lora Lingrey [26], Robert T. Miller [27], Michele Morris,[28] Shawn N. Murphy,[29] Karthik Natarajan [30], Matvey B. Palchuk [26], Usman Sheikh,[21] Harold Solbrig [19], Shyam Visweswaran [28], Anita Walden [1,6], Kellie M. Walters,[9] Griffin M. Weber [31], Xiaohan Tanner Zhang [19], Richard L. Zhu [19], Benjamin Amor,[32] Andrew T. Girvin [32], Amin Manna,[32] Nabeel Qureshi,[32] Michael G. Kurilla,[33] Sam G. Michael,[34] Lili M. Portilla [35], Joni L. Rutter [36], Christopher P. Austin [34], Ken R. Gersing [21] and the N3C Consortium

[1]Oregon Clinical and Translational Research Institute, Oregon Health and Science University, Portland, Oregon, USA, [2]Translational and Integrative Sciences Center, Department of Molecular Toxicology, Oregon State University, Corvallis, Oregon, USA, [3]Schools of Medicine, Public Health, and Nursing, Johns Hopkins University, Baltimore, Maryland, USA, [4]Section of Informatics and Data Science, Department of Pediatrics, University of Colorado School of Medicine, University of Colorado, Aurora, Colorado, USA, [5]School of Library and Information Science, The University of Iowa, Iowa City, Iowa, USA, [6]Sage Bionetworks, Seattle, Washington, USA, [7]Duke University, Durham,North Carolina, USA, [8]Institute for Informatics, Washington University in St. Louis, Saint Louis,Missouri, USA, [9]North Carolina Translational and Clinical Sciences Institute (NC TraCS), University of North Carolina at Chapel Hill, Chapel Hill,North Carolina, USA, [10]Jackson Laboratory, Bar Harbor, Maine, USA, [11]Department of Biomedical Informatics, Stony Brook University, Stony Brook, New York, USA, [12]University of Texas Medical Branch, Galveston, Texas, USA, [13]University of Washington, Seattle, Washington, USA, [14]Tufts Medical Center Clinical and Translational Science Institute, Tufts Medical Center, Boston,Massachusetts, USA, [15]Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, California, USA, [16]Janssen Research and Development, LLC, Raritan, New Jersey, USA, [17]University of Alabama-Birmingham, Birmingham, Alabama, USA, [18]Department of Pharmaceutical Outcomes and Policy, University of North Carolina at Chapel Hill, Chapel Hill,North Carolina, USA, [19]Johns Hopkins University School of Medicine, Baltimore, Maryland, USA, [20]University of Iowa Institute for Clinical and Translational Science, The University of Iowa, Iowa City, Iowa, USA, [21]National Center for Advancing Translational Science, Bethesda, Maryland, USA, [22]Department of Biomedical Informatics, Columbia University, New York, New York, USA, [23]Harvard Medical School, Boston,Massachusetts, USA, [24]IQVIA, Durham,North Carolina, USA, [25]University of North Carolina at Chapel Hill, Chapel Hill,North Carolina, USA, [26]TriNetX, Cambridge,Massachusetts, USA, [27]Tufts Clinical and Translational Science Institute, Tufts University, Boston,Massachusetts, USA, [28]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh,Pennsylvania, USA, [29]Mass General Brigham, Boston,Massachusetts, USA, [30]Irving Medical Center, Columbia University, New York, New York, USA, [31]Department of Biomedical Informatics, Harvard Medical School, Boston,Massachusetts, USA, [32]Palantir Technologies, Palo Alto, California, USA, [33]Division of Clinical Innovation, National Center for Advancing Translational Science, Bethesda, Maryland, USA, [34]National Center for Advancing Translational Sci-

ences, National Institutes of Health, Bethesda, Maryland, USA, [35]Office of Strategic Alliances, National Center for Advancing Translational Sciences, National Institutes of Health, Bethesda, Maryland, USA and [36]Office of the Director, National Center for Advancing Translational Science, Bethesda, Maryland, USA

Co-authors:
Please see attached supplemental files for masthead and Contributing authors.

Corresponding Authors: Melissa A. Haendel, Linus Pauling Science Center, Corvallis, OR 97331, USA (melissa@tislab.org); Christopher G. Chute, Johns Hopkins University, 2024 E Monument St, Baltimore, MD 21287, USA (chute@jhu.edu)

Received 14 July 2020; Editorial Decision 29 July 2020; Accepted 14 August 2020

ABSTRACT

Objective: Coronavirus disease 2019 (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.

Materials and Methods: The Clinical and Translational Science Award 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.

Results: Organized in inclusive workstreams, 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.

Conclusions: The N3C has demonstrated that a multisite collaborative learning health network can overcome barriers to rapidly build a scalable infrastructure incorporating multiorganizational 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.

Key words: COVID-19, open science, clinical data model harmonization, EHR data, collaborative analytics, SARS-CoV-2

INTRODUCTION

Rationale

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) had infected 12.6 million people—and the novel coronavirus disease 2019 (COVID-19) had caused 562 000 deaths—worldwide as of July 11, 2020, according to Johns Hopkins University.[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 are limited within given organizations or consortia of local and regional organizations. Research consortia such as Accrual to Clinical Trials (ACT) Network,[4] National Patient-Centered Clinical Research Network (PCORnet),[5] Observational Health Data Sciences and Informatics (OHDSI),[6] the Food and Drug Administration'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 prespecified, 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 discov-

ery, 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 on the substantial resources (eg, CDM-specific data quality tools) 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 in which analyses must be fast, verifiable, and based on the latest data.[22]

National COVID Cohort Collaborative overview

The National COVID Cohort Collaborative (N3C) (covid.cd2h.org) aims to aggregate and harmonize EHR data across clinical organizations in the United States, and is a novel partnership that includes the Clinical and Translational Science Awards (CTSA) Program hubs (60 institutions), the National Center for Advancing Translational Science (NCATS), the Center for Data to Health (CD2H) and the community.[23] 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 4 CDMs, and development of a collaborative analytics platform to support deployment of novel algorithms of data aggregated from the United States. 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 on 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 US organization that wishes to contribute data. N3C also welcomes registered researchers from any country 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.

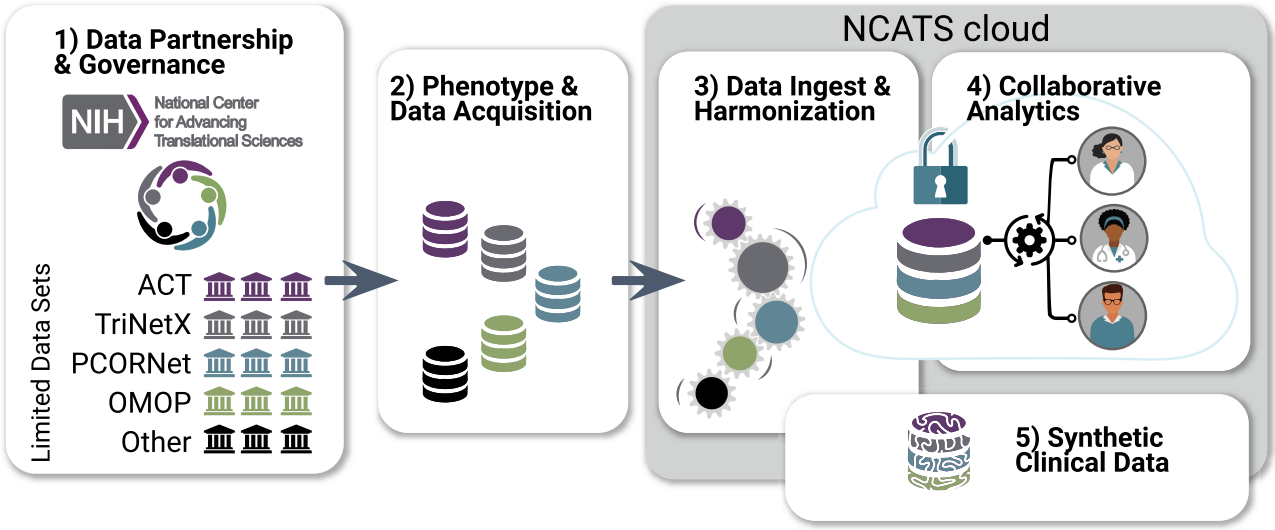
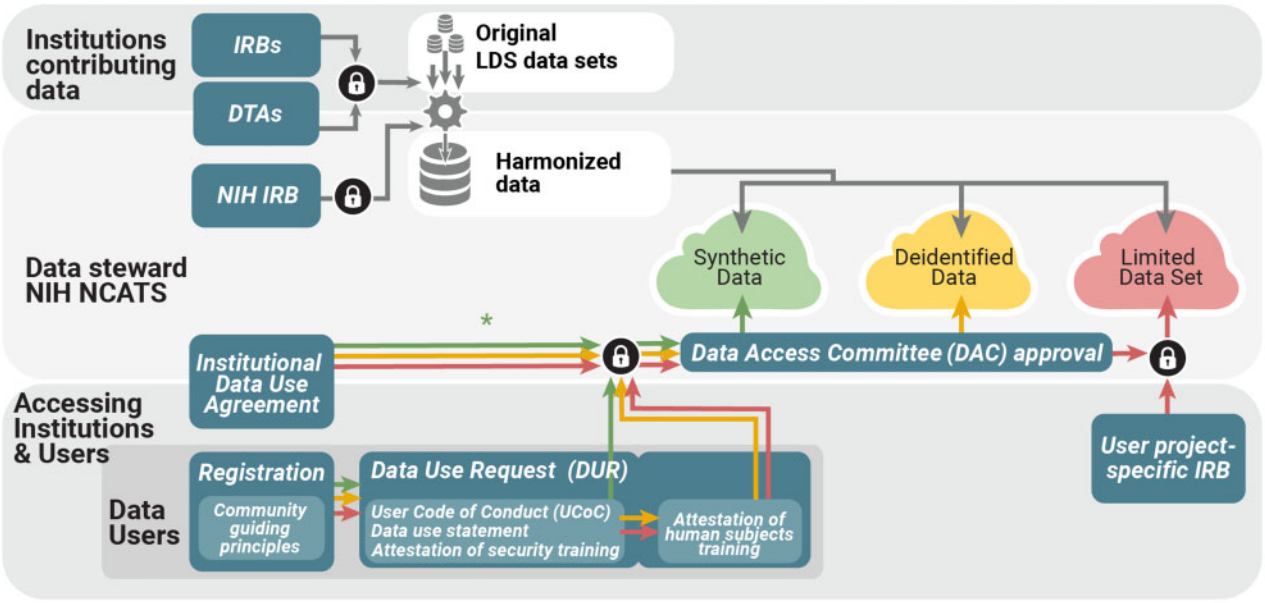


Figure 1. Establishing National COVID Cohort Collaborative (N3C) sociotechnical processes and infrastructure via community workstreams. Each workstream includes representatives from National Center for Advancing Translational Sciences (NCATS),[25] the Clinical and Translational Science Awards hubs,[23] the Center for Data to Health,[26] sites contributing data, and other members of the research community. (1) Data Partnership and Governance: This workstream designs governance and makes regulatory recommendations to National Institutes of Health (NIH) for their execution. Organizations sign a Data Transfer Agreement (DTA) with NCATS and may use the central institutional review board. (2) Phenotype and Data Acquisition: The community defines inclusion criteria for the N3C COVID-19 (coronavirus disease 2019) cohort and supports organizations in customized data export. (3) Data Ingest and Harmonization: Data reside within different organizations in different common data models. This workstream quality-assures and harmonizes data from different sources and common data models 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 Data Set are able to approximate analyses derived from original data, while enhancing shareable data outside the N3C Enclave. ACT: Accrual to Clinical Trials; OMOP: Observational Medical Outcomes Partnership; PCORnet: National Patient-Centered Clinical Research Network.

A



B

Data Type	Level 1 Synthetic Data (pending pilot)	Level 2 De-identified	Level 3 Limited Data Set
Description	Computational data derivative that statistically resembles the original data	Data stripped of 17 direct identifiers called out in the HIPAA Privacy Rule (with longitudinal data date-shifted to safeguard privacy)	Data stripped of 16 direct identifiers called out in the HIPAA Privacy Rule except dates and zip code
Capabilities			
Downloadable data	No*	No	No
Access Prerequisites			
Investigator Affiliation Requirement	Any academic or commercial research organizations*	Any academic or commercial research organizations	US academic or commercial research organizations
Data Use Agreement Signed by Home Org.	Required	Required	Required
Human Subjects Training	Not required	Required	Required
NIH Security Training	Required	Required	Required
Request Submission and Approval Steps			
Data Use Request	Required*	Required	Required
Rationale for Accessing the Data at Requested Level	Required	Required	Required
General description of research project**	Yes	Yes	Yes
Public abstract of the research project	Yes	Yes	Yes
Approval Process	DAC	DAC (+IRB approval from the accessing institution, if they require)	DAC + IRB approval from the accessing institution

Figure 2. Panel A. 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 National Center for Advancing Translational Sciences (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 institutional review board (IRB) approval. The DAC reviews the DUR and upon approval, grants access to the appropriate data level within the National COVID Cohort Collaborative (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. HIPAA: Health Insurance Portability and Accountability Act; LDS: Limited Data Set; NIH: National Institutes of Health. Panel B. Features and requirements for each level of data in the N3C Enclave: Synthetic,[35,36] De-identified data [33,34,37], and Limited Data Set, [34].

- **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 Sciences (NCATS)–controlled cloud environment, includes clinical data from patients who meet criteria in the N3C COVID-19 phenotype from sites across the United States 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 datasets. The N3C data is available to researchers to conduct a broad range of COVID-19–related analyses. N3C activities are divided into 5 workstreams as shown in Figure 1.

DATA PARTNERSHIP AND 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,[27] Human Tumor Atlas Network,[28] the Synapse platform[27–32]—and consulted governance frameworks of other networks—Global Alliance for Genomics and Health,[30] International Cancer Genome Consortium,[31] ACT Network.[32] 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 principle, and (5) an attribution and publication policy. The regulatory steps for organizations and users are shown in Figure 2A, 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 3 levels of data: Health Insurance Portability and Accountability Act (HIPAA) limited data, de-identified data, and synthetic data (see Figure 2B).[33,34]

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 dataset, is protected by a Certificate of Confidentiality.[38] This prohibits disclosure of identifiable, sensitive research information to anyone not connected to the research except when the subject consents or in a few other specific situations. 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, 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 2 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,[34] 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 reidentify 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.

Institutional review board 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,[39] we established a central IRB at Johns Hopkins University School of Medicine via the SMART IRB[40] 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. In addition, the N3C Data Enclave also requires ongoing IRB oversight. Because NCATS is the steward of the repository, data received by NCATS for the N3C Data Enclave from collection (post-DTA), maintenance, and storage is covered under an NIH IRB-approved protocol to make EHR-derived data available for the clinical and research community to use for studying COVID-19 and for identifying potential treatments, countermeasures, and diagnostics.

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 approach to data access allows us to reduce regulatory burden on investigators, while ensuring appropriate regulatory approvals are in place. There are 3 tiers of access—Synthetic, De-identified, and Limited Dataset—as described in Table 1.

Investigators wishing to access the data must have an N3C user profile linked to a public ORCID (Open Researcher and Contributor Identifier).[41] Access requirements and approval processes vary depending on the level of access requested. For each project for which a user wishes to access data, they must submit a DUR with their intended data use statement and include a nonconfidential 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 the National Institutes of Health, will review the DUR and verify that the conditions for access (see Table 1) are met. The DAC will regularly engage with the N3C community members and other stakeholders to provide an opportunity for feedback and dialogue. The DAC's role is to evaluate DURs; it does not exist to evaluate the scientific merit of the project.

Attribution and publication policy

N3C community members share a commitment to the dissemination of scientific knowledge for the public good. The Attribution and Publication Policy extends FAIR[22,42] to encompass all contributions to the N3C. Analyses posted within the N3C Enclave leverages the Contributor Attribution Model[43] to track the transitive credit[44] 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, a privacy-preserving strategy has been established to enable linkages within and external to the N3C dataset. 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 3-fold: (1) to determine the data inclusion and 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 4 CDMs—ACT, Observational Medical Outcomes Partnership (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 code sets from a number of organizations—for example, Centers for Disease Control and Prevention coding guidance,[45,46] PCORnet,[47] OHDSI,[48] LOINC,[49] 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[24] 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[24] for detailed definitions of these categories.

To encourage maximal community input into the phenotype definition, we chose to use GitHub[50] to host all versions of the phenotype definition in both machine-readable (SQL) format and human-readable descriptions.[24] The phenotype is updated approximately every 2 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.

Table 1. Scale comparison of 3 sites' positive COVID-19 cases, their N3C-relevant cohort, and their denominator (number of patients seen in a 1-year period)

	Site 1	Site 2	Site 3
COVID-19–positive patients as publicly reported by site[a]	2550	5540	390
N3C-relevant cohort[b]	67 350	46 500	12 000
Denominator[c]	1 271 510	1 259 330	172 000

All numbers rounded to nearest 10.

COVID-19: coronavirus disease 2019; N3C: National COVID Cohort Collaborative.

[a]The number of COVID-19–positive patients publicly reported by this site as of the week of June 8, 2020.

[b]The number of patients qualifying for the N3C COVID-19–relevant phenotype at this site as of the week of 6/8/2020.

[c]The number of unique patients seen in a 1-year period at this site.

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 4 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 1 or more CDMs. In particular, the following 4 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)[51] platform, focused on real-time querying across sites.[4]
- **PCORnet**: The official federated network and data model for the Patient-Centered Outcomes Research Institute[52] is a U.S.-based network of networks focusing on patient-centered outcomes.
- **OHDSI**: A multistakeholder, open science collaborative focused on large-scale analytics in an international network of researchers and observational health databases maintaining and using the OMOP CDM.[53]
- **TriNetX**: An international network of clinical sites coordinated by a commercial entity (TriNetX, Cambridge, MA) providing clinical data for cohort identification, site selection, and research to investigators in health care and life sciences.[8,54]

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

N3C's SQL scripts serve 2 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

Table 2. 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.)	1. Run phenotype and extract scripts. 2. Extract results to individual files, following N3C naming and structure conventions. 3. sFTP extract to N3C.
Python Package	1. Download the Python and SQL code. 2. Configure local variables (DB connection, schema names, etc.)	1. Run phenotype and extract scripts. 2. Extract results to individual files, following N3C naming and structure conventions. 3. 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

DB: database; N3C: National COVID Cohort Collaborative.

data models can be found in each model's documentation.[55–57] Like the phenotype definition, all scripts are publicly posted on GitHub[24] 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 4 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.

DATA INGESTION AND HARMONIZATION

N3C aims to support consistency in the data acquisition process across the 4 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.[58,59]

Target data model selection

A single data model enables scalable analytics. The emergent Health Level Seven International Fast Healthcare Interoperability Resources (FHIR)[60] standard may form a pluripotent research data model in complete synchrony with EHR source data.[61] The CD2H[62] has been working through its Next Generation Data Sharing core and catalyzing the formation of the Vulcan FHIR Accelerator for Translational Research[63] to advance this strategic goal. However, FHIR is not sufficiently mature in its specification and, more pertinently, its development of "bulk" multipatient research data transfers. The most expedient alternative was to select among the 4 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[64] 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.[65,66]

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 ETL (extract-transform-load) processes. Fortunately, as part of the Common Data Model Harmonization[67] project, CD2H and related federal projects had initiated mapping from each CDM to the BRIDG[68] and FHIR standards. N3C was able to leverage this previous work to jump-start the required mappings between each CDM and OMOP 5.3.1.

N3C worked with contractors and colleagues from the Common Data Model Harmonization project to build 2 sets of harmonization data for each source CDM:

1. Syntactic mapping for each CDM field to a corresponding table or field in OMOP with conversion logic
2. Semantic mapping of in which 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.[69]

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,[70] a cloud-based Platform as a Service on the secure NCATS Amazon Web Services 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 subsequently, 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, in which many sources combine to form a larger dataset, there are issues caused by the data heterogeneity, which impact data quality (DQ).[71,72] DQ measures, including consistency, correctness, concordance, currency, and plausibility, are important to support analysis and computation.[73,74] Many large-scale data aggregation projects benefit from focusing on a set of contextual use cases or a defined population research domain.[75–77] 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.

N3C ingestion and harmonization data quality checks

The Data Ingestion and 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 are 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. Owing to the criticality of these data for N3C, we corrected erroneous coding using text data indicating COVID-19 status, which would otherwise be lost.[85]

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.

COLLABORATIVE ANALYTICS AND 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).

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.[86] 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 FedRAMP (Federal Risk and Authorization Management Program) Moderate,[100] 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

Table 3. Data quality tools and methods evaluated

Tool Type		Tool
Native CDM DQ Processes	*PCORnet*	Data Check Scripts (v8.0)[78]
	ACT	"Smoke" tests[79]
	TriNetX	Focused Curation Process
	Adeptia Platform Processes	Process automation support[80]
		Data & Map validation functions[81]
OHDSI Collaborative Tools	*Data Quality Dashboard*	Data quality tests of OMOP databases[77]
	Atlas	Design/execute analytics on OMOP databases[82]
	Achilles	Data characterization of OMOP databases[83]
	White Rabbit	ETL preparation and support[84]
	Custom Scripts	SQL, R

ACT: Accrual to Clinical Trials; CDM: common data model; DQ: data quality; ETL: extract-transform-load; OHDSI: Observational Health Data Sciences and Informatics; OMOP: Observational Medical Outcomes Partnership; PCORnet: National Patient-Centered Clinical Research Network.

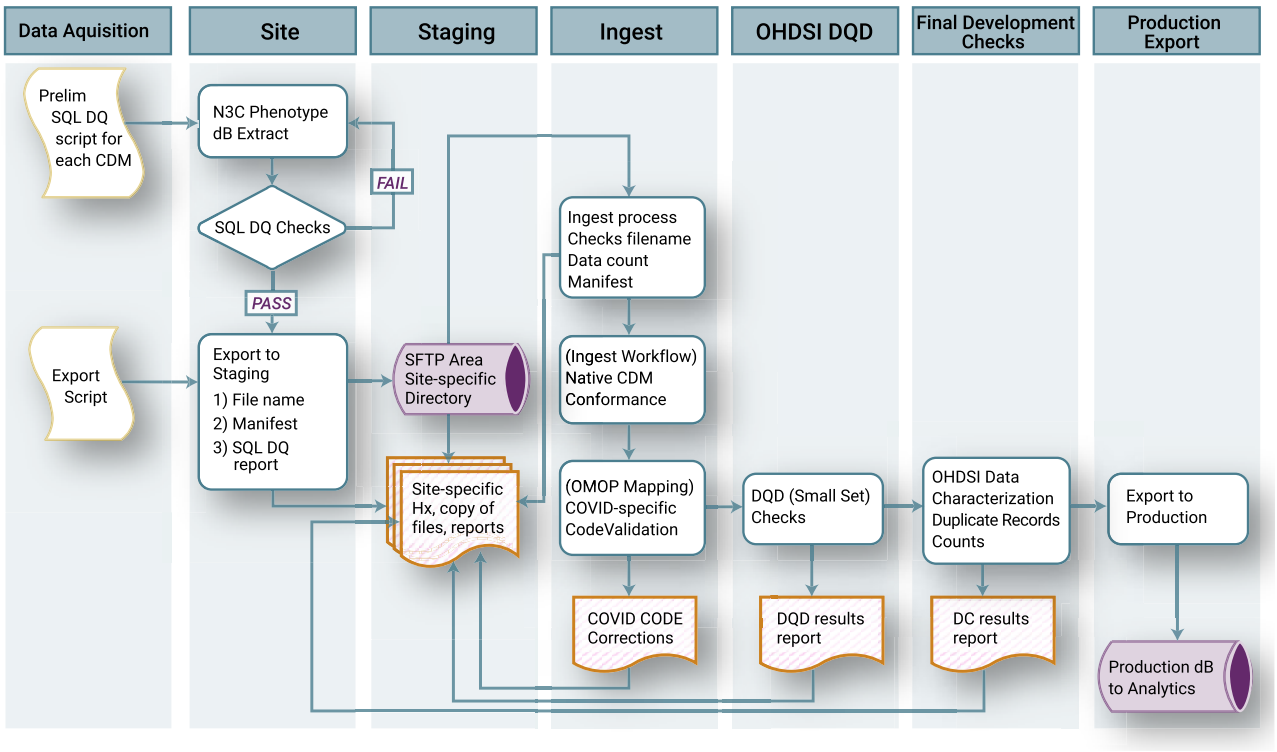


Figure 3. National COVID Cohort Collaborative (N3C) Data Quality Checks. At the sites, the extraction script performs a check for duplicate primary keys; if duplicate keys are found, the extraction fails until the site resolves the error. When extraction is successfully completed, a data "manifest" is 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 common data model (CDM), which ensures source data model conformance. Next, a series of data quality checks including a subset of coronavirus disease 2019 (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 Observational Medical Outcomes Partnership (OMOP) 5.3.1 using the validated maps from the payload's native CDM. Once in OMOP 5.3.1, a subset of the Observational Health Data Sciences and Informatics (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. DC: Data Characterization; DQD: Data Quality Dashboard.

specifically designated data and resources such as Limited Data Set (LDS) and de-identified data. Additional security and auditing mechanisms include the ability to limit patient-level data access; read and write access to datasets; 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 <10 patients. Through this enclave functionality, secure analyses of data containing limited Personal Health Information (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.

Transparency and reproducibility are fundamental to the prescribed use of the N3C Enclave.[101] 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 the following:

- **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).

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:

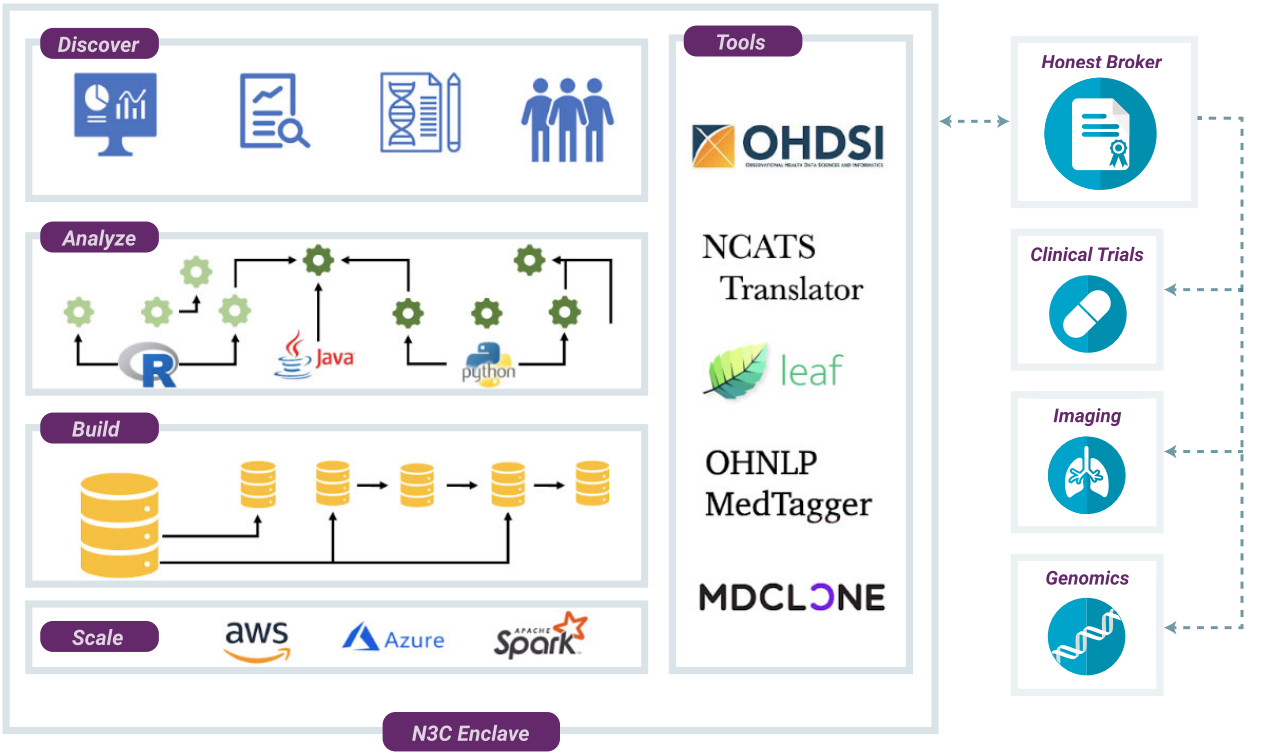


Figure 4. National COVID Cohort Collaborative (N3C) Enclave. The analytical environment for N3C is a secure, virtualized, cloud-based platform. Within the N3C Enclave, researchers have access to raw data, as well as transformed and harmonized datasets derived by other researchers. Analytical tools hosted within the environment support complex ETL (extract-transform-load), generation of coronavirus disease 2019 (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 Observational Health Data Sciences and Informatics (OHDSI) tools and the Leaf patient cohort builder. N3C is developing methods for integration of genomic, imaging, and other data modalities.

Table 4. 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 and reliable phenotype development[87]; phenotype performance evaluation[88]; integration of validated phenotypes definitions into study skeletons that learn and validate predictive models[89]; and execute a variety of comparative cohort study designs using empirically validated best practices.[90–92]
LOINC2HPO	**Mapping of LOINC-encoded laboratory test results to 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.[93]
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[94–96]; suite of resources include the Biolink Model,[97] a distributed API architecture, and a variety of KGs covering a range of biological entities such as genes, biological processes, and diseases; the KG-COVID-19[98] knowledge graph also includes literature annotation.
Leaf	**Web-based cohort builder.** Study feasibility for clinician investigators with limited informatics skills[99]; hierarchical concepts and ontologies to construct SQL query building blocks, exposed by a simple drag-and-drop user interface.

API: application programming interface; CDM: common data model; HPO: Human Phenotype Ontology; KG: knowledge graph; N3C: National COVID Cohort Collaborative; NCATS: National Center for Advancing Translational Sciences; OHDSI: Observational Health Data Sciences and Informatics; OMOP: Observational Medical Outcomes Partnership.

- **Statistical simulation:** Statistical models or profiles of normal human physiology or disease states are created based on real-world data. The ensuing simulated patients and their data are generally consistent with population-level norms.[102–104]
- **Computational derivation:** Computational models of real-world data are produced on demand, which can be used to produce novel data in a multidimensional space (eg, features) that adhere to the quantitative distributions and covariance

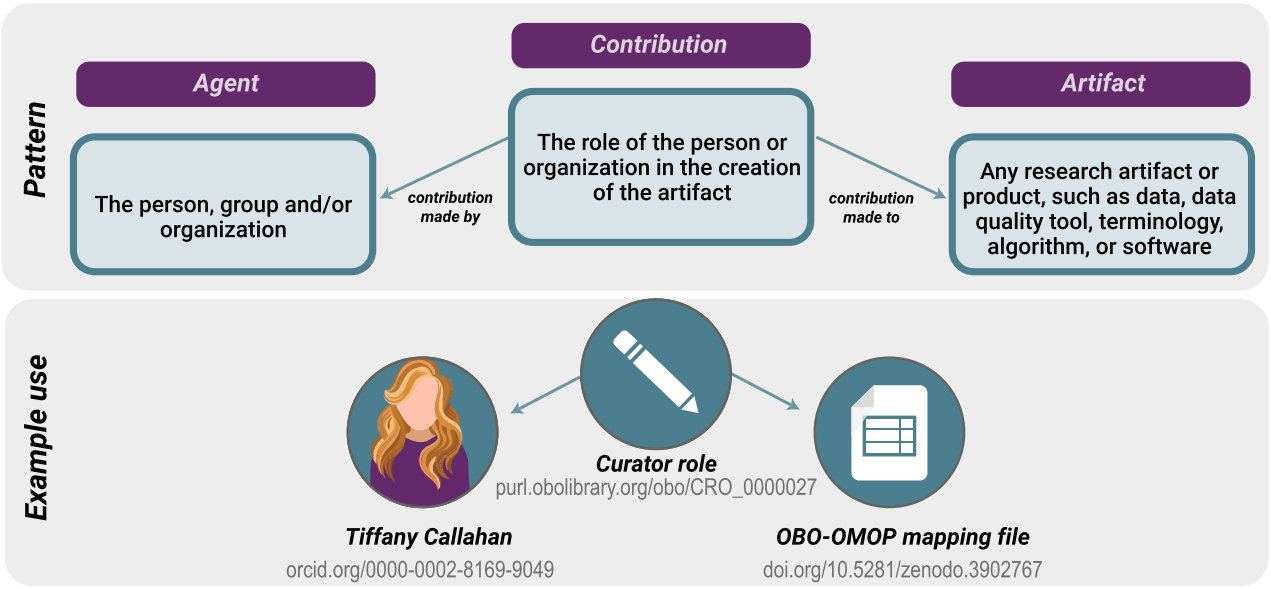


Figure 5. The Contributor Attribution Model. In the National COVID Cohort Collaborative Enclave, the Contributor Attribution Model is used to aggregate all contributions to any given workflow or report generated with a specific declaration of what exactly each person contributed, supporting the notion of transitive credit.[44] ORCID identifiers are used to identify users. An example contributor to an artifact used in the National COVID Cohort Collaborative is shown on the lower panel.

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, enabling the production of multiple derivative synthetic datasets. Further, such computationally derived synthetic datasets do not share mutual information with source data, minimizing the potential for reidentification.[35,36,105–107]

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 or 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 on 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?"

The N3C Attribution Policy[130] 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.[20,21] Attribution is native to the system, and supports the notion of transitive credit[44] for all contributors. Investigators are encouraged to prespecify 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.[108] 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 are 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

As of November 11, 2020, 72 sites have now executed data transfer agreements (DTA's); of these sites, 40 have deposited data in the N3C Pipeline (10 OMOP, 13 PCORnet, 10 TriNetX, 7 ACT. Of these 40 sites, 27 have Data Released in N3C Enclave. Additionally, researchers from over 120 institutions can begin analyzing these data as their institutional data use agreement (DUA) is in place. Collectively these released data now contain more than 1.4 billion rows and more than 200,000 COVID-positive patients.

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? How effective is convalescent plasma in COVID-19 treatment? What are the markers for and best practices to prevent COVID-19–related clotting disorders? What are the best practices for inflammatory monitoring prior to cytokine storm syndrome? The first 3 of these might be answerable in a federated network, but the last 2 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 intensive care unit 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.[109] To characterize patient subtypes, tensor factorization approaches have been shown to be quite effective for similar tasks.[110] 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,[111] 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.[112] Investigators can develop tools to predict treatment response based on these analyses. Clinicians could match a patient's phenotype to 1 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.[113] 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.[114,115]

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 (eg, clinical, utilization) allows more efficient and effective resource allocation and interventions[116] as well as enable healthcare providers to measure and balance the risk of COVID-19 complications vs other clinical conditions and morbidities. For example, identifying patients who will benefit the most from the anticipated COVID-19 vaccination is of utmost importance.[117] Assessing heterogeneity of treatment and 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.[118–121] N3C-based studies can support healthcare providers to identify clinical outcome disparities and SDoH, as well as to help public health officials and policy makers to identify inequalities on a systemic level (eg, 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 (eg, 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[122]; thus, collectively aiming for predictive models promoting equity for all minorities[123] 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.[61,124] 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.[125] EHR data captured in real time are often wrong (eg, 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.[126]

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 oppor-

tunities for novel approaches. N3C is driving new standards in openness for collaboration on sensitive clinical data, and builds on 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 health care. While N3C is focused on the United States, 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.[127,128]

FUNDING

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

AUTHOR CONTRIBUTIONS

Contribution summary (see appendix for details):

Melissa A. Haendel,[1,4,7,8,10,13,14,52,78,101] Christopher G. Chute,[1,4,8,10,13,14,52,78,100,101] Tellen D. Bennett,[9,10,13,14,52,100,101] David A. Eichmann,[4,9,10,13,78,101] Justin Guinney,[4,9,10,14,78,101] Warren A. Kibbe,[9,10,52,78,101] Philip R.O. Payne,[4,9,10,78,101] Emily R. Pfaff,[9,10,13,15,52,78] Peter N. Robinson,[4,9,10,15,52,78,100] Joel H. Saltz,[10,13,14,15,52,78,101] Heidi Spratt,[9,10,100] Christine Suver,[10,78,101] John Wilbanks,[10,78,101] Adam B. Wilcox,[10,101] Andrew E. Williams,[10,13,78] Chunlei Wu,[9,13,14,78] Clair Blacketer,[15,52] Robert L. Bradford,[9,52] James J. Cimino,[10,14,101] Marshall Clark,[9,15,52] Evan W. Colmenares,[9,15,52] Patricia A. Francis,[78] Davera Gabriel,[9,10,13,14,15,52] Alexis Graves,[7,9,78] Raju Hemadri,[9,15,52] Stephanie S. Hong,[9,15,52] George Hripscak,[10,52] Dazhi Jiao,[9,15,52] Jeffrey G. Klann,[14,52,101] Kristin Kostka,[9,15,52] Adam M. Lee,[9,15,52] Harold P. Lehmann,[9,15,52] Lora Lingrey,[9,15,52] Robert T. Miller,[9,15,52] Michele Morris,[9,15,52] Shawn N. Murphy,[9,15,52] Karthik Natarajan,[9,15,52] Matvey B. Palchuk,[9,15,52] Usman Sheikh,[9,78] Harold Solbrig,[9,15,52] Shyam Visweswaran,[10,15,52,101] Anita Walden,[7,10,13,14,52,101] Kellie M. Walters,[10,14,101] Griffin M. Weber,[10,101] Xiaohan Tanner Zhang,[9,15,52] Richard L. Zhu,[9,15,52] Benjamin Amor,[78] Andrew T. Girvin,[15,78] Amin Manna,[78] Nabeel Qureshi,[15,78] Michael G. Kurilla,[10,78] Sam G. Michael,[10,78] Lili M. Portilla,[101] Joni L. Rutter,[1,101] Christopher P. Austin,[101] Ken R. Gersing,[78] Shaymaa Al-Shukri,[4,15] Adil Alaoui,[101] Ahmad Baghal,[15] Pamela D. Banning,[15,100] Edward M. Barbour,[8,15] Michael J. Becich,[15,52,101] Afshin Beheshti,[14] Gordon R. Bernard,[8,15] Sharmodeep Bhattacharyya,[100] Mark M. Bissell,[9,15] L. Ebony Boulware,[14,100] Samuel Bozzette,[100,101] Donald E. Brown,[101] John B. Buse,[14] Brian J. Bush,[8,101] Tiffany J. Callahan,[14,52] Thomas R. Campion,[8,15] Elena Casiraghi,[9,15] Ammar A. Chaudhry,[13,14] Guanhua Chen,[9] Anjun Chen,[13] Gari D. Clifford,[8,15] Megan P. Coffee,[14,100] Tom Conlin,[14] Connor Cook,[7,78] Keith A. Crandall,[9,14,101] Mariam Deacy,[78] Racquel R. Dietz,[78] Nicholas J. Dobbins,[8,9] Peter L. Elkin,[15,52,100] Peter J. Embi,[52,101] Julio C. Facelli,[8,15] Karamarie Fecho,[13] Xue Feng,[9] Randi E. Foraker,[8,13,15] Tamas S. Gal,[8,15] Linqiang Ge,[14] George Golovko,[15,101] Ramkiran Gouripeddi,[14,15] Casey S. Greene,[13,14] Sangeeta Gupta,[52,101] Ashish Gupta,[13,101] Janos G. Hajagos, [9,15] David A. Hanauer,[15,52] Jeremy Richard Harper,[9,14,52] Nomi L. Harris,[14] Paul A. Harris,[101] Mehadi R. Hassan,[9] Yongqun He,[15,52,100] Elaine L. Hill,[9,14] Maureen E. Hoatlin,[14] Kristi L. Holmes,[4,101] LaRon Hughes,[14] Randeep S. Jawa,[14] Guoqian Jiang,[14] Xia Jing,[7,14] Marcin P. Joachimiak,[8,15] Steven G. Johnson,[9,14,101] Rishikesan Kamaleswaran,[9,15,78] Thomas George Kannampallil,[15,101] Andrew S. Kanter,[15,52] Ramakanth Kavuluru,[9,13,14] Kamil Khanipov,[8,14] Hadi Kharrazi,[9,14] Dongkyu Kim,[15,52] Boyd M. Knosp,[8,15] Arunkumar Krishnan,[9] Tahsin Kurc,[9,15] Albert M. Lai,[101] Christophe G. Lambert,[52,101] Michael Larionov,[14] Stephen B. Lee,[1,14] Michael D. Lesh,[9] Olivier Lichtarge,[14] John Liu,[9] Sijia Liu,[8,9,101] Hongfang Liu,[9,15] Johanna J. Loomba,[1,15,78,101] Sandeep K. Mallipattu,[9,14,15] Chaitanya K. Mamillapalli,[14] Christopher E. Mason,[15] Jomol P. Mathew,[8,15,52] James C. McClay,[101] Julie A. McMurry,[1,4,7,9,13,14,78] Paras P. Mehta,[14] Ofer Mendelevitch,[9] Stephane Meystre,[8,14,15] Richard A. Moffitt,[9,13,15] Jason H. Moore,[8,9] Hiroki Morizono,[13,14,15,52] Christopher J. Mungall,[15,52] Monica C. Munoz-Torres,[7,10,78] Andrew J. Neumann,[78] Xia Ning,[14] Jennifer E. Nyland,[13,14] Lisa O'Keefe,[78] Anna O'Malley,[78] Shawn T. O'Neil,[78] Jihad S. Obeid,[10,14,15] Elizabeth L. Ogburn,[13] Jimmy Phuong,[9,15,52,100,101] Jose D Posada, [8,15] Prateek Prasanna,[14,52] Fred Prior,[9,14,15] Justin Prosser,[9,78] Amanda Lienau Purnell,[101] Ali Rahnavard,[9,52] Harish Ramadas,[9,52,78] Justin T. Reese,[9,10] Jennifer L. Robinson,[14,100] Daniel L. Rubin,[101] Cody D. Rutherford,[9,101] Eugene M. Sadhu,[8,15] Amit Saha,[9] Mary Morrison Saltz,[15,52,101] Thomas Schaffter,[78] Titus KL Schleyer,[14] Soko Setoguchi,[8,14,15] Nigam H. Shah,[8,14] Noha Sharafeldin,[14] Evan Sholle,[15,52] Jonathan C. Silverstein,[15,52,101] Anthony Solomonides,[101] Julian Solway,[14,101] Jing Su,[101] Vignesh Subbian,[9,52,101] Hyo Jung Tak,[15] Bradley W. Taylor,[9,14] Anne E. Thessen,[14,101] Jason A. Thomas,[15] Umit Topaloglu,[15,52] Deepak R. Unni,[8,9,15,52] Joshua T. Vogelstein,[14] AndrÕa M. Volz,[7] David A. Williams,[14,15] Kelli M. Wilson,[9,78] Clark B. Xu,[8,9,15] Hua Xu,[9,10,14] Yao Yan,[9,15,52] Elizabeth Zak,[8,15] Lanjing Zhang,[101] Chengda Zhang,[14] Jingyi Zheng[14]

[1]CREDIT_00000001 (Conceptualization) [4]CREDIT_00000004 (Funding acquisition) [7]CRO_0000007 (Marketing and Communications) [8]CREDIT_00000008 (Resources) [9]CREDIT_00000009 (Software role) [10]CREDIT_00000010 (Supervision role) [13]CREDIT_00000013 (Original draft) [14]CREDIT_00000014 (Review and editing) [15]CRO_0000015 (Data role) [52]CRO_0000052 (Standards role) [78]CRO_0000078 (Infrastructure role) [100]Clinical Use Cases [101]Governance

ETHICS APPROVAL

While no IRB review is required for the work presented in this manuscript, we describe the creation of a central IRB at Johns Hopkins University for use by member organizations as well as the NIH IRB for the Enclave itself. The protocols have been made public.[129]

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

We acknowledge the Oregon Clinical and Translational Research Institute for their guidance and review of N3C plans and regulatory processes as they unfolded. The work described in this publication were conducted with data or tools accessed through the NCATS N3C Data Enclave (ncats.nih.gov/n3c/about). This research was possible because of the patients whose information is included within the data and the organizations and scientists who have contributed to the on-going development of this community resource.

Other N3C Consortial Authors:

Shaymaa Al-Shukri,[37] Adil Alaoui,[38] Ahmad Baghal,[37] Pamela D. Banning,[39] Edward M. Barbour,[40] Michael J. Becich,[41] Afshin Beheshti,[42] Gordon R. Bernard,[43] Sharmodeep Bhattacharyya,[44] Mark M. Bissell,[32] L. Ebony Boulware,[7] Samuel Bozzette,[34] Donald E. Brown,[45] John B. Buse,[46] Brian J. Bush,[47] Tiffany J. Callahan,[48] Thomas R. Campion,[49] Elena Casiraghi,[50] Ammar A. Chaudhry,[51] Guanhua Chen,[52] Anjun Chen,[53] Gari D. Clifford,[54] Megan P. Coffee,[55] Tom Conlin,[2] Connor Cook,[1] Keith A. Crandall,[56] Mariam Deacy,[34] Racquel R. Dietz,[1] Nicholas J. Dobbins,[13] Peter L. Elkin,[57,58,59] Peter J. Embi,[60,61] Julio C. Facelli,[62] Karamarie Fecho,[25,63] Xue Feng,[64] Randi E. Foraker,[65] Tamas S. Gal,[47] Linqiang Ge,[66] George Golovko,[67] Ramkiran Gouripeddi,[68] Casey S. Greene,[69] Sangeeta Gupta,[70] Ashish Gupta,[66] Janos G. Hajagos,[71] David A. Hanauer,[72] Jeremy Richard Harper,[60] Nomi L. Harris,[73] Paul A. Harris,[43] Mehadi R. Hassan,[13] Yongqun

He,[74] Elaine L. Hill,[75] Maureen E. Hoatlin,[76] Kristi L. Holmes,[77] LaRon Hughes,[78] Randeep S. Jawa,[79] Guoqian Jiang,[80] Xia Jing,[81] Marcin P. Joachimiak,[73] Steven G. Johnson,[82] Rishikesan Kamaleswaran,[83] Thomas George Kannampallil,[8] Andrew S. Kanter,[84] Ramakanth Kavuluru,[85] Kamil Khanipov,[12] Hadi Kharrazi,[86] Dongkyu Kim,[87] Boyd M. Knosp,[20] Arunkumar Krishnan,[88] Tahsin Kurc,[89] Albert M. Lai,[8] Christophe G. Lambert,[90] Michael Larionov,[91] Stephen B. Lee,[92] Michael D. Lesh,[93,94] Olivier Lichtarge,[95] John Liu,[96] Sijia Liu,[97] Hongfang Liu,[80] Johanna J. Loomba,[98] Sandeep K. Mallipattu,[71] Chaitanya K. Mamillapalli,[99] Christopher E. Mason,[49] Jomol P. Mathew,[100] James C. McClay,[101] Julie A. McMurry,[2] Paras P. Mehta,[102] Ofer Mendelevitch,[94] Stephane Meystre,[103] Richard A. Moffitt,[11] Jason H. Moore,[104] Hiroki Morizono,[87] Christopher J. Mungall,[73] Monica C. Munoz-Torres,[2] Andrew J. Neumann,[44] Xia Ning,[105] Jennifer E. Nyland,[106] Lisa O'Keefe,[107] Anna O'Malley,[32] Shawn T. O'Neil,[44] Jihad S. Obeid,[108] Elizabeth L. Ogburn,[109] Jimmy Phuong,[110] Jose D Posada,[111] Prateek Prasanna,[71] Fred Prior,[37] Justin Prosser,[13] Amanda Lienau Purnell,[112] Ali Rahnavard,[56] Harish Ramadas,[32] Justin T. Reese,[73] Jennifer L. Robinson,[66] Daniel L. Rubin,[111] Cody D. Rutherford,[113] Eugene M. Sadhu,[40] Amit Saha,[114] Mary Morrison Saltz,[79] Thomas Schaffter,[6] Titus KL Schleyer,[60] Soko Setoguchi,[115] Nigam H. Shah,[111] Noha Sharafeldin,[116] Evan Sholle,[49] Jonathan C. Silverstein,[41] Anthony Solomonides,[117] Julian Solway,[118] Jing Su,[119] Vignesh Subbian,[120] Hyo Jung Tak,[121] Bradley W. Taylor,[122] Anne E. Thessen,[44] Jason A. Thomas,[13] Umit Topaloglu,[119] Deepak R. Unni,[73] Joshua T. Vogelstein,[19] AndrÕa M. Volz,[1] David A. Williams,[72] Kelli M. Wilson,[34] Clark B. Xu,[52] Hua Xu,[123] Yao Yan,[124] Elizabeth Zak,[125] Lanjing Zhang,[126,127] Chengda Zhang,[128] and Jingyi Zheng[66]

[37]University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA, [38]Georgetown University, Washington, District of Columbia, USA, [39]3M Health Information Systems, St. Paul, Minnesota, USA, [40]University of Illinois at Chicago, Chicago, Illinois, USA, [41]Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA, [42]KBR, Space Biosciences Division, NASA Ames Research Center, Moffett Field, California, USA, [43]Vanderbilt University Medical Center, Nashville, Tennessee, USA, [44]Oregon State University, Corvallis, Oregon, USA, [45]School of Data Science, University of Virginia, Charlottesville, Virginia, USA, [46]University of North Carolina School of Medicine, Chapel Hill, North Carolina, USA, [47]Virginia Commonwealth University, Richmond, Virginia, USA, [48]Computational Bioscience, University of Colorado Anschutz Medical Campus, Boulder, Colorado, USA, [49]Weill Cornell Medicine, Cornell University, New York, New York, USA, [50]Computer Science Department, UniversitÁ degli Studi di Milano, Milano, Milan, Italy, [51]City of Hope National Medical Center, Duarte, California, USA, [52]University of Wisconsin-Madison, Madison, Wisconsin, USA, [53]Web2express.org, [54]Emory University and Georgia Institute of Technology, Atlanta, Georgia, USA, [55]Grossman School of Medicine, New York University, New York, New York, USA, [56]Computational Biology Institute and Department of Biostatistics and Bioinformatics, Milken Institute School of Public Health, The George Washington University, Washington, District of Columbia, USA, [57]Department of Biomedical Informatics, University at Buffalo, Buffalo, New York, USA, [58]Department of Veterans Affairs, Western New York, New York, USA, [59]Faculty of Engineering, University of Southern Denmark, Odense, Denmark, [60]Regenstrief Institute, Indianapolis, Indiana, USA, [61]Indiana University School of Medicine, Indianapolis, Indiana, USA, [62]Center for Clinical and Transnational Science, The University of Utah, Salt Lake City, Utah, USA, [63]Copperline Professional Solutions, LLC, Chapel Hill, North Carolina, USA, [64]Department of Biomedical Engineering, University of Virginia, Charlottesville, Virginia, USA, [65]Washington University in St. Louis School of Medicine, St. Louis, Missouri, USA, [66]Auburn University, Auburn, Alabama, USA, [67]The University of Texas Medical Branch, Galveston, Texas, USA, [68]University of Utah, Salt Lake City, Utah, USA, [69]Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA, [70]Delaware State University, Dover, Delaware, USA, [71]Stony Brook University, Stony Brook, New York, USA, [72]University of Michigan, Ann Arbor, Michigan, USA, [73]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA, [74]University of Michigan Medical School, Ann Arbor, Michigan, USA, [75]University of Rochester Medical Center, Rochester, New York, USA, [76]Hoatlin Biomedical Consulting, Portland, Oregon, USA, [77]Northwestern University Feinberg School of Medicine, Chicago, IL, USA, [78]Center for Translational Data Science, University of Chicago, Chicago, IL, USA, [79]Renaissance School of Medicine, Stony Brook University, Stony Brook, New York, USA, [80]Mayo Clinic, Rochester, Minnesota, USA, [81]Clemson University, Clemson, South Carolina, USA, [82]University of Minnesota, Minneapolis, Minnesota, USA, [83]Emory University School of Medicine, Atlanta, Georgia, USA, [84]Columbia University, New York, New York, USA, [85]University of Kentucky, Lexington, Kentucky, USA, [86]Johns Hopkins School of Public Health, Baltimore, Maryland, USA, [87]Children's National Hospital, Washington, District of Columbia, USA, [88]Division of Gastroenterology and Hepatology, Johns Hopkins School of Medicine, Baltimore, Maryland, USA, [89]Biomedical Informatics, Stony Brook University, Stony Brook, New York, USA, [90]Department of Internal Medicine, Center for Global Health, Division of Translational Informatics, University of New Mexico Health Sciences Center, Albuquerque, New Mexico, USA, [91]Spok, Inc., Springfield, Virginia, USA, [92]University of Saskatchewan (Regina), Saskatoon, SK, Canada, [93]University of California San Francisco, San Francisco, California, USA, [94]Syntegra.io, San Carlos, California, USA, [95]Baylor College of Medicine, Houston, Texas, USA, [96]Optum, Eden Prairie, Minnesota, USA, [97]Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA, [98]University of Virginia, Charlottesville, Virginia, USA, [99]Springfield Clinic, Springfield, Illinois, USA, [100]School of Medicine and Public Health, University of Wisconsin-Madison, Madison, Wisconsin, USA, [101]University of Nebraska Medical Center, Omaha, Nebraska, USA, [102]College of Medicine, The University of Arizona, Tucson, Arizona, USA, [103]Medical University of South Carolina, Charleston, South Carolina, USA, [104]Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, USA, [105]The Ohio State University, Columbus, Ohio, USA, [106]Penn State College of Medicine, Hershey, Pennsylvania, USA, [107]Northwestern University, Chicago, Illinois, USA, [108]Department of Public Health Sciences, Medical University of South Carolina, Charleston, South Carolina, USA, [109]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA, [110]School of Medicine, Division of Biomedical and Health Informatics, University of Washington, Seattle, Washington, USA, [111]Stanford University, Stanford, California, USA, [112]VHA Innovation Ecosystem, Washington, District of Columbia, USA, [113]Noblis, Inc., Reston, Virginia, USA, [114]Wake Forest Baptist Medical, Winston Salem, North Carolina, USA, [115]Biomedical and Health Sciences, Rutgers University, New Brunswick, New Jersey, USA, USA, [116]School of Medicine, University of Alabama at Birmingham, Birmingham, Alabama, USA, [117]Research Institute, NorthShore University HealthSystem, Evanston, Illinois, USA, [118]University of Chicago, Chicago, Illinois, USA, [119]School of Medicine, Wake Forest University, Winston Salem, North Carolina, USA, [120]College of Engineering, The University of Arizona, Tucson, Arizona, USA, [121]Department of Health Services Research and Administration, University of Nebraska Medical Center, Lincoln, Nebraska, USA, [122]Medical College of Wisconsin, Wauwatosa, Wisconsin, USA, [123]The University of Texas Health Science Center at Houston, Houston, Texas, USA, [124]Molecular Engineering and Sciences Institute, University of Washington, Seattle, Seattle, Washington, USA, [125]University of Iowa, Iowa City, Iowa, USA, [126]Rutgers University, New Brunswick, New Jersey, USA, USA, [127]Princeton Medical Center, Plainsboro, New Jersey, USA and [128]Oregon Health & Science University, Portland, Oregon, USA

CONFLICT OF INTEREST STATEMENT

N3C includes a number of commercial partners, without whom N3C would not be possible; they are: Adeptia, TriNetX, Palantir Technologies, Microsoft Corporation, MDClone, IQVIA, and Amazon. MAH and JAM have a founding interest in Pryzm Health. KK is an employee of IQVIA. ATG, AM, HR, BA, and NQ are employees of Palantir Technologies. MB and LL are employees of TriNetX. CB is an employee of Janssen Research & Development. Cody Rutherford is an employee of Noblis. JL is an employee of Optum. ML is an employee of Spok. OF and MDL are founders and shareholders of Syntegra USA. Andrew Kanter is the CMO of Intelligent Medical Objects. HX has research-related financial interests in Melax Technologies.

REFERENCES

1. Johns Hopkins Coronavirus Resource Center. COVID-19 Map. https://coronavirus.jhu.edu/map.html Accessed July 12, 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 (6493): 860–8.
3. Williamson EJ, Walker AJ, Bhaskaran K, *et al.* Factors associated with COVID-19-related death using OpenSAFELY. *Nature* 2020; 584: 430–6.
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 (2): 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 (4): 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. Health Policy Brief. *Health Affairs* 2015. https://www.healthaffairs.org/do/10.1377/hpb20150604.936915/full/healthpolicybrief_139.pdf Accessed June 7, 2020.
8. Topaloglu U, Palchuk MB. Using a federated network of real-world data to optimize clinical trials operations. *JCO Clin Cancer Inform* 2018; 2 (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. *npj Digit Med* 2020; 3: 109.
10. Carton TW, Marsolo K, Block JP. *PCORnet COVID-19 common data model design and results. Zenodo* 2020 Jun 16; doi: 10.5281/zenodo.3897398.
11. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019; 380 (14): 1347–58.
12. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng* 2018; 2 (10): 719–31.
13. Kramer WG, Perentesis G, Affrime MB, *et al.* Pharmacokinetics of dilevalol in normotensive and hypertensive volunteers. *Am J Cardiol* 1989; 63 (19): 7I–11I.
14. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med* 2016; 375 (13): 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 (3): 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 (4): 184–200.
18. Liu, P Qi, J, *et al.* Federated machine learning: concept and applications. *ACM Trans Intell Syst Technol* 2019; 10 (6): 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, Desai SS, Ruschitzka F, Patel AN. Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. Lancet 2020 May 22; doi: 10.1016/S0140-6736(20)31180-6.
21. Mehra MR, Desai SS, Kuy S, *et al.* Retraction: cardiovascular disease, drug therapy, and mortality in Covid-19. *N Engl J Med* 2020; 382 (25): e102.
22. Wilkinson MD, Dumontier M, Aalbersberg IJJ, *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016; 3 (1): 160018.
23. CTSA Program Hubs. National Center for Advancing Translational Sciences. 2015. https://ncats.nih.gov/ctsa/about/hubs Accessed June 13, 2020.
24. *Phenotype_Data_Acquisition*. GitHub https://github.com/National-COVID-Cohort-Collaborative/Phenotype_Data_Acquisition Accessed June 20, 2020.
25. National Center for Advancing Translational Sciences. https://ncats.nih.gov/ Accessed June 7, 2020.
26. CD2H. https://ctsa.ncats.nih.gov/cd2h/ Accessed June 7, 2020.
27. All of Us Research Hub. https://www.researchallofus.org/ Accessed June 18, 2020.
28. Human Tumor Atlas Network. https://humantumoratlas.org/ Accessed June 18, 2020.
29. 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. *SSRN* 2019 Jan 19 [E-pub ahead of print]. doi: 10.2139/ssrn.3502410.
30. Global Alliance for Genomics and Health. Regulatory & Ethics Toolkit. https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/ Accessed June 18, 2020.
31. Data Access Compliance Office. https://icgc.org/daco Accessed June 18, 2020.
32. i2b2: Informatics for Integrating Biology and the Bedside. https://www.i2b2.org/ Accessed June 18, 2020.
33. US Code of Federal Regulations. govinfo . https://www.govinfo.gov/app/details/CFR-2011-title45-vol1/CFR-2011-title45-vol1-part164 Accessed June 7, 2020.
34. HIPAA Privacy Rule and Its Impacts on Research. https://privacyruleandresearch.nih.gov/pr_08.asp Accessed June 18, 2020.
35. Raab GM, Nowok B, Dibben C. Guidelines for Producing Useful Synthetic Data. *arXiv*: 1712.04078; 2017.
36. Snoke J, Raab GM, Nowok B, *et al.* General and specific utility measures for synthetic data. *J R Stat Soc A* 2018; 181 (3): 663–88.
37. Office for Civil Rights. Methods for De-identification of PHI. 2015. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html Accessed June 20, 2020.
38. Certificates of Confidentiality (CoC)—Human Subjects. https://grants.nih.gov/policy/humansubjects/coc.htm Accessed June 15, 2020.
39. Office for Human Research Protections. The Revised Common Rule's Cooperative Research Provision (45 CFR 46.114). 2019. https://www.hhs.gov/ohrp/regulations-and-policy/single-irb-requirement/index.html Accessed June 20, 2020.
40. SMART IRB. National IRB Reliance Initiative. https://smartirb.org/ Accessed April 14, 2020.
41. Sprague ER. ORCID. *J Med Libr Assoc* 2017; 105: 207.
42. Haendel M, Su A, McMurry J. *FAIR-TLC: Metrics to Assess Value of Biomedical Digital Repositories: Response to RFI NOT-OD-16-133. Zenodo* 2016 Dec 15; doi: 10.5281/zenodo.203295.
43. Welcome to the Contributor Attribution Model—Contributor Attribution Model Documentation. https://contributor-attribution-model.readthedocs.io/en/latest/ Accessed June 20, 2020.
44. Katz DS, Smith AM. Transitive credit and JSON-LD *J Open Res Soft* 2015; 3: 14.
45. 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 Accessed June 7, 2020.
46. Centers 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 Accessed June 7, 2020.
47. PCORnet. COVID-19 Common Data Model Launched, Enabling Rapid Capture of Insights on Patients Infected with the Novel Coronavirus. 2020. https://pcornet.org/news/pcornet-covid-19-common-data-model-launched-enabling-rapid-capture-of-insights/ Accessed June 8, 2020.

48. 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* 2020. doi: 10.1101/2020.04.22.20074336.
49. SARS-CoV-2 and COVID-19 Related LOINC Terms—LOINC. LOINC. https://loinc.org/sars-cov-2-and-covid-19/ Accessed June 8, 2020.
50. National COVID Cohort Collaborative. GitHub. https://github.com/National-COVID-Cohort-Collaborative Accessed June 14, 2020.
51. 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 (5): 624–30.
52. Patient-Centered Outcomes Research Institute (PCORI). https://www.pcori.org/ Accessed April 12, 2020.
53. Observational Health Data Sciences and Informatics. https://ohdsi.org/ Accessed April 12, 2020.
54. TriNetX. https://www.trinetx.com/ Accessed April 12, 2020.
55. 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 Accessed June 7, 2020.
56. University of Pittsburgh. Box. https://pitt.app.box.com/s/qoj5afssw4oz3v27ipmfidhitmgya9nt Accessed June 21, 2020.
57. *CommonDataModel*. GitHub https://github.com/OHDSI/CommonDataModel Accessed June 21, 2020.
58. Chute CG. Clinical classification and terminology: some history and current observations. *J Am Med Inform Assoc* 2000; 7 (3): 298–303.
59. Haendel MA, Chute CG, Robinson PN. Classification, ontology, and precision medicine. *N Engl J Med* 2018; 379 (15): 1452–62.
60. Health Level 7 (HL7). Fast Healthcare Interoperability Resources (FHIR). https://www.hl7.org/fhir/ Accessed May 21, 2020.
61. Chute CG, Huff SM. The pluripotent rendering of clinical data for precision medicine. *Stud Health Technol Inform* 2017; 245: 337–40.
62. Center for Data to Health (CD2H). https://ctsa.ncats.nih.gov/cd2h/ Accessed April 12, 2020.
63. Health Level 7 (HL7). Vulcan Accelerator Home—Vulcan Accelerator—Confluence. https://confluence.hl7.org/display/VA/Vulcan-Accelerator+Home Accessed May 21, 2020.
64. CDM v5.3.1. https://ohdsi.github.io/CommonDataModel/cdm531.html Accessed June 21, 2020.
65. Kahn MG, Batson D, Schilling LM. Data model considerations for clinical effectiveness researchers. *Med Care* 2012; 50: S60–7.
66. 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 (8 Suppl 3): S45–52.
67. 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 June 7, 2020.
68. CDISC. BRIDG. https://www.cdisc.org/standards/domain-information-module/bridg Accessed April 13, 2020.
69. Data-Ingestion-and-Harmonization. GitHub https://github.com/National-COVID-Cohort-Collaborative/Data-Ingestion-and-Harmonization Accessed June 14, 2020.
70. Banga J, Tyagi MR, Hans S. B2B Integration Platform for Next-gen Business Connectivity | Adeptia. https://adeptia.com/ Accessed April 13, 2020.
71. Kahn MG, Brown JS, Chun AT, *et al.* Transparent reporting of data quality in distributed data networks. *EGEMS (Wash DC)* 2015; 3 (1): 1052.
72. 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 (6): 1072–9.
73. Weiskopf NG, Hripcsak G, Swaminathan S, *et al.* Defining and measuring completeness of electronic health records for secondary use. *J Biomed Inform* 2013; 46 (5): 830–6.
74. 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 (1): 144–51.
75. Zozus M. *The Data Book: Collection and Management of Research Data*. Boca Raton, FL: CRC Press; 2017.
76. Kahn MG, Eliason BB, Bathurst J. Quantifying clinical data quality using relative gold standards. *AMIA Annu Symp Proc* 2010; 2010: 356–60.
77. Execute and View Data Quality Checks on OMOP CDM Database. https://ohdsi.github.io/DataQualityDashboard/ Accessed June 20, 2020.
78. 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 June 20, 2020.
79. Wikipedia Contributors. Smoke Testing (software). Wikipedia, the Free Encyclopedia. 2020. https://en.wikipedia.org/w/index.php?title=Smoke_testing_(software)&oldid=962025059 Accessed July 12, 2020.
80. Hans S. Adeptia. Explore B2B Process Automation Solutions for Integration Needs. https://adeptia.com/solutions/b2b-process-automation Accessed June 20, 2020.
81. ETL Data Integration Software for Connecting Business Data. https://adeptia.com/products/etl-data-integration Accessed June 20, 2020.
82. ATLAS. https://atlas.ohdsi.org/#/home Accessed June 20, 2020.
83. Creates Descriptive Statistics Summary for an Entire OMOP CDM Instance. https://ohdsi.github.io/Achilles/ Accessed June 20, 2020.
84. Eagleton MJ, Kashyap VS. Introduction. *J Vasc Surg* 2020; 72 (1): e4–5.
85. Dong X, Li J, Soysal E. COVID-19 TestNorm—a tool to normalize COVID-19 testing names to LOINC codes. *J Am Med Inform Assoc* 2020; 27 (9): 1437–42.
86. 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 (5p2): 1456–67.
87. Hripcsak G, Shang N, Peissig PL, *et al.* Facilitating phenotype transfer using a common data model. *J Biomed Inform* 2019; 96: 103253.
88. Swerdel JN, Hripcsak G, Ryan PB. PheValuator: development and evaluation of a phenotype algorithm evaluator. *J Biomed Inform* 2019; 97: 103258.
89. 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 (8): 969–75.
90. Schuemie MJ, Cepede MS, Suchard MA, *et al.* How confident are we about observational findings in health care: a benchmark study. *Harv Data Sci Rev* 2020; 2 (1); doi: 10.1162/99608f92.147cc28e
91. 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; 376:20170356.
92. 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 (11): 2571–7.
93. 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; 2:32.
94. Biomedical Data Translator Consortium. Toward a universal biomedical data translator. *Clin Transl Sci* 2019; 12: 86–90.
95. Biomedical Data Translator Consortium. The biomedical data translator program: conception, culture, and community. *Clin Transl Sci* 2019; 12(2): 91–4.
96. Austin CP, Colvis CM, Southall NT. Deconstructing the translational tower of babel. *Clin Transl Sci* 2019; 12 (2): 85.

97. Biolink Model. https://biolink.github.io/biolink-model Accessed June 21, 2020.
98. *kg-covid-19*. GitHub. https://github.com/Knowledge-Graph-Hub/kg-covid-19 Accessed June 20, 2020.
99. 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 (1): 109–18.
100. FedRAMP.gov. https://www.fedramp.gov/ Accessed June 21, 2020.
101. Brito JJ, Li J, Moore JH, *et al.* Recommendations to enhance rigor and reproducibility in biomedical research. *GigaScience* 2020; 9 (6): giaa056
102. 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 (3): 230–8.
103. 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 (3): 228–41.
104. 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 (1): 44.
105. Hayes J, Melis L, Danezis G, *et al.* LOGAN: membership inference attacks against generative models. *arXiv*: 1705.07663; 2017.
106. Erez L. Computer system of computer servers and dedicated computer clients specially programmed to generate synthetic non-reversible electronic data records based on real-time electronic querying and methods of use thereof. U.S. Patent 10,235,537. 2019. https://patents.google.com/patent/US10235537B2/en Accessed June 7, 2020.
107. Foraker R, Mann DL, Payne PRO. Are synthetic data derivatives the future of translational medicine? *J Am Coll Cardio Basic Trans Sci* 2018; 3 (5): 716–8.
108. Head ML, Holman L, Lanfear R, *et al.* The extent and consequences of p-hacking in science. *PLoS Biol* 2015; 13 (3): e1002106.
109. 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 (5): 1589–604.
110. Luo Y, Wang F, Szolovits P. Tensor factorization toward precision medicine. *Brief Bioinform* 2017; 18 (3): 511–4.
111. Thompson AE, Ranard BL, Wei Y, *et al.* Prone positioning in awake, nonintubated patients with COVID-19 hypoxemic respiratory failure. *JAMA Intern Med* 2020 Jun 17 [E-pub ahead of print]; doi: 10.1001/jamainternmed.2020.3030.
112. Mehta P, McAuley DF, Brown M, *et al.* COVID-19: consider cytokine storm syndromes and immunosuppression. *Lancet* 2020; 395 (10229): 1033–4.
113. Suo Q, Ma F, Yuan Y, *et al.* Deep patient similarity learning for personalized healthcare. *IEEE Trans Nanobiosci* 2018; 17 (3): 219–27.
114. 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* 2020; 142: 429–36. doi: 10.1161/CIRCULATIONAHA.120.048360.
115. 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.
116. 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 (1): 2–12.
117. Kharrazi, H Chi, WChang 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 (8): 789–96.
118. Williams DR, Cooper LA. COVID-19 and health equity-a new kind of 'herd immunity'. *JAMA* 2020; 323 (24): 2478.
119. 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* 2020; 128: 35–48.
120. 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* 2020; 382 (26): 2534–43.
121. Millett GA, Jones AT, Benkeser D, *et al.* Assessing differential impacts of COVID-19 on Black communities. *Ann Epidemiol* 2020; 47: 37–44.
122. Gamache R, Kharrazi H, Weiner JP. Public and population health informatics: the bridging of big data to benefit communities. *Yearb Med Inform* 2018; 27 (1): 199–206.
123. Obermeyer Z, Powers B, Vogeli C, *et al.* Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; 366 (6464): 447–53.
124. 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.
125. 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 (8 Suppl 3): S30–7.
126. Hersh W, Cimino J, Payne PRO, *et al.* Recommendations for the use of operational electronic health record data in comparative effectiveness research. *EGEMS (Wash DC)* 2013; 1 (1): 1018.
127. EMBL-EBI Launches COVID-19 Data Portal. https://www.ebi.ac.uk/about/news/press-releases/embl-ebi-launches-covid-19-data-portal Accessed June 21, 2020.
128. ELIXIR Support to COVID-19 Research | ELIXIR. https://elixir-europe.org/services/covid-19 Accessed June 21, 2020.
129. Chute CG. *National COVID Cohort Collaborative (N3C) institutional review board protocol. Zenodo* 2020 Apr 22. doi: 10.5281/zenodo.3902948.
130. N3C Consortium. *Attribution and Publication Principles for N3C (National Covid Cohort Collaborative). Zenodo* 2020 August 25; doi: 10.5281/zenodo.3992394.