National COVID Cohort Collaborative (N3C)

A national resource for shared analytics

Version 6
18 September 2023

This following link will never expire, and will always redirect to the most recent version of this document: doi:10.5281/zenodo.5120812

Study Overview and Goals

The National COVID Cohort Collaborative (N3C) established a central registry of patients who have been tested for COVID or have a clinical diagnosis of COVID. It is derived by harmonizing COVID clinical data extracted from the federated clinical repositories associated with the Common Data Model (CDM) programs, enumerated below. These data ultimately are extracted from the electronic health records (EHRs) of the medical centers who will contribute data. The requested data constitutes a HIPAA limited data set[1] in that it will contain dates and zip codes; both are necessary for the epidemiologic tracking of the COVID pandemic over time and space. The data will be updated from the contributing sites as frequently as practical, ideally twice a week.

The data extracted at each contributing site is transferred by secure FTP to a FedRAMP[2] certified staging area on the NIH National Center for Advancing Translational Science (NCATS) server. There it is harmonized into a single data model, OMOP 5.3.1[3] after multiple data quality checks. The merged dataset is transferred to the Palantir analytic platform[4], which can allow sophisticated data science analytics while preventing data download or human browsing of individual-level data. Access to this limited dataset will be determined by the N3C Data Access Committee.

Additionally, a synthetic dataset will be derived from the limited dataset, created by sampling from the statistical distribution of the underlying data. The synthetic dataset, or "synthetic derivative" does not contain data of real patients and cannot be traced to real patients; therefore, it can be used free of HIPAA regulations. This dataset is proposed to enable more open data science investigation, and would be accessible through a simple registration and attestation process.

The NIH IRB has established its own IRB to cover data use and access (attached in Section 20). This IRB is now restricted to the Data Transfer process, now subject to the NIH Data Transfer Agreement which replaces our earlier DTA (also in Section 20).

Background and Significance

The Center for Data to Health (CD2H)[5] is the coordinating center for informatics activities throughout the CTSA network of academic medical centers[6]. Leadership at NIH/NCATS requested CD2H to coordinate the contributions by CTSA hub sites to a central commons of clinical data about COVID tested and diagnosed persons. Subsequently, this proposed commons was adopted by NIH overall, with many NIH centers, institutes, and offices partnering with NCATS in its design and approval. Additionally, many federal agencies including CDC, FDA, and HHS leadership contributed to enabling this initiative. Many CTSA hub sites have expressed enormous interest in contributing to this effort.

Most CTSA sites participate in one or more federated data cooperatives. All of these efforts involve academic medical centers creating a local repository of clinical data extracted from their

EHRs, that conforms to the collaborative data model (called a Common Data Model (CDM)). Queries are sent to the collaborating sites, where an analysis is conducted locally to address the query. Answers are returned to the coordinating center for that CDM network, where they are aggregated into a sophisticated meta-analyses.

The four major CDM collaboratives are:

Accelerating Clinical Trials (ACT)[7]: A CTSA supported collaborative of most CTSA sites operating with an i2b2 data model conformant to the ACT ontology structure. Federated queries occur through the SHRINE system.

PCORNet: The official federated network of the Patient-Centered Outcomes Research Institute[8] is a US-based network of networks focusing on patient centered outcomes.

Observational Health Data Sciences and Informatics[9] (OHDSI): is an international federation of clinical data sites, originating from the pharmacoepidemiology community.

TriNetX[10]: ia an international network of clinical sites coordinated by a commercial entity established to facilitate case finding for pharmaceutical funded clinical trials.

These collaboratives excel in providing case counts, cohorts, and in many cases, answers to specific questions. The nature of federated analytics typically limits such questions to hypotheses testing: for example, whether a specific drug helps or hurts COVID patient courses. While it is possible to conduct federated logistic regression across sites, this is rarely done in practice. It is impractical to conduct many machine learning algorithms across federated datasets, as many algorithms require access to all analyzed data in the same computing environment.

Significance

Creating the N3C registry of individual-level (containing information specific to individual patients, sometimes called row-level) data as a limited--albeit protected--dataset of EHR data at a national level will be unprecedented in US clinical research. It will support novel machine learning analytics and discovery of important predictors associated with emergency visits, hospitalizations, ICU transfer, ventilator dependency, and death, amongst a myriad of related outcomes. It will have the scale, statistical power, and computing platform to address most questions the clinical and research communities seek to answer. Ideally it will be used by multidisciplinary teams engaging clinicians, statisticians, epidemiologists, biologists, and data scientists, who together can leverage this unique resource at this critical time in the COVID pandemic. Furthermore, these activities can be conducted with minimal risk of data breach or inadvertent disclosure, as any effort to view or download individual level data is generally disabled.

Research Design and Methods

CD2H established four Workstreams[11] to undertake this work:

- Data Partnership & Governance
- Phenotype & Data Acquisition
- Data Ingestion & Harmonization
- Collaborative Analytics

Data Partnership and Governance

CD2H, in partnership with NCATS, has convened a community to leverage mechanisms from recent dataset integration efforts, particularly the All of Us research initiative[12] at NIH. As will all workstreams, public meetings, list serves, Slack channels, and document repositories comprise efforts to engage as many CTSA hubs, stakeholders, and representatives from all four of the CDM communities.

The committee partnered with NIH to create the final Data Use Agreement for enclave access, now covered by the NIH IRB. They also partnered in the creation of the final NIH authored Data Transfer Agreement which is now the focus of this IRB.

The project was originally planned to exist for two years. The new NIH PASC proposals are likely to extend this use, and a CIR will be filed when this is clear.

Data Use Agreement

NIH authored the final Data Use Agreement (DUA), which is attached to this application. It describes data contribution, processing, access terms and conditions. It clarifies confidentiality agreements, intellectual property, warranties and liabilities, conflict resolution, and expiration or termination. This portion of the DUA would be signed by the clinical data contributor (such as a CTSA hub) and NCATS. The document also includes a proposed data access application for the limited dataset signed by the investigator team requesting access and NCATS. The access request stipulates the data is to be used only for research, cannot be downloaded, and all results must be shared with the community. Investigators must attest that they will not attempt to reidentify data in any way, and acknowledge the contributions of the communities that made the N3C resource possible.

Single IRB

To lower any burden associated with contributing data to the N3C, this protocol is submitted to establish a central IRB. N3C and NCATS leadership accept the offer of Johns Hopkins Medicine to operate in this capacity. Sites that contribute data to N3C are not obligated to use the central IRB, and may opt to operate through their own institutional IRB. Note, this IRB pertains only to the data transfer component of this project. Data access is now superseded by the NIH IRB (attached).

It is the preference of Johns Hopkins Medicine IRB to use the SMART IRB[13] reliance agreement as the basis of reliance. The SMART IRB master reliance agreement was created in 2016 to harmonize and streamline the IRB review process for multisite studies. It enables reliance on a study-by-study basis, clearly defines roles and responsibilities of relying institutions and reviewing IRBs, and eliminates the need to sign reliance agreements for each study [e.g., a non-SMART IRB agreement]. 700+ institutions have already signed onto this agreement and are actively using it as the basis of reliance for multisite projects.

Phenotype & Data Acquisition

Clinical Phenotype (Query)

CD2H and the Phenotype team have drafted a prototype query for clinical data extraction from the CDM platforms at the contributing sites; this document is attached to this proposal. It has been widely vetted among clinical experts, CDM representatives, and CTSA sites. While it serves as an initial consensus document we recognize that modification and updates, released as N3C phenotyping versions, will need to be made as tests, data, and clinical practice evolves through this pandemic. For example, the recently released serology tests to establish whether an individual has previously been infected with COVID may impact phenotyping strategies.

At present, there are four phenotype categories defined by the document.

Inclusion criteria:

- Patients of all ages
- Encounter on or after 1/1/2020
- Meeting any of the following criteria:
    - Lab Confirmed Negative
        - LOINC codes Negative result, sampled controls
        - Note: The version 3.2 phenotype in present use sample two patients who match demographic criteria for positive cases, to reduce the number of patients enrolled. Otherwise, a large fraction of EHR patients have been tested and would be sent.
    - Lab Confirmed Positive
        - LOINC codes Positive result (based on enumerated LOINC codes)
    - Likely Positive
        - COVID Dx Code (from short list of ICD codes) only
    - Possible Positive
        - Two or more suggestive ICD codes only

Data transmitted:

The CDMs targeted contain information on the following categories:

- Demographics, including zip code
- A hashed identifier to support future linkage with imaging, viral RNA, or other major national data resources with a shared hash identifier (below).
- Laboratory tests and results, including dates
- Medications, including start and where available stop dates
- Vital signs, with dates
- Diagnoses, with dates

- Procedures, with dates
- Admission, Discharge, Transfer information with dates
    - Including death with dates

The information will be requested retrospectively to 1 Jan 2018, at the contributing sites discretion, prior to the query start date (1/1/2020), and through the present with serial data updates for a given patient.

Hashed identifier:

To support future linkage with other datasets, the N3C team will develop a strategy for sites to develop a unique, encrypted hashed identifier. The goal of a hashed identifier is to allow data to be linked using identifiers, such as name, date of birth, without actually disclosing those identifiers. The hashed identifier will be constructed using industry standards[14].

CDM Translations

To minimize burden on contributing sites, the CD2H team has consulted with representatives from each CDM community to translate the proto-query into reproducibly executable code to operate against the sites CDM repository. Four queries are generated, one each for each CDM in the format native to each CDM. The queries will be validated on test data, and reviewed with CDM subject matter experts. These queries are available here: . https://github.com/National-COVID-Cohort-Collaborative/Phenotype_Data_Acquisition

Algorithm Execution at Contributing Sites

Each site will select one and only one CDM at a time as the basis for its data contribution to minimize record duplication in the N3C repository. They may opt to change their sourcing CDM during participation. Sites may access the CDM specific phenotyping query, and execute at their site. Local modifications may be required, if for example not all COVID tests are not yet LOINC coded at their site. The CD2H will maintain a “white glove” helpdesk for each of the four CDM involved in the program. This helpdesk will be the primary support resource of the contributing sites. Each helpdesk in turn will have access to CDM subject matter experts supported on the N3C COVID supplement to CD2H. The helpdesk will help each site adopt the CDM query code, and facilitate the transfer of extracted data to the NCATS staging area via sFTP.

Data Workflow

The Data Workflow will consist of three key components: Authentication, Data Ingestion, and Collaborative Analytics.

Authentication

The Unified NCATS Authentication (UNA) system will be used to authenticate Data Platform users. The UNA system is an identity broker that facilitates collaboration by enabling networked applications to easily support multiple identity providers (IdPs) and authentication protocols. UNA is dynamically configurable, multi-tenant authorization and federation service for applications. It allows applications to authenticate users via multiple identity providers using standard protocols. The protocols are implemented as Passport.js[15] strategies and include Open ID Connect[16], SAML[17], and WSFederation[18]. The authentication protocol allows users to access the platform. By utilizing a Federated solution, and if the organization is Federated with the NIH, users of various organizations can access the platform by using their own organization credentials. This authentication solution is dynamically configurable, which allows users to be authenticated using alternative identity providers (such as login.gov) in case an organization is not Federated with the NIH. This not only brings more security to the application, but provides flexibility in administration for system administrators and users alike. For additional security, this authentication solution utilizes a layered approach, where users and their credentials are verified by multiple authentication protocols and methods. Utilizing the layered structure with multiple identity providers allows the authentication solution to coordinate a seamless and undisturbed login for users. This authentication solution also serves as an Identity Broker, which allows client applications to support one or more identity providers.

Data Ingestion

The NCATS ingestion platform will be a tool called Adeptia[19], which is a cloud based Platform as a Service. Adeptia will be highly restricted and not directly accessible by sites or investigators. Users are added through invitation only and are monitored by system administrators. Strict Password policies are enforced, which include limited entry retries and password expiration options are available for administrators to enforce. Adeptia uses certificates in order to ensure security while data is transmitted across the platform. Adeptia is protected by protocols such as FTPS, SFTP, HTTPS, and many others that require the use of certificates in order to encrypt the data and to verify the digital signature of the application and the user sending the data.

Collaborative Analytics

The data analyses platform will be accessed through a portal, which invokes Palantir[4]. Palantir provides a variety of tools, which support common data analytics resources such as R, Python, and Jupyter. These tools are configured to disallow any downloading, and to prohibit individual-level data viewing by default. Palantir Federal Cloud Service has FedRAMP[2] authorization at the Moderate level which is deployed on Amazon Web Services (AWS) GovCloud[20] that will be leveraged for this Data Platform. Palantir utilizes a multi-tiered approach to security and compliance. Palantir utilizes automated audit and logging tools, and provides automatic alerting for high priority events. Palantir also utilizes continuous deployment methods to ensure patches and updates are delivered seamlessly without user or system-wide downside or effects. Palantir contains HIPAA compliance protocols, and enforces global standards such as Findability, Accessibility, Interoperability, and Reusability (FAIR)[21] and General Data Protection Regulation (EU GDPR)[22].

Data Ingestion & Harmonization

Data Staging Quality Control

Data entering the NCATS staging area by sFTP from a site will be managed in a separate partition space for each contributing site. The data will be unpacked from the transmitted tables, and reincarnated as a database under the CDM model and version forwarded by the contributing site. There will be three stages of data quality checking and validation.

First-phase: The live CDM replication of the data will undergo the suite of data quality metrics and dashboards made available by their corresponding CDM; all CDM communities support suit suites. Additionally, value set verification will be undertaken to ensure that coded data does not include obsolete or deprecated codes (widely present in laboratory and medication coding systems). Any discrepancies will be iterated with the contributing site at this phase.

Second-phase: After conversion of the CDM data into OMOP 5.3.1 (below), the transformed data will be examined for systematic transform errors. We will also run test queries on the native CDM version of the data and the transformed OMOP 5.3.1 dataset, and compare answers. Discrepancies will be addressed by the data transform team.

Third-phase: After OMOP transformation, the standard suite of data quality and dashboarding tooling from the OHDSI consortium will be run. Discrepancies and findings at this level will be shared with contributing sites, but not addressed until the next data refresh cycle.

Data Harmonization into Common Data Model

There are two phases of data harmonization work: creating the transform mappings, and executing them on contributed mappings.

The CD2H and NCATS have been working for years on methods, tooling, and strategies for model to model transformation. Several maps have been created to go from CDMs to the CDISC BRIDG[23] metamodel; this allows transforms between and among each model. We are building on these foundations to create a series of pairwise data transforms, from the major versions of each CDM to OMOP 5.3.1. These transform mappings include not only structure to structure mappings, but also the value sets that are bound to those structures. For example some models use SNOMED codes for diagnoses, while others use differing versions of ICD codes. We will conform to the OMOP 5.3.1 data target with one exception, we will use ICD-10-CM codes for diagnoses to avoid any intellectual property concerns and obviate information loss engendered through data value transforms from the native ICD-10-CM sources in all US EHR data today.

The data transform mappings will be reviewed and validated by subject matter experts from each CDM. The mappings will be uploaded to a commercial extraction/transform/load (ETL) utility procured and used by NCATS, Adeptia.[19]

We will use the data transform mappings on the Adeptia platform to conduct the ETL from the source CDM into OMOP 5.3. The platform also has intrinsic data quality metrics to supplement the public tools that will form the core elements of our data quality efforts.

Collaborative Analytics

The main goal of this project is to enable investigators to conduct data discovery on the COVID cohort data. Thus, the analytics environment is key. **The analytics and data access are no longer covered under this IRB, but rather the NIH/NCATS IRB.** This information is retained for context..

Data Enclave Environment

We will establish a secure data enclave, which is defined as:

> *A data enclave is a secure network through which confidential data, such as identifiable information from census data, can be stored and disseminated. In a virtual data enclave a researcher can access the data from their own computer but cannot download or remove it from the remote server.*[24]

The Palantir product can enforce "enclave" conditions. The platform will include the standard suite of R, Python, and Jupyter analytic resources, and can support the uploading of custom libraries.

Modes of Access

We anticipate three modes of access to N3C data:

Public data scientists: This category will have a low bar for entry, requiring only investigator registration. However, access will only be granted to the synthetic derivative (see below) of the data. We expect most teams will initially use this access path. All teams doing software development that requires viewing of individual-level data must use this mode. Should compelling findings be identified, investigators may apply to validate their findings on the limited data set enclave. Even though this data is synthetic, it will remain on the Palantir data enclave, and cannot be downloaded.

Limited data set access: This category permits investigators to run software in the data enclave against the limited dataset, though not to see the individual-level data. Software development requiring any viewing of individual-level data must be done on the synthetic derivative. Access via this mechanism will require DAC review and approval. All the conditions and stipulations of the data access DUA would pertain at this level.

Direct Data Access: In rare cases investigators may petition for direct access to the limited dataset within the enclave. They will still be prohibited from downloading data. Access at this level would require IRB approval from their home institution and approval by the DAC.

Synthetic Derivative

Synthetic data is a term of art in clinical and translational research referencing clinical data that has been "synthesized" from real clinical data in a manner that obviates any reidentification. The explicit goal is to create a non-human subjects research dataset that statistically resembles the original data, but cannot be used to reidentify any subjects. The technique analyses the true source data to describe its contents, for example laboratory data, as a mathematical distribution of those results. Data is "synthesized" by randomly sampling from that mathematical distribution. The same is done for date and demographic information. Relationships between elements, for example drug use and diagnoses, is preserved by calculating a series of correlations and associations; these are then used to inform the statistical sampling in a manner that preserves these statistical associations.

Ideally, synthetic data are nearly identical to original PHI data, and can be analyzed as if they were original data but without any privacy concerns. Not only is there significant potential to protect patient privacy through analysis of data as a synthetic derivative, but synthetic derivatives of data can enable data sharing and accelerate discovery. Once real patient data are synthesized, the resulting data set no longer contains data on individual patients, but rather is a collection of observations which maintain the statistical properties of the original data set. Since the data set no longer contains data on real patients, synthetic derivatives can be shared between researchers at different institutions.[25]

The MDClone[26] synthetic data engine will be used to generate the synthetic data derivative in this project. MDClone's synthetic data engine is able to take any given data set, analyze its statistical properties, and create a brand-new set of data for research and clinical decision-making. This new data is synthetic, and since it does not contain data of real patients and cannot be traced to real patients, it can be used free of HIPAA regulations.

Human Subjects

Protection of Human Subjects

We understand the principle of minimum necessary data. The core challenge with the COVID pandemic is that society does not have the luxury of leisurely proposing hypotheses that can be addressed through the more conventional federated data networks. N3C intends to allow the research community the opportunity to leverage a harmonized clinical dataset in support of a variety of secondary data analyses, including machine learning algorithms. Among the characteristics of machine learning algorithms is their capacity to discover features and predictors that might otherwise be unanticipated. We are not proposing "black box" prediction algorithms, but rather an emphasis on looking at all the data to discover interpretable elements that can help medicine fight this condition, armed with information.

These aspirations require the creation of a central registry, and of making it a limited dataset. How the pandemic moves and behaves through time and over geographic areas requires dates and zip codes ideally at the five digit level.

Given these realities, it is our obligation to protect this large, limited data set to the best of our ability, minimizing risk of reidentification, breach, or inadvertent disclosure. Moreover, use of these data is limited to secondary data analyses; there will be no participant contact.

The following four topics are now covered by the NIH IRB, but are retained for context.

Enclave Computing Platform

We believe the most secure action we can take to balance the need for legitimate access with patient privacy and confidentiality is to restrict access to a data enclave environment. We are fortunate that NCATS has been establishing such an environment with the Palantir resource. The enclave will disallow any data downloading. In addition, except for extremely limited use cases (above in Mode of Access), it will prohibit the individual-level viewing of the limited dataset.

Data Access Committee

All access to the limited data set must be approved by the DAC, described above. Their role is to identify bone fide researchers with compelling questions. Investigators are expected to have explored the synthetic derivative prior to requesting access to the limited dataset. Investigators requesting such access must also agree to all conditions and stipulations in the data access DUA.

Synthetic Derivative

To maximize the public benefit of this resource, a synthetic derivative of the data, which by its very nature will not contain any identifiers, will be generated, as described above. The synthetic derivative does not contain data of real patients and cannot be traced to real patients; therefore, it can be used free of HIPAA regulations. While barriers to this data will be lower, investigators are still expected to attest that they will use this data only for COVID research and disclose all results for the public good. They will not be permitted to download the data, and must do analyses and software development on the Palantir platform.

Results Sharing

Any risk associated with a large, aggregated dataset must be offset by benefit. To ensure timely benefit in the face of this urgent pandemic, all investigators will be required to share intermediate and final results, as well as the computer code to generate these. This is easily enforced on the Palantir platform, which can record all analyses made. This would include software generated on the synthetic copy, and of course all software run on the limited dataset.

Inclusion of Women

The distribution of women in the N3C repository will correspond to the distribution of women in the clinical contributing sites fulfilling the phenotyping query. No effort will be made to exclude or otherwise select for gender.

Inclusion of Minorities

The distribution of minorities in the N3C repository will correspond to the distribution of minorities in the clinical contributing sites fulfilling the phenotyping query. No effort will be made to exclude or otherwise select for race or ethnicity.

Inclusion of Children

The distribution of children in the N3C repository will correspond to the distribution of children in the clinical contributing sites fulfilling the phenotyping query. No effort will be made to exclude or otherwise select for patients under 21.

Inclusion of Prisoners

The clinical records of prisoners were inadvertently included in the data uploads of one of our data contributors. That contributor manages the EHRs systems at many locations, some of which are prisoners. The volume of prisoners is estimated to have been much less than 1% of their total volumes. When the protocol violation was discovered, a protocol event report was submitted and ultimately acknowledges (PE00011447). The data contributor was able to eliminate all prisoners in subsequent data refreshes. No prisoners are now included to our knowledge, and we are now vigilant about avoiding such violations in future.

The previously uploaded prisoners remain in our historical dataset "snapshots" made on a weekly basis. In our protocol event report, we outlined our reasons for not removing them from those historical snapshots, recapitulated here:

- There are no indicators in these datasets that those persons are prisoners
- Maintaining historical accuracy and validity of those datasets for study reproducibility was prioritized
- There are very few prisoners (<100) in the multi-million person datasets comprising the snapshots
- Retroactively removing those people would effectively identify them as prisoners
- It is logistically challenging to track down all data derivatives of those snapshots, and remove those individuals from all derivative

We "acknowledge … that the prisoner data that was inadvertently collected may not be used until OHRP's determination and that no new prisoner data may be collected by the study."

References

1 Johns Hopkins Office of Human Subjects Research. HIPAA - Definition of Limited Data Set. 2018; published online March 6. https://www.hopkinsmedicine.org/institutional_review_board/hipaa_research/limited_data_set.html (accessed April 12, 2020).

2 US General Services Administration. Understanding Baselines and Impact Levels in FedRAMP | FedRAMP.gov. https://www.fedramp.gov/understanding-baselines-and-impact-levels/ (accessed April 19, 2020).

3 Observational Health Data Sciences and Informatics (OHDSI). Definition and DDLs for the OMOP Common Data Model (CDM) I Version 5.3. Github, 2018 https://github.com/OHDSI/CommonDataModel (accessed April 13, 2020).

4 Palantir Technologies. Palantir. https://www.palantir.com/ (accessed April 12, 2020).

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

6 Clinical and Translational Science Awards (CTSA) Program. National Center for Advancing Translational Sciences. 2017; published online Oct 31. https://ncats.nih.gov/ctsa (accessed April 19, 2020).

7 Visweswaran S, Becich MJ, D'Itri VS, *et al.* Accrual to Clinical Trials (ACT): A Clinical and Translational Science Award Consortium Network. *JAMIA Open* 2018; **1**: 147–52.

8 Patient-Centered Outcomes Research Institute (PCORI). https://www.pcori.org/ (accessed April 12, 2020).

9 OHDSI – Observational Health Data Sciences and Informatics. https://ohdsi.org/ (accessed April 12, 2020).

10 TriNetX. TriNetX. https://www.trinetx.com/ (accessed April 12, 2020).

11 NCATS Center for Data to Health (CD2H). National COVID Cohort Collaborative (N3C). https://covid.cd2h.org/clinical_evidence (accessed April 12, 2020).

12 National Institutes of Health (NIH). 'All of Us' Research Program. https://allofus.nih.gov/ (accessed April 17, 2020).

13 SMART IRB | National IRB Reliance Initiative. https://smartirb.org/ (accessed April 14, 2020).

14 Github. Public software repositories implementing sha3-256 hashing algorithms. Github https://github.com/topics/sha3-256 (accessed April 19, 2020).

15 Passport Consortium. Passport.js. http://www.passportjs.org/ (accessed April 19, 2020).

16 OpenID Foundation. Specifications | OpenID. https://openid.net/developers/specs/ (accessed April 19, 2020).

17 Organization for the Advancement of Structured Information Standards (OASIS). Security Assertion Markup Language (SAML) V2.0 Technical Overview. http://docs.oasis-open.org/security/saml/Post2.0/sstc-saml-tech-overview-2.0.html.

18 Organization for the Advancement of Structured Information Standards (OASIS). Web Services Federation Language (WS-Federation) Version 1.2. http://docs.oasis-open.org/wsfed/federation/v1.2/os/ws-federation-1.2-spec-os.html (accessed April 19, 2020).

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

20 Amazon Web Services (AWS). AWS GovCloud (US). https://aws.amazon.com/govcloud-us/?whats-new-ess.sort-by=item.additionalFields.postDateTime&whats-new-ess.sort-order=desc.

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

22 Privacy Europe. General Data Protection Regulation (GDPR) – Official Legal Text. https://gdpr-info.eu/ (accessed April 19, 2020).

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

24 Data Enclave | NNLM. https://nnlm.gov/data/thesaurus/data-enclave (accessed April 13, 2020).

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

26 MDClone. Synthetic Data Generation Platform. MDClone. https://www.mdclone.com/ (accessed April 19, 2020).