Journal of the American Medical Informatics Association, 29(4), 2022, 609–618
doi: 10.1093/jamia/ocab217
Advance Access Publication Date: 2 November 2021
Research and Applications

Research and Applications

Synergies between centralized and federated approaches to data quality: a report from the national COVID cohort collaborative

Emily R. Pfaff[1], Andrew T. Girvin[2], Davera L. Gabriel[3], Kristin Kostka[4], Michele Morris[5], Matvey B. Palchuk[6], Harold P. Lehmann[7], Benjamin Amor[2], Mark Bissell[2], Katie R. Bradwell[2], Sigfried Gold[3], Stephanie S. Hong[3], Johanna Loomba[8], Amin Manna[2], Julie A. McMurry[9], Emily Niehaus[2], Nabeel Qureshi[2], Anita Walden[10], Xiaohan Tanner Zhang[11], Richard L. Zhu[11], Richard A. Moffitt[12], Melissa A. Haendel[13], and Christopher G. Chute[14]; The N3C Consortium

[1]Department of Medicine, UNC Chapel Hill School of Medicine, Chapel Hill, North Carolina, USA, [2]Palantir Technologies, Denver, Colorado, USA, [3]Section of Biomedical Informatics and Data Science, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA, [4]The OHDSI Center at the Roux Institute, Northeastern University, Portland, Maine, USA, [5]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA, [6]TriNetX LLC, Cambridge, Massachusetts, USA, [7]Department of Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland, USA, [8]University of Virginia, Charlottesville, Virginia, USA, [9]Center for Health AI, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA, [10]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon, USA, [11]Johns Hopkins University School of Medicine, Baltimore, Maryland, USA, [12]Department of Biomedical Informatics, Stony Brook University, Stony Brook, New York, USA, [13]University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA and [14]Schools of Medicine, Public Health, and Nursing, Johns Hopkins University, Baltimore, Maryland, USA

Corresponding Author: Emily R. Pfaff, PhD, MS, Department of Medicine, UNC Chapel Hill School of Medicine, 160 N Medical Drive, Chapel Hill, NC 27599, USA; epfaff@email.unc.edu

Received 15 July 2021; Revised 19 August 2021; Editorial Decision 12 September 2021; Accepted 23 September 2021

ABSTRACT

Objective: In response to COVID-19, the informatics community united to aggregate as much clinical data as possible to characterize this new disease and reduce its impact through collaborative analytics. The National COVID Cohort Collaborative (N3C) is now the largest publicly available HIPAA limited dataset in US history with over 6.4 million patients and is a testament to a partnership of over 100 organizations.

Materials and Methods: We developed a pipeline for ingesting, harmonizing, and centralizing data from 56 contributing data partners using 4 federated Common Data Models. N3C data quality (DQ) review involves both automated and manual procedures. In the process, several DQ heuristics were discovered in our centralized context, both within the pipeline and during downstream project-based analysis. Feedback to the sites led to many local and centralized DQ improvements.

Results: Beyond well-recognized DQ findings, we discovered 15 heuristics relating to source Common Data Model conformance, demographics, COVID tests, conditions, encounters, measurements, observations, coding completeness, and fitness for use. Of 56 sites, 37 sites (66%) demonstrated issues through these heuristics. These 37 sites demonstrated improvement after receiving feedback.

Discussion: We encountered site-to-site differences in DQ which would have been challenging to discover using federated checks alone. We have demonstrated that centralized DQ benchmarking reveals unique opportunities for DQ improvement that will support improved research analytics locally and in aggregate.

Conclusion: By combining rapid, continual assessment of DQ with a large volume of multisite data, it is possible to support more nuanced scientific questions with the scale and rigor that they require.

Key words: electronic health records, data accuracy, COVID-19

INTRODUCTION

COVID-19 has precipitated a worldwide public health emergency requiring responsive action from all branches of medical science, including informatics. The National COVID Cohort Collaborative (N3C), sponsored by the NIH National Center for Advancing Translational Sciences (NCATS), is a data-driven response to this challenge. Through leading-edge technology, N3C uses harmonized electronic health record (EHR) data to support pioneering collaborative research that spans the full COVID disease cycle—from risk factors, to disease progression, to treatment decisions. The result of this collaboration is a research environment that addresses technical, legal, and policy barriers to rapid discovery and dissemination of actionable clinical findings to optimize the acute and long-term health outcomes of diverse populations nationwide.

At the heart of the N3C collaborative is a centralized enclave of EHR data assembled from Clinical and Translational Sciences Award (CTSA) hubs, Institutional Development Award (IDeA) Networks for Clinical and Translational Research (IDeA-CTR) hubs, and the OCHIN network.[1,2] As of this writing, it is the largest ever assembly of harmonized EHR data for research in the United States, comprising 6.4 million patients from 56 sites including 7.2 billion rows of data.[3] N3C ingests and harmonizes patient-level EHR data from participating sites for patients with positive COVID-19 tests or whose symptoms are consistent with COVID-19. Additional records collected include persons who have tested negative for COVID-19 (and have never tested positive) to support comparative studies.[4] Data harmonization is made possible through the efforts of the National Center for Data to Health (CD2H) and subject matter experts from Observational Health Data Sciences and Informatics (OHDSI), the Patient-Centered Clinical Research Network (PCORnet), the Accrual to Clinical Trials (ACT) network, and TriNetX.

Precursors to N3C: federated data networks

Clinical data repositories from EHR sources have evolved over the decades. Ad hoc database designs from the early period of EHR adoption had limited generalization across sites. The earliest federated models[5] in the 1990s ultimately gave rise to Sentinel[6] and to Common Data Models (CDMs) such as PCORnet,[7] the ACT network,[8] TriNetX,[9–11] and Observational Medical Outcomes Partnership (OMOP), which later became OHDSI.[12–14] In a federated data network, each participating site's data stay behind its institutional firewall, but are structured according to a CDM. This enables queries and results to be shared across sites rather than raw data.

The CDMs have been key to using EHR data for research; however, secondary analytic uses—particularly involving multiple contributing sites—require resource-intensive quality control to achieve the required uniformity and specificity to support open-ended research.[15–18] Despite their importance, data quality (DQ) checks are inconsistently applied and implementation methodologies are largely not evaluated.[17,19] Further, evaluations of DQ often fail to accurately determine the data's "fitness for use," which evaluates both its intrinsic and intentional aspects.[20–22] DQ evaluations must take into account the initial purpose of the data in their source systems, as well as the intended use of these data once harmonization processes have transformed them for use in secondary research.[19,23,24]

Each of the aforementioned federated networks has methods to promote local DQ and adherence to data model conventions; these methods vary in maturity. As examples, OHDSI offers its DQ Dashboard[25] tool for sites to run against their local CDM in order to evaluate adherence to OMOP CDM convention and diagnose common issues; PCORnet requires a quarterly data "curation" and quality check that uses prepackaged SAS scripts[26]; ACT has a "smoke test" to ensure federated network query response and has a DQ Dashboard of its own under development; and TriNetX employs a growing library of DQ metrics and visualizations with site benchmarking as the basis for evaluating the results. Because the data are federated, DQ evaluation and remediation are performed locally at the site-by-site personnel. These local checks ensure that data conform to the specifications of the chosen CDM, and may also check for data anomalies such as statistical outliers, invalid dates, biological implausibility, abundant missing data, and other common clinical data issues. Such approaches support the alignment of data; however, with the data remaining behind the institutional firewalls, it can be difficult to assess conformance or determine overall variability across sites, especially if the CDM does not have the capability of executing ad hoc DQ-related queries across the data network.

N3C's centralized approach

In contrast to the federated approach, N3C pools data from each partner site in accordance with its signed Data Transfer Agreement[4] and harmonizes all submitted data to the OMOP CDM. By the time, data are submitted to N3C, sites have already applied a layer of local DQ checks, and are submitting data that are "clean" by the local definition. Once data are merged across sites, additional opportunities for improvement may become apparent due to the ability to efficiently compare and benchmark among similar sites.[27] N3C centralized DQ methods complement and augment the foundation accomplished by sites in the context of the federated networks.

Data quality

DQ emerged in the 2010s as an explicit subject of attention and research in informatics.[22,28,29] Notably, the Patient-Centered Outcomes Research Institute (PCORI) funded an effort to define DQ, resulting in the "Harmonization" framework of Kahn and colleagues.[20] The dimensions of DQ defined there (internal verification, external validation, conformance, completeness, and plausibility) are the basis of the PCORnet[26] and OHDSI[30] conception and framework for DQ checking. N3C continued to build on this foundation for our centralized DQ approach.

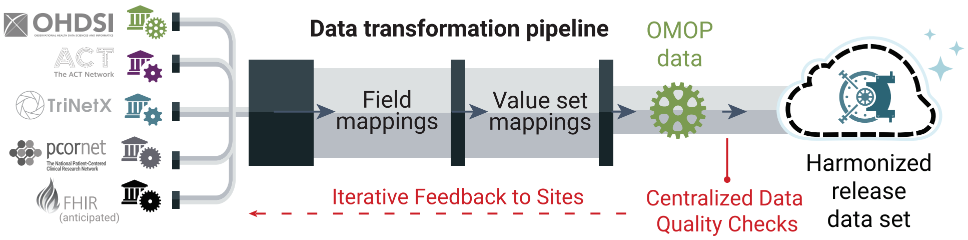
OHDSI
ACT
The ACT Network
TriNetX
pcornet
The National Patient-Centered Clinical Research Network
FHIR
(anticipated)
Data transformation pipeline
Field mappings
Value set mappings
OMOP data
Harmonized release data set
Iterative Feedback to Sites
Centralized Data Quality Checks

Figure 1. The N3C data ingestion and harmonization pipeline. Participating sites regularly submit data in their native CDM format to an ingest server. A parsing step validates whether the data are formatted properly and check the contents of the payload against its package description, or "manifest." The pipeline then transforms the submitted data to the OMOP model; data provenance is automatically maintained such that transformed data can be traced back to source at any time. The transformed data are then reviewed for DQ by a team of subject matter experts using a suite of data characterization and visualization tools. Every week, the latest data from all sites passing DQ checks are published as a versioned "release" for use by investigators. DQ: data quality; N3C: National COVID Cohort Collaborative; OMOP: Observational Medical Outcomes Partnership.

OBJECTIVE

To describe N3C's approach to DQ curation and explore the value that a centralized data architecture and approach to DQ adds to what sites can accomplish locally, either alone or in the context of a federated data research network.

MATERIALS AND METHODS

N3C's data harmonization pipeline

N3C has engaged in a precedent-setting endeavor to centrally harmonize the 4 major CDMs (OMOP, PCORnet, ACT, and TriNetX) into OMOP (Figure 1). This enables N3C's EHR dataset to span 56 institutions at the time of this writing while putting minimal burden on sites themselves. The EHR data spans multiple data domains, including patient demographics, encounter details, diagnoses, procedures, medications, lab tests, and clinical observations; a detailed list of supported fields can be found in the OMOP 5.3.1 common data model specification.[28] The pipeline evolved over 3 implementation environments and is presently running in Palantir Foundry.

Centralized DQ assessment

After transformation to OMOP, each site's inaugural N3C submission is loaded into the N3C Data Quality Portal (DQP), which performs a series of automated DQ checks prior to manual evaluation by the Data Ingestion & Harmonization (DI&H) team (see "Acknowledgments" for members). The DQP is built on queries similar to OHDSI's DQ Dashboard and provides a series of targeted visualizations that allow the team to assess each site's DQ in areas of importance for COVID research (see Table 1). The DQP supports site data review both in the context of the site's source CDM and in comparison to the other 3 CDMs.

N3C purposely uses a light touch during these DQ checks, placing a high value on including as much submitted data as possible, with the understanding that each site's data likely contains local idiosyncrasies and inconsistencies that are acceptable so long as they are known. Such issues (eg, a site is able to provide only outpatient data, only supports a subset of vital signs, or is frequently missing units of measure) can be reported in N3C's release notes, but would not prevent a site from "passing" on to inclusion in that week's release. Issues that can prevent sites from passing ("Must Pass"), and thus make up our minimum data standards, are detailed in Table 1. Issues that do not prevent passage but are still of concern are labeled "Heads Up."

Providing feedback to sites

As previously described,[1] N3C's signature "white glove," or one-on-one DQ support, provides feedback and individualized source model-specific help to sites. Members of N3C's Phenotype and Data Acquisition (P&DA) team (see "Acknowledgments" for members) serve as liaisons between sites and the DI&H team. The P&DA team is composed of subject matter experts in each of N3C's 4 supported data models. Each site is assigned an expert in their source model as their P&DA point of contact. After a site's initial payload is evaluated using the DQP, the site's P&DA contact compiles a list of data issues in the "must pass" and "heads up" categories and emails that list to the site. Corrections are generally iterative in nature, and correction cycles will continue until the site passes all "must pass" checks. Some corrections are simple, while others require individual troubleshooting meetings with the P&DA team, attendance at P&DA office hours, or sharing code snippets.

More recently, we have started an initiative to provide some of the benchmarking data generated by the DQP directly to sites, in visual format. Our centralized architecture gives us the unique ability to provide this type of benchmarking data and may reveal opportunities for DQ improvement of which sites were previously unaware. One of the visualizations we sent, a heatmap illustrating "coverage" of different vital signs for COVID inpatients across a variety of (anonymized) sites, is shown in Figure 2. These visualizations allow sites to compare their DQ to that of other sites that are using the same CDM. Hierarchical clustering was used to bring together sites with similar profiles of vital reporting.

Assessing N3C's DQ impact

To assess the impact of N3C DQ feedback on sites' local DQ, we reviewed all submitted DQ issues filed on site data in the N3C Data Enclave and performed a qualitative analysis to extract DQ heuristics, as well as the number of sites to which each heuristic applied. Only "released" sites (ie, sites whose data are available for research in the Enclave) were included in the analysis; a denominator of 56 sites. Issue instances were counted if they were in the "Must Pass" category. Sites that are still working through data issues (and are thus not yet released) are not included in the denominator, and "Heads Up"-type issues or simple formatting errors (eg, incorrect delimiters, missing headers, etc.) were not included in the count of issues.

RESULTS

Table 2 provides an accounting of the DQ issues found and improvements made by N3C-participating sites based on our feed-

Table 1. Data quality issue types

Check type	Data checks
Source CDM conformance	*Must Pass:* All tables required by the native CDM specs are present, with all CDM-required fields populated; fields that use a controlled value set (eg, "M" for male, "F" for female, etc.) are populated with valid values
Demographics	*Must Pass*: Count of patients qualifying for COVID phenotype is reasonable when compared with sites of similar size; sex, race, and ethnicity distributions reasonable for the site's population; month of birth evenly distributed throughout the calendar year *Heads Up*: >20% of race or ethnicity is missing or "No Matching Concept"
COVID tests	*Must Pass*: All COVID tests must be coded with an OMOP standard concept (or, for non-OMOP source data, the LOINC equivalent); all COVID test results must be coded with an OMOP standard concept (or, for non-OMOP source data, the equivalent controlled vocabulary term); numbers of negative and positive COVID tests are reasonable when compared with sites of similar size *Heads Up*: High numbers of COVID tests with *null* results
Conditions	*Must Pass:* Clinical encounters are present that are coded with U07.1 (ICD-10 code for COVID), and those encounters are distributed across various visit types (eg, outpatient, inpatient, emergency)
Encounters	*Must Pass*: Clinical encounters are distributed across a variety of standard visit types (eg, outpatient, inpatient, emergency); the distribution of visit types is reasonable when compared with similar sites; the majority of inpatient visits have valid end dates; the mean duration of visits of various types is reasonable for that type of visit; the vast majority of visit end dates are later than or equal to the visit start date
Measurements/observations	*Heads Up*: The site supports only a small number (eg, 5–10) of unique measurement or observation types
Coding completeness	*Must Pass*: No more than 20% of records in any domain are coded with nonstandard OMOP concept IDs without further explanation (OMOP sites only); no more than 20% of records in any domain are coded with "0—No Matching Concept" without further explanation (affects OMOP sites only); the PERSON_ID attached to all records in domain tables must exist in the PERSON table; primary keys are valid (ie, no duplicate rows in any table); if applied by the site, date shifting is consistent within each patient across all domains
Fitness for use	Use of the data by researchers often reveals additional DQ issues for one or more sites (eg, sparsely populated body mass index data, in the context of a study of obesity and COVID). In these cases, we report the findings to sites so that they can take action in their local data if they wish to have their site's data included in the study

"Must Pass" and "Heads Up" data check for release into the N3C Data Enclave.
DQ: data quality; N3C: National COVID Cohort Collaborative; OMOP: Observational Medical Outcomes Partnership.

back. These improvements map to the checks we perform against each site's first payload, detailed earlier in Table 1.

These heuristics revealed DQ issues in 37 (66%) of the 56 sites, all of which demonstrated improvement after receiving feedback from N3C. Selected examples are detailed here.

Example of heuristic #1: not using (or improperly using) source CDM's controlled vocabulary in one or more fields

There are numerous examples of fields that require controlled vocabularies among the 4 source CDMs. Nearly a quarter of all N3C sites have violated the use of these vocabularies on one or more occasions within their source CDM. Examples of this issue include nonuse or improper use of the ACT race and ethnicity vocabulary (eg, using local codes rather than the controlled vocabulary's value sets), incorrect DX and DX_TYPE agreement (eg, labeling an ICD-10-CM diagnosis code as type "Other") in PCORnet, or using standard concepts in an inappropriate domain (eg, filing conditions wrongly in the OBSERVATION table) in OMOP. These types of errors often create the illusion of missing data. Having centralized access to sites' source data gives the N3C team the ability to diagnose the issue in detail and offer sites ways to remedy the problem.

Example of heuristic #2: COVID test result values not standardized or null

Eleven sites submitted nonstandard, null, or otherwise unusable COVID test result values in their initial submissions. Quality issues included:

- Nonharmonizable COVID test results (eg, submitting a free-text result rather than the source CDM's controlled vocabulary equivalent).
- Null COVID test results in excess of a reasonable number of pending results.
- COVID test results that used the source CDM's controlled vocabulary, but mapped to an unusual concept (eg, one site mapped results to OMOP concept ID 45877980, "Not," presumably for a negative test).

Even where it is not a source CDM requirement to map every qualitative test result to the CDM's controlled vocabulary, for N3C's use case, harmonized COVID tests are essential. We worked with sites to prioritize these mappings, even as new COVID test codes continued to emerge over the course of the pandemic. Figure 3 shows the improvements made by these 11 sites over time.

Example of heuristic #4: implausible distribution of visit types

Each source CDM has its version of an encounter or visit table, as well as a controlled vocabulary to assign a "type" to visits, such as inpatient, outpatient, or emergency. Different models' vocabularies have different levels of specificity for visit types; the N3C DQ process accounts for this by aggregating multiple valid visit type codes to a higher-level category (eg, "Inpatient Hospital" and "Inpatient Visit" in the OMOP vocabulary can roll up to an overall category of *Inpatient* for the purposes of quality analysis). Even when these roll-ups are taken into account, however, 7 N3C sites had implausible distributions of visit types, where *implausibility* was defined by the overall distribution across sites. Such implausibility is illustrated here for 5 of the 7 sites that specifically had an implausible proportion of inpatient visits. One of the 5 sites started with inpatient visits significantly above the mean (Figure 4A), while the other 4 had significantly fewer (Figure 4B). As shown in Figure 4, after feedback from N3C, all 5 sites improved the quality of their visit type mappings to bring their inpatient visit proportion closer to the mean. Of note, the proportions at times worsened with subsequent payloads, pointing out the need for recurring vigilance: Fixing a problem once does not mean it stays fixed.

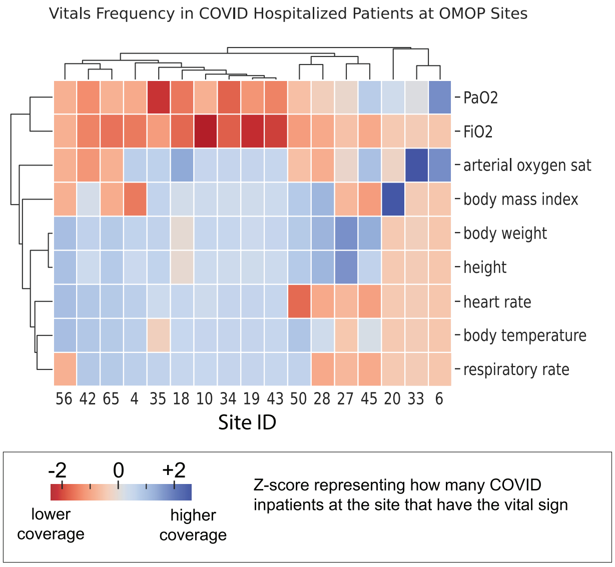

Figure 2. Vital sign coverage visualization, N3C OMOP sites. This heatmap is representative of those that we sent to sites to provide them with benchmarked lab and vital coverage information. The rows represent concept sets for vital signs and the columns are individual sites. The cell colors reflect the z-score of the percentage of COVID inpatients at each site that have at least 1 lab or vital of that type recorded during their hospitalization. The bluer the color, the higher the percentage of COVID inpatients that have that vital sign at that site—redder shades mean a lower percentage of patients with that vital. Rows and columns are hierarchically clustered, bringing similar sites closer together, and similar vitals closer together. This visualization enables sites to compare their data coverage with other sites using the same data model. (Site numbers are anonymized and have been changed from the site numbers used inside the N3C Enclave.) N3C: National COVID Cohort Collaborative; OMOP: Observational Medical Outcomes Partnership.

Example of heuristic #15: data utility challenges

A number of N3C DQ findings have come from analysts using the data, spotting inconsistencies, and submitting issue tickets. One such example involves data on mortality, the most commonly investigated endpoint in N3C. Generally, sites document patient deaths in their source CDM with pairs of patient IDs and death dates. However, because not all CDMs require a death date to be present to note a patient as deceased, numerous sites provide patients IDs, but no dates. In one instance, a site's Death data table included *all* of their patient IDs (for both living and deceased patients); in their definition, a null date denoted a living patient, and a populated date denoted a death. Taken individually, each of these structures makes sense. However, aggregate analyses that either ignore missing dates or require dates to be present would come to drastically different conclusions, either over- or underestimating mortality. In these cases, we worked individually with sites to standardize where possible, and in other cases, provided user education and analytical workarounds. Other researcher-identified site-level issues include those tied to a particular type of measurement, such as a misrepresented unit of measure for a specific type of value (discovered when comparing height/weight calculated body mass index [BMI] to reported BMI) or mismapped measurements (discovered when reviewing mean SpO_2 by site). In each case, the DQ concern was referred for further "upstream" remediation in the pipeline beyond the project that discovered it and the site was informed.

Table 2. Data quality heuristics

No.	Heuristic	Type	No. of sites	sites (%)[a]
1	Not using (or improperly using) source CDM's controlled vocabulary in one or more fields	Source CDM conformance	13	23.2
2	COVID test result values not standardized or null	COVID tests	11	19.6
3	Lacking/incorrectly populating field(s) required by source CDM	Source CDM conformance	9	16.1
4	Implausible distribution of visit types (eg, 75% inpatient)	Encounters	7	12.5
5	Large number of "No Matching Concept" records (OMOP source only)	Coding completeness	6	10.7
6	Lacking table(s) required by source CDM	Source CDM conformance	5	9.0
7	Many or all inpatient visits lacking valid end dates	Encounters	5	9.0
8	Few or no clinical encounters coded with U07.1	Conditions	5	9.0
9	Implausible count of patients qualifying for phenotype	Demographics	3	5.4
10	Small number of unique measurement/observation types	Measurement/observation	2	3.6
11	PERSON_IDs in fact tables that are not in the PERSON table	Coding completeness	2	3.6
12	Primary keys are not unique	Coding completeness	2	3.6
13	Inconsistent local date shifting causing implausible timelines	Coding completeness	2	3.6
14	Implausible demographics (eg, 100% male patients)	Demographics	2	3.6
15	Data utility challenges (eg, missing mortality data)	Fitness for use	N/A	N/A

Items compiled here are from a qualitative analysis of the "Must Pass" data issues filed on any one of the 56 currently released N3C sites that resulted in a fix by the site. Fitness for Use is an additional heuristic that applies to all sites and is thus also included here. Simple formatting errors (eg, incorrect delimiters) and noncritical "Heads Up" issues are excluded from this analysis.

[a] Denominator: 56 sites; 37 unique sites are represented across these categories.

N3C: National COVID Cohort Collaborative; OMOP: Observational Medical Outcomes Partnership.

DISCUSSION

We have demonstrated that centralized DQ assessment reveals unique opportunities for iterative quality improvement for submitting sites. N3C's DQ checks take into account various dimensions introduced in prior DQ work, such as conformance to source data models, density and completeness (eg, of COVID test results), and plausibility (eg, of percentage of inpatient visits).[20,29] By participating in a consortium like N3C, sites receive routine feedback on their overall quality with tactical information on ways to address local issues. Moreover, N3C's dedicated team of analysts with protected time to concentrate on DQ, deep subject matter expertise, and access to powerful visualization tools enable efficient support for participating sites in making rapid improvements in high-priority areas. This process can be transformative; by combining efficient, continual assessment of DQ with a large volume of multisite data, it is possible to support more nuanced scientific questions with the scale and rigor that they require.

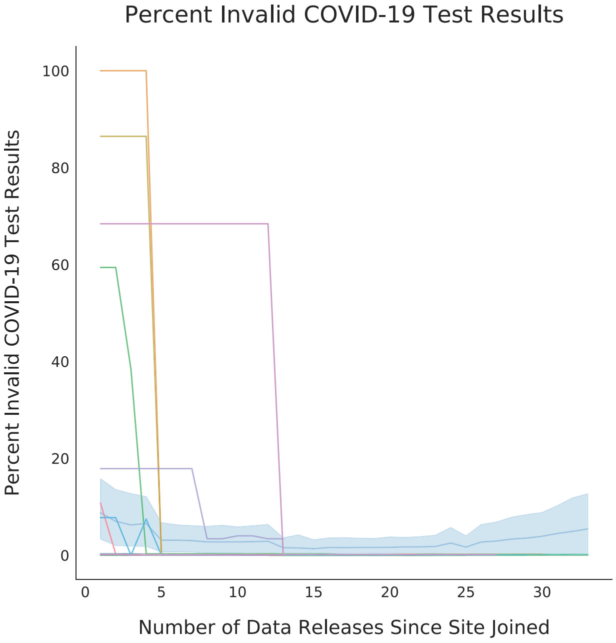

Figure 3. Improved percentages of valid COVID-19 test results across 11 N3C sites. The 11 sites shown here each had initial N3C submissions with high numbers of invalid (null, nonstandard) COVID test results. As time moves forward (left to right on the x-axis), drastic improvements are made following feedback from N3C. The blue line and shaded area represent the mean and standard deviation across all sites. N3C: National COVID Cohort Collaborative.

Centralized data enable site-to-site comparisons

When examining data across N3C partner sites, the centralized approach revealed significant site-to-site DQ differences that would have been challenging to discover in isolation. Many of these "data issues" are indeed errors, but others arise from differences in interpretation or adoption of CDM components. Sites' use of the encounter data domain is an example, where the definition of "one visit" can vary widely depending on the site's EHR, organizational structure, or billing practices (see Figure 5). The site's definitions of various visit types are generally not erroneous and likely comply with the rules of their source CDM. Problems surface, however, when data are combined across sites for multisite projects, leading to a need for data users to (1) understand that the issue exists and (2) develop consistent analytic workarounds and harmonization strategies. Benchmarking, or the ability to compare sites with their peers, is an efficient way to catch such issues.

Assessing the quantity and variety of data available per patient is another use case for benchmarking. Because instantiation and maintenance of CDMs are resource-intensive, it is common for sites to take a minimalistic approach to CDM data curation, particularly in the early days of implementing a new CDM. This may entail purposely

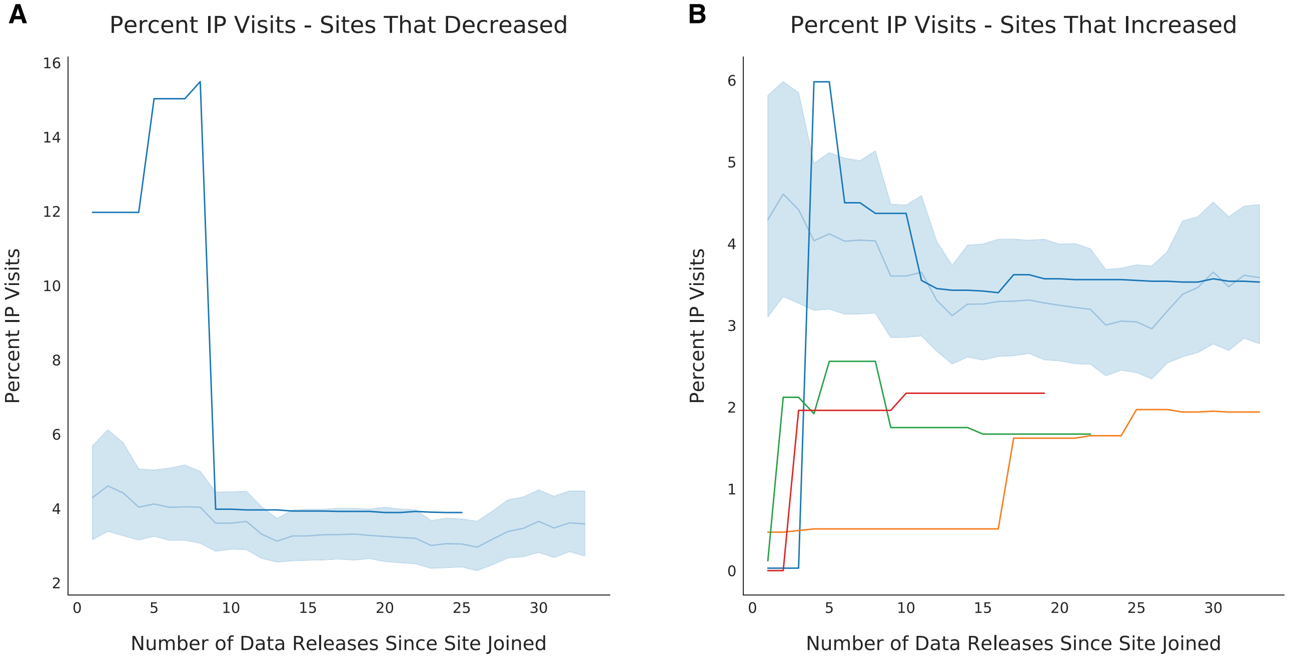

Figure 4. In A, one site's initial N3C submission had a proportion of visits of type inpatient far above that of similar sites; in B, 4 sites' initial submissions had no (or nearly no) inpatient visits. Our feedback encouraged the sites to re-examine and remap their source-to-CDM visit type mappings. In these cases, proportions improved. The shaded area reflects the mean and standard deviation of all sites. N3C: National COVID Cohort Collaborative.

We have 43 qualifying inpatient visits
We have 27 qualifying inpatient visits.
We have 806 qualifying inpatient visits
Site 1
Site 2
Site 3

Clearly, sites differ in how they define "a visit."

Site	Patient	Visit Type	Adm. Date	Disc. Date
1	123	IP	7/4/2020	7/8/2020
1	456	IP	5/6/2020	5/20/2020
2	987	IP	8/2/2019	8/7/2019
2	654	IP	9/3/2019	9/14/2019
3	234	IP	**1/26/2021**	1/26/2021
3	234	IP	**1/26/2021**	1/29/2021
3	234	IP	**1/26/2021**	1/30/2021
3	234	IP	**1/26/2021**	1/27/2021

Figure 5. Comparing sites within centralized data. One of the most stark differences we have observed among different sites is the different ways that a "visit" (or encounter) can be defined. Indeed, inpatient visits at several N3C sites are made up of a number (at times hundreds) of "microvisits"—consults with different specialists, imaging, infusions, et cetera. Because sites define inpatient visits so differently, they are difficult to harmonize. Centralized data make it easier to compare how sites define visits and develop derivative variables to enable harmonization. N3C: National COVID Cohort Collaborative.

choosing to support only a small number of lab tests, vitals, or other observations (eg, the top 50 most common labs; only blood pressure, weight, and height out of all possible vital measurements, etc.). However, this sparsity can have a big impact on what research questions are possible to answer using the data. Because N3C aims to be a multipurpose resource for COVID research, we use benchmarking to spot sites supporting fewer facts per patient (or a limited variety of clinical concepts) and encourage them to gradually add on as they can.

Benchmarking data can also serve as a clear and persuasive DQ communication tool. As shown in Figure 2, our ability to compare sites' coverage of lab and vital concepts across sites gives sites a "report card" to see which concepts have more or less coverage at their peer sites. This information can help sites prioritize bringing in new types of data, or may spur an investigation of a data issue of which the site was unaware. It should be noted that federated networks (particularly PCORnet and TriNetX) are also capable of site-to-site benchmarking, but at an aggregated rather than detailed level. To use Figure 5 as an example, aggregate checks could produce the inpatient visit counts reported on the left side, but would not enable the row-level deep dive shown on the right. Though row-level checks enable more detailed benchmarking, both aggregate and row-level comparisons are highly valuable.

Centralized data enable crowd sourced DQ evaluation

The N3C repository not only pools data but also brings together multidisciplinary teams of clinicians, researchers, and statisticians from across the network of organizations with N3C Data Use Agreements. N3C Enclave researchers are invited to join any of more than 25 Clinical Domain Teams, where they can share domain-specific knowledge with a diverse set of peers. Researchers who join a Domain Team have access to a shared workspace where they can create sets of derived variables specific to their research question and conduct-related analyses. As detailed dataset reviews are an essential step in this process, these teams may reveal resolvable data issues such as hard-to-harmonize variations among contributing CDMs or incorrect mappings at the individual record level. Sometimes these quality errors, inconsistencies, or omissions are remedied within the Enclave by N3C, such as unit of measure harmonization, whereas for other issues these discoveries are referred back to the contributing site for remediation.

The ability for research teams to review row-level data in the N3C Enclave also helps end-users write more accurate analytic code by avoiding blind assumptions related to how representation of clinical facts varies by site or CDM. For example,[30] when creating a flag to indicate the co-occurrence of a positive COVID test and a diagnosis code representing a comorbidity, date logic is applied in the code. Review of row-level data, which can be filtered to one or more sites whose variable distribution does not match other sites, may reveal that some measurement dates are representative of test *result* date instead of test *order* date–either of which may be acceptable in the source CDM. Though the data are not "wrong" in this case and do not need to be corrected by the contributing site, this ambiguity is a DQ issue nonetheless and requires resolution by data users during analysis. In general, identification of DQ issues that are highly dependent on the context of use[31,32] is, by design, left for analysts pursuing specific questions to discover. In such cases, our centralized DQ team serves as a liaison to the sites and passes feedback from analysts to sites for local discussion and remediation, if deemed necessary and high value by the site. Issues of this type are extremely challenging or impractical to identify without access to the row-level data, in addition to the ability to compare across sites.

Centralization increases DQ efficiency

While centralization alone is not a recipe for improved DQ, it does present opportunities to implement generalized solutions at scale. N3C adopted 2 postprocessing DQ processes that illustrate this: (1) interrogating information "loss" in standardized terminology mappings and (2) performing post hoc evaluation of harmonized clinical information to ensure analytical utility. These solutions are more practical to achieve in a centralized environment where economy of scale provides the ability to easily see strings or terms that could be supported in value sets, or logic that could be modified to bring clinical information to the right analytical domain. Moreover, a centralized data ingestion pipeline allows for bidirectional improvement as the target model (OMOP, in this case) can evolve to include conventions that capture the heterogeneity of source system data.

Finally, centralized review simply puts additional trained eyes on a site's data. The advantages to this are demonstrated by the fact that N3C's centralized review process found errors and room for improvement even in data that had passed either local review or one of the federated networks' required checks. Indeed, this combination of local checks, by personnel who intimately understand their site's data, and centralized checking, which can take advantage of economies of scale, may represent an ideal model for DQ assessment.

Limitations

N3C is devoted to a single disease, so some of our particular DQ checks derive from that focus. Though many of the issues identified here (eg, encounter definitions, unit harmonization, plausibility) are not limited to COVID-related data, the emphasis of our suite of DQ checks is on variables required for COVID research. As an example, we ensure careful harmonization of qualitative COVID lab test results, but do not perform checks at the same level of detail on non-COVID labs. Still, many of our checks listed in Table 1 can apply to clinical data more generically.

Centralization does mean that data are further from the EHR source, and we rely on local staff to be final arbiters of their own DQ. In addition, as sites may have already executed one of the federated networks' DQ checking protocols prior to submitting to N3C, the data that N3C receive may have already undergone a prior round DQ improvements. Yet, based on our results, we feel that the insights that sites receive from the composite experience of the entire Enclave add value even to previously improved data.

Despite the efficiency that N3C's centralized Enclave enables for the assessment of DQ issues, much of the heavy lift of definitive DQ remediation lies with the submitting sites. We acknowledge that consortial data resources such as N3C are only possible because of the efforts of local teams and believe that centralized DQ supports collaboration and knowledge exchange that also helps improve local DQ. In informatics research, there is generally limited funding available to specifically support local DQ, which makes N3C sites' engagement with our DQ process all the more impressive. This may be an indicator of demand for more centralized DQ in the future, given ongoing funding to do so. The ability to take some of the DQ workload off of local sites may incentivize site participation in future centralized repositories.

CONCLUSION

Federated data repositories, where the data remain at the generating site, offer the advantages of local curation by personnel deeply familiar with the data. Central repositories enable efficient DQ benchmarking at scale, and the generation of derivative, harmonized variables and units of measure for comparable and consistent analytics. Together, these advantages can synergize to a best of both worlds approach for DQ improvement and enhancement in clinical data repositories. Cooperation and communication between these complementary environments, as illustrated by the common data model communities and N3C, promise mutual advantage and maximal DQ.

FUNDING

This work was supported by NCATS U24TR002306, in addition to individual sites' NCATS grants:

Stony Brook University—U24TR002306 • University of Oklahoma Health Sciences Center—U54GM104938: Oklahoma Clinical and Translational Science Institute (OCTSI) • West Virginia University—U54GM104942: West Virginia Clinical and Translational Science Institute (WVCTSI) • University of Mississippi Medical Center—U54GM115428: Mississippi Center for Clinical and Translational Research (CCTR) • University of Nebraska Medical Center—U54GM115458: Great Plains IDeA-Clinical & Translational Research • Maine Medical Center—U54GM115516: Northern New England Clinical & Translational Research (NNE-CTR) Network • Wake Forest University Health Sciences—UL1TR001420: Wake Forest Clinical and Translational Science Institute • Northwestern University at Chicago—UL1TR001422: Northwestern University Clinical and Translational Science Institute (NUCATS) • University of Cincinnati—UL1TR001425: Center for Clinical and Translational Science and Training • The University of Texas Medical Branch at Galveston—UL1TR001439: The Institute for Translational Sciences • Medical University of South Carolina—UL1TR001450: South Carolina Clinical & Translational Research Institute (SCTR) • University of Massachusetts Medical School Worcester—UL1TR001453: The UMass Center for Clinical and Translational Science (UMCCTS) • University of Southern California—UL1TR001855: The Southern California Clinical and Translational Science Institute (SC CTSI) • Columbia University Irving Medical Center—UL1TR001873: Irving Institute for Clinical and Translational Research • George Washington Children's Research Institute—UL1TR001876: Clinical and Translational Science Institute at Children's National (CTSA-CN) • University of Kentucky—UL1TR001998: UK Center for Clinical and Translational Science • University of Rochester—UL1TR002001: UR Clinical & Translational Science Institute • University of Illinois at Chicago—UL1TR002003: UIC Center for Clinical and Translational Science • Penn State Health Milton S. Hershey Medical Center—UL1TR002014: Penn State Clinical and Translational Science Institute • The University of Michigan at Ann Arbor—UL1TR002240: Michigan Institute for Clinical and Health Research • Vanderbilt University Medical Center—UL1TR002243: Vanderbilt Institute for Clinical and Translational Research • University of Washington—UL1TR002319: Institute of Translational Health Sciences • Washington University in St. Louis—UL1TR002345: Institute of Clinical and Translational Sciences • Oregon Health & Science University—UL1TR002369: Oregon Clinical and Translational Research Institute • University of Wisconsin-Madison—UL1TR002373: UW Institute for Clinical and Translational Research • Rush University Medical Center—UL1TR002389: The Institute for Translational Medicine (ITM) • The University of Chicago—UL1TR002389: The Institute for Translational Medicine (ITM) • University of North Carolina at Chapel Hill—UL1TR002489: North Carolina Translational and Clinical Science Institute • University of Minnesota—UL1TR002494: Clinical and Translational Science Institute • Children's Hospital Colorado—UL1TR002535: Colorado Clinical and Translational Sciences Institute • The University of Iowa—UL1TR002537: Institute for Clinical and Translational Science • The University of Utah—UL1TR002538: Uhealth Center for Clinical and Translational Science • Tufts Medical Center—UL1TR002544: Tufts Clinical and Translational Science Institute • Duke University—UL1TR002553: Duke Clinical and Translational Science Institute • Virginia Commonwealth University—UL1TR002649: C. Kenneth and Dianne Wright Center for Clinical and Translational Research • The Ohio State University—UL1TR002733: Center for Clinical and Translational Science • The University of Miami Leonard M. Miller School of Medicine—UL1TR002736: University of Miami Clinical and Translational Science Institute • University of Virginia—UL1TR003015: iTHRIV integrated Translational Health Research Institute of Virginia • Carilion Clinic—UL1TR003015: iTHRIV integrated Translational

Health Research Institute of Virginia • University of Alabama at Birmingham—UL1TR003096: Center for Clinical and Translational Science • Johns Hopkins University—UL1TR003098: Johns Hopkins Institute for Clinical and Translational Research • University of Arkansas for Medical Sciences—UL1TR003107: UAMS Translational Research Institute • Nemours—U54GM104941: Delaware CTR ACCEL Program • University Medical Center New Orleans—U54GM104940: Louisiana Clinical and Translational Science (LA CaTS) Center • University of Colorado Denver, Anschutz Medical Campus—UL1TR002535: Colorado Clinical and Translational Sciences Institute • Mayo Clinic Rochester—UL1TR002377: Mayo Clinic Center for Clinical and Translational Science (CCaTS) • Tulane University—UL1TR003096: Center for Clinical and Translational Science • Loyola University Medical Center—UL1TR002389: The Institute for Translational Medicine (ITM) • Advocate Health Care Network—UL1TR002389: The Institute for Translational Medicine (ITM) • OCHIN—INV-018455: Bill and Melinda Gates Foundation grant to Sage Bionetworks • The Rockefeller University—UL1TR001866: Center for Clinical and Translational Science • The Scripps Research Institute—UL1TR002550: Scripps Research Translational Institute • University of Texas Health Science Center at San Antonio—UL1TR002645: Institute for Integration of Medicine and Science • The University of Texas Health Science Center at Houston—UL1TR003167: Center for Clinical and Translational Sciences (CCTS) • NorthShore University HealthSystem—UL1TR002389: The Institute for Translational Medicine (ITM) • Yale New Haven Hospital—UL1TR001863: Yale Center for Clinical Investigation • Emory University—UL1TR002378: Georgia Clinical and Translational Science Alliance • Weill Medical College of Cornell University—UL1TR002384: Weill Cornell Medicine Clinical and Translational Science Center • Montefiore Medical Center—UL1TR002556: Institute for Clinical and Translational Research at Einstein and Montefiore • Medical College of Wisconsin—UL1TR001436: Clinical and Translational Science Institute of Southeast Wisconsin • University of New Mexico Health Sciences Center—UL1TR001449: University of New Mexico Clinical and Translational Science Center • George Washington University—UL1TR001876: Clinical and Translational Science Institute at Children's National (CTSA-CN) • Stanford University—UL1TR003142: Spectrum: The Stanford Center for Clinical and Translational Research and Education • Regenstrief Institute—UL1TR002529: Indiana Clinical and Translational Science Institute • Cincinnati Children's Hospital Medical Center—UL1TR001425: Center for Clinical and Translational Science and Training • Boston University Medical Campus—UL1TR001430: Boston University Clinical and Translational Science Institute • The State University of New York at Buffalo—UL1TR001412: Clinical and Translational Science Institute • Aurora Health Care—UL1TR002373: Wisconsin Network For Health Research • Brown University—U54GM115677: Advance Clinical Translational Research (Advance-CTR) • Rutgers, The State University of New Jersey—UL1TR003017: New Jersey Alliance for Clinical and Translational Science • Loyola University Chicago—UL1TR002389: The Institute for Translational Medicine (ITM) • UL1TR001445: Langone Health's Clinical and Translational Science Institute • Children's Hospital of Philadelphia—UL1TR001878: Institute for Translational Medicine and Therapeutics • University of Kansas Medical Center—UL1TR002366: Frontiers: University of Kansas Clinical and Translational Science Institute • Massachusetts General Brigham—UL1TR002541: Harvard Catalyst • Icahn School of Medicine at Mount Sinai—UL1TR001433: ConduITS Institute for Translational Sciences • Ochsner Medical Center—U54GM104940: Louisiana Clinical and Translational Science (LA CaTS) Center • HonorHealth—None (Voluntary) • University of California, Irvine—UL1TR001414: The UC Irvine Institute for Clinical and Translational Science (ICTS) • University of California, San Diego—UL1TR001442: Altman Clinical and Translational Research Institute • University of California, Davis—UL1TR001860: UCDavis Health Clinical and Translational Science Center • University of California, San Francisco—UL1TR001872: UCSF Clinical and Translational Science Institute • University of California, Los Angeles—UL1TR001881: UCLA Clinical Translational Science Institute • University of Vermont—U54GM115516: Northern New England Clinical & Translational Research (NNE-CTR) Network • Arkansas Children's Hospital—UL1TR003107: UAMS Translational Research Institute—UL1TR001857: University of Pittsburgh CTSA.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

AUTHOR CONTRIBUTIONS

Database/information systems admin: JAM, ATG, EAN, BA, and AM. Manuscript drafting: HPL, RAM, MBP, CGC, DLG, ERP, ATG, KK, MM, JL, and AW. Clinical subject matter expertise: MBP, CGC, and SG. Clinical data model expertise: HPL, MBP, CGC, DLG, ERP, ATG, MAH, KK, and TZ. Data analysis: JAM, HPL, RAM, CGC, ERP, ATG, SSH, ATG, MMB, KB, and JL. Data curation: CGC, DLG, ERP, ATG, SSH, ATG, EAN, KK, MM, and KB. Data quality assurance: HPL, MBP, CGC, ERP, ATG, SSH, MMB, EAN, KK, MM, RZ, and TZ. Data integration: MBP, CGC, DLG, ERP, ATG, SSH, MMB, EAN, KK, MM, BA, RZ, and TZ. Funding acquisition: CGC, NQ, and MAH. N3C Phenotype definition: MBP, CGC, ERP, KK, and MM. Project management: JAM, CGC, DLG, NQ, MAH, and AW. Biological subject matter expertise: ATG and MAH. Statistical analysis: ATG Governance: CGC and MAH. Marketing and communications: NQ. Regulatory oversight/admin: CGC and MAH. Data security: ATG, NQ, MMB, EAN, and BA. Critical revision of the manuscript for important intellectual content: JAM, HPL, RAM, CGC, DLG, ERP, ATG, KK, MM, SG, and AW.

ACKNOWLEDGMENTS

We gratefully acknowledge contributions from the following N3C core teams (leads indicated with an asterisk):

- Principal Investigators: MAH*, CGC*, Kenneth R. Gersing, and AW.
- Workstream, subgroup, and administrative leaders: MAH*, Tellen D. Bennett, CGC, David A. Eichmann, Justin Guinney, Warren A. Kibbe, Hongfang Liu, Philip R.O. Payne, ERP, Peter N. Robinson, Joel H. Saltz, Heidi Spratt, Justin Starren, Christine Suver, Adam B. Wilcox, Andrew E. Williams, and Chunlei Wu.
- Key liaisons and regulatory staff at data partner sites.
- Individuals at the sites who are responsible for creating the datasets and submitting data to N3C.
- Data Ingest and Harmonization Team: CGC*, ERP*, D.L.G., SSH, KK, HPL, RAM, MM, MBP, XTZ, and RLZ.
- Phenotype Team (Individuals who create the scripts that the sites use to submit their data, based on the COVID and Long COVID definitions): ERP*, Marshall Clark, KK, Adam M. Lee, Robert T. Miller, Karthik Natarajan, Shyam Visweswaran, MM, MBP, and KMW.
- Project Management and Operations Team: AW*, Yooree Chae, Connor Cook, Alexandra Dest, Racquel R. Dietz, Thomas Dillon, Patricia A. Francis, Rafael Fuentes, Alexis Graves, JAM, Andrew J. Neumann, Shawn T. O'Neil, Usman Sheikh, Andréa M. Volz, and Elizabeth Zampino.

- Partners from NIH and other federal agencies: CPA*, Kenneth R. Gersing*, Samuel Bozzette, Mariam Deacy, Nicole Garbarini, Michael G. Kurilla, Sam G. Michael, Joni L. Rutter, Meredith Temple-O'Connor.
- Analytics Team (Individuals who build the Enclave infrastructure, help create codesets, variables, and help Domain Teams and project teams with their datasets): BA*, MB, KRB, ATG, AM, and NQ.
- Publication Committee Management Team: Mary Morrison Saltz*, Christine Suver*, CGC, MAH, JAM, AMV, and AW.
- Publication Committee Review Team: Carolyn Bramante, Jeremy Richard Harper, Wenndy Hernandez, Farrukh M. Koraishy, Federico Mariona, Amit Saha, and Satyanarayana Vedula.

We also acknowledge the contributions of the following partners and collaborators at data contributing sites: Adil Alaoui, Satyanarayana Vedula, Jason Womack, David Randolph, Melissa Basford, Alejandro Araya, and R. Joseph Applegate.

CONFLICT OF INTEREST STATEMENT

JAM and MAH are cofounders of Pryzm Health. BA, MB, KRB, ATG, AM, NQ, and EN are employees of Palantir Technologies. MBP is an employee of TriNetX, LLC.

DATA AVAILABILITY

The N3C data transfer to NCATS is performed under a Johns Hopkins University Reliance Protocol # IRB00249128 or individual site agreements with NIH. The N3C Data Enclave is managed under the authority of the NIH; information can be found at https://ncats.nih.gov/n3c/resources. Enclave data are protected and can be accessed for COVID-related research with an approved (1) IRB protocol and (2) Data Use Request (DUR). A detailed accounting of data protections and access tiers is found in Ref.[1] Enclave and data access instructions can be found at https://covid.cd2h.org/for-researchers.

REFERENCES

1. Haendel MA, Chute CG, Bennett TD, *et al.*; N3C Consortium. The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc* 2020; 28 (3): 427–43.
2. Bennett TD, Moffitt RA, Hajagos JG, *et al.* Clinical Characterization and Prediction of Clinical Severity of SARS-CoV-2 Infection Among US Adults Using Data From the US National COVID Cohort Collaborative. *JAMA Netw Open* 2021; 4 (7): e2116901. doi:10.1001/jamanetworkopen.2021.16901
3. National COVID Cohort Collaborative. N3C Cohort Exploration. https://covid.cd2h.org/dashboard/ Accessed Jun 28, 2021.
4. NCATS. NIH COVID-19 Data Warehouse Data Transfer Agreement. 2020. https://ncats.nih.gov/files/NCATS_Data_Transfer_Agreement_05-11-2020_Updated%20508.pdf Accessed July 15, 2021.
5. Vogt TM, Lafata JE, Tolsma DD, *et al.* The role of research in integrated health care systems: the HMO research network. *Perm J* 2004; 8: 10–7.
6. Behrman RE, Benner JS, Brown JS, *et al.* Developing the Sentinel System—a national resource for evidence development. *N Engl J Med* 2011; 364 (6): 498–9.
7. 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.
8. 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.
9. Stacey J, Mehta M. Using EHR data extraction to streamline the clinical trial process. *Clin Res* 2017; 4: 2–7.
10. Stapff M. Use of electronic health data in clinical development. *Pharm Ind* 2017; 79: 204–10.
11. Stapff MP. Using real world data to assess cardiovascular outcomes of two antidiabetic treatment classes. *World J Diabetes* 2018; 9 (12): 252–7.
12. Reisinger SJ, Ryan PB, O'Hara DJ, *et al.* Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. *J Am Med Inform Assoc* 2010; 17 (6): 652–62.
13. Stang PE, Ryan PB, Racoosin JA, *et al.* Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann Intern Med* 2010; 153 (9): 600–6.
14. 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.
15. Adams L, Kennedy S, Allen L, *et al.* Innovative solutions for state medicaid programs to leverage their data, build their analytic capacity, and create evidence-based policy. *EGEMS (Wash DC)* 2019; 7 (1): 41.
16. Gillespie BW, Laurin L-P, Zinsser D, *et al.* Improving data quality in observational research studies: report of the Cure Glomerulonephropathy (CureGN) network. *Contemp Clin Trials Commun* 2021; 22: 100749.
17. Bian J, Lyu T, Loiacono A, *et al.* Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. *J Am Med Inform Assoc* 2020; 27 (12): 1999–2010.
18. Kahn MG, Raebel MA, Glanz JM, *et al.* A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Med Care* 2012; 50 (Suppl): S21–9.
19. Khare R, Utidjian LH, Razzaghi H, *et al.* Design and refinement of a data quality assessment workflow for a large pediatric research network. *EGEMS (Wash DC)* 2019; 7 (1): 36.
20. Kahn MG, Callahan TJ, Barnard J, *et al.* A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS (Wash DC)* 2016; 4 (1): 1244.
21. Holve E, Kahn M, Nahm M, *et al.* A comprehensive framework for data quality assessment in CER. *AMIA Jt Summits Transl Sci Proc* 2013; 2013: 86–8.
22. 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.
23. Henley-Smith S, Boyle D, Gray K. Improving a secondary use health data warehouse: proposing a multi-level data quality framework. *EGEMS (Wash DC)* 2019; 7 (1): 38.
24. Johnson KE, Kamineni A, Fuller S, *et al.* How the provenance of electronic health record data matters for research: a case example using system mapping. *EGEMS (Wash DC)* 2014; 2 (1): 1058.
25. OHDSI. Data Quality Dashboard. Github https://github.com/OHDSI/DataQualityDashboard Accessed Jul 8, 2021.
26. Qualls LG, Phillips TA, Hammill BG, *et al.* Evaluating foundational data quality in the national patient-centered clinical research network (PCORnet®). *EGEMS (Wash DC)* 2018; 6 (1): 3.
27. Sengupta S, Bachman D, Laws R, *et al.* Data quality assessment and multi-organizational reporting: tools to enhance network knowledge. *EGEMS (Wash DC)* 2019; 7 (1): 8.
28. OMOP 5.3.1 Specification. https://ohdsi.github.io/CommonDataModel/cdm531.html Accessed Aug 17, 2021.
29. 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.
30. NCATS_N3C_Data_Use_Agreement.pdf. https://ncats.nih.gov/files/NCATS_N3C_Data_Use_Agreement.pdf Accessed July 15, 2021.
31. Jiang G, Dhruva SS, Chen J, *et al.* Feasibility of capturing real-world data from health information technology systems at multiple centers to assess cardiac ablation device outcomes: a fit-for-purpose informatics analysis report. *J Am Med Inform Assoc* 2021; 28 (10): 2241–50.
32. Nahm M. Data quality in clinical research. In: Rachel L, Richesson JEA, ed. *Clinical Research Informatics*. New York, NY: Springer; 2012: 175–201.