scientific data

Check for updates

OPEN

ARTICLE

Multimodal data curation via interoperability: use cases with the Medical Imaging and Data Resource Center

Weijie Chen[1,9]✉, Heather M. Whitney[2,9]✉, Seyed Kahaki[1], Christopher Meyer[3], Hui Li[2], Rui Carlos Sá[4,8], Diane Lauderdale[5], Sandy Napel[6], Kenneth Gersing[7], Robert L. Grossman[3] & Maryellen L. Giger[2]

Interoperability (the ability of data or tools from non-cooperating resources to integrate or work together with minimal effort) is particularly important for curation of multimodal datasets from multiple data sources. The Medical Imaging and Data Resource Center (MIDRC), a multi-institutional collaborative initiative to collect, curate, and share medical imaging datasets, has made interoperability with other data commons one of its top priorities. The purpose of this study was to demonstrate the interoperability between MIDRC and two other data repositories, BioData Catalyst (BDC) and National Clinical Cohort Collaborative (N3C). Using interoperability capabilities of the data repositories, we built two cohorts for example use cases, with each containing clinical and imaging data on matched patients. The representativeness of the cohorts is characterized by comparing with CDC population statistics using the Jensen-Shannon distance. The process and methods of interoperability demonstrated in this work can be utilized by MIDRC, BDC, and N3C users to create multimodal datasets for development of artificial intelligence/machine learning models.

Introduction

Multimodal data integration of radiological and histological imaging, clinical data, and molecular diagnostics has the potential to advance medicine beyond the current standard of care[1]. Artificial intelligence (AI), machine learning (ML), and big-data analytic technologies are expected to play a critical role in this advancement, yet these techniques are data hungry and require large amounts of multimodal data to train and validate AI/ML models[2]. Fortunately, there are now many data commons and repositories that are publicly available thanks to significant efforts by many organizations including the National Institutes of Health (NIH)[3], which has supported several repositories, such as The Cancer Imaging Archive (TCIA, https://www.cancerimagingarchive.net/)[4,5], The Imaging Data Commons (IDC, https://datacommons.cancer.gov/repository/imaging-data-commons)[6], the National Clinical Cohort Collaborative (N3C, https://ncats.nih.gov/research/research-activities/n3c/overview)[7], the BioData Catalyst (https://biodatacatalyst.nhlbi.nih.gov/)[8], The Database of Genotypes and Phenotypes (dbGaP, https://www.ncbi.nlm.nih.gov/gap/)[9,10], and the Medical Imaging and Data Resource Center (MIDRC, https://data.midrc.org/)[11]. Each was initiated and sustained for different purposes. Some, like TCIA, IDC, and MIDRC, contain de-identified medical images. Others, such N3C and dbGaP, contain medical record information related to medical images, such as clinical measurements taken near or at the time of imaging. The theme of this paper is the interoperability of different data commons with MIDRC.

[1]Division of Imaging, Diagnostics, and Software Reliability, Office of Science and Engineering Laboratories, CDRH, US FDA, Silver Spring, MD, USA. [2]Department of Radiology, University of Chicago, Chicago, IL, USA. [3]Center for Translational Data Science, University of Chicago, Chicago, IL, USA. [4]National Institute of Biomedical Imaging and Bioengineering, National Institutes of Health, Bethesda, MD, USA. [5]Department of Public Health Sciences, University of Chicago, Chicago, IL, USA. [6]Department of Radiology, Stanford University School of Medicine, Stanford, CA, USA. [7]Office of the Director, National Center for Advancing Translational Sciences, National Institutes of Health, Bethesda, MD, USA. [8]Present address: Department of Medicine, University of California, San Diego, CA, USA. [9]These authors contributed equally: Weijie Chen, Heather M. Whitney. ✉e-mail: weijie.chen@fda.hhs.gov; hwhitney@uchicago.edu

Interoperability is one of the key guiding principles for scientific data management and stewardship outlined in the FAIR principles (Findability, Accessibility, Interoperability, and Reusability), in which interoperability is defined as "the ability of data or tools from non-cooperating resources to integrate or work together with minimal effort"[12]. To be interoperable, according to the FAIR principles, the following must be true:

- The (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- The (meta)data use vocabularies that follow FAIR principles.
- The (meta)data include qualified references to other (meta)data.

The importance of interoperability in digital healthcare systems has been well recognized. Lehne *et al.*[13] stressed the importance of interoperability for AI and big-data analytics, medical communication, medical research, and international cooperation. Perlin[14] emphasized the central role of interoperability in realizing the economic and clinical benefits of big data. The authors[14] suggested improving patient identification and data matching as one of the priorities in advancing health information technology interoperability, as errors in patient data matching can result in suboptimal care and medical errors.

Despite its well-recognized value, interoperability has not been commonly implemented in practice due to both technical and governance challenges. Lack of interoperability has resulted in isolated data silos and incompatible systems that prevented the linking of data from multiple sources, which is particularly critical in multimodal data applications where data are scarce. Intentional interoperability efforts are needed to create multimodal datasets to address this scarcity.

MIDRC is a multi-institutional collaborative initiative driven by the medical imaging community that was initiated in late summer 2020 to help combat the global COVID-19 health emergency. Leveraging the existing and developing infrastructure provided by the participating organizations, MIDRC serves as a linked-data commons that coordinates access to data and harmonizes data management activities. MIDRC was designed to follow the FAIR principles, including interoperability. Now, MIDRC continues to expand with data ingestion of imaging studies acquired for diseases beyond COVID-19 such as oncology. Since its inception, MIDRC has also fostered FAIR principles by hosting a simple-to-use data portal (http://data.midrc.org) for exploration and cohort building, by sharing data as well as associated algorithms openly and freely.

The purpose of this study was to demonstrate the interoperability between MIDRC and two other data repositories, BioData Catalyst (BDC, Fig. 1) and N3C (Fig. 2) by describing the datasets that result from interoperability efforts between these repositories. The focus was to create cohorts for two example use cases: (1) a cohort for multi-omics association studies and data fusion (demonstrated on BDC) and (2) a cohort for developing AI computer vision models for medical images (demonstrated on N3C).

Results

We developed methods of interoperability between MIDRC and BDC and between MIDRC and N3C (see "Methods" for details) and used these methods to curate two multimodal datasets. The first dataset, the Repository of Electronic Data COVID-19 Observational Study (RED CORAL, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002363.v1.p1, dbGaP Study Accession: phs002363.v1.p1) dataset[15], was collected by the Prevention and Early Treatment of Acute Lung Injury (PETAL) Network for investigation of demographics, clinical characteristics, risk factors, care practices, outcomes and resource utilization of patients hospitalized with severe acute COVID-19. The clinical data of RED CORAL (demographics, laboratory test results, medical history and blood pressure tests) is hosted by BDC with 1,480 unique patients and the imaging data is hosted by MIDRC. Via interoperability, we identified 1,477 unique patients with matches between BDC and MIDRC, with 1,223 patients having images acquired March1 to April 1, 2020 that are currently hosted on MIDRC. The images include chest X-ray images for 1,200 patients and chest CT images for 226 patients. Table 1 summarizes the clinical and demographic characteristics of the dataset of 1,223 unique patients with matches between BDC and MIDRC.

The second dataset, the National Clinical Cohort Collaborative (N3C)[7], is a collection of data from over 80 institutions originally created to expand knowledge and treatment of COVID-19. N3C holds a wide range of clinical data, including clinical observations, lab results, medication records, procedure descriptions, and visits. As of March 2024, N3C holds data for 22.1 million unique persons, including over 8.7 million COVID-19 positive cases. Via interoperability, we identified 2,124 unique patients with matches between N3C and MIDRC, with images having been acquired between March 2020 and June 2021. Table 2 summarizes the demographic information of the patients in the matched dataset of 2,124 unique patients. Additional patient characteristics including history of COVID and smoking status were identified using the N3C Logic Liaison tools which serve as value sets that map concepts with values (https://covid.cd2h.org/dashboard/concept-sets).

In order to measure the representativeness of the cohorts to the actual COVID-19 positive patient populations, we used the Jensen-Shannon Distance (JSD) metric[16] to characterize the representativeness of the specific datasets relative to cumulative COVID-19 positive patient case counts from the Centers for Disease Control and Prevention (CDC)[17] (Tables 3–5). Measuring representativeness enables user to assess how similar patients in the data subset are to the broader population, and the JSD provides a summary measure that takes into account multiple subgroups. The JSD indicated varying levels of difference in the similarity of the patient demographics in RED CORAL for which imaging data exists in MIDRC and the patients in N3C for which imaging data exists in MIDRC, compared to the cumulative COVID-19 positive patient case counts (Fig. 3). It appears that the two datasets represent the patient population (as defined by the CDC statistics) well in terms of sex distributions with both JSD values below 0.2. In contrast, the raw JSD values for race and ethnicity distributions are in the range of 0.45–0.6 for both datasets, and the race JSD value for the N3C dataset is still above 0.5 after adjustment

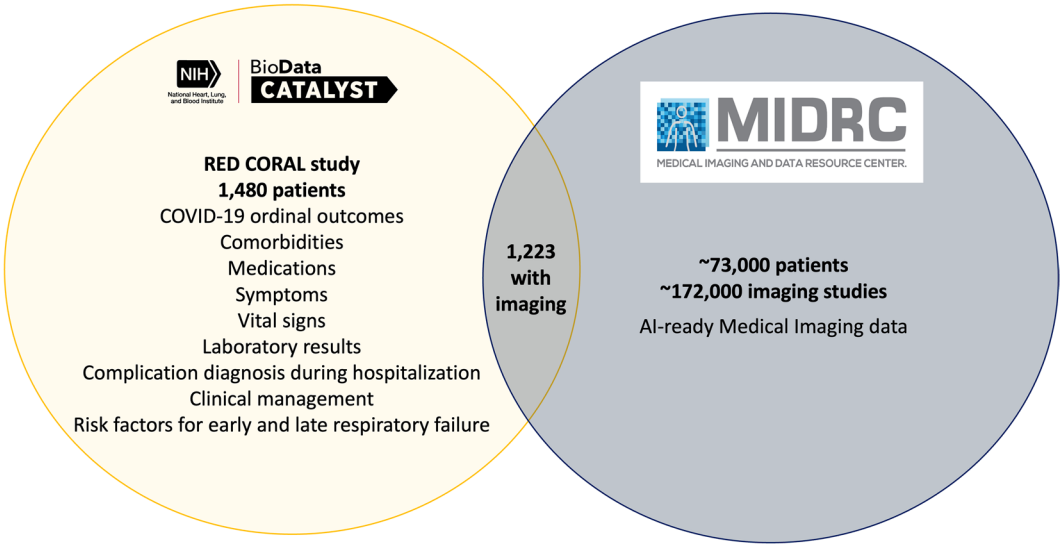

Fig. 1 Overview of interoperability between MIDRC and BioData Catalyst. Relative sizes of the datasets are not shown to scale.

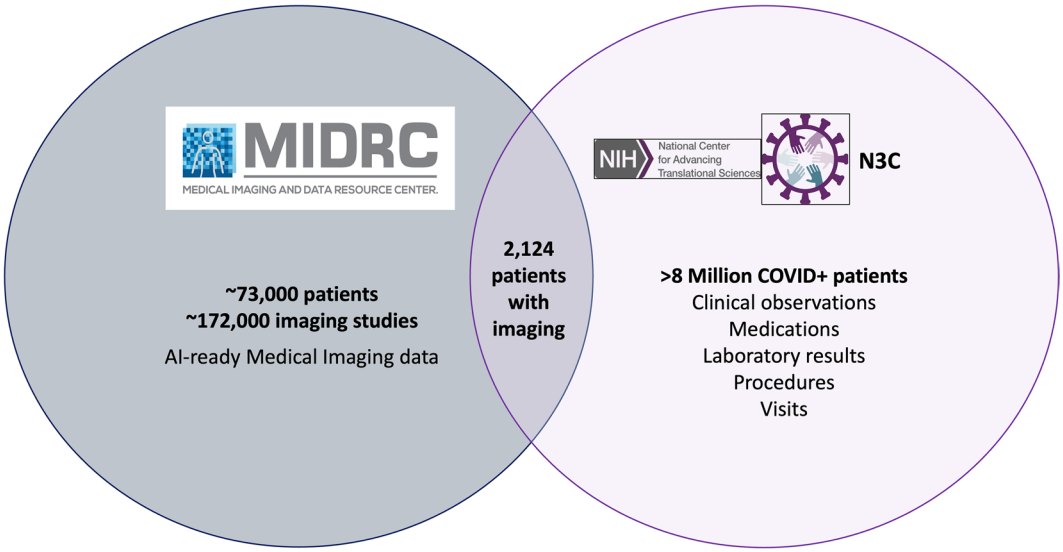

Fig. 2 Overview of interoperability between MIDRC and N3C. Relative sizes of the datasets are not shown to scale.

assuming missing at random. A close examination of the race distributions indicates substantial difference between the curated MIDRC-N3C overlapping subset and the CDC population statistics. For example, the Black and White subpopulations occupy 76.3% and 15.9% respectively in the N3C dataset whereas the corresponding figures in the CDC population are 8.9% and 46.1% (or 15.1% and 78.6% respectively after adjustment for missing at random). Such a characterization has important implications in the development and assessment of AI/ML models. Subgroup analysis would be necessary to examine the effect of the mismatch and mitigation may be necessary to avoid biased performance[18] on the two subgroups.

Discussion

The curation of these datasets was conducted with the goal of demonstrating interoperability between data repositories funded and created with separate mechanisms and aims. Through collaboration and cooperation between governance organizations, the datasets demonstrate the potential gains that can be made through multimodal research by identifying cohorts of patients with data held at different repositories. Such multimodal datasets may enrich research initiatives, with the potential for greater information compared to using single modality datasets on their own.

The multimodal data interoperability sets described here were intentionally crafted using FAIR principles, which notably include the goal of "minimal effort." One outstanding factor in interoperability between these

Number of patients		N = 1223
Age(years) (Standard Deviation)		60.55 ± 16.33
Sex	Female	539 (44.1%)
	Male	680 (55.6%)
	Missing	4 (0.4%)
Smoking	Never	764 (62.5%)
	Former	338 (27.6%)
	Current	53 (4.3%)
	Missing	68(5.6%)
COVID severity	Not ICU	719 (58.8%)
	ICU	500 (40.9%)
	Missing	4 (0.3%)
Ethnicity	Not Hispanic or Latino	933 (76.3%)
	Hispanic or Latino	208 (17.0%)
	Unknown	78 (6.4%)
	Missing	4 (0.3%)
Race	American Indian or Alaska Native	0 (0.0%)
	Asian	61 (5.0%)
	Black or African American	373 (30.5%)
	Native Hawaiian or Other Pacific Islander	0 (0.0%)
	White	589 (48.2%)
	Multiple	10 (0.8%)
	Other	145 (11.9%)
	Not reported	41 (3.4%)
	Missing	4 (0.3%)

Table 1. Characteristics of patients in the RED CORAL dataset with clinical data in BDC matched with imaging data in MIDRC.

Number of patients		N = 2124
Age (years)* (Standard Deviation)		57.65 ± 25.99
Sex*	Female	1141 (53.7%)
	Male	983 (46.3%)
Smoking†	False	1414 (66.6%)
	True	212 (10.0%)
	Unknown	498 (23.4%)
COVID severity†	False	1825 (85.9%) (No clinical record evidence of COVID-19 related ECMO or invasive ventilation)
	True	299 (14.1%) (Ever received COVID-19 related ECMO or invasive ventilation)
Ethnicity*	Not Hispanic or Latino	1968 (92.7%)
	Hispanic or Latino	132 (6.2%)
	Not reported	24 (1.1%)
Race*	American Indian or Alaska Native	<20 (<1%)
	Asian	26 (1.2%)
	Black or African American	1620 (76.3%)
	Native Hawaiian or Other Pacific Islander	<20 (<1%)
	White	338 (15.9%)
	Not reported	134 (6.3%)

Table 2. Characteristics of patients with clinical data in N3C matched with imaging data in MIDRC. *Patient characteristics in the matching MIDRC-N3C dataset as recorded in MIDRC. †Patient characteristics in the matching MIDRC-N3C dataset as of 30 May 2024 as recorded in N3C. Smoker status was identified from the Covid-19 Patient Summary Facts Table Logic Liaison available at N3C via variable TOBACCOSMOKER_before_or_day_of_covid_indicator. Note that the smoker status as determined from the Covid-19 Patient Summary Facts Table Logic Liaison does not differentiate between current or former tobacco use. COVID severity status was identified from the Invasive Respiratory Support Logic Liaison available at N3C, which is determined from the following datasets: critical_covid_ecmo_by_procedure, critical_covid_invasive_vent_by_observation, critical_covid_invasive_vent_by_condition, critical_covid_inasive_vent_by_procedure.

Time period	Female	Male	Not reported
March 2020 (N_c = 381,963)	45.3% (50.1%)	45.2% (49.9%)	9.6%
March 2020 – June 2021 (N_c = 248,769,785)	50.3% (52.4%)	45.7% (47.6%)	4.1%

Table 3. Proportion of COVID-19 positive case counts by sex calculated based on CDC data over the comparison periods used in this study. N_c: Number of counts reported by the CDC. Reported as % from raw data (% after adjustment assuming data missing at random). Percentages may not add to 100% due to rounding. Note that the CDC data represent individual tests reported to the CDC.

Time period	American Indian or Alaska Native	Asian	Black	Native Hawaiian or Other Pacific Islander	White	Multiple or Other	Not reported
March 2020	0.07% (0.15%)	2.5% (4.9%)	14.8% (29.6%)	0.012% (0.02%)	32.3% (64.4%)	0.47% (0.93%)	50.0%
March 2020 – June 2021	0.61% (1.0%)	2.1% (3.6%)	8.9% (15.1%)	0.075% (0.13%)	46.1% (78.6%)	0.9% (1.5%)	41.4%

Table 4. Proportion of COVID-19 positive case counts by race over the comparison periods used in this study calculated based on CDC data. Reported as % from raw data (% after adjustment for data missing at random). Percentages may not add to 100% due to rounding.

Time period	Hispanic or Latino	Not Hispanic or Latino	Not reported
March 2020	8.1% (17.8%)	37.4% (82.2%)	54.5%
March 2020 – June 2021	11.6% (21.6%)	42.3% (78.4%)	46.0%

Table 5. Proportion of COVID-19 positive case counts by ethnicity over the comparison periods used in this study calculated based on CDC data. Reported as % from raw data (% after adjustment for data missing at random). Percentages may not add to 100% due to rounding.

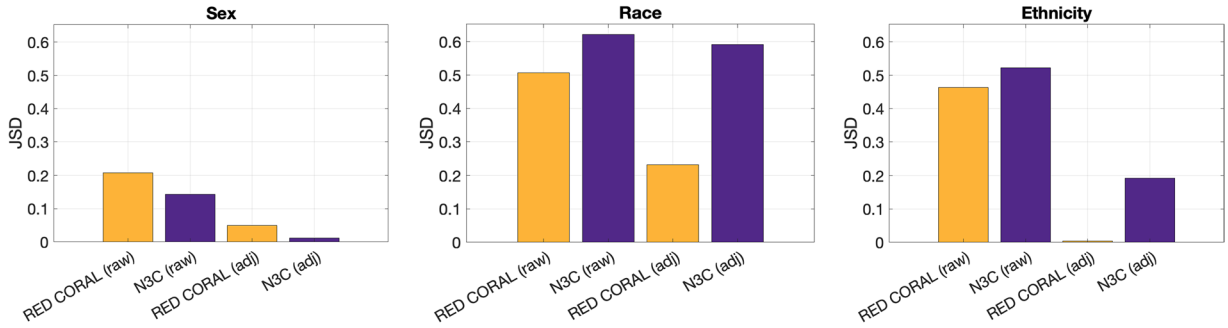

Fig. 3 Representativeness of patient cohorts relative to the CDC cumulative COVID-19 + case counts over their respective time periods. Results are shown when calculating the JSD using (1) the raw CDC data and (2) all data adjusted when assuming missing-at-random data. Lower JSD values indicated more similarity between the dataset and its comparison group (cumulative COVID-19 positive patient case counts from the CDC).

repositories that increases the amount of effort for interoperability is the differences in governance systems. By design, the data held in MIDRC undergoes extensive de-identification as they are ingested and thus are freely available to all who register for access. In contrast, N3C places restrictions on data access, does not allow data download, and requires an application process for data uploads, due to the nature of the controlled data it holds. This means that users who wish to interoperate between the two repositories are limited to using N3C as the computational enclave for interacting with the connected data. On the other hand, BDC users can download and compute on the data locally (on their own computer or cloud-based computational spaces) after authorization is granted by dbGaP, and then can access the data during the active period of the authorized project following a Data Use Certification Agreement, which specifies terms such as "for research use", user responsibilities, non-identification, and so on. It takes considerable administrative efforts to implement interoperability due to the differences in governance systems. Technologically, interoperability between MIDRC and BDC can be automated to a large extent; for this use case, matching of data and linking of identifiers across the two repositories leveraged the Gen3 crosswalk service. However, as implemented here, the Gen3 crosswalk service was not totally automatic because the governance of the two data commons is different. It is possible that early planning of coordination between data repositories with different governance models could broaden the availability of multimodal data through interoperability, reducing time and effort needed.

In the future, interoperability among the many data commons may be facilitated by further development and adoption of multimodal healthcare data standards, which refer to "methods, protocols, terminologies, and

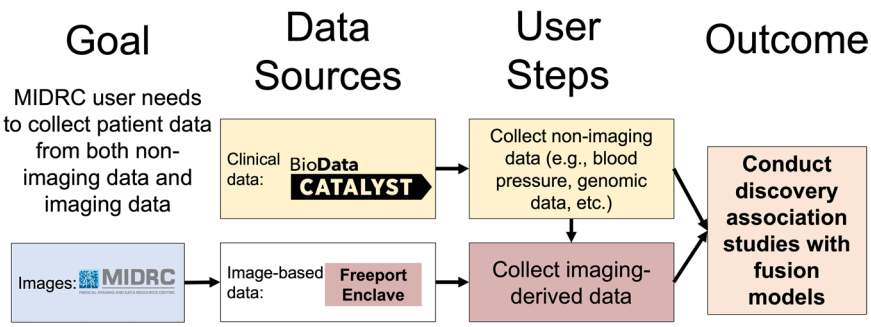

Fig. 4 Overview of example use case of interoperability between MIDRC and BioData Catalyst for multimodal data fusion.

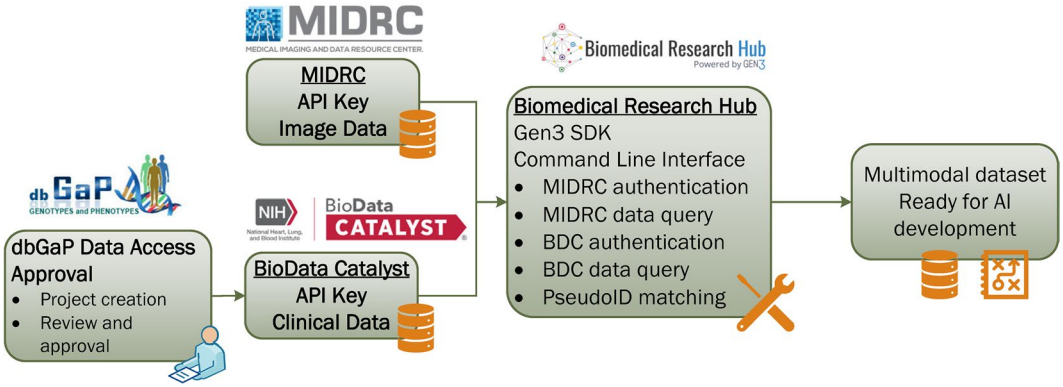

Fig. 5 Workflow for interoperability between MIDRC and BioData Catalyst.

specifications for the collection, exchange, storage, and retrieval of information associated with health care applications, including medical records, medications, radiological images, payment and reimbursement, medical devices and monitoring systems, and administrative processes"[19]. An example of a data standard is the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), which is an open community data standard designed to standardize the structure and content of observational data and to enable efficient analyses that can produce reliable evidence (https://www.ohdsi.org/data-standardization/). Moreover, the use of common data elements, defined as data collection units comprising one or more questions together with a set of valid values, can play a valuable role for widescale interoperability[20]. These efforts are especially important for addressing the challenges of combining imaging data with other clinical information.

Methods

Interoperability between MIDRC and BDC. The first example use case benefitting from interoperability between data repositories is to acquire multimodal data from multiple sources in order to enable association, correlation, and fusion analyses. For example, for the PETAL RED CORAL dataset, the BDC repository has clinical metadata of COVID-19 patients whereas MIDRC has the medical images of the same patients. Here the use case is to investigate the associations of multi-modality data (e.g., medical images and clinical laboratory testing) or integrate the multimodal data using AI/ML fusion models for a certain clinical task such as the characterization of COVID severity to tailor patient treatments (Fig. 4).

The demonstration of interoperability between MIDRC and BDC is to match patients with clinical data in BDC to corresponding patients with imaging data in MIDRC. To implement this, we follow a few steps as illustrated (in a simplified presentation) in Fig. 5.

Both MIDRC and BDC operate upon the Gen3 Data Platform (https://gen3.org/). Gen3 is an open-source data platform for managing, analyzing, and sharing biomedical data. It supports open Application Programming Interfaces (APIs) so that the data it manages are findable, accessible, interoperable and reusable (FAIR)[12]. The FAIR APIs that both MIDRC and BDC provide are the essential foundation for the interoperability described in this study. Gen3 supports data objects, such as image files; structured data, such as clinical data; and semi-structured data, such as JSON-based metadata. Gen3 based platforms, including MIDRC and BDC, support a variety of identity providers for authenticating users, including InCommon (https://incommon.org/) and ORCID (Open Researcher and Contributor Identifier, https://orcid.org/). Gen3 based platforms also support several methods for managing authorization information specifying which users are authorised to access which data. In particular, Gen3 can interoperate with the NIH dbGaP system[21], which is used by BDC to manage authorization information.

Data linkage across MIDRC and BDC is performed through Gen3's opaque Globally Unique IDentifiers (GUID) that follow the DRS GA4GH standard[22]. An DRS identifier has a prefix specifying a particular data platform, such as MIDRC or BDC, followed by a GUID. Opaque in this context means that the GUID does not contain a name, medical record number or any other string that has semantic information.

Accessing data from both MIDRC and BDC requires authentication. In addition, access to the BDC clinical data used for this study is controlled access and requires authorization through dbGaP. The required authentication and authorization is handled by MIDRC and BDC through the Gen3 Fence service. Once a user is authenticated and authorized and a cohort or dataset is specified, the DRS identifiers for the data in the cohort or dataset can be accessed in a workspace or downloaded. In particular, a list of DRS identifiers for the MIDRC PETAL RED CORAL Imaging dataset can be obtained in this way and a list of DRS identifiers for the BDC PETAL RED CORAL dataset can also be obtained. Note that for a user to be authorized to access the RED CORAL clinical dataset on BDC, the user must register on dbGaP, create a project, submit a Data Request Form, and have the Data Request Form approved by the Data Access Review Committee[21]. After approval, the dbGaP authorization information for the user is updated, and this information is available for systems that are approved for interoperating with dbGaP, such as Gen3.

Gen3 has a privacy preserving record linkage (PPRL) service called the Crosswalk Service that, given DRS identifiers for images from a data commons, can provide the associated DRS identifiers for matching data in another commons. For example, given a list of DRS identifiers for images in MIDRC, the Crosswalk service can provide the DRS identifiers for associated clinical data in BDC, and vice versa. Note that since both MIDRC and BDC use privacy preserving opaque identifiers, the Crosswalk service must be provided with a mapping or cross linking of these opaque identifiers. This mapping is usually provided when data are submitted, but can be done at any time. It is important to note that even with this cross linking of identifiers, all the information is still private since all the identifiers are opaque and contain no PII.

With these lists of DRS identifiers, the image data and corresponding clinical data can be easily exported from the commons and imported into any analysis environment that is approved for managing controlled access data and is authorized to interoperate with the commons[23]. Sometimes these are called authorized environments, computational enclaves, or freeports[23]. For this study, the image and corresponding clinical data were imported into workspaces that were part of Gen3's Biomedical Research Hub[24], and which are approved analysis environments for analyzing controlled access BDC data.

In summary, the demonstration of interoperability between MIDRC and BDC demonstrates the process for meeting the objective of collecting patient data from both an imaging and non-imaging data source.

Interoperability between MIDRC and N3C. The second example use case benefitting from interoperability between data repositories is to aggregate data in one to create cohorts in another in order to enable development of AI models that incorporate data across modalities. In this example use case, the goal is to develop an algorithm based on medical images to predict severity of COVID-19 disease, defined as the admission of a COVID-19 positive patient to the intensive care unit (ICU) or intubation within 24 hours of chest radiography. The chest radiographs are to be collected from MIDRC, while the clinical data (i.e., information on ICU admission and/or intubation or lack thereof) are to be collected from N3C, which ingests data and transforms the associated data models to a harmonized Observational Medical Outcomes Partnership (OMOP) analytics dataset (https://ncats.nih.gov/research/research-activities/n3c/covid-enclave/data-overview; https://www.ohdsi.org/data-standardization/). Therefore, a MIDRC user developing the algorithm would aim to create two cohorts of images: (1) patients with severe COVID and (2) patients with mild COVID (Fig. 6).

The first step for interoperability is to identify and characterize the subjects with relevant data in each data repository, based upon the task for which the AI is to be developed. MIDRC and N3C have different governance (including models of access), which impact the interoperability workflow. Conducting this use case requires that users have separate log-in accounts at MIDRC Open Data Commons and at N3C. Any registered user at the MIDRC Open Data Commons can download images, while registered users at N3C must complete a Data Use Request (DUR) for the specific study they are conducting and agree to keep the individual N3C data within an N3C computational enclave (freeport). At this time, the matching of patients with data in both N3C (clinical data) and MIDRC (imaging data) is conducted using Privacy Preserving Record Linkage via an honest broker. These matches are produced on request and held as a table within N3C, which for now operates as the freeport enclave. Additionally, it is recommended that users use this list as the starting point and then download from MIDRC only those images relevant for the study and associated with the patients in the match table.

Subsequently, the user can calculate an imaging-derived measure (such as the severity index[25]) on the images from the MIDRC cohort using their local computer, and then import into the freeport enclave (which is for now limited to the N3C enclave) via an upload request. Figure 7 outlines the steps required for interoperability between N3C and MIDRC.

In summary, the demonstration of interoperability between MIDRC and N3C demonstrates the process for meeting the objective of creating patient cohorts of imaging data based upon characteristics from clinical data.

Characterization of the representativeness of the curated datasets. To characterize the representativeness of the two datasets we curated via interoperability, we evaluated the demographic characteristics of patients in each of the two datasets for the categories of sex, race, and ethnicity. Similar to our previous work in this area[26], we compared each of these demographic characteristics to the cumulative COVID-19 positive case counts as reported by the Centers for Disease Control and Prevention[17] over the period of image collection, using the Jensen-Shannon Distance (JSD)[16] as a measure of similarity. The JSD is bounded between 0 and 1 when log2 is

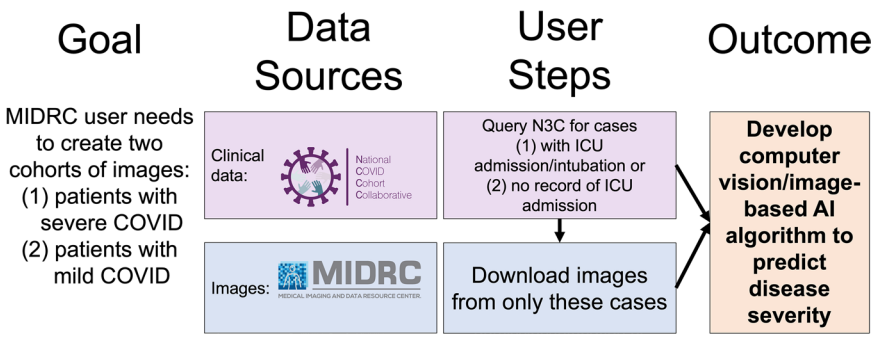

Fig. 6 Overview of example use case for cohort building between MIDRC and N3C for developing AI algorithms that incorporate image-based data.

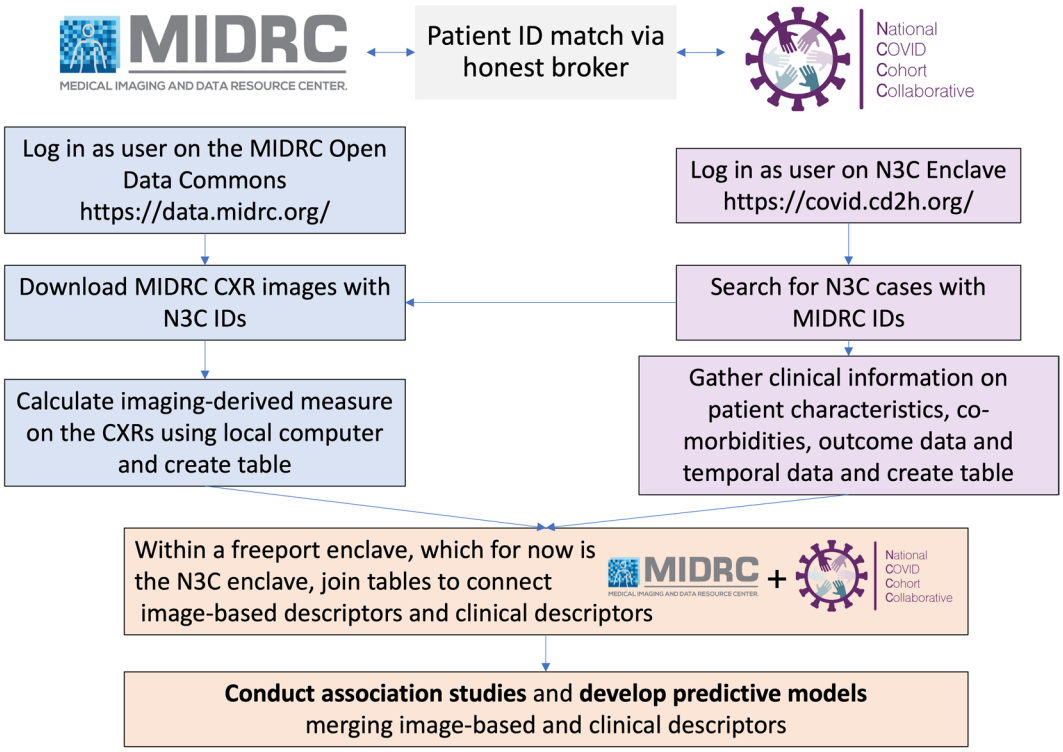

Fig. 7 Workflow for interoperability between MIDRC and N3C.

used within the analytical expression for this measure, where JSD = 0 indicates complete similarity between two distributions and JSD = 1 indicates no similarity.

A practically important issue is that it is common to have missing data in the collected demographic characteristics. As a hypothetical example, we assume the sex distribution of a curated dataset is as follows: 45% female, 45% male, 10% missing, while the population distribution is as follows: 35% female, 35% male, and 30% missing. A raw JSD score using the three categories (female, male, missing) is 0.18. However, if the missing information can be assumed to be distributed at random (i.e., not associated with sex), then the adjusted distribution would be 50% female and 50% male for both the curated dataset and the population, thereby yielding a JSD score of 0. This means that, if we assume missing data is randomly distributed, the different proportion of missing data would impact the JSD metric. In this study, we provided both the raw and adjusted JSD values.

Data availability

The imaging data are freely available at https://data.midrc.org. The clinical data of the PETAL RED CORAL dataset are freely available after approval/authorization of dbGaP (https://www.ncbi.nlm.nih.gov/gap/). The clinical data of the N3C dataset were available to the authors by Data Use Request DUR-BB40587 and are available to the public with a valid Data Use Request. The dbGaP data can be accessed at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002363.v1.p1, dbGaP Study Accession: phs002363.v1.p1. The N3C Data Use Agreement is available at https://ncats.nih.gov/sites/default/files/NCATS_N3C_Data_Use_

Agreement.pdf. Users must be approved and can only analyze data within the platform; data cannot be removed or downloaded. Requests for access to the MIDRC-N3C dataset described in this study can be made directly to N3C (https://n3c.ncats.nih.gov/education/dur).

Code availability

Custom code for processing the PETAL RED CORAL dataset as described is available at https://github.com/MIDRC. Code for processing the N3C dataset is not publicly available due to restrictions on the Limited Dataset, as the N3C Data Use Agreement specifies that N3C COVID Enclave data will be used only for clinical and translational research and public health surveillance of COVID-19. Please contact the authors with requests for more information. In general, all code shareable by MIDRC is available at https://github.com/MIDRC.

Received: 23 June 2024; Accepted: 23 July 2025;
Published online: 01 August 2025

References

1. Boehm, K. M., Khosravi, P., Vanguri, R., Gao, J. & Shah, S. P. Harnessing multimodal data integration to advance precision oncology. *Nature Reviews Cancer* **22**, 114–126, https://doi.org/10.1038/s41568-021-00408-3 (2022).
2. Chen, W. *et al*. Machine learning with multimodal data for COVID-19. *Heliyon* **9**, e17934, https://doi.org/10.1016/j.heliyon.2023.e17934 (2023).
3. Hinkson, I. V. *et al*. A Comprehensive Infrastructure for Big Data in Cancer Research: Accelerating Cancer Research and Precision Medicine. *Front Cell Dev Biol* **5**, 83, https://doi.org/10.3389/fcell.2017.00083 (2017).
4. Clark, K. *et al*. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging* **26**, 1045–1057, https://doi.org/10.1007/s10278-013-9622-7 (2013).
5. The Cancer Imaging Archive (TCIA). https://www.cancerimagingarchive.net/ (2024).
6. *Imaging Data Commons | CRDC*. https://datacommons.cancer.gov/repository/imaging-data-commons.
7. National COVID Cohort Collaborative (N3C). https://ncats.nih.gov/research/research-activities/n3c (2024).
8. National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services. The NHLBI BioData Catalyst. (National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services, 2020).
9. Tryka, K. A. *et al*. NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res* **42**, D975–979, https://doi.org/10.1093/nar/gkt1211 (2014).
10. NCBI's Database of Genotypes and Phenotypes: dbGaP. https://www.ncbi.nlm.nih.gov/gap/ (2024).
11. *The Medical Imaging and Data Resource Center. MIDRC* https://www.midrc.org.
12. Wilkinson, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* **3**, 1–9 (2016).
13. Lehne, M., Sass, J., Essenwanger, A., Schepers, J. & Thun, S. Why digital medicine depends on interoperability. *NPJ Digit Med* **2**, 79, https://doi.org/10.1038/s41746-019-0158-1 (2019).
14. Perlin, J. B. Health Information Technology Interoperability and Use for Better Care and Evidence. *JAMA* **316**, 1667–1668, https://doi.org/10.1001/jama.2016.12337 (2016).
15. Peltan, I. D. *et al*. Characteristics and Outcomes of US Patients Hospitalized With COVID-19. *American Journal of Critical Care* **31**, 146–157, https://doi.org/10.4037/ajcc2022549 (2022).
16. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory* **37**, 145–151 (1991).
17. Centers for Disease Control and Prevention of USA (CDC). *COVID-19 Case Surveillance Public Use Data*, https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4/about_data (2024).
18. Karen, D. *et al*. Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. *Journal of Medical Imaging* **10**, 061104, https://doi.org/10.1117/1.JMI.10.6.061104 (2023).
19. Institute of Medicine (US) Committee on Data Standards for Patient Safety. in *Patient Safety: Achieving a New Standard for Care* (eds P. Aspden, J. M. Corrigan, J. Wolcott, & S. M. Erickson) (National Academies Press (US), Washington (DC), 2004).
20. Kush, R. D. *et al*. FAIR data sharing: The roles of common data elements and harmonization. *J Biomed Inform* **107**, 103421, https://doi.org/10.1016/j.jbi.2020.103421 (2020).
21. Mailman, M. D. *et al*. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* **39**, 1181–1186, https://doi.org/10.1038/ng1007-1181 (2007).
22. Rehm, H. L. *et al*. GA4GH: International policies and standards for data sharing across genomic research and healthcare. *Cell Genom* **1**, https://doi.org/10.1016/j.xgen.2021.100029 (2021).
23. Grossman, R. L. *et al*. A Framework for the Interoperability of Cloud Platforms: Towards FAIR Data in SAFE Environments. *Sci Data* **11**, 241, https://doi.org/10.1038/s41597-024-03041-5 (2024).
24. Barnes, C. *et al*. The Biomedical Research Hub: a federated platform for patient research data. *J Am Med Inform Assoc* **29**, 619–625, https://doi.org/10.1093/jamia/ocab247 (2022).
25. Li, H. *et al*. Predicting intensive care need for COVID-19 patients using deep learning on chest radiography. *J Med Imaging (Bellingham)* **10**, 044504, https://doi.org/10.1117/1.JMI.10.4.044504 (2023).
26. Whitney, H. M. *et al*. Longitudinal assessment of demographic representativeness in the Medical Imaging and Data Resource Center open data commons. *J Med Imaging (Bellingham)* **10**, 61105, https://doi.org/10.1117/1.JMI.10.6.061105 (2023).

Acknowledgements

MIDRC is supported by the National Institutes of Health, National Institute of Biomedical Imaging and Bioengineering contract 75N92020D00021 and by ARPA-H. The authors are grateful to Emily Townley, program manager at the American Association of Physicists in Medicine, for her support of this work. The authors wish to acknowledge the contributions of the consortium working on the development of the NHLBI BioData Catalyst® (BDC) ecosystem. We gratefully acknowledge the following core contributors to N3C: Adam B. Wilcox, Adam M. Lee, Alexis Graves, Alfred (Jerrod) Anzalone, Amin Manna, Amit Saha, Amy Olex, Andrea Zhou, Andrew E. Williams, Andrew Southerland, Andrew T. Girvin, Anita Walden, Anjali A. Sharathkumar, Benjamin Amor, Benjamin Bates, Brian Hendricks, Brijesh Patel, Caleb Alexander, Carolyn Bramante, Cavin Ward-Caviness, Charisse Madlock-Brown, Christine Suver, Christopher Chute, Christopher Dillon, Chunlei Wu, Clare Schmitt, Cliff Takemoto, Dan Housman, Davera Gabriel, David A. Eichmann, Diego Mazzotti, Don Brown, Eilis Boudreau, Elaine Hill, Emily Carlson Marti, Emily R. Pfaff, Evan French, Farrukh M Koraishy, Federico Mariona, Fred Prior, George Sokos, Greg Martin, Harold Lehmann, Heidi Spratt, Hemalkumar Mehta, J.W. Awori Hayanga, Jami

Pincavitch, Jaylyn Clark, Jeremy Richard Harper, Jessica Islam, Jin Ge, Joel Gagnier, Johanna Loomba, John Buse, Jomol Mathew, Joni L. Rutter, Julie A. McMurry, Justin Guinney, Justin Starren, Karen Crowley, Katie Rebecca Bradwell, Kellie M. Walters, Ken Wilkins, Kenneth R. Gersing, Kenrick Dwain Cato, Kimberly Murray, Kristin Kostka, Lavance Northington, Lee Allan Pyles, Lesley Cottrell, Lili Portilla, Mariam Deacy, Mark M. Bissell, Marshall Clark, Mary Emmett, Matvey B. Palchuk, Melissa A. Haendel, Meredith Adams, Meredith Temple-O'Connor, Michael G. Kurilla, Michele Morris, Nasia Safdar, Nicole Garbarini, Noha Sharafeldin, Ofer Sadan, Patricia A. Francis, Penny Wung Burgoon, Philip R.O. Payne, Randeep Jawa, Rebecca Erwin-Cohen, Rena Patel, Richard A. Moffitt, Richard L. Zhu, Rishi Kamaleswaran, Robert Hurley, Robert T. Miller, Saiju Pyarajan, Sam G. Michael, Samuel Bozzette, Sandeep Mallipattu, Satyanarayana Vedula, Scott Chapman, Shawn T. O'Neil, Soko Setoguchi, Stephanie S. Hong, Steve Johnson, Tellen D. Bennett, Tiffany Callahan, Umit Topaloglu, Valery Gordon, Vignesh Subbian, Warren A. Kibbe, Wenndy Hernandez, Will Beasley, Will Cooper, William Hillegass, Xiaohan Tanner Zhang. Details of contributions available at covid.cd2h.org/core-contributors. The N3C Publication committee confirmed that this manuscript msid: 1956.271 is in accordance with N3C data use and attribution policies; however, this content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the N3C program. The mention of commercial or open-source products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services. This is a contribution of the U.S. Food and Drug Administration and is not subject to copyright. The analyses described in this publication were conducted with data or tools accessed through the NCATS N3C Data Enclave https://covid.cd2h.org and N3C Attribution & Publication Policy v 1.2-2020-08-25b supported by NCATS Contract No. 75N95023D00001, and Axle Informatics Subcontract: NCATS-P00438-B. This research was possible because of the patients whose information is included within the data and the organizations (https://ncats.nih.gov/n3c/resources/data-contribution/data-transfer-agreement-signatories) and scientists who have contributed to the on-going development of this community resource [https://doi.org/10.1093/jamia/ocaa196]. The N3C Enclave is available for public research use. To access data including that used in this manuscript, institutions must have a signed Data Use Agreement executed with the U.S. National Center for Advancing Translational Sciences (NCATS) and their investigators must complete mandatory training and must submit a Data Use Request (DUR) to N3C. To request N3C data access, researchers must follow instructions at https://covid.cd2h.org/onboarding. More than 4,000 researchers currently have access to data in N3C; together they represent more than 300 US research institutions. Reviewers can confidentially request access by these same means. The original study "PETAL Repository of Electronic Data COVID-19 Observational Study (RED CORAL)" was supported by grants from the NHLBI (3U01HL123009-06S1, U01HL123009, U01HL122998, U01HL123018, U01HL123023, U01HL123008, U01HL123031, U01HL123004, U01HL123027, U01HL123010, U01HL123033, U01HL122989, U01HL123022, and U01HL123020). Study data were collected and managed using REDCap electronic data capture tools hosted by Partners HealthCare Research Computing, Enterprise Research Infrastructure & Services (ERIS) group, which is supported by Harvard Catalyst (NIH Award #UL1 RR 025758 and financial contributions from Harvard University and its affiliated academic health care centers). The original data can be found in BioData Catalyst phs002363.

Author contributions

Conceptualization: W.C., H.M.W., S.K., C.M., H.L., R.C.S., D.L., S.N., K.G., R.L.G., M.L.G.; data curation: W.C., H.M.W., S.K.; formal analysis: W.C., H.M.W., S.K., H.L.; writing (original draft): W.C., H.M.W.; writing (review and editing): all authors. Authorship was determined by I.C.M.J.E. requirements.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to W.C. or H.M.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.