RADx Data Hub: Enables Capabilities to Find, Access, Analyze, and Reuse COVID-19 RADx Study Data in a Cloud-based Platform

MARCH 2024

Table of Contents

Executive Summary

The RADx Data Hub was established to support researchers in accessing curated and de-identified COVID-19 data, allowing them to find, aggregate, and perform data analyses in a cloud-enabled platform. The centralized cloud resources offered by the Data Hub provide study data and analytic capacities for researchers to better understand diagnostic COVID-19 testing tools, technologies, and future pandemics, especially for underserved populations and those who are disproportionately impacted by COVID-19. Additionally, these resources support rapid review of data requests from authorized users through the NIH RADx Data Access Committee to ensure proposed research purposes are consistent with participant protections and the data use agreement. Access to the Data Hub allows users to query a wide range of health data including clinical, behavioral, social determinants of health, survey, interview, diagnostic and testing results, viral sequences, output from smart sensors, self-reported symptoms, and imaging data.

The RADx Data Hub was built as a modernized, cloud-based data repository by curating collected data elements and creating anonymous COVID-19 research data optimized for cost, performance, and streamlined data access. This platform acts as a centralized data and metadata collection point across the four RADx programs that collect and submit the data (RADx-Digital Health Technologies, RADx-rad, RADx Tech, RADx-UP). Once the data is collected, it is systematically mapped, standardized, and curated across the various RADx studies to ensure consistent data collection and alignment to designated RADx Common Data Elements (CDEs).

The RADx Data Hub was also designed to optimize engagement with the research community. The Data Hub website allows users to explore the various data variables and study information for each registered RADx study. The RADx Data Hub support team also provides community outreach webinars that discuss the RADx data and the Data Hub functionality.

To date, the RADx Data Hub contains 138 studies from across the four RADx programs that are available to researchers. Each study contains data that is harmonized to 12 RADx CDEs and up to 133 mapped variables that are study-dependent. The RADx programs continue to deposit data

into the Data Hub and approximately 60 additional studies are expected to be included once data collection and harmonization is complete.

As a portion of these studies include American Indian/Alaska Native (AI/AN) communities, the RADx Data Hub team consulted with the Office of Data Science Strategy (ODSS), the National Institute of Minority Health and Health Disparities (NIMHD), and Tribal Leadership to hold a Tribal Consultation on July 30, 2021. The goal of this consultation was to address the unique cultural, governance, sovereignty needs, and expectations to support and share data collected from the AI/AN communities. This led to the establishment of the RADx Tribal Data Repository (RADx TDR) for responsible data sharing and access of RADx AI/AN research data. This repository is the first NIH research repository designed to implement sovereignty-based governance to promote tribal scientific values for researchers working with AI/AN data.

Program Overview and Structure

Each of the four RADx programs (RADx-Digital Health Technologies, RADx-rad, RADx Tech, RADx-UP) was established to address different aspects of the COVID-19 pandemic, resulting in a wide breadth of data types that needed to be consolidated within a centralized data depository. The RADx Data Hub was established to collate data from the various programs and to act as a cloud-based data repository of COVID-19 research data – including clinical, behavioral, social determinants of health, survey, interview, diagnostic and testing results, viral sequences, output from smart sensors, self-reported symptoms, and imaging data. The Data Hub provides access to anonymous RADx study research data, algorithms, and other analytic capabilities to expand testing and identify effective testing implementation strategies, especially for underserved populations and those disproportionately impacted by COVD-19.

As each RADx study has a unique scientific perspective and purpose, the data collected can vary widely across the individual studies. For the Data Hub to act as a centralized data depository, the data must undergo a standardized procedure to 1) ensure that all protected health information (PHI) is removed from the data sets and 2) harmonize the data so that it can be queried and analyzed across the studies. As study data is submitted to the RADx Data Hub, it first undergoes

validation that the 18 identifiers of PHI are removed to protect participant privacy in the research data. Once that is verified, the data is analyzed to identify, curate, standardize, and map the collected data elements and metadata to ensure user-friendly platform for scientific discovery. Within the RADx Data Hub platform, advanced technologies, such as AI and machine learning, are applied to ensure that the data/metadata can be findable, accessible, interoperable, and reusable across a multidisciplinary research community. This includes performing vertical and horizontal mapping variables across various scientific domain areas to enable a public browser searchable by variable, study, and file. Additionally, this data collation standardizes the metadata schema for data submission from the four RADx Collection and Data Coordination Centers (C)DCCs.

Program Accomplishments

Since launching in December 2022, the RADx Data Hub has provided a secure workspace to combine authorized data use and analytics tools, enabled researcher collaborations, ensured the ability to share analyses results, and created a framework for generating artificial intelligence-ready datasets. The RADx Data Hub, as of November 27, 2023, contains 138 studies available to researchers. Additionally, the Data Hub includes 1,129 data files and 287 metadata files, including data dictionaries, README files, and schema. This accumulates to a total of 45,467 data variables across all studies, including the 12 RADx CDEs and 133 mapped study-specific variables. The Data Hub also contains 1,896 viral samples with genomic sequencing data and 2,519 viral images for researchers to access.

Community Engagement

The RADx Data Hub was formed as a collaborative effort across the RADx programs and continues to actively engage with these scientific communities. Nationally, the RADx Data Hub has accepted studies from 46 US states and territories, with 24 states conducting research in two RADx programs, six states contributing to three RADx programs, and one state contributing to all four RADx programs. The RADx Data Hub community, as of November 27, 2023, consists of 421 individuals, 25 user-focus groups, and 575 subscribers. The RADx Data Hub leadership

team continues to hold regular, bi-weekly touchpoints with each of the four (C)DCCs, as well as quarterly, program-wide meetings. They have also hosted five webinars to engage the research community and have developed publicly available information for these researcher engagement activities (RADx Data Hub Events | RADx (nih.gov)).

In addition to these engagement activities, the RADx Data Hub has aided in community outreach for the multiple funding opportunities for researchers interested in secondary analysis of data in the RADx Data Hub. These funding opportunities (NOT-OD-24-026, the renewal of NOT-OD-23-040) aims to support advance scientific inquiry related to COVID-19 through the existing data resources in the RADx Data. NIH also issued a Notice of Special Interest (NOT-OD-23-041) to enhance the diversity of the data science workforce to address questions and scientific inquires to COVID-19 through the existing data resources in the RADx Data Hub. Both research pathways would utilize the RADx Data Hub to answer relevant, public health-centric scientific inquiries.

One of the RADx Data Hub's most important community partnerships is the development of the RADx Tribal Data Repository (TDR) for sovereignty-based governance of AI/AN RADx data. This was done in a partnership with Tribal Leadership and represents the first NIH research repository designed to implement sovereignty-based governance to promote Tribal scientific values for researchers working with RADx AI/AN data. The RADx TDR addresses the unique cultural, governance, sovereignty needs, and expectations to support and share this data and will likely be used as a template for future collaborations.

Impact and the Path Moving Forward

Through the design, development, and deployment of the RADx Data Hub, the leadership team has identified various aspects of the process that could be improved for future, similar, program developments. Primarily, it is vital for the data to be as consistent as possible across the programs, so implementing CDEs as early as possible will mitigate data collection/harmonization complications. Furthermore, early adoption of a complete and consistent data dictionary is crucial to collecting data that can be applied across multiple

programs. The team found that the program-specific data dictionaries sometimes contained weak or missing descriptions for variables when these were merged into a common data dictionary. However, a data dictionary browser designed at the onset of data collection could be a significant improvement to the data collection and harmonization process. Program teams continue to optimize this process, both for uses in the current RADx Data Hub and for the development of future data repositories.

For the process of CDE selection and maintenance, the RADx Data Hub team recommends that the CDEs cover a wide range of health data and should be automated, when possible. The collected data for the RADx Data Hub often required manual mapping and path review, both of which could benefit from automation. If algorithms were to be developed at the onset of data collection, it could simplify data collection and could broaden the evaluation of more variables.

In the case that automation is unattainable, then researchers should consider the scalability and the amount of manual data harmonization/curation that will be required in the data collection process. If these considerations can be applied early in the process, it can improve the efficiency of data processing and can help improve any harmonization algorithms required.

Lastly, early community outreach is critical to promoting a newly data-rich resources that will be adopted and used by researchers. By reaching these communities through webinars, community meetings, social media, etc., the RADx Data Hub Team can inform resources, stimulate scientific research interest and promote early adoption of the platform.

Conclusion

The RADx Data Hub represents one of the largest collections of NIH COVID-19 testing data available to researchers that allows researchers to explore, access, and analyze COVID-related data developed through NIH RADx programs. The Data Hub will provide curated and harmonized data across ~ 200 RADx studies and all data/metadata will be findable, accessible, interoperable, and reusable. These data will be of interest to researchers focused on testing approaches, diagnostic and surveillance testing through wearable sensors and non-traditional approaches, and implementation strategies to enable an enhance testing of underserved, under-

resourced, and/or vulnerable populations. Furthermore, the development of the RADx Tribal Data Repository has created a pathway for future collaboration with American Indian/Alaska Native communities. Through a close collaboration with the four RADx (C)DCCs, the RADx Data Hub capitalized on the speed and flexibility of these programs to streamline the processing of complicated data submissions and making those de-identified and harmonized data available to the larger research community.