

This is a “preproof” accepted article for *Journal of Clinical and Translational Science*.

This version may be subject to change during the production process.

10.1017/cts.2024.530

FAIR Enough: Building an Academic Data Ecosystem to Make Real-World Data Available for Translational Research

Isabella Chu¹, Rebecca Miller¹, Ian Mathews², Ayin Vala¹, Lesley Sept¹, Ruth O’Hara^{3,4}, David H. Rehkopf^{1,5,6,7,8,9}

¹The Stanford Center for Population Health Sciences, Stanford School of Medicine, Stanford University, Palo Alto, CA

²Redivis, Inc, Oakland, CA

³Psychiatry and Behavioral Sciences, Stanford School of Medicine, Stanford University, Stanford, CA

⁴Veterans Administration Palo Alto Health Care System, Sierra Pacific Mental Illness, Research, Education, and Clinical Center, Palo Alto, CA, US

⁵Department of Epidemiology and Population Health, Stanford School of Medicine, Stanford University, Stanford, CA

⁶Department of Medicine, Division of Primary Care and Population Health, Stanford School of Medicine, Stanford University, Stanford, CA

⁷Department of Pediatrics, Stanford School of Medicine, Stanford University, Stanford, CA

⁸Department of Health Policy, Stanford School of Medicine, Stanford University, Stanford, CA

⁹Department of Sociology, Stanford University, Stanford, CA

Corresponding Author: David H. Rehkopf, ScD, MPH, Department of Epidemiology and Population Health, Stanford University School of Medicine, Phone: 650-725-0356, Email: drehkopf@stanford.edu, ORCID: 0000-0002-7597-6513

This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is unaltered and is properly cited. The written permission of Cambridge University Press must be obtained for commercial re-use or in order to create a derivative work.

Abstract

The genesis of the Stanford Population Health Sciences Data Ecosystem was to facilitate the use of large datasets containing health records from hundreds of millions of individuals. This necessitated technical solutions optimized for an academic medical center to manage and share high-risk data at scale. Through collaboration with external and internal partners, we have built a Data Ecosystem to host, curate and share data with hundreds of users in a secure and compliant manner. This platform has enabled us to host unique data assets and serve the needs of researchers across Stanford University. The technology and approach were designed to be replicable and portable to other institutions. Through the experience of creating the Data Ecosystem we have found that the technical advances that we made in developing this are necessary but not sufficient. Challenges around making data findable, accessible, interoperable, and reusable (FAIR) remain. Our experience has demonstrated there is high demand for access to real-world data and that if the appropriate tools and structures are in place, translational research can be advanced considerably. Together, technological solutions, management structures and education to support researcher, data science and community collaborations offer more impactful processes over the longer-term for supporting translational research with real-world data.

The Stanford Center for Population Health Sciences (S-PHS) was founded in 2015, with the mission to “improve the health of populations by bringing together diverse disciplines and data to understand and address social, environmental, behavioral, and biological factors on both a domestic and global scale.”¹ One of the central means to achieve that mission was to advance translational science through providing access to real-world data (RWD). We sought to create structures and processes that promote impactful translational science using RWD, while acknowledging all the complexities of the data and the new knowledge and skills needed to use RWD effectively. In this article we first briefly describe some of the primary benefits of RWD that motivated this effort, along with the major challenges of using RWD for translational research. We then describe the process of addressing the challenges through developing a data ecosystem for utilizing RWD in the academic setting. We set out to solve these problems not just for our own organization but to create a solution replicable and affordable for large research groups and other academic and non-profit organizations seeking to utilize high-value, high-risk data in a secure, scalable, and compliant manner. The platform developed at S-PHS, Redivis, has been adopted by Stanford University Libraries and several other university libraries and data centers. Thirdly, we discuss new directions of interest based on lessons learned and what we see as critical next steps for increasing the impact of RWD to solve pressing population health and health equity problems. We share what we have learned over the past eight years about how best to facilitate the use of these data for translational science, and what we think still needs to happen to maximize RWD impact for improving population health and reducing health inequities. We hope this past pathway and our observations on looking forward will be of interest to research groups and institutions that are interested in utilizing RWD to advance translational science.

Rationale for a focus on real-world data

Real-world data can be defined most generally as data that was collected for a purpose other than for research. As this data relates to health, this most notably includes data from wearable devices, billing data, and data from electronic health records. These data have several advantages over data used for traditional research studies. Many of these advantages are because these data typically include a much larger number of individuals than is possible to collect data from when a study is designed specifically for research purposes. Thus, there is greater statistical power in RWD to identify small but important impacts of exposures and treatments, both advantageous and adverse. They are a better fit to support a new generation of neural network models, like

transformer models, that fit massive numbers of nodes to data and require large sample sizes to produce meaningful findings. Real-world data with large numbers of individuals also are much better for precision health approaches to treatment², given the ability to have stable findings within subgroups of the population. Similarly, although from a different motivation, real-world data also allow for more quantitative intersectional approaches to studying health, because there are enough individuals in subgroups of the population to estimate heterogeneity of treatment effects.³ Collecting data from this many individuals for a research project would be prohibitively expensive, as the costs of collecting the RWD are typically already paid for by the group for which the data was used for the intended non-research purpose. A further advantage of RWD is the multipurpose nature of types of data that make the study of a wide range of health topics feasible. When permitted, linkage with other datasets can enable tracking of patients across many years, something that can be done instantaneously and retrospectively, without waiting decades as is the case with the traditional prospective, longitudinal study design structures. Many RWD sets, although not all, contain more diverse populations that are closer to being representative of a target population since enrollment and inclusion in the data is often the default. Finally, RWD typically is continuously and constantly being updated, allowing more up to date insights into health risks and protective factors. This also makes RWD critical for adapting and improving patient care in real time in the context of learning health systems.

Challenges of using real-world data

Three major challenges of using RWD are: 1) the computational challenges of the large size of the data, 2) the need to keep proprietary data collected for another use private and secure, and 3) the fact that RWD have very different forms of bias and measurement error than data that is collected for research purposes. At many terabytes each, RWD typically exceed the capacity of any personal computer. Like almost all data for health, RWD has a risk of re-identification of individuals, so data must be kept secure. What is unique about RWD, however, is that there are often also proprietary risks of disclosure because the data was not collected by researchers themselves, but by another entity, often a private entity. Many datasets have specific stipulations on who may use the data and how, such as limiting what research questions may be asked, or on the affiliations and roles of researchers, such as requiring affiliation with the licensing organization or prohibiting the use of data for trainee projects. The selection and measurement errors in RWD have been well described, most notably that with claims data and EHR data

patients will typically only be observed when they have health problems, identifying the population that is at risk is often a challenge in open cohorts, and clinical definitions of health outcomes are often not included since that is not the purpose of many RWD types.⁴

Our approaches to addressing real-world data access and computation challenges

To effectively meet the needs of researchers using RWD, we focused on bringing together the following:

- Data that allow data use agreements that enable a large number of people to use and reuse the data for a wide range of projects;
- Data administration and governance that is transparent, tractable and that protects the privacy and confidentiality of individuals in datasets and the researchers from violating policies, regulations or contractual obligations;
- Data management and expertise to ensure data are curated in a way that is most useful for researchers;
- Computational environments and software that are secure, powerful, flexible and scalable.

Wilkinson et al. set a series of standards to ensure maximum utility of data—namely that they must be Findable, Accessible, Interoperable, and Reusable.⁵ We have reinterpreted these FAIR standards for high-risk data as follows:

- **Findable**: Entities should issue a unique digital object identifier (DOI) for each iteration of a dataset. The data should be annotated, tagged and indexed in such a way that they are optimized for commonly used search engines. Specifically, metadata should be presented in a machine-readable format utilizing standard schemata and vocabularies.
- **Accessible**: There should be a transparent, standardized process for obtaining data access and using data that is closely monitored and managed by a data custodian that meets all relevant regulatory and security requirements, including those specific to individual datasets and/or high-risk datasets. Meta-data must be retrievable by their identifier using a standardized communications and sufficient to enable prospective users can evaluate the data's utility for their research questions in the instance data have a high barrier to access.

- Interoperable: Data and meta-data use a formal, accessible, shared, and broadly applicable language for knowledge representation, vocabularies that follow FAIR principles and include qualified references to other data.
- Reusable: Data and meta-data should be well described and released with clear parameters around provenance and licensure and meet domain relevant standards.

At the time of S-PHS's inception, there were no products optimized or affordable to academic markets for achieving these objectives. S-PHS also needed to manage complex data access requirements and maintain visibility and direct control of S-PHS hosted data. S-PHS worked with a startup, Redivis⁶, to build a platform that could handle data discovery, access, curation, annotation, sharing, exploration and visualization. In parallel, the Stanford Research Computing Center built a secure, customizable, and scalable computational environment based in the Google Cloud Platform. Data curation, annotation, and administration were handled by S-PHS staff. Together, these three pillars of data management form the core essentials (Figure 1) of what is needed for the ability to use RWD by clinical researchers, epidemiologists and statisticians who do not have advanced large database data science skills.

In the time between S-PHS's initial investment and now, other platforms have emerged that provide some similar features. Many of these have been funded or are directly operated by the NIH. Examples include Vivli, N3C, and All of Us.⁷⁻⁹ A team from Harvard built Aetion which specializes in models for clinical research.^{10,11} These systems share several features in common with the S-PHS system. Scalability of the system is achieved with commercial cloud. In almost all cases, partnership with private entities was also used to build the system and, in many cases, for ongoing operation. For example, the All of Us Data and Research Center was achieved through a collaboration with Verily Life Sciences¹² and the National COVID Cohort Collaborative (N3C) collaborated with Palantir.¹³ In the private sector, platforms such as Snowflake and TriNetX also make data available and owners of claims data such as IQVIA, Optum and MarketScan use proprietary, in house platforms. Similar to Redivis, these platforms offer a secure, scalable, customizable computational environment with tiered access depending on the researcher's expertise and need. Many of these platforms have lowered barriers to access by making only de-identified data generally available to investigators, though Research Identifiable Files may be available to smaller groups of researchers.¹²

Perhaps the most important distinction of our data platform is that it was optimized for academic data centers and designed to be replicable: an affordable, off the shelf software¹⁴. A feature particularly important to academic use, is that it is implemented as a multi-tenant environment, which in practice means that researchers can discover and combine any dataset that they have access to across institutions that are utilizing this software. This has the potential to encourage sharing of high-risk data, since the data custodian can share the data without having to relinquish branding, credit, or control.

Redivis is unique in its particularly robust permissioning system and the level of complexity that it supports. Most labs work with multiple datasets with varying risk levels and different access requirements. For example, a lab with eight researchers may have a study with four high-risk datasets, two of which require proof of project approval from data owners, three of which require IRB and all of which require proof of encryption to access. The Redivis system can manage this complexity and allow team members to share projects, data, and code in the project tool. It automatically restricts access to each dataset to the tier (view, meta-data, 1% sample, full dataset) for which the team member is authorized. This enables us to remain compliant without having to homogenize the risk profile or access requirements of datasets within a project and for teams to work together in an organic way.

Incorporating FAIR principles into our technical infrastructure

Findable

The entry point to data is the S-PHS Data Portal (Figure 2) which enables data search within the platform using either natural language or common ontologies, manages access as described above, and enables investigators to explore data. In addition to managing data discovery and access, it has tools that enable researchers to explore of a wide array of dataset types across a federated system of data lakes and enclaves. It also indexes all documentation, tags, variables, and variable metadata across a dataset, allowing the system to return the dataset(s) relevant to a researcher's line of inquiry. For example, if a medical diagnosis variable is annotated as containing ICD-X diagnosis codes, and the dataset includes observations for this variable of the value "I26" (pulmonary embolism), any searches for "pulmonary" will return this dataset and highlight the specific variable whose metadata matched the search term.

Moreover, we want to ensure that datasets hosted on the S-PHS Data Portal can be found via other discovery tools. To this end, all datasets are indexed by Google Dataset Search through the presence of well-formatted web crawler metadata, as well as by DataCite, through the issuance of DOIs and accompanying metadata. Both Google and DataCite use open and well-documented schema for their metadata, expanding the possibility for future tools to further leverage this metadata in aiding discovery.

Redivis also enables researchers to discover and explore a wide variety of data types, including geospatial, and unstructured data. Built-in previews are available for several unstructured data types (images, text, PDF, HDF5, DICOM), and any arbitrary file type can be uploaded and stored. Redivis automatically captures and displays metadata for uploaded data, such as univariate statistics for tabular data, geospatial dispersion for GIS data, and file metadata and checksums for unstructured data.

Accessible

Access requirements are complex and often cumbersome. The system has allowed us to streamline and standardize these steps, eliminate duplicative effort for both the researcher and administrator and ensure compliance. For tabular data, the S-PHS Data Portal automatically pulls and visualizes meta-data. Meta-data can also be loaded into the portal as a separate file. The system allows an administrator to set the “least necessary” requirements so that a researcher can see as much up front information as possible to assess the suitability of the data for their question.

At S-PHS’s inception, the task of tracking and verifying that data access requirements were complete and current was managed in a tabular database (REDCap).^{15,16} Although excellent for data collection, this type of software was not designed to manage data access. Because this approach was tedious, time-consuming, and often required duplicate efforts for both the data administrator and the researcher, data access was one of the first tasks S-PHS sought to automate. The system tracks which requirements an investigator has already completed so that application for an additional dataset only requires completing any new requirements specific to that dataset. Requirements for access are made transparent to investigators, and the system is designed to be user-friendly to minimize friction for investigators and administrators.

Data access is managed with a customizable, tiered, granular permissioning system that allows data custodians to set specific requirements for different tiers of data visibility and access

(Figure 3). These tiers are overview, which includes detailed descriptions of data provenance and methods for collection and other narrative information; metadata view which includes variable descriptions and characteristics; data sample access, for example, for national datasets with tens of millions of records, data owners may elect to make a 1 or 5% sample available prior to releasing the full dataset; and full data access. It is also possible to directly upload a dataset and share it with individual collaborators. For especially sensitive datasets, data custodians can hide them from search and specify which individuals will be able to discover the data. All of these tiers and settings may be combined as needed to comply with security regulations or contractual requirements. It is also straightforward to make low risk datasets publicly available.

The S-PHS Data Portal meta-data view enables researchers to assess the utility of a dataset for their research question, and includes information such as variable names, descriptions, summary statistics, join fields, and other population-level information (Figure 4). Searches can be performed on metadata such as tags, variable names, timeframe, size of dataset, and any other information included in the meta-data. Once a dataset is identified, researchers can view a more detailed description of the data provenance, methodology, a list of all variables, visualize key summary statistics (e.g., count, missingness, range, mean, median, histogram, and box-plot) for each variable, and learn which data access requirements are necessary to gain access to the dataset.

Once researchers have completed all data access requirements necessary for the 1% or 5% sample or the full dataset, they can use graphical and SQL interface tools (Figure 5) to identify their analytic cohort. They can select tables, variables and records of interest and reshape, clean, and prepare the data for research. Because the Redivis is a software layer built on the Google Cloud Platform, all computations on both Redivis and Nero are performed through the Google Cloud Platform and interfaced through a user's browser.¹⁷ All projects on Redivis can be shared with collaborators and—similarly to collaborative work tools such as Google Docs—multiple users can communicate within a project.

Once an investigator has selected the appropriate files, variables and records using a GUI or SQL tools within Redivis, data can be ported to any of several secure computational environments, which have a wide range of analytic tools.

Interoperable: built for collaboration

A collaborator accessing a dataset from another institution can authenticate using their institutional ID if they have an account and the hosting institution permits view or access of data from external collaborators. Redivis authenticates users with InCommon^{18,19} which is an access and identity management service specializing in academic and government organization authentications. Because it is built on the Google Cloud Platform, it is also possible to compute on data hosted in two different institutions' instances without moving the data, provided that all data access requirements have been met. This addresses a major barrier to data sharing: institutions and investigators often do not want to let high-value data out of their direct control, and in many cases are not permitted to do so, preventing computation from other institutions. Other organizations, such as TriNetX²⁰ and Google Data Commons²¹, also used federated data lakes, but the S-PHS approach is somewhat different in that other platforms tend to have binary data access (all or nothing), only offer this capability for one type of data (such as EHRs or claims), and are more likely to make only low to moderate-risk data available.

Prior to loading on the portal, data undergo curation for research use. A dedicated data management team at S-PHS coordinates all data transfers from data proprietors and is responsible for removal of person-readable identifiers, appending encrypted linkage keys, cleaning, curation, and all linkages which require access to explicit identifiers. This team also coordinates de-identification, conversion into common data models such as OMOP, and other curation, annotation, and documentation activities in a dedicated, secure computational environment. For data managers and custodians, there are tools for data ingestion and curation, data annotation, and other data management tasks. It automatically tracks contributors to a dataset and ties those contributions (including ORCID) to the DOI metadata. This means that when a paper cites a dataset, we can link back to the people that curated and contributed to the dataset. This will enable S-PHS to pull a list of all papers derived from particular datasets (any dataset that a data manager or researcher has contributed to), which could in turn provide credit to contributors of data and career advancement-related to data sharing.

Re-usable: computational environments

Once the files, variables, and records for an analytic cohort have been selected, the cohort can be moved to any of several secure computational environments depending upon the needs of the research team and the data's risk profile. The Redivis platform offers the ability for investigators to initialize Jupyter Notebooks in R, Python, and Stata. To accommodate larger data derivatives, as well as the growing prevalence of computationally intensive AI and ML methodologies, Redivis notebooks allows investigators to customize their notebook computational environment to the needs of their data and analysis, while billing investigators directly for these computational resources.

In addition to the Redivis native Jupyter notebook, Stanford has three high-performance computational environments for high-risk data: a secure, on-premise cluster of Windows servers (i.e. the S-PHS Windows Servers); Carina, which is an on-premise cluster of secure Linux servers; and Nero, which is based in the Google Cloud Platform. Administrators can specify the IP addresses of permissible environments for analyses, and all computational environments are designed for the highest risk data.

The secure computational environments present a suite of analytical tools, including (but not limited to) Jupyter, R, Python, SAS, Stata, and MATLAB, all of which are accessible through a Jupyter Lab interface and the command-line. Jupyter Lab allows researchers to switch between preferred analytic tools, annotate data, and document data curation steps in a notebook format, which can easily be shared with other investigators or subsequent data users, providing practical versatility.

Finally, we have several resources available to users including Slack channels, a centralized ticketing system for resolving technical issues (e.g. requests for additional software packages or other system updates), in-person and online office hours. For the most frequently used datasets, S-PHS also hosts research consortia for researchers to meet others working on similar data or topics, ask detailed data or methodology questions, and present proposed or current projects.

Collaborating with partners to translate real-world data into real-world impact

S-PHS partnerships have played a central role in ensuring RWD is translated into real-world evidence used to reduce inequities in the development, approval, and adoption of effective medical and public health interventions—a capacity we believe is critical for centers offering access to data. An important component of reducing disparities is engaging the communities which have contributed their data, particularly communities which have historically been excluded from data sources. With all RWD we work with, it is strictly forbidden to contract patients whose information is in the data. This, however, does not mean that community perspectives and interests can't be part of projects from their inception. To do this, S-PHS works with the Stanford Office of Community Engagement to advise study teams on how to engage communities around their topics of study. A number of our projects engage with key groups who are working to implement health promoting policies and practices for patients and communities. In collaboration with the American Board of Family Medicine, S-PHS data is being used to track racial and ethnic differences in long-COVID. S-PHS and the San Joaquin Valley Public Health Consortium, comprising eleven local public health jurisdictions dedicated to advancing health equity, jointly assessed the impact of COVID-19 on preventive screening rates in low-income, rural areas of California's Central Valley. S-PHS used claims data to obtain historic information for the area, and EHR data from primary care—which are updated much more frequently than claims datasets—to conduct a timely assessment. We are also conducting community engaged work on the effects of structural racism on health and have a project studying the impact of cash transfers on persistent poverty. Learnings from these projects will be used to inform retrospective analyses on structural racism and persistent poverty using datasets hosted by S-PHS. These and other partnerships have enabled S-PHS to implement many of the Diversity, Inclusion, and Equity recommendations developed during the 2020 NCATS meeting.²²

Critical further new areas needed for working with real-world data

As we look to the future, we are considering multiple new approaches for how RWD can be used to support translational science, with a focus on addressing several different barriers that we have encountered while supporting over 1000 projects over the last eight years. These include how to best harmonize across datasets with differing infrastructure and format, developing synthetic data, providing more numerous geographic overlays to data, supporting training across intuitions in the use of RWD, and supporting RWD analysis with interdisciplinary research teams.

Creating data structures that allow for harmonizing data from across platforms as well as more easily allow for replication of analyses across data is increasingly a focus of our work. Our current focus in this area is working to convert data to Observational Medical Outcomes Partnership (OMOP) data model, in line with the choice of many large collaborative studies.²³ This has so far resulted in great efficiencies for our work, but we have also realized that OMOP conversion is a process, not a destination. Subject matter experts using our data have continued to provide feedback on how best to capture different variables in the data, and the OMOP framework is well adapted for being updated to reflect expert opinion and changes that occur in coding in different EHR platforms.

One area of new work for us that we think is particularly promising is the development of synthetic data for our most used datasets. Synthetic data refers to fake data that is generated based on real data. The goal is to retain the properties of the original data as much as possible, such that the original and synthetic data are indistinguishable in terms of distributions of variables relationships between variables. In the context of healthcare data, the synthetic copy of the data is not subject to HIPAA/GDPR rules as it does not include real personal information. This provides the potential to simplify the process of sharing and collaborating with RWD while preserving patient privacy. There are useful new approaches to creating synthetic data that capitalize on recent advancements in deep learning methods to achieve synthetic copies of complex data with high fidelity to the original data source. Because the software that constructs these datasets is computationally intensive, they must be run on a commercial cloud or similar computational environment, especially if the original data are large. We are mindful of the limitations of the underlying data, and high-fidelity synthetic data are likely to recreate these biases as the benchmark for high-fidelity synthetic data is that they perform similarly to the original data in models. While the optimal approach is reducing selection and measurement error in the collection of the data itself, we can also implement approaches to reducing some of these sources of error in synthetic data. Strategies for addressing these issues include incorporating sampling weights and other adjustments into synthetic populations and adding guidance on addressing these limitations in data documentation.

A further area of development is to focus on datasets that have geographic information to address the fact that RWD rarely includes original data collection on social and environmental determinants of health. If some level of geographic information is included in RWD, geographic

overlays can be done which greatly expands the types of determinants of health that can be studied. Even levels of geography as large as the state can be beneficial, for example, for looking at the impacts of state differences in Medicaid expansion on health outcomes. Smaller levels of geography (e.g. ZIP code or census tract) allow for examining neighborhood effects on health and contaminants in the physical environment. Because data on health policies, social context and environmental exposures is also often difficult to find and also takes time to curate, focusing on providing these measures and crosswalk files for overlay can make including these factors in translational research projects much easier for researchers. We recently, for example, have calculated several area level deprivation measures at multiple levels of geography for capturing the local social environment of patients, and have made these available to researchers on our platform. We are also working to use geographic overlay data on the demographics of small areas to reweight RWD to reflect the underlying population.

There are many efficiencies that can be gained by supporting cross-institution collaborations on training in the use of RWD. There are currently few established curricula on the use of RWD at academic medical centers, yet there is a growing need for this unique focus in instruction. Rather than developing coursework that is institution based, we believe it will be important to partner to develop courses with an entirely or primarily remote based structure, to allow more efficient instruction across institutions. Both asynchronous online coursework as well as in-person short courses that bring people together across institutions will help build critical capacity in the skillsets needed to do translational research with RWD.

We are also increasingly focused on supporting interdisciplinary teams of staff and faculty to support RWD projects. Up until now, the traditional model has been that of a subject matter expert working with a statistician to analyze RWD to answer a particular research question. What we have found is that there is also an important role for a team member with deep knowledge of the data itself, which often can rapidly increase the speed and quality of translational research using RWD that can occur. In addition to these three types of expertise, we want to move towards a team that also includes individuals with expertise in study design, community and patient engagement, and health equity. Key to successfully implementing these teams to support the use of RWD includes educating researchers about the value of these teams, increasing capacity of individuals with expertise in these areas, and connecting researchers to an assembled team at the inception of a project.

Our experiences over the past eight years have shown us how much demand there is for access to real-world data, and that if the appropriate tools and structures are in place successful use of that data can be achieved. While the technological solutions and management structures to support researcher, data science and community collaborations take time to develop and implement, the longer-term gains for translational science have proven to be substantial.

Author Contributions: The concept for the paper originated with IC and DHR. All authors gave substantive input to the scope and concept for the paper. The original draft of the manuscript was done by IC and DHR. All authors added substantial content to the paper and revised for content. DHR takes responsibility for the manuscript as a whole.

Acknowledgments: The authors would like to thank the Stanford Research Computing Center, the architects of the Stanford computational environments. Initial funding of the S-PHS Data Portal (Redivis) was provided by grant G-2016-7232 by the Alfred P Sloan Foundation and S-PHS is part of the NIH Stanford Center for Clinical & Translational Education and Research (Spectrum) UL1TR001085 and 5UL1TR003142.

Disclosures: Ian Mathews is the founder of Redivis. No other authors have any conflict of interest.

Conflicts of Interest:

Ian Mathews is the CEO and Founder of Redivis.

None of the other authors have any conflicts of interest.

References

1. Stanford Center for Population Health Sciences. Accessed January 30, 2023. <https://med.stanford.edu/phs.html>
2. Minor LB, Rees M. *Discovering Precision Health: Predict, Prevent, and Cure to Advance Health and Well-Being*. Wiley-Blackwell; 2020.
3. Bauer GR, Churchill SM, Mahendran M, Walwyn C, Lizotte D, Villa-Rueda AA. Intersectionality in quantitative research: A systematic review of its emergence and applications of theory and methods. *SSM - Popul Health*. 2021;14:100798. doi:10.1016/j.ssmph.2021.100798
4. Chen J, Ho M, Lee K, et al. The Current Landscape in Biostatistics of Real-World Data and Evidence: Clinical Study Design and Analysis. *Stat Biopharm Res*. 2023;15(1):29-42. doi:10.1080/19466315.2021.1883474
5. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3(1):160018. doi:10.1038/sdata.2016.18
6. Mathews I, McIntyre S, DeLaney E. Redivis, Inc. Published online January 2023. <https://redivis.com/>
7. All of Us Research Program, National Institutes of Health. All of Us Researcher Workbench. <https://www.researchallofus.org/data-tools/workbench/>
8. National Center for Advancing Translational Sciences. National COVID Cohort Collaborative (N3C). Accessed January 30, 2023. <https://ncats.nih.gov/n3c>
9. Vivli Center for Global Research Data. Accessed January 30, 2023. <https://vivli.org/>
10. Rassen J, Schneeweiss S. Aetion Inc., Database System for Analysis of Longitudinal Data Sets. <https://aetion.com/>

11. Wang S, Verpillat P, Rassen J, Patrick A, Garry E, Bartels D. Transparency and Reproducibility of Observational Cohort Studies Using Large Healthcare Databases.: Transparency and reproducibility in healthcare databases. *Clin Pharmacol Ther.* 2016;99(3):325-332. doi:10.1002/cpt.329
12. Kerawala RP, King PF, Mayo KR, Vogel JM, the All of Us Research Program. *Analysis of the All of Us Research Program Researcher Workbench Workspaces.* Public and Global Health; 2022. doi:10.1101/2022.05.12.22274998
13. Haendel MA, Chute CG, Bennett TD, et al. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc.* 2021;28(3):427-443. doi:10.1093/jamia/ocaa196
14. Silva F, Souza R, Machado I. Taming and unveiling software reuse opportunities through white label software in startups. In: *2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA).* IEEE; 2020:302-305. doi:10.1109/SEAA51224.2020.00057
15. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform.* 2009;42(2):377-381. doi:10.1016/j.jbi.2008.08.010
16. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. *J Biomed Inform.* 2019;95:103208. doi:10.1016/j.jbi.2019.103208
17. Law W, Soderquist N, Marinshaw R. Stanford Nero - Jupyter Notebooks. Accessed January 28, 2023. <http://nero-docs.stanford.edu/jupyter.html>
18. Childs D, Christ B, (Vcie-Chair) ME, et al. InCommon. <https://www.incommon.org/>
19. Jokl J, Basney J, Caskey P, et al. *InCommon Certification Practices Statement for Server Certificates.*; 2010.

20. TriNetX, Inc. TriNetX. Published January 30, 2023. <https://trinetx.com/>
21. Guha RV. Data Commons. Published January 28, 2023. Accessed January 28, 2023. <https://www.datacommons.org/>
22. Boulware LE, Vitale A, Ruiz R, et al. Diversity, equity and inclusion actions from the NCATS Clinical and Translational Science awarded programs. *Nat Med*. 2022;28(9):1730-1731. doi:10.1038/s41591-022-01863-7
23. Klann JG, Joss MAH, Embree K, Murphy SN. Data model harmonization for the All Of Us Research Program: Transforming i2b2 data into the OMOP common data model. Lovis C, ed. *PLOS ONE*. 2019;14(2):e0212463. doi:10.1371/journal.pone.0212463

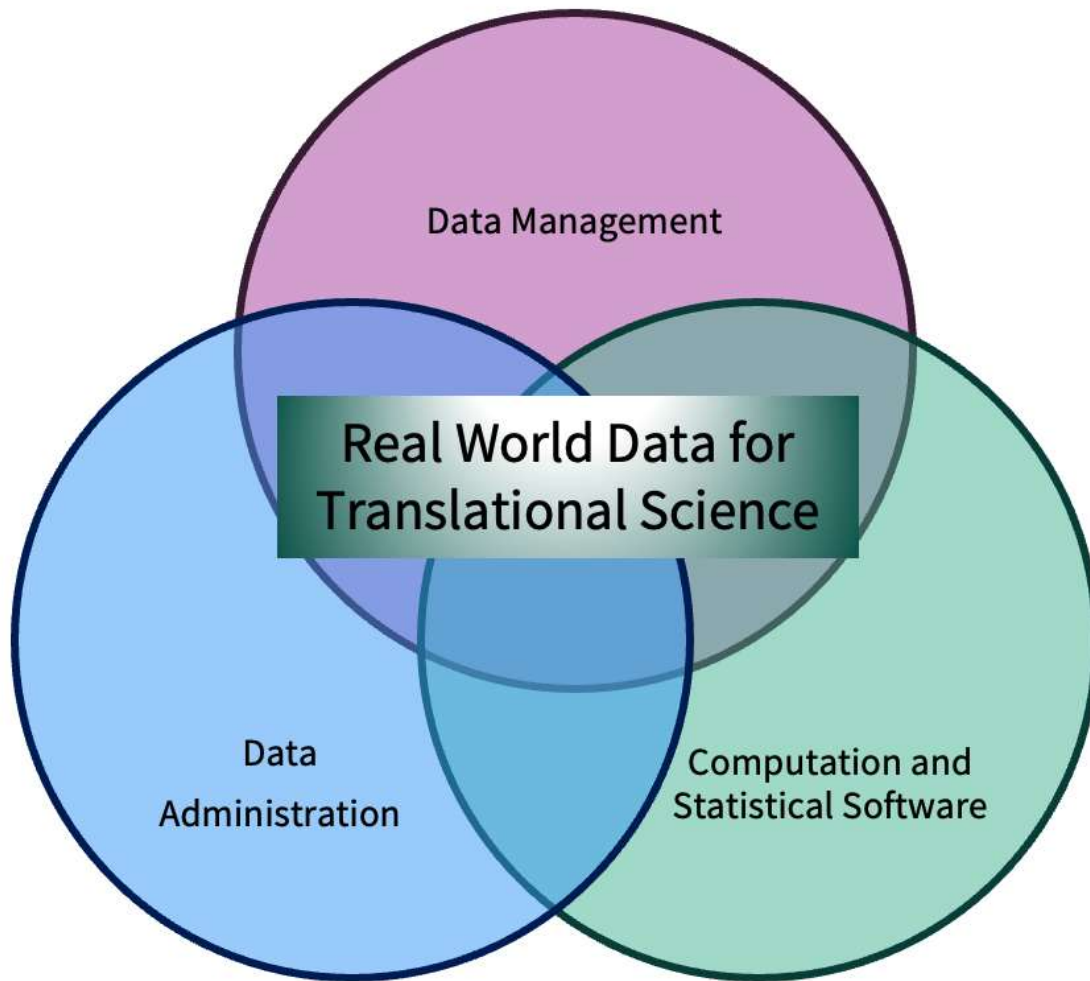


Figure 1: Three pillars of making real-world data useful

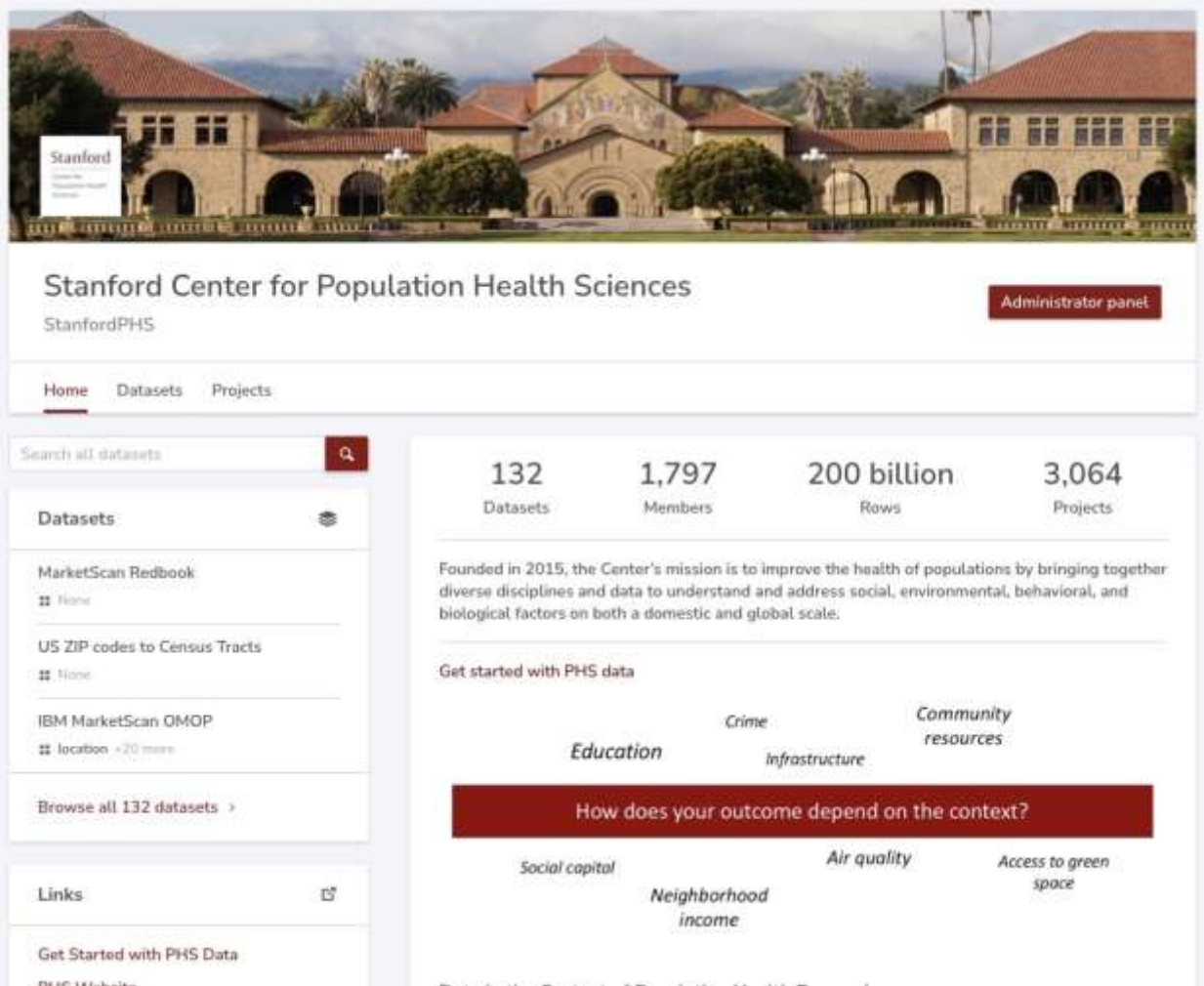


Figure 2: Landing page of the S-PHS Data Portal

Dashboard

MANAGE

Members 12

Studies 3

Projects

CREATE

Datasets

Permission groups

Requirements

AUDIT

Reports

Logs

Billing

Settings

Help

Stanford
Stanford Cente...
StanfordPHS

Permission groups > AFC RIF

10 datasets

Access

 Separate sample access

Overview access (142)

Direct access

142 users with direct access to one or more datasets

Metadata access (10)

Requirements

Affiliation - Academic or Govt

Attestation - COI

DUA - PHS

Encryption Verification

Membership form

+ Add new requirement

Data access: 1% sample (10)

Requirements

In addition to above requirements

Approval - AFC RIF

Approval - IRB RIF per dataset

Funding and Project Info per dataset

PI - RIF Collaborator per dataset

ORCID Authentication

Training - CITI

Training - HIPAA

Training - PHS Data Security

Affiliation - SU Collaborator

+ Add new requirement

Data access (10)

Anyone with sample access

Usage rules

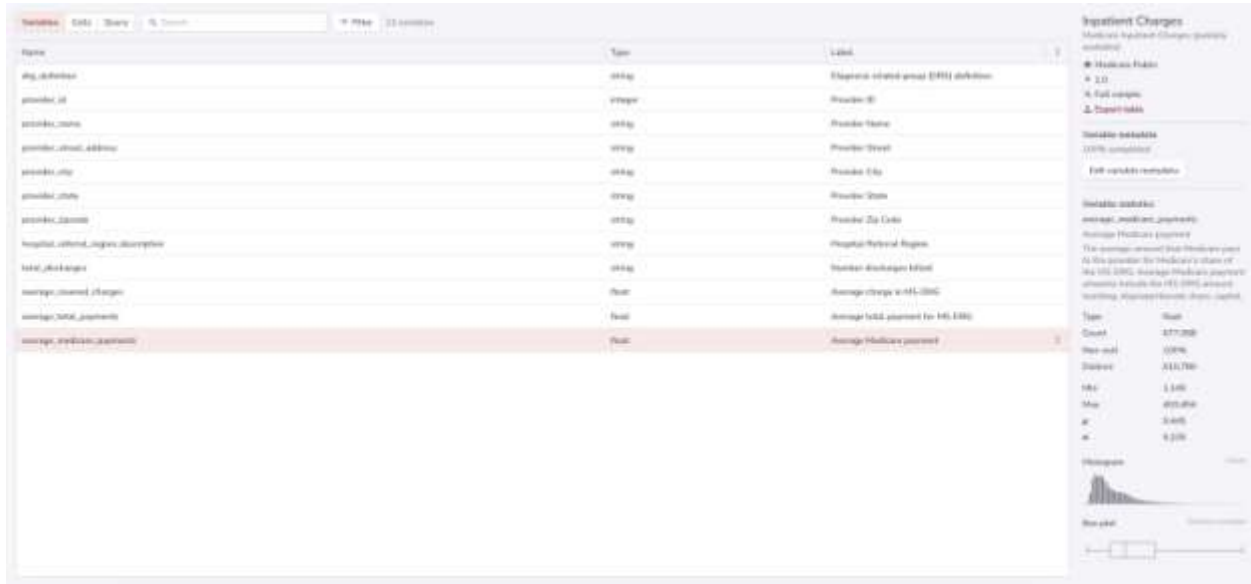


Figure 4: Meta-data view on the S-PHS Data Portal

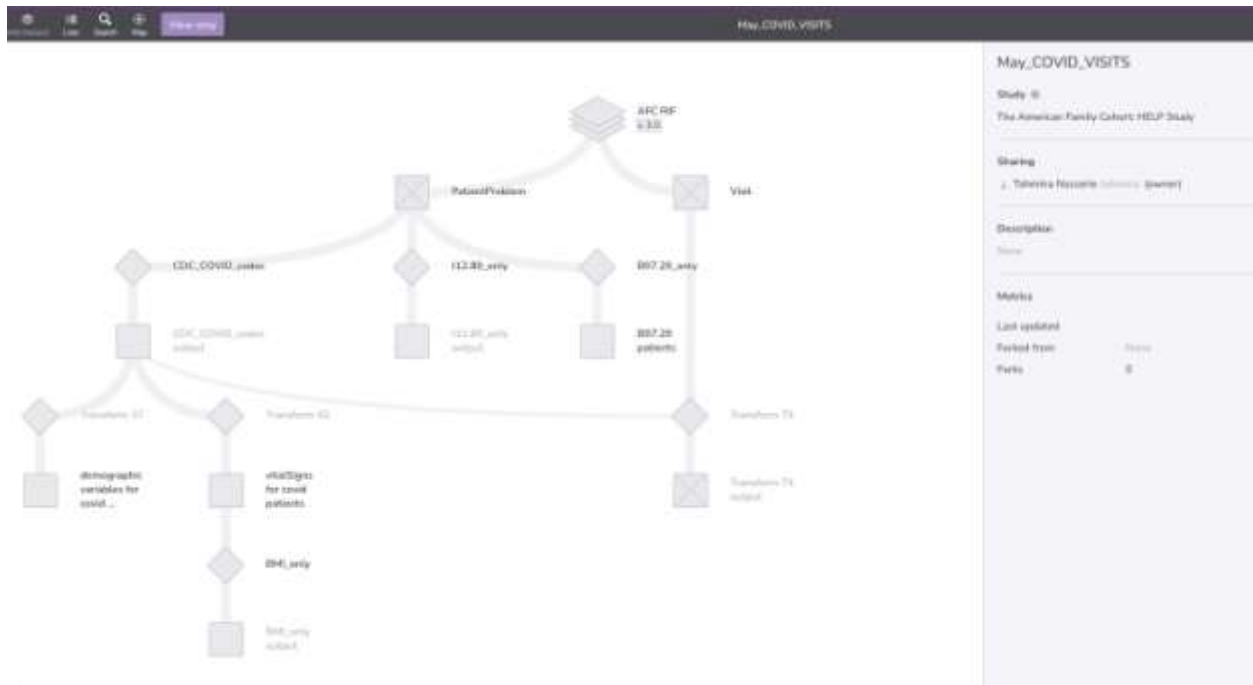


Figure 5: Project tool for data curation and cohort selection