UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Data Standardization, Federated Learning, and Informed Consent Algorithms and Tools to Honor Patient Privacy and Preferences in Clinical Research

Permalink

https://escholarship.org/uc/item/6mf5c2wc

Author

Kim, Jihoon

Publication Date

2023

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

Data Standardization, Federated Learning, and Informed Consent Algorithms and Tools to Honor Patient Privacy and Preferences in Clinical Research

A Dissertation submitted in partial satisfaction of
the requirements for the degree
Doctor of Philosophy

in

Bioinformatics and Systems Biology
with a Specialization in
Biomedical Informatics

by

Jihoon Kim

Committee in charge:

Professor Lucila Ohno-Machado, Chair
Professor Tsung-Ting Kuo, Co-Chair
Professor Jane C Burns
Professor Terry Gaasterland
Professor Michael Hogarth

2023

Copyright

Jihoon Kim, 2023

All rights reserved.

The Dissertation of Jihoon Kim is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To my parents, family, and friends

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to Dr. Lucila Ohno-Machado who provided me unceasing support, guidance, and encouragement in the successful completion of this dissertation. She always listened to me, trusted me, and pushed me to be the best version of myself. I would not be here today it if were not for her.

This doctoral dissertation would not have been possible without my committee members. I want to recognize Dr. Tsung-Ting Kuo, who always answered my questions about data analysis and machine learning and provided insights to the research as a co-chair of my dissertation committee. A special thank goes to Dr. Terry Gaasterland, who always stood by me and all other BISB students in academic, administrative, and emotional matters. I listened to her advice (initially reluctantly, to be honest) that was given to the first-year students during orientation to take diverse courses about the fundamentals outside the BISB core curriculum; these courses paved the way to the completion of my thesis. I am also grateful for the insightful teachings of Dr. Michael Hogarth about all things data and computer systems in the context of clinical informatics, which contributed to two chapters of this thesis. Finally, I want to show gratitude to Dr. Jane Burns, who helped me develop a scientific mind to develop, test, and validate a hypothesis with tenacity and creativity.

I am indebted to the current and former Department of Biomedical Informatics (DBMI) members, who provided me invaluable support throughout the course of my career. I learned both technical and soft skills from the bright minds at DBMI. I would like to thank DBMI members Tyler Bath, Sally Baxter, Elizabeth Bell, Michelle Day, Robert El-Kareh,

Nancy Herbst, Andrew Hodges, Xiaoqian Jiang, Hyeoneui Kim, Wentao Li, Larissa Neumann, Paulina Paul, Anh Pham, Kai Post, Siddharth Singh, Amy Sitapati, and Hai Yang.

I also place on record, my sense of gratitude to the students Caitlin Guccione, Jonathan Lam, Lauryn Bruce and the 2020 BISB/BMI cohort students, who have lent their hands in this venture. I am also grateful to my collaborators Pravina Kota, Elisa T Lee, Yu Rang Park, Chisato Shimizu, Ying Zhang, and Kai Zheng.

Finally, I would like to thank my wife Yookyoung and my daughter Chloe for their unconditional love and constant support. I would not have finished this journey without it.

Chapter 2, in full, is a reprint of the material as it appears in "Privacy-protecting, reliable response data discovery using COVID-19 patient observations" by Jihoon Kim, Larissa Neumann, Paulina Paul, Michele E Day, Michael Aratow, Douglas S Bell, Jason N Doctor, Ludwig C Hinske, Xiaoqian Jiang, Katherine K Kim, Michael E Matheny, Daniella Meeker, Mark J Pletcher, Lisa M Schilling, Spencer SooHoo, Hua Xu, Kai Zheng, Lucila Ohno-Machado, and the R2D2 Consortium, *J Am Med Inform Assoc*. 2021 Jul 30;28(8):1765-1776. Doi: 10.1093/jamia/ocab054, PMID: 34051088. I was the primary investigator and the first author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in "VERTIcal Grid lOgistic regression with Confidence Intervals (VERTIGO-CI)" by Jihoon Kim, Wentao Li, Tyler Bath, Xiaoqian Jiang, and Lucila Ohno-Machado, *AMIA Jt Summits Transl Sci Proc* 2021 May 17;2021:355-364. eCollection 2021, PMID: 34457150. I was the primary investigator and the first author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in "Transforming Cohort Data Collected over the Span of 30-years into a Common Data Model Without External

Access to Individual Records: the Strong Heart Study" by Jihoon Kim, Paulina Paul, Pravina Kota, Tyler Bath, Kai W Post, Yu Rang Park, Ying Zhang, Elisa T Lee, and Lucila Ohno-Machado, currently under review for a publication. I was the primary investigator and the first author of this paper.

Chapter 5, in full, is a reprint of the material as it appears in "Patient Perspectives About Decisions to Share Medical Data and Biospecimens for Research" by Jihoon Kim, Hyeoneui Kim, Elizabeth Bell, Tyler Bath, Paulina Paul, Anh Pham, Xiaoqian Jiang, Kai Zheng, and Lucila Ohno-Machado, *JAMA Network Open*. 2019 Aug 2;2(8):e199550. Doi: 10.1001/jamanetworkopen.2019.9550, PMID: 31433479. I was the primary investigator and the first author of this paper.

VITA

1998	B.S. in Statistics, Seoul National University, Seoul, Republic of Korea
1998-2001	Software Developer Officer, Central Computing Center, Korean Navy, Daejeon, Republic of Korea
2004	M.S. in Bioinformatics, Seoul National University, Seoul, Republic of Korea
2007	M.S. in Statistics, University of Wisconsin, Madison
2007-2009	Staff Statistician, Brigham and Women's Hospital / Harvard Medical School
2009-2011	Senior Statistician, University of California San Diego Health
2011-2023	Principal Statistician, University of California San Diego Health
2023	Ph.D. in Bioinformatics and Systems Biology with a Specialization in Biomedical Informatics, University of California San Diego

PUBLICATIONS

Jihoon Kim, Wentao Li, Tyler Bath, Xiaoqian Jiang, Lucila Ohno-Machado."VERTIcal Grid lOgistic regression with Confidence Intervals (VERTIGO-CI)", *AMIA Jt Summits Transl Sci Proc*. 2021 May 17;2021:355-364. eCollection 2021. PMID: 34457150

Jihoon Kim, Larissa Neumann, Paulina Paul, Michele E Day, Michael Aratow, Douglas S Bell, Jason N Doctor, Ludwig C Hinske, Xiaoqian Jiang, Katherine K Kim, Michael E Matheny, Daniella Meeker, Mark J Pletcher, Lisa M Schilling, Spencer SooHoo, Hua Xu, Kai Zheng, Lucila Ohno-Machado, R2D2 Consortium. "Privacy-protecting, reliable response data discovery using COVID-19 patient observations", *J Am Med Inform Assoc*. 2021 Jul 30;28(8):1765-1776. doi: 10.1093/jamia/ocab054. PMID: 34051088

Jihoon Kim, Hyeoneui Kim, Elizabeth Bell, Tyler Bath, Paulina Paul, Anh Pham, Xiaoqian Jiang, Kai Zheng, Lucila Ohno-Machado. "Patient Perspectives About Decisions to Share Medical Data and Biospecimens for Research", *JAMA Network Open*. 2019 Aug 2;2(8):e199550. doi: 10.1001/jamanetworkopen.2019.9550. PMID: 31433479

Jihoon Kim, Paulina Paul, Pravina Kota, Tyler Bath, Kai W Post, Yu Rang Park, Ying Zhang, Elisa T Lee, Lucila Ohno-Machado. "Transforming Cohort Data Collected over the Span of 30-years into a Common Data Model Without External Access to Individual Records: the Strong Heart Study", submitted for publication.

The following publications were not included as part of this dissertation, but were also significant byproducts of my doctoral training.

Chisato Shimizu, **Jihoon Kim**, Ming He, Adriana H Tremoulet, Hal M Hoffman, John Y-J Shyy, Jane C Burns. "RNA Sequencing Reveals Beneficial Effects of Atorvastatin on Endothelial Cells in Acute Kawasaki Disease", J Am Heart Assoc. 2022;11: e025408. doi:10.1161/JAHA.122.025408. PMID: 35861833

Pradipta Ghosh, Gajanan D Katka, Chisato Shimizu, **Jihoon Kim**, Soni Khandelwal, Adriana H Tremoulet, John T Kanegaye, Pediatric Emergency Medicine Kawasaki Disease Research Group, Joseph Bocchini, Soumita Das, Jane C Burns, Debashis Sahoo. "An Artificial Intelligence-guided signature reveals the shared host immune response in MIS-C and Kawasaki disease", Nat Commun. 2022;13: 2687. doi:10.1038/s41467-022-30357-w. PMID: 35577777

Clive Hoggart, Chisato Shimizu, Rachel Galassini, Victoria J Wright, Hannah Shailes, Evan Bellos, Jethro A Herberg, Andrew J Pollard, Daniel O'Connor, Shing Wan Choi, Eleanor G Seaby, Stephanie Menikou, Martin Hibberd, Neneh Sallah, David Burgner, Paul Brogan, Harsita Patel, **Jihoon Kim**, Adriana H Tremoulet, Eeva Salo, Diana van Stijn, Taco Kuijpers, Jane C Burns, Michael Levin, International Kawasaki Disease Genetics Consortium; UK Kawasaki Disease Genetics Consortium; EUCLIDS Consortium. "Identification of novel locus associated with coronary artery aneurysms and validation of loci for susceptibility to Kawasaki disease", Eur J Hum Genet. 2021;29: 1734–1744", doi:10.1038/s41431-021-00838-5. PMID: 33772158

Tsung-Ting Kuo, Anh Pham, Maxim E Edelson, **Jihoon Kim**, Jason Chan, Yash Gupta, Lucila Ohno-Machado. "R2D2 Consortium Blockchain-Enabled Immutable, Distributed, and Highly Available Clinical Research Activity Logging System for Federated COVID-19 Data Analysis from Multiple Institutions", *J Am Med Inform Assoc*. 2023 Mar4 14; doi.org/10.1093/jamia/ocad049. PMID: 36916740

Tsung-Ting Kuo, Xiaoqian Jiang, Haixu Tang, XiaoFeng Wang, Arif Harmanci, Miran Kim, Kai Post, Diyue Bu, Tyler Bath, **Jihoon Kim**, Weijie Liu, Hongbo Chen, Lucila Ohno-Machado. "The evolving privacy and security concerns for genomic data analysis and sharing as observed from the iDASH competition", *J Am Med Inform Assoc*. 2022 Nov 14;29(12):2182-2190. doi: 10.1093/jamia/ocac165. PMID: 36164820

Tsung-Ting Kuo, Tyler Bath, Shuaicheng Ma, Nicholas Pattengale, Meng Yang, Yang Cao, Corey M Hudson, **Jihoon Kim**, Kai Post, Li Xiong, Lucila Ohno-Machado. "Benchmarking blockchain-based gene-drug interaction data sharing methods: A case study from the iDASH 2019 secure genome analysis competition blockchain track", *Int J Med Inform.* 2021 Oct;154:104559. doi: 10.1016/j.ijmedinf.2021.104559. PMID: 34474309

ABSTRACT OF THE DISSERTATION

Data Standardization, Federated Learning, and Informed Consent Algorithms and Tools to
Honor Patient Privacy and Preferences
in Clinical Research

by

Jihoon Kim

Doctor of Philosophy in Bioinformatics and Systems Biology

with a Specialization in Biomedical Informatics

University of California San Diego, 2023

Professor Lucila Ohno-Machado, Chair
Professor Tsung-Ting Kuo, Co-Chair

There is growing public awareness and concern about patient privacy and the potential risks of sharing clinical data for research. While investigators try to collect large amounts of data to increase the statistical power and diversity of the population, patients strive to gain

more control of their data, in a way that respects their preferences; they want assurances about protecting their privacy. In this dissertation, I show how we can satisfy both researchers and patients through the novel use and development of data standardization, federated learning, and tiered informed consent algorithms and tools to foster clinical research while protecting patient privacy and preferences.

Chapter 1 is an introduction that consists of research background, significance, problem statement, objectives, and thesis organization. Chapter 2 illustrates a distributed, federated network of 12 health systems that harmonized their electronic health records to a common data model to answer clinical questions related to COVID-19 and post these answers online in a privacy-preserving manner. This network is composed of horizontally partitioned data (i.e., complete data about a set of patients are located in different sites and these sites cannot share data at the individual level). Chapter 3 presents a new algorithm and implementation of distributed logistic regression model for vertically partitioned data (i.e., partial data about a patient are located in different participating sites and these sites cannot share those data at the individual level). Chapter 4 delineates how a source database of medical records can be transformed to a destination database following a common data model, under the constraint that an external expert team cannot access to individual level data. While these chapters describe how various institutions could collaborate without sharing individual level data, Chapter 5 explores whether it is feasible for patients to describe their sharing preferences so that a healthcare system can share data according to these preferences, and different ways to elicit these preferences. Chapter 6 provides a conclusion and future directions for this work.

Chapter 1 Introduction

The wide adoption of electronic health record in healthcare organizations and the advance of software tools, smart devices, and computing power are generating large and diverse data that meet the requirements of clinical researchers[1]. More data are now stored, accessed, and shared with more individuals, organizations, and compute devices than any other time in history. On the patient side, however, there is a rising concern about patient privacy and confidentiality and increasing incidents of healthcare data breaches that could cause stigmatization, job dismissal, or denial of insurance coverage[2,3].

A traditional model of clinical research is to set up a central server into which each collaborating site transfers its de-identified individual-level data and to conduct the analysis about a specific disease such as cancer[4] or eye disease[5] on a combined dataset. This approach has a high risk of patient privacy breaches and is often challenging or even prohibited due to institutional, federal, or nation level regulation and policy[6]. For example, data egress out of a dedicated compute environment is not allowed in large research programs such as the All of Us Research Program (AoU)[7] and the Million Veterans Program (MVP)[8]. Additionally, data transfer outside of the country of origin is prohibited in some European countries[9].

A federated data network is an alternative model that aims to overcome these limitations by broadcasting clinical queries or specific requests for calculations to participating sites, with each site running analyses locally and then transferring local site-level aggregate results into a coordinating site to combine as a global result, while keeping the patient-level data within the compute environment of each site[10]. The Patient-Centered Outcomes Research Network (PCORnet)[11] and the Accrual to Clinical Trials (ACT) networks[12] are two large federated data networks[10]. Disease-specific federated data networks for asthma[9] and pulmonary hypertension[13] are emerging. The number of participating health systems joining federated data networks is expected to grow[14].

Federated learning extends beyond distributed count queries suitable for data characterization or exploratory analysis, to advanced multivariate analyses. Federated learning is an approach in which statistical models are built in a distributed fashion over participating sites in a federated data network, by iteratively communicating aggregate results such as statistics (e.g., counts or variances) and model parameters, while keeping the patient-level data within each site. Figure 1.1 illustrates a schematic view of federate learning, where four health systems cooperate to build a statistical model for prediction by transferring summary statistics or parameters while keeping individual-level data within each local site. These data are kept on a harmonized database that is compliant with a common data model.

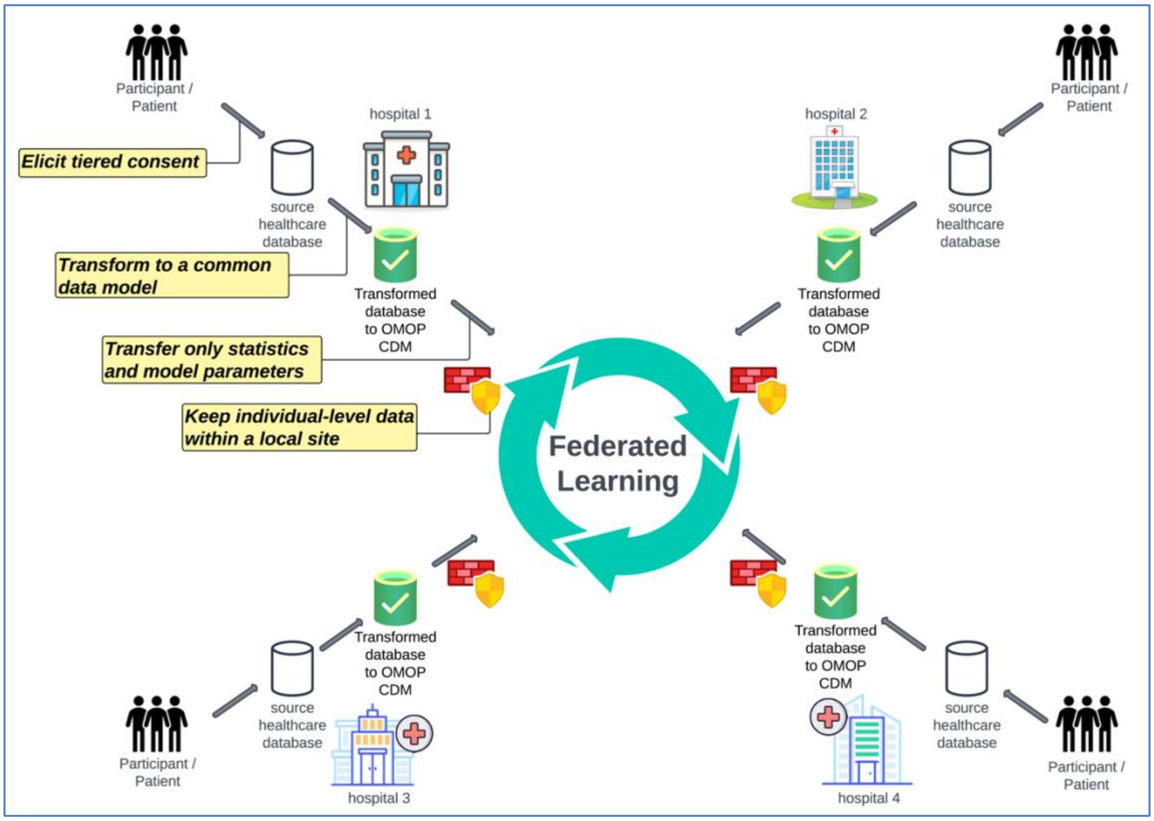

Participant / Patient
Elicit tiered consent
hospital 1
source healthcare database
Transform to a common data model
Transformed database to OMOP CDM
Transfer only statistics and model parameters
Keep individual-level data within a local site
Federated Learning
hospital 2
Participant / Patient
source healthcare database
Transformed database to OMOP CDM
Transformed database to OMOP CDM
source healthcare database
Participant / Patient
hospital 3
Transformed database to OMOP CDM
source healthcare database
Participant / Patient
hospital 4

Figure 1.1: Overview of federated learning on the harmonized data with tiered consent

Figure 1.2 shows the organization of core chapters of this thesis. Figure 1.2A-C refers to analyzing the data without allowing external people access the individual level data (i.e., analyzing data “in-situ”). I consider two different settings of data partitioning.

Figure 1.2A represents a setting of horizontally partitioned data, where each row represents a patient, each column represents a feature, and every party or health system has the same features. For horizontally partitioned data, many federated learning algorithms exist[15–18] but their deployment on the real world dataset still has practical challenges such as data standardization or harmonization[19]. I hypothesize that federated learning on horizontally partitioned data on real world data is practical (possible and fast) in a large network.

Figure 1.2B illustrates a setting of vertically partitioned data, where every party has the same individual on the row and their features are stored separately in three different sites. I hypothesize that a federated learning algorithm for logistic regression based on convex optimization and ring-structure learning can produce nearly identical results (both coefficients and their standard errors) as a centralized one.

Figure 1.2C depicts data standardization while protecting privacy of patients. Input data need to be in a standardized format in federated analyses. For American Indian data, this meant experts helping a local team that is authorized to work with the data standardize their data, without having these external experts access individual-level data. I hypothesize that source data can be mapped to a common data model successfully if the teams are tightly integrated.

Chapters 2 and 3 dealt with how to compute with distributed data considering data controlled only at the institutional level, but did not explore how data could be controlled at the patient level. The traditional all-or-nothing informed consent for long-term general use of data for research that is used today seems antiquated in the era of digital health, where the role of

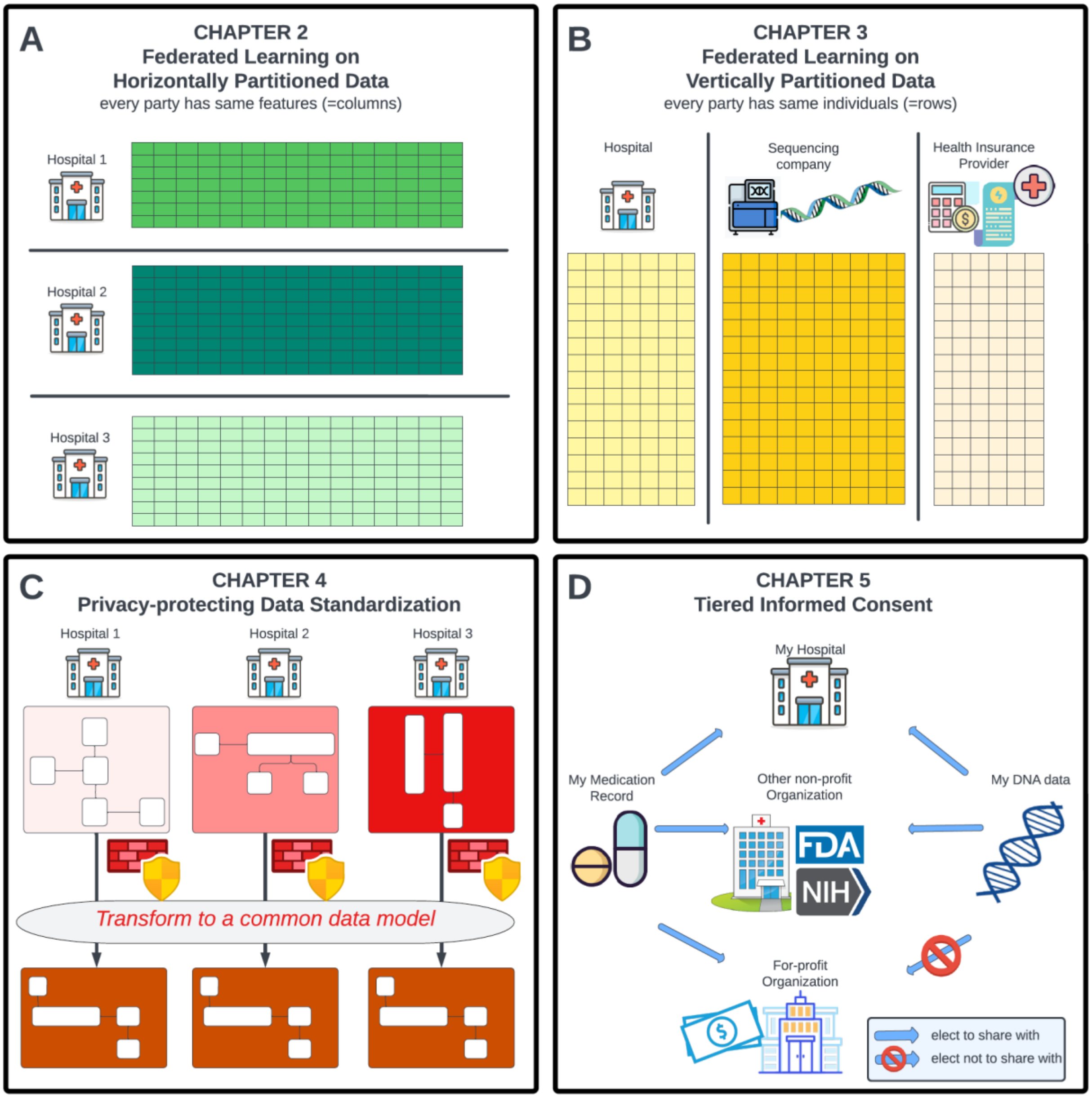


Figure 1.2: Organization of core chapters. A: Horizontally partitioned data setting. Each row is a patient, each column is a feature, and each party has the same features. **B**: Vertically partitioned data setting. Each row is a patient, each column is a feature, and each party has the same patients. **C**: Before data transformation, each hospital has its own source data model represented as different shapes and colors. After data transformation conducted in a privacy protecting manner, all three hospitals have the same data model in compliant with a common data model. **D**: A patient has the ability to elect to share which data elements with which organizations including my hospital, other non-profit organizations (other hospitals or government agencies), or for-profit organizations.

patients and their demands for autonomy and active engagement are increasing. Additionally, some researchers have in the past not respected agreements on the use of data for research. For example, the Havasupai tribe of American Indians filed a lawsuit against an academic institute when they found out that their DNA samples initially collected for genetic studies on type 2 diabetes had been used in other genetic studies[20]. The National Center for Vital and Health Statistics has recommended that the future Nationwide Health Information Network adopt an informed consent model that allows a patient to control the disclosure of predetermined sensitive data elements of health information[21]. The data scope of federated learning and data standardization covered in previous chapters could be limited to the consented data by patients. In our study, we randomly assigned patients to 1 of 4 types of preference elicitation forms to examine whether the tiered informed consent form layout and opting-in or opting-out method were associated with patients' sharing preferences. (Figure 1.2D) I hypothesized that patient decisions about sharing their EHRs and biospecimen for research would vary depending on researchers' affiliations, patient characteristics, the user interface design format of the consent form in which data sharing preferences were elicited, including the type of default format (i.e., opt-in or opt-out).

Chapter 2 Privacy-Protecting, Reliable Response Data Discovery Using COVID-19 Patient Observations

2.1 ABSTRACT

To utilize an individual- and institutional-privacy-preserving manner, electronic health record (EHR) data from 202 hospitals were analyzed to answer questions related to COVID-19 and post these answers online.

We developed a distributed, federated network of 12 health systems that harmonized their EHRs and submitted aggregate answers to consortia questions posted at https://www.covid19questions.org. Our consortium developed processes and implemented distributed algorithms to produce answers to a variety of questions. We were able to generate counts, descriptive statistics, and build a multivariate, iterative regression model without centralizing individual-level data.

Our public web site contains answers to various clinical questions, a web form for users to ask questions in natural language, and a list of items that are currently pending responses. The results show, for example, that patients who were taking Angiotensin-Converting Enzyme Inhibitors and Angiotensin II Receptor Blockers, within the year before admission, had lower unadjusted in-hospital mortality rates. We also showed that, when adjusted for age, sex and ethnicity were not significantly associated with mortality. We demonstrated that it is possible to answer questions about COVID-19 using EHR data from systems that have different policies and must follow various regulations, without moving data out of their health systems.

We present an alternative or complement to centralized COVID-19 registries of EHR data. We can use multivariate distributed logistic regression on observations recorded in the process of care to generate results without transferring individual-level data outside the health systems.

2.2 INTRODUCTION

The COVID-19 pandemic represents a watershed event in public health and has highlighted numerous opportunities and needs in clinical and public health informatics infrastructure[22–24]. One of the key challenges is the rapid response of analyses and interpretation of observational data to inform clinical decision making and patient expectations, understanding, and perceptions[25–29].

Several initiatives are building COVID-19 registries or consortia to analyze electronic health record (EHR) data[28]. The expectation is that these resources will provide researchers and clinicians access to a rich source of observational data to understand the clinical progression of COVID-19, to estimate the impact of therapies, and to make predictions regarding outcomes. Registries may contain limited data for patients diagnosed with COVID-19: the barriers for having more data are based on both privacy concerns and also on what elements have been deemed valuable by health professionals and researchers at a particular point in time. The problem with a new and evolving disease like COVID-19 is that we do not know what data or information will be most valuable. For example, in the pandemic's early stages, the dermatological and hematological findings were not evident, and those data were not included in registries or reports[30]. Interest in specific laboratory markers (e.g., D-dimer, troponin) for these disturbances and additional medications (e.g., antihypertensive drugs) or phenotypes (e.g., diabetes, blood type) has increased over time[31–33]. Additionally, it is challenging for researchers and clinicians to understand the structure and quality of the data in data repositories, and to formulate queries to consult the data in their institution and in others.

Thus, the utilization of EHRs to characterize COVID-19 disease progression and outcomes is challenging. However, EHR data may be useful when a randomized clinical trial cannot be conducted. Observational data may also help determine if results from a randomized clinical trial (RCT) replicate after relaxing eligibility criteria for real-world applications. While the scientific community has raised concerns about the reproducibility of findings, data provenance, and proper utilization of observational data, resulting in some COVID-19 articles being retracted[34], there remains a clear need to responsibly, ethically, and transparently analyze observational data to provide hypothesis generation and guidance in the pursuit of evidence-based healthcare.

In this study, we focus on using novel decentralized data governance and methods to analyze EHR-derived data.

2.3 METHODS

2.3.1 Distributed research network of 12 health systems

Researchers' questions posed in natural language are answered by distributed data maintained in 12 health systems, covering 202 hospitals located in all U.S. states and two territories, and one international academic medical center (Table 2.1). This collaboration provides the capability for comparisons with historical data from over 45 million patients and uses a dynamic approach to account for an evolving awareness of the most impactful COVID-19 questions to answer and hypotheses to explore. All sites have transformed or are actively transforming data into OMOP, but some of them only use data from COVID-19 registry patients (i.e., do not transform the full EHR-based data warehouse), and some only have in OMOP the items required by the query. The ability to build and evaluate multivariate models across a large number of health systems and to integrate results from registries differentiates our approach from most federated clinical data research network approaches.

Table 2.1: Participating sites. Cedars Sinai Medical Center (CSMC), University of Colorado Anschutz Medical Campus (CU-AMC), Ludwig Maximillian University of Munich (LMU), San Mateo Medical Center (SMMC), University of California (UC) Davis (UCD), Irvine (UCI), San Diego (UCSD), San Francisco (UCSF), University of Southern California (USC), University of Texas Health Science Center at Houston and Memorial Hermann Health System (UTH), Veterans Affairs Medical Center (VAMC). *Available data on hospital characteristics from 2018. **Two additional sites joined the consortium and will begin answering queries in 2021.

Institution**	Hospitals	Beds	Discharges/yr	EHR system	Data Source
CSMC	2	1,019	61,386	Epic	EHR
CU-AMC	12	1,829	106,325	Epic	EHR
LMU*	12	1,964	78,673	SAP/i.s.h.med, QCare, IMESO	COVID-19 registry
SMMC	1	62	1,951	Harris Software (Pulsecheck), Cerner (Soarian), eClinicalworks	EHR
UCD	1	620	32,248	Epic	EHR
UCI	1	417	21,656	Epic	EHR
UCLA	2	786	47,491	Epic	EHR
UCSD	3	808	29,895	Epic	EHR
UCSF	3	796	48,120	Epic	EHR
USC	2	1,511	23,454	Cerner	EHR
UTH	17	4,164	233,890	Cerner	COVID-19 registry
VAMC	146	13,000	676,402	ViSTa/CPRS	EHR
Total	202	26,976	1,361,491		

2.3.2 Rapid Q&A system about COVID-19 inpatients

The development of our Q&A system involved the inclusion of new concept codes in local repositories, agreement on concept definitions (e.g., what constitutes a COVID-19 hospitalization, what codes should be included in the definition of History of Coronary Heart Disease, and how to map laboratory test records into LOINC, for which we developed a mapping tool)[35]. Instead of a singular control of a coordinating center, the R2D2 consortium allows participating institutions to “own” the development and testing of queries across various sites, which promotes a balanced division of workload and increases the ability of individual sites to develop generalizable queries and manage responses with help from the whole consortium. The translation of questions into code relies on members of the Reliable Response Data Discovery for COVID-19 (R2D2) Consortium. The analyses performed on data transformed into the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) from relevant patient cohorts do not require data transfer outside the participating institutions and reduce the risk of individual or institutional privacy breaches. After a partially automated quality control process, which is carefully reviewed by multiple consortium members, only the results of calculations (e.g., counts, statistics, coefficients, variance-covariance matrices) are released from the healthcare institutions; no individual patient-level data are shared[15].

2.3.3 Detailed workflow

Figure 2.1 shows our general workflow, including human interpretation and clarification of questions, human quality control of answers, using graphs and related visualizations as much as possible. The responsibility of the Lead Site--to create a template query for all responding sites to use for rapid response--rotates among institutions (i.e., Health Systems). A more detailed workflow is illustrated in Figure 2.2 using a swim-lane format with an emphasis on roles. The Q&A process starts when a user creates a request through the public web-site, https://covid19questions.org. Next, the data scientist at the Consortium Hub checks if this question had been answered before and pass it on to the clinician at the Consortium Hub to assess the feasibility (i.e., if the received question is answerable from the local data mart) and

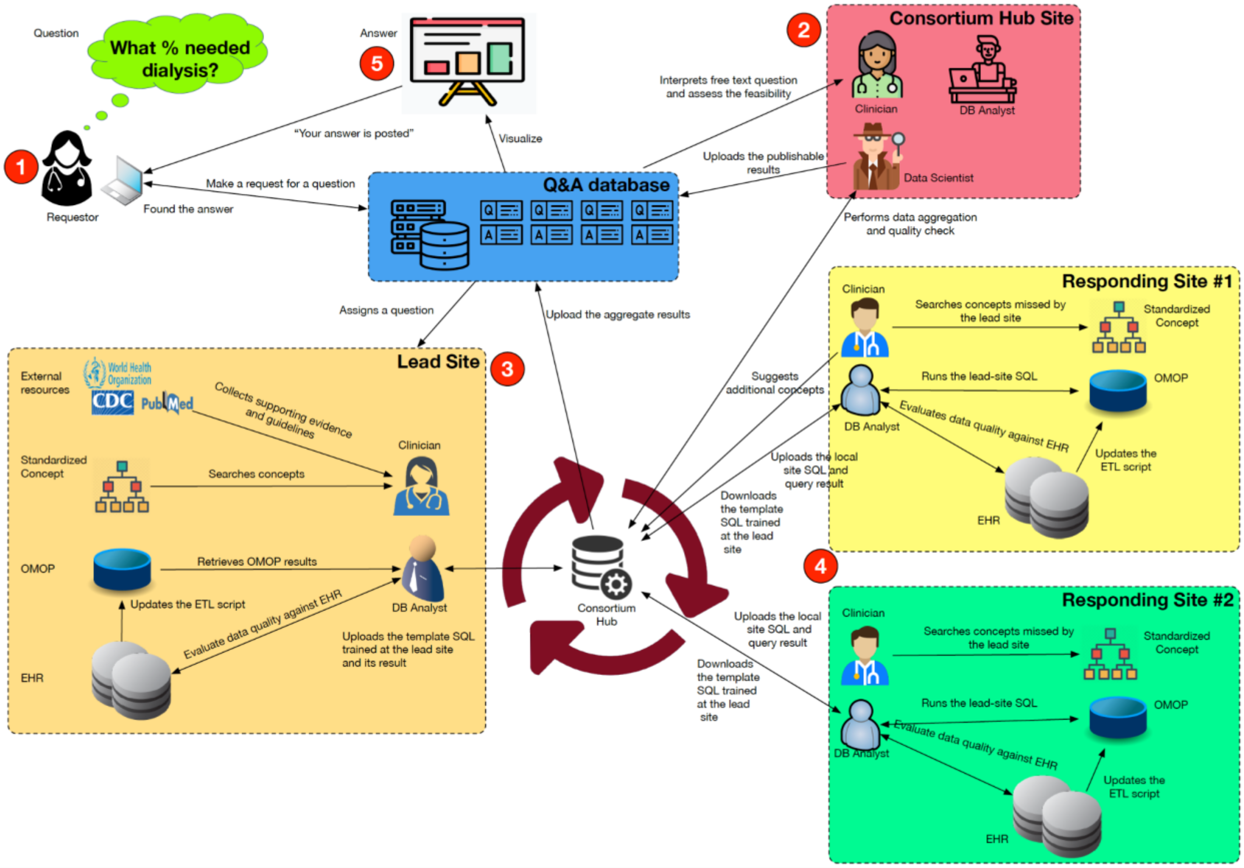

Question
What % needed dialysis?
1
Requestor
Answer
5
"Your answer is posted"
Make a request for a question
Found the answer
Visualize
Q&A database
Q
A
2
Consortium Hub Site
Interprets free text question and assess the feasibility
Clinician
DB Analyst
Uploads the publishable results
Data Scientist
Performs data aggregation and quality check
Assigns a question
Upload the aggregate results
3
Lead Site
External resources
World Health Organization
CDC
PubMed
Collects supporting evidence and guidelines
Clinician
Standardized Concept
Searches concepts
OMOP
Retrieves OMOP results
Updates the ETL script
Evaluate data quality against EHR
DB Analyst
Uploads the template SQL trained at the lead site and its result
EHR
Consortium Hub
Suggests additional concepts
Uploads the local site SQL and query result
Downloads the template SQL trained at the lead site
Responding Site #1
Clinician
Searches concepts missed by the lead site
Standardized Concept
Runs the lead-site SQL
OMOP
DB Analyst
Evaluates data quality against EHR
Updates the ETL script
EHR
4
Uploads the local site SQL and query result
Downloads the template SQL trained at the lead site
Responding Site #2
Clinician
Searches concepts missed by the lead site
Standardized Concept
Runs the lead-site SQL
OMOP
DB Analyst
Evaluate data quality against EHR
Updates the ETL script
EHR

Figure 2.1: What happens behind the scenes: From questions to answers. The workflow of the question-answer system is shown in five steps. Step 1. Users access a public web-portal to post a new question if they cannot find a posted answer. Step 2. The questions get triaged to a Consortium Hub Clinical Informatician who determines their general interest and assigns the edited version of the question to a Lead Site. Step 3. At the Lead Site, the Clinical Informatician and the Database Analyst works together to create concept sets, design a query, and check local results. Step 4. The Responding Site runs the released SQL code and uploads its results to the Consortium Hub. During this step, the Clinical Informatician and the Responding Site Data Analyst adjust the concept set, inclusion logic, and database query code in structured query language (SQL) for local implementation, obtain and quality controls the site-level results, and submit results to the Consortium Hub. Step 5. The Consortium Hub aggregates the Site-level results, generates the visualizations and posts the answer on the web-portal.

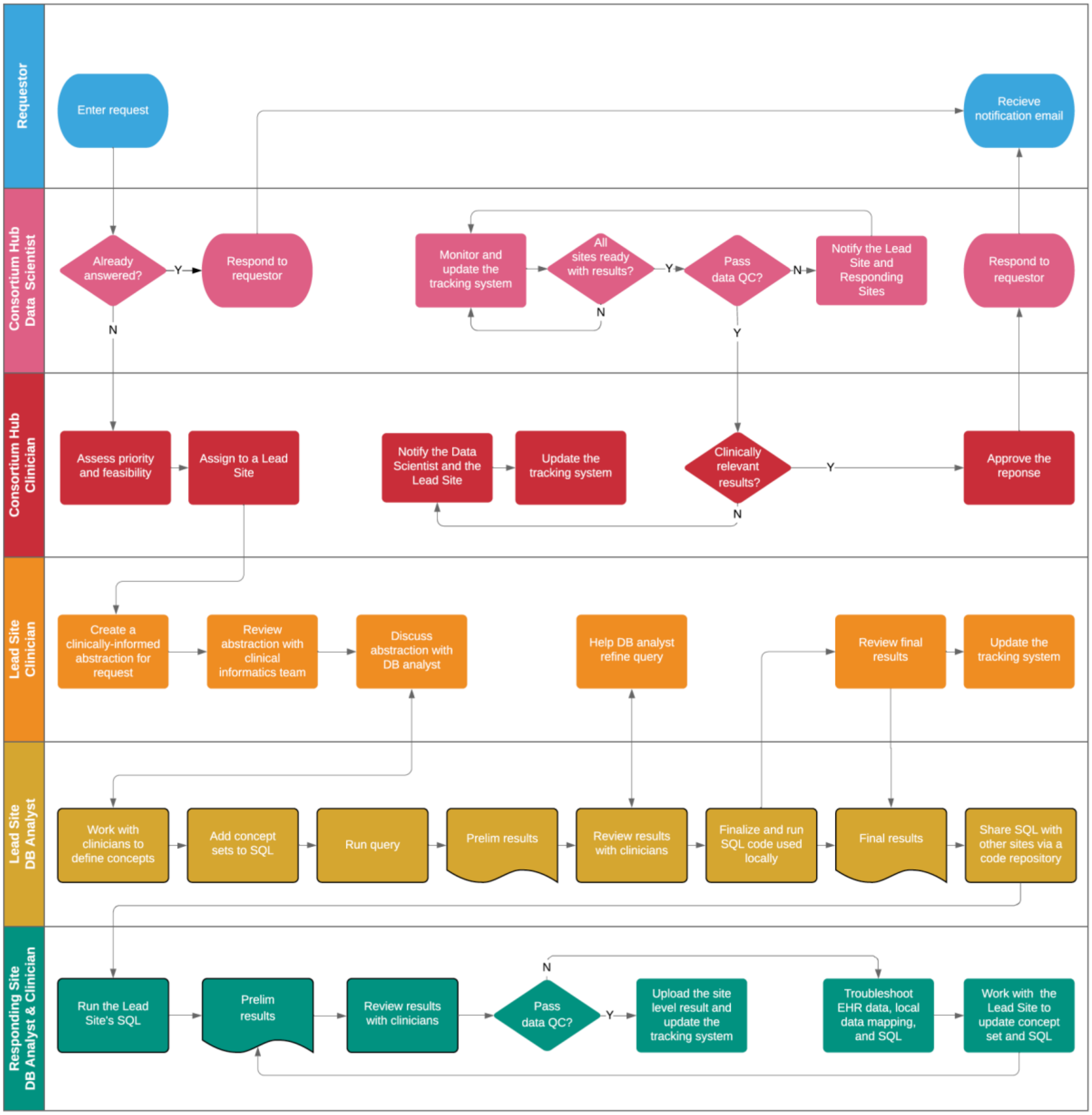
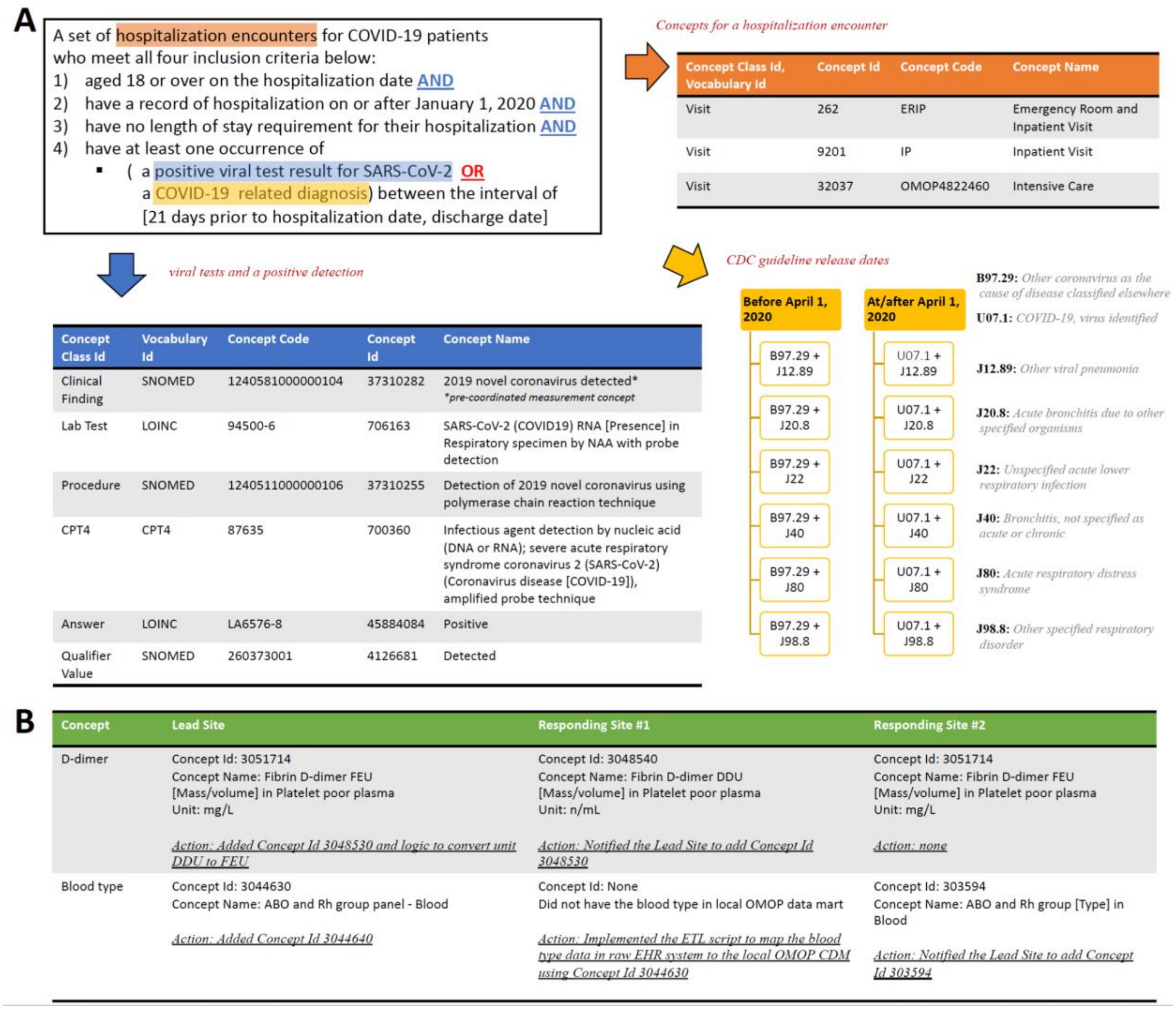

Requestor
Consortium Hub Data Scientist
Consortium Hub Clinician
Lead Site Clinician
Lead Site DB Analyst
Responding Site DB Analyst & Clinician
Enter request
Recieve notification email
Already answered?
Y
N
Respond to requestor
Monitor and update the tracking system
All sites ready with results?
Y
N
Pass data QC?
N
Y
Notify the Lead Site and Responding Sites
Respond to requestor
Assess priority and feasibility
Assign to a Lead Site
Notify the Data Scientist and the Lead Site
Update the tracking system
Clinically relevant results?
Y
N
Approve the reponse
Create a clinically-informed abstraction for request
Review abstraction with clinical informatics team
Discuss abstraction with DB analyst
Help DB analyst refine query
Review final results
Update the tracking system
Work with clinicians to define concepts
Add concept sets to SQL
Run query
Prelim results
Review results with clinicians
Finalize and run SQL code used locally
Final results
Share SQL with other sites via a code repository
Run the Lead Site's SQL
Prelim results
Review results with clinicians
Pass data QC?
N
Y
Upload the site level result and update the tracking system
Troubleshoot EHR data, local data mapping, and SQL
Work with the Lead Site to update concept set and SQL

Figure 2.2: Swimlane diagram. A Q&A process flow starts from a user entering a request and ends with the user receiving email notification about a response. At the Consortium Hub, the Data Scientist is responsible for aggregating Site-level results and for data quality checks. The Clinician at the Consortium Hub is responsible for feasibility assessment of the question, triaging to a Lead Site, and for the approval of the aggregate answer. At the Lead Site, the Clinician reviews the assigned question text and works with the database analyst to translate the question into SQL and to make sure the results are clinically relevant. The Database Analyst at the Lead Site writes the SQL code, runs it, verifies the results, and releases the code to the Consortium Hub. At the Responding site, the Database Analyst runs the Lead Site's SQL code, reviews the results together with local clinicians and uploads the Site-level results to the Consortium Hub, through an iterative process of ETL update, local data mapping and concept set development, all led by the Lead Site.

assigns it to one of the 12 institutions as a Lead Site. Throughout the whole process, the tracking system is used to report an issue to assignees, respond to the issue, update the code and results, and prompt to re-run the updated SQL. Next, another clinician at the Lead Site will work with the local database analyst to review and develop the concept set. This is an iterative process within the Lead Site, to develop concept sets, create SQL, generate results, and evaluate the results against the EHR records including chart reviews. The outputs of Lead Site level process are a template query (.sql format) and a template output (.csv format), which are uploaded to the shared code repository.

Once 11 Responding Sites get notification emails about the template query and format for the results, their database analysts will run the template SQL to get preliminary results, and review these against their EHR data with clinicians. This part of the process is where the Responding Site most frequently runs into errors, challenges, and requires troubleshooting. For example, when missing concepts like D-Dimer or blood type, illustrated in Figure 2.3, are discovered, the database analyst at the Responding Site creates an issue in the tracking system and resolves this with the database analyst and the clinician at the Leading Site. Since there are 11 Responding Sites, this means the Lead Site coordinates the concept set and SQL development through one-on-one sessions between the Lead Site and Responding Site.

Through this iterative process among 12 sites, the concept set and SQL keep being updated and keep improving their sensitivity and specificity to identify the right patients and hospitalization encounter records. This involves rewriting and updating existing Extract-Transform-Load (ETL) scripts to map source EHR data to target the common data model (CDM, which in our case is the Observational Medical Outcomes Partnership -- OMOP)[36]. The institutions with the same EHR system or database management system share common


A
Concepts for a hospitalization encounter
viral tests and a positive detection
CDC guideline release dates
B

A set of hospitalization encounters for COVID-19 patients who meet all four inclusion criteria below:

1) aged 18 or over on the hospitalization date AND
2) have a record of hospitalization on or after January 1, 2020 AND
3) have no length of stay requirement for their hospitalization AND
4) have at least one occurrence of
   - (a positive viral test result for SARS-CoV-2 OR a COVID-19 related diagnosis) between the interval of [21 days prior to hospitalization date, discharge date]

Concept Class Id, Vocabulary Id	Concept Id	Concept Code	Concept Name
Visit	262	ERIP	Emergency Room and Inpatient Visit
Visit	9201	IP	Inpatient Visit
Visit	32037	OMOP4822460	Intensive Care

Concept Class Id	Vocabulary Id	Concept Code	Concept Id	Concept Name
Clinical Finding	SNOMED	1240581000000104	37310282	2019 novel coronavirus detected* *pre-coordinated measurement concept*
Lab Test	LOINC	94500-6	706163	SARS-CoV-2 (COVID19) RNA [Presence] in Respiratory specimen by NAA with probe detection
Procedure	SNOMED	1240511000000106	37310255	Detection of 2019 novel coronavirus using polymerase chain reaction technique
CPT4	CPT4	87635	700360	Infectious agent detection by nucleic acid (DNA or RNA); severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Coronavirus disease [COVID-19]), amplified probe technique
Answer	LOINC	LA6576-8	45884084	Positive
Qualifier Value	SNOMED	260373001	4126681	Detected

Before April 1, 2020	At/after April 1, 2020
B97.29 + J12.89	U07.1 + J12.89
B97.29 + J20.8	U07.1 + J20.8
B97.29 + J22	U07.1 + J22
B97.29 + J40	U07.1 + J40
B97.29 + J80	U07.1 + J80
B97.29 + J98.8	U07.1 + J98.8

B97.29: *Other coronavirus as the cause of disease classified elsewhere*

U07.1: *COVID-19, virus identified*

J12.89: *Other viral pneumonia*

J20.8: *Acute bronchitis due to other specified organisms*

J22: *Unspecified acute lower respiratory infection*

J40: *Bronchitis, not specified as acute or chronic*

J80: *Acute respiratory distress syndrome*

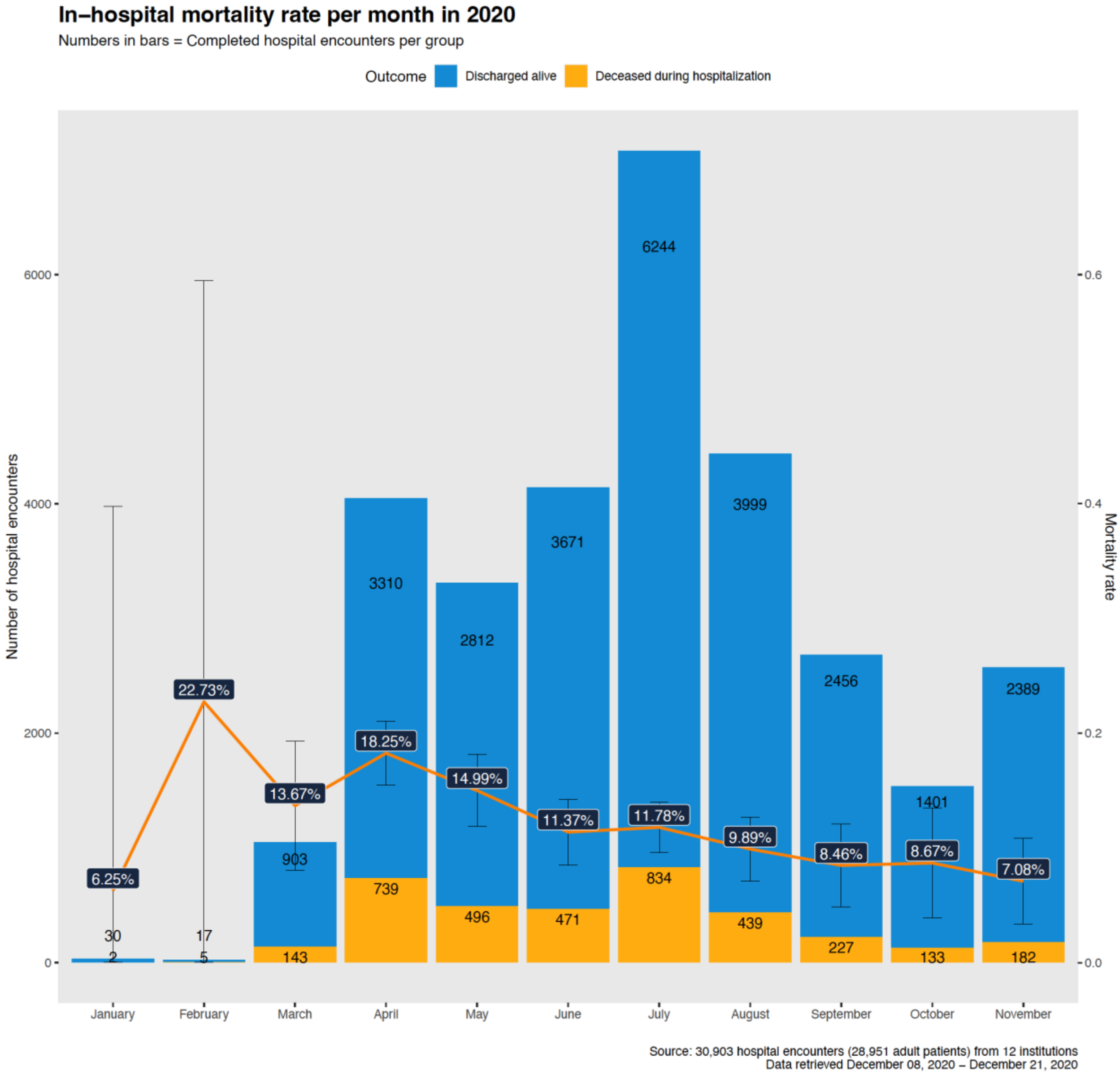
J98.8: *Other specified respiratory disorder*

Concept	Lead Site	Responding Site #1	Responding Site #2
D-dimer	Concept Id: 3051714 Concept Name: Fibrin D-dimer FEU [Mass/volume] in Platelet poor plasma Unit: mg/L *Action: Added Concept Id 3048530 and logic to convert unit DDU to FEU*	Concept Id: 3048540 Concept Name: Fibrin D-dimer DDU [Mass/volume] in Platelet poor plasma Unit: n/mL *Action: Notified the Lead Site to add Concept Id 3048530*	Concept Id: 3051714 Concept Name: Fibrin D-dimer FEU [Mass/volume] in Platelet poor plasma Unit: mg/L *Action: none*
Blood type	Concept Id: 3044630 Concept Name: ABO and Rh group panel - Blood *Action: Added Concept Id 3044640*	Concept Id: None Did not have the blood type in local OMOP data mart *Action: Implemented the ETL script to map the blood type data in raw EHR system to the local OMOP CDM using Concept Id 3044630*	Concept Id: 303594 Concept Name: ABO and Rh group [Type] in Blood *Action: Notified the Lead Site to add Concept Id 303594*

Figure 2.3: Cohort definition and Concept set development. Defining a cohort of patients that is frequently used to answer questions helps us reuse code. In this example, defining the cohort of patients hospitalized with COVID-19 involves use of SARS-CoV-2 test results or diagnosis codes (A). In (B), we illustrate how a laboratory test is defined differently at two sites, and how blood type has yet to be harmonized into OMOP at one site.

experience and knowledge to help each other develop ETL scripts together and evaluate the OMOP query results against EHRs.

When all Responding Sites have uploaded their site level results, the data scientist at Consortium Hub merges these results into a single file. A generic and extensible format for a site-level summary result is used to answer general epidemiology and clinical research questions (Figure 2.4). Then a data quality check is conducted. While use of a CDM in a large clinical data research network is a widely used approach to enable interoperable query development, a query formulated in one institution may not return accurate results in another due to variations in data integration and data quality differences. Several rounds of confirmations and checks with data analysts and clinical informaticians at each institution are often necessary to answer questions with confidence. There are many potential sources of errors and Table 2.2 displays selected examples of data quality checks. The check types are based on the PEDSnet framework[37] and revised to fit our project's specific needs. The data scientist resolves issues together with the Lead Site and the Responding Site. When the aggregate results pass the quality control (QC) test, the Consortium Hub Clinician conducts the final review to ensure its clinical relevance. During several rounds of code releases and responses among the Lead Site and the Responding Sites, database developers rewrote their ETL scripts to increase the accuracy of the query results. Finally, if the clinician approves the release of the result, the data scientist uploads the answer to the public web-site (https://covid19questions.org), notifies the requestor via an email, and this completes the workflow. Quality improvement related steps and data visualization are either semi-automated or manually conducted. ETL refresh, initial data quality check, and data aggregation are automated with scheduling scripts.



Source: 30,903 hospital encounters (28,951 adult patients) from 12 institutions
Data retrieved December 08, 2020 – December 21, 2020

Figure 2.4: An example of a COVID-19 Question: Monthly mortality. The in-hospital mortality rate per month (red line) is shown as a percentage, with its 95% confidence interval between January and November in 2020. The observed counts for the deceased during hospitalization (orange) and the discharged alive (blue) are shown in bar plots. The unit of analysis is the hospital encounter.

Table 2.2: Data quality checks and issues. Different data quality check types are enumerated together with real issues identified with this COVID-19 project.

Check Type	**Example of data quality issue**
Datetime reversal	A condition/observation was recorded after discharge date
Extreme outlier	The hospital length of stay was greater than 80 days. The median length of stay ranged between 11 and 15 days in China and US studies
Gaps in data transformation	Discharge disposition and ICU departments were not transformed to OMOP
Loss of granularity during mapping	Invasive and non-invasive mechanical ventilation mapped to the same concept
Impossible events	Multiple death events occurred in different time points from multiple hospital encounters
Non-compliance to the output format	Header was missing in the predefined output .csv format, missing columns, shifted columns, and duplicate rows
Unexpected proportion	The percentage of current smokers was 65% at a certain site. The national percentage of smoking was 15.6% among male adults in 2018 US CDC data
Unexpected zero count	The number of patients who were taking any anti-hypertensives was zero
Unmatched group sum	The total sums of patient count in age groups and race groups were different even when all cell counts were greater than 10
Version mismatch	The version of the template query was revised after the query result was uploaded

2.3.4 Federated regression

In addition to count queries, we also applied Grid Binary LOgistic REgression (GLORE)[15] to compute the effect of the exposure variable on the outcome, adjusted for confounders, without sharing patient-level data, as this would increase the risk for a privacy breach. We rewrote the Newton-Raphson (NR) method to find the maximizer of the likelihood function of the parameters in logistic regression for horizontally partitioned datasets. Since the first and the second derivatives of the log likelihood functions are separable (i.e., they can be partially calculated at each site), in each NR iteration, each client institute calculated a (p+1) dimensional vector of parameters, where p is the number of features in the model such as age, sex, and race and a (p+1) by (p+1) variance-covariance matrix, and sent JSON files containing these two objects to the Consortium Hub. At each iteration, the Consortium Hub automatically updated the global coefficient vector and the variance-covariance matrix and sent them back to clients.

2.4 RESULTS

2.4.1 Answered questions

Between 12/11/2020 and 8/31/2020, our consortium had 928,255 tested patients for SARS-CoV-2, 59,074 diagnosed with COVID-19, with 19,022 hospitalized and 2,591 deceased. Our public Questions and Answers portal (https://covid19questions.org) provides answers to research questions using several univariate or multivariate analyses, including potential associations between mortality and comorbidities; pre-hospitalization use of medications; laboratory values and hospital events. For each question, we report on the number of participating institutions and the time period within which local queries were run. Figure 2.4 - 2. 6 illustrate the answers.

Example 1. “Many adult COVID-19 patients who were hospitalized did not get admitted to the ICU and were discharged alive. How many returned to the hospital within a week, either to the Emergency Room or for another hospital stay?” This question is both important from the standpoint of understanding the natural course of disease and planning for needed resources. Although efforts are underway to understand post-discharge outcomes in COVID-19 infected patients, to date they have been limited to case series[38], modest sample sizes[39], or single-center or geographically concentrated health systems[40]. These extant studies may also be hampered by fixed inclusion/exclusion criteria[41].

Example 2. “Among adults hospitalized with COVID-19, how does the in-hospital mortality rate compare per subgroup (age, ethnicity, sex and race)?” The answers from univariate analyses indicate that age, ethnicity and sex are significant. A distributed logistic regression (Figure 2.6) shows, among these, that only age is significant. There is great interest and growing peer-reviewed literature on risk factors for COVID-19 mortality: the agility of our approach allows us to quickly re-run queries and rebuild models as new predictors become relevant and the understanding of the disease evolves[40,42,43].

2.4.2 Cohort and concept set

As questions frequently refer to the same subsets of patients, we developed electronic cohort definitions that facilitate our answers. We followed the United States Centers for Disease

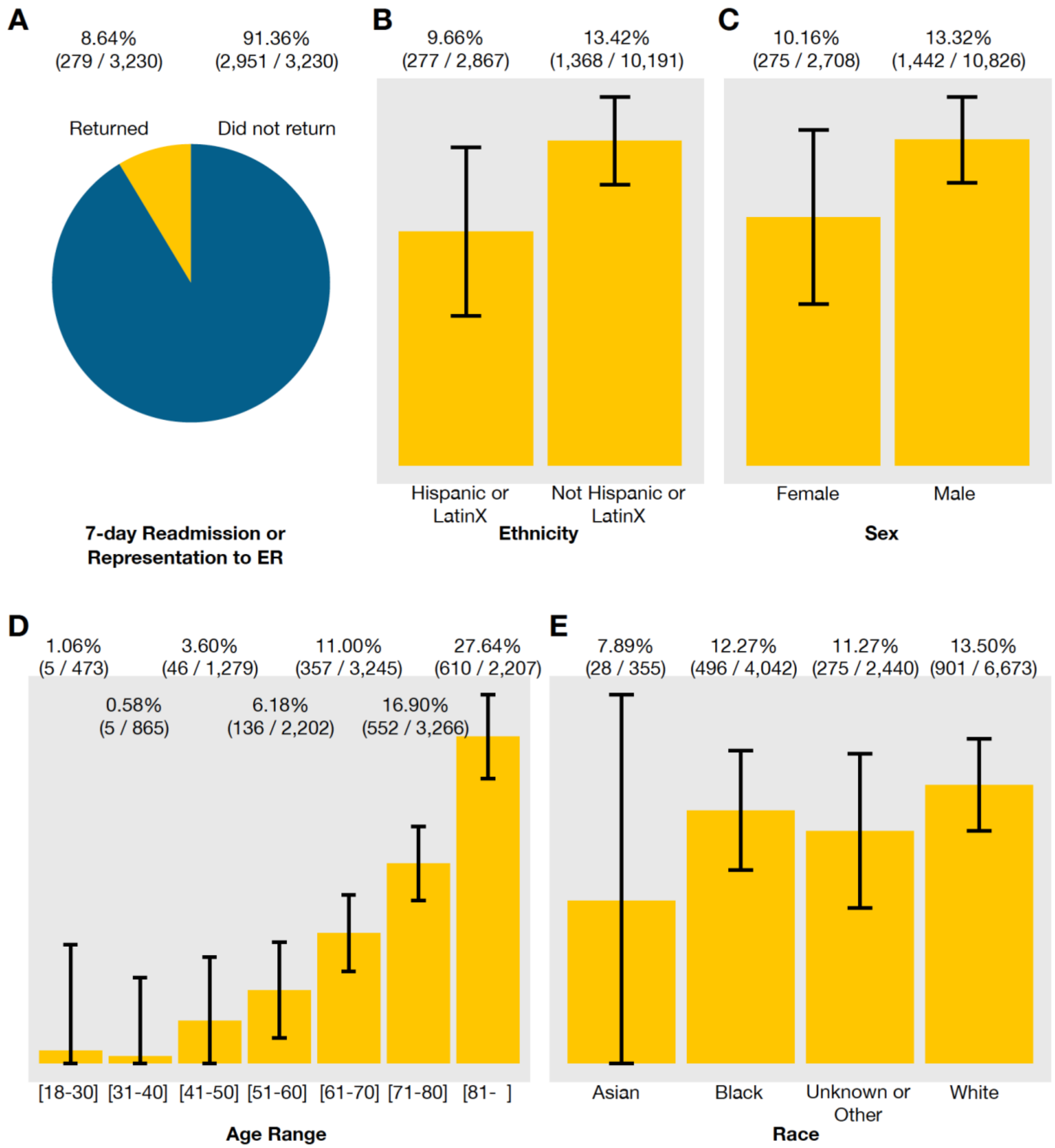
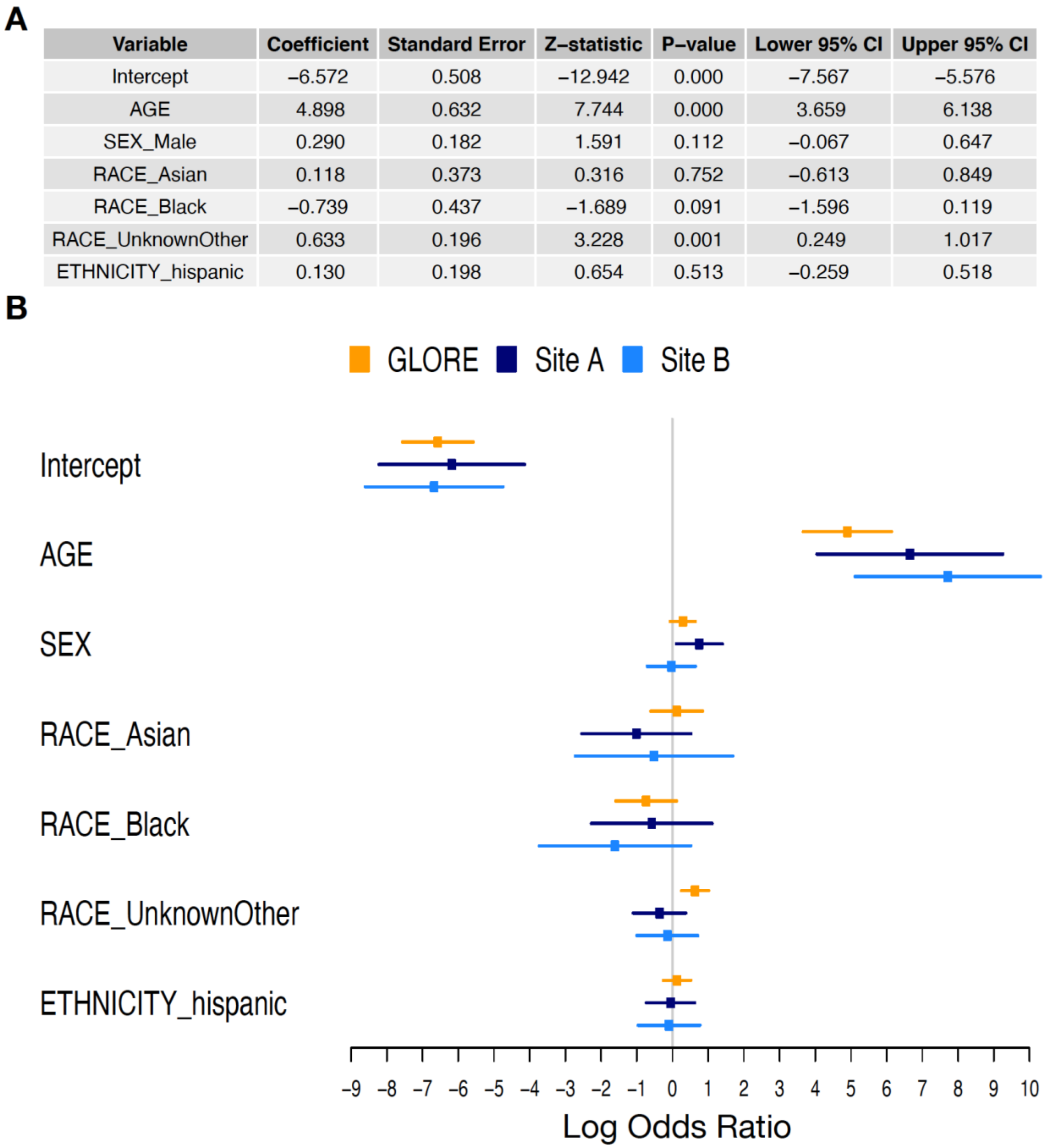


Figure 2.5: Examples of two COVID-19 Questions and Answers: Return to hospital and mortality. (A) 8.6% of hospitalizations without an ICU admission resulted in the patient presenting to the Emergency Room or a hospital readmission within seven days (data from ten health systems). (B-E) Unadjusted mortality rates from aggregated results are shown with 95% confidence intervals (data from ten health systems). Univariate analyses indicate that lower age, Hispanic ethnicity, and female sex (as recorded in the EHR) are associated with lower mortality for adult hospitalized COVID-19 patients.



A

Variable	Coefficient	Standard Error	Z–statistic	P–value	Lower 95% CI	Upper 95% CI
Intercept	–6.572	0.508	–12.942	0.000	–7.567	–5.576
AGE	4.898	0.632	7.744	0.000	3.659	6.138
SEX_Male	0.290	0.182	1.591	0.112	–0.067	0.647
RACE_Asian	0.118	0.373	0.316	0.752	–0.613	0.849
RACE_Black	–0.739	0.437	–1.689	0.091	–1.596	0.119
RACE_UnknownOther	0.633	0.196	3.228	0.001	0.249	1.017
ETHNICITY_hispanic	0.130	0.198	0.654	0.513	–0.259	0.518

Figure 2.6: Regression Results. (A) Adjusted effects from the Grid Binary LOgistic REgression (GLORE) (11) federated logistic regression model (3,146 patients from eight health systems). The baselines were SEX=female, RACE=white, ETHNICITY=non-Hispanic. AGE (in years) was divided by 100. After adjustment via distributed logistic regression, AGE remains significant. (B) Results from local logistic regression performed at two sites are also shown for comparison with GLORE results.

Control and Prevention (CDC) guideline[44], and the National COVID-19 Cohort Collaborative[45], and Observational Health Data Sciences and Informatics (OHDSI) approaches[46] to develop a cohort of hospitalization encounters for COVID-19 as a base for all inpatient questions. Through an iterative process among multiple sites, we developed a canonical SQL whose result matches with that of the ground truth cohort definition. The intersection of the R2D2 canonical SQL, the private reference (i.e., EHR- or Registry based) and the universal reference (i.e., a positive polymerase chain reaction test for SARS-COV-2) was maximized for existing and new sites.

Figure 2.3A displays the electronic phenotyping of adults hospitalized with COVID-19 derived by the canonical SQL and stored procedure SQL scripts. Hospitalization encounters were identified by using the following concepts stored in the OMOP <VISIT_OCCURRENCE> table: Emergency Room and Inpatient Visit (Concept Id 262), Inpatient Visit (Concept Id 92021) or Intensive Care (Concept Id 32037). To enter the COVID-19 hospitalization cohort, all four inclusion criteria needed to be met:

1) a minimum age of 18 years at the date of hospitalization,

2-3) a hospitalization without a length of stay requirement on or after January 1, 2020 and

4) at least one occurrence of

a) a positive viral test for SARS-CoV-2, or

b) a COVID-19 related diagnosis between the interval of 21 days prior to hospitalization and hospital encounter discharge.

The following concepts of the OMOP <MEASUREMENT> table for the definition of a positive viral test for SARS-COV-2 were used:

1) the occurrence of the pre-coordinated measurement Concept (Concept Name:

2019 novel coronavirus detected, Concept Id: 37310282), or

2) the occurrence of at least one concept for a SARS-CoV-2 viral test (e.g. Concept Name: SARS-CoV-2 (COVID19) RNA [Presence] in Respiratory specimen by NAA with probe detection, Concept Id: 706163) and at least one value_as_concept_id for a positive result (e.g. Concept Name: Positive, Concept Id: 45884084).

For identification of COVID-19 related diagnoses, we included the following ICD-10-CM Codes: Other coronavirus as the cause of diseases classified elsewhere (B97.29), COVID-19, virus identified (U07.1), Pneumonia (J12.89), Acute Bronchitis (J20.8), Lower Respiratory Infection (J22, J98.8) and Acute Respiratory Distress Syndrome (J80). Following two ICD-10-CM Official Coding and Reporting Guidelines released by CDC before and at/after April 1, 2020, we used diagnosis code aggregations to define a COVID-19 related diagnosis. An illness due to COVID-19 was specified if one of the ICD-10-CM codes (J12.89, J20.8, J22, J98.8, J80) was recorded in combination with either B97.29 (before April 1, 2020), or in combination with U07.1 (on/after April 1, 2020). These joint diagnosis codes needed to occur during the same hospitalization encounter, with a look-back period of 21 days prior to hospitalization. We applied the same logic for mapped SNOMED Concepts (261326, 260139, 4307774, 256451, 4195694, 320136, 4100065, 37311061). More ICD codes are detailed in Table 2.3. Pre-coordinated diagnoses codes (SNOMED, OMOP Extension) are shown in the Supplementary Tables A.1-A.3. Refinement of phenotypes was guided by chart review.

Table 2.3: Concept Relationships between ICD10CM and SNOMED Concepts. ICD10CM Concepts and their mapped SNOMED Concepts from the CONDITION_OCCURRENCE table. In OMOP CDM, ICD10CM Concepts are non-Standard Concepts. Therefore, ICD10CM Concepts are mapped to SNOMED-based Standard Concepts. These relationships are stored in the OMOP CDM CONCEPT_RELATIONSHIP table. In this case each ICD10CM Concept got the relationship_id = ‘Maps to’, which directs to one SNOMED Concept.

Concept Code 1 (ICD10CM)	**Concept Name 1**	**Concept ID 1**	**Relationship ID**	**Concept ID 2**	**Concept Name 2**	**Concept Code 2 (SNOMED)**
J12.89	Other viral pneumonia	45572161	'Maps to'	261326	Viral pneumonia	75570004
J20.8	Acute bronchitis due to other specified organisms	35207965	'Maps to'	260139	Acute bronchitis	10509002
J22	Unspecified acute lower respiratory infection	35207970	'Maps to'	4307774	Acute lower respiratory tract infection	195742007
J40	Bronchitis, not specified as acute or chronic	35208013	'Maps to'	256451	Bronchitis	32398004
J80	Acute respiratory distress syndrome	35208069	'Maps to'	4195694	Acute respiratory distress syndrome	67782005
J98.8	Other specified respiratory disorders	35208108	'Maps to'	320136	Disorder of respiratory system	50043002
B97.29	Other coronavirus as the cause of diseases classified elsewhere	45600471	'Maps to'	4100065	Disease due to Coronaviridae	27619001
U07.1	Emergency use of U07.1 | Disease caused by severe acute respiratory syndrome coronavirus 2	702953	'Maps to'	37311061	Disease caused by 2019-nCoV	840539006

Use cases of concept set are shown in Figure 2.3B. As the Responding Sites' OMOP databases are not accessible to the Lead Site, a query developed at the Lead Site might miss a concept used in other sites. In such a case, the database analyst at the Responding Site notifies the Lead Site by creating a GitHub issue, with zero or unexpectedly low count or proportion in the results generated by the initial template query authored by the Lead Site. For example, in Figure 2.3B, during the Concept Set development for the quantitative laboratory measurement D-dimer, the responding site notified the Lead site about using another concept for D-dimer (Concept ID: 3048540 instead of Concept ID: 3051714), returning values with a different measurement unit than the ones of the Lead Site (n/L instead of mg/L). Therefore, the Lead site had to add the missing concept to Concept Set and implemented logic to cover a measurement unit transformation. In the case of the Concept Set development for blood type, a responding site was missing concepts for blood type in its local OMOP CDM database. An ETL script was implemented to map EHR data to OMOP CDM. Sources of discrepancy were diverse; examples included unit differences in measurement values, differently mapped concepts, and non-compliance to the coding guideline. All SQL codes and concept sets for answered questions are publicly available from the GitHub repository: https://github.com/DBMI/R2D2-Public. The public repo is updated whenever a new question and its answer get posted on the public web site. The similarities and the differences of our approach to other consortia are detailed in the Supplementary Text A.1.

Supplementary Figure A.1 shows the screen shot of the real example JSON file used during the GLORE run to answer the in-hospital mortality question. No patient level information was shared or transferred between institutions. All clients repeatedly sent the updated JSON file to the Consortium Hub, until the estimates stabilized or reached a predefined number of iterations. To enhance the security, the Consortium Hub server allowed (i.e., “whitelisted”) only the pre-registered IP addresses of client machines and opened the port only during the scheduled time window.

Several other questions and answers are shown in the portal. A novel governance structure (Figures 2.1-2.2) allows us to distribute the workload across various teams without relying on a traditional coordinating center, instead including a Consortium Hub. This approach keeps patient data in-house, simplifies data use agreements, avoids delegation of control of patient data to another institution, and allows any institution to benchmark its results to those produced by the consortium, since all questions and respective final, aggregated answers, database query code, concept definitions and analytics code are made public. It complies with HIPAA[47], the Common Rule[48], the GDPR[49], and the California Consumer Privacy Act[50] with regards to handling of patient data. Code sharing and public answers promote transparency and reproducibility without disclosing patient information or institutional information.

2.5 DISCUSSION

2.5.1 Privacy-aware distributed research network

Our approach is practical and generalizable: The network can be repurposed to any other disease of interest, as it is not based exclusively on data elements deemed relevant for COVID-19. Because privacy protection is at the core of our network, a wide range of institutions can participate. We provide a rapidly deployable and reproducible alternative or complement to centralized registries of EHR data that allows healthcare institutions to stay in control of their data.

2.5.2 Limitations

This study has advantages but also some limitations. The advantages are that we can, in relatively short time, publicly post answers, using data from a spectrum of institutions with different levels of information technology baselines and expertise in standardized data models and vocabularies, institutional policies, state and federal regulations. Because we keep data locally and only consult data elements that are necessary to answer specific questions, this approach has a very low risk of privacy breach. However, for this reason, our approach does not provide answers in real-time. We made this practical decision to quickly collect aggregate counts and statistics near real-time within existing institutional policies and OMOP implementation to meet the clinical need of a rapidly spreading pandemic while preserving patient privacy. A real-time query with a fully automated process would be ideal, but this necessitates a long process of inter-institution agreement, amendments to the institutional policies, and a complete harmonization of EHR data across all sites. The use of OMOP CDM data is dependent on recurring ETL processes on each site, which presents a challenge to presenting real-time data. Additionally, as opposed to registries that typically focus on a single disease or condition, we have comparator data from other patients. Institutional privacy is also preserved because all public answers combine the aggregate data from at least three Responding Sites. Making concept definitions, query code, and results publicly available enhances reproducibility. A major advantage is that existing registries of consortia can serve as additional sites to help answer certain questions. The limitations are inherent from considering all sites equal when formulating a final answer. Regional or institutional practice variations are not represented in the answers. Additionally, the distributed nature of the consortium adds a requirement for educating some system leaders on distributed analytics. A specific limitation of our current consortium is the preponderance of institutions based in California: 67%, or 17.5% of COVID-19 patients. This was a convenience sample of organizations that had a history of collaboration. We are currently adding two new large health systems. One system is in the Northeast and another in the Southeast US. To display changes over time, and to help users compare our results to public results, new SQL code has been developed. Additionally, the increasing use of automated stored procedures will help us to reduce the manual process.

2.5.3 Implication and future work

We believe that our ‘Covid-19 Clinical Data Consult’ is a tool to achieve rapid and robust responses to COVID-19 questions, submitted by the public or by researchers. We can achieve those efforts by combining a transparent, privacy-preserving code sharing workflow with the use of harmonized distributed data. A vision for the future in which there is convergence of data services would include interoperability with other efforts, including federated multivariate analyses across different consortiums (e.g., R2D2, 4CE, and N3C).

2.6 CONCLUSION

Instead of centralizing data at the Consortium Hub, we focus on interpreting and clarifying the research questions in order to determine the data elements required. Our teams analyze these data elements to generate aggregate statistics at the multiple institutions, documenting the specific version of structured query language (SQL) code executed at a specific time point to generate their answers. In addition to basic counts and proportions, to adjust for confounders, we use distributed multivariate analyses to estimate risk-adjusted odds ratios. This is done in a synchronized fashion for iterative federated algorithms, such as one previously reported for building a logistic regression model. We have shown previously that a model obtained this way is identical to one built using data that are centralized in a single location. We made SQL codes, cohort definition and concept sets publicly available at https://github.com/DBMI/R2D2-Public. We invite other institutions, consortia and registries worldwide to join us at https://covid19questions.org

2.7 ACKNOWLEDGEMENTS

This work was supported by the Gordon and Betty Moore Foundation #9639. The distributed analytics algorithm was funded by NIH-R01GM118609. Trainees were funded by NIH-T15LM011271. LN was funded by DIFUTURE (BMBF grant 01ZZ1804C).

Chapter 2, in full, in a reprint of the material as it appears in “Privacy-protecting, reliable response data discovery using COVID-19 patient observations.” Jihoon Kim, Larissa Neumann, Paulina Paul, Michele E Day, Michael Aratow, Douglas S Bell, Jason N Doctor, Ludwig C Hinske, Xiaoqian Jiang, Katherine K Kim, Michael E Matheny, Daniella Meeker, Mark J Pletcher, Lisa M Schilling, Spencer SooHoo, Hua Xu, Kai Zheng, Lucila Ohno-Machado, R2D2 Consortium. *J Am Med Inform Assoc.* 2021 Jul 30;28(8):1765-1776. doi: 10.1093/jamia/ocab054. The dissertation author was the primary investigator and the first author of this paper.

Chapter 3 VERTIcal Grid lOgistic regression with Confidence Intervals (VERTIGO-CI)

3.1 ABSTRACT

Federated learning of data from multiple participating parties is getting more attention and has many healthcare applications. We have previously developed VERTIGO, a distributed logistic regression model for vertically partitioned data. The model takes advantage of the linear separation property of kernel matrices of a dual space model to harmonize information in a privacy-preserving manner. However, this method does not handle the variance estimation and only provides point estimates: it cannot report test statistics and associated P-values. In this work, we extended VERTIGO by introducing a novel ring-structure protocol to pass on intermediary statistics among clients and successfully reconstructed the covariance matrix in the dual space. This extension, VERTIGO-CI, is a complete protocol to construct a logistic regression model from vertically partitioned datasets as if it is trained on combined data in a centralized setting. We evaluated our results on synthetic and real data, showing the equivalent accuracy and tolerable performance overhead compared to the centralized version. This novel extension can be applied to other types of generalized linear models that have dual objectives.

3.2 INTRODUCTION

With the adoption of electronic health records (EHRs) in the US and advances in health information technology (HIT), a vast amount of health data is being generated rapidly. These data come from different sources (e.g., hospitals, cohort studies, disease registries, health insurance providers, and DNA/RNA sequencers). The conventional solution is first to gather datasets from multiple sources at a central site and then conduct analyses to answer a clinical/research question. However, such a centralized approach is not always viable because of potential harm to patient privacy, regulations, and policies, mistrust among participants, etc. If analyses could be conducted with data that are maintained in different places, this would greatly mitigate these factors.

A dataset can be partitioned in two ways: horizontally or vertically. Datasets are horizontally partitioned if all participating sites have the same set of features from different individuals. For example, a risk score model for coronary heart disease collects demographic, cholesterol, blood pressure, diabetes and smoking status from different institutions to develop or validate the model. Horizontally partitioned datasets[15] occur in multi-site clinical trials, clinical data research networks (CDRNs), registries, and risk prediction models with non-overlapping development and validation sites[51]. On the other hand, a dataset can be partitioned vertically in two or more different features from the same individual and the subset of features of it can be stored in different sites. For example, a Strong Heart Study, the largest epidemiology study of cardiovascular disease in American Indians, stores the genotype data in one institution and the phenotype data in another institution, allowing access only to approved researcher[52]. Booming direct-to-consumer (DTC) genetic testing[53] companies keep the individual's genetic data in their storage server, but the clinical information of the patients are stowed in the patient registry or EHR system. However, the association test of these genetic data can be performed only when they are linked with phenotypes, typically EHRs, thus physically separated from the genotype data. While healthcare claims data are saved in health insurance companies, detailed patient data are located in hospital EHR systems. Current protocol of data access involves a lengthy process of request, review, approval, and monitoring limiting the opportunity of clinical research. Even when datasets can be centralized, transferring these to a central site is not trivial with genomic, imaging, and patient-generated health data from mobile phones and devices because these datasets can be very large in size. For this reason, commercial cloud computing platforms provide commonly used public genomics datasets such as 1000 genome[54] or The Cancer Genomics Atlas (TCGA)[55] datasets so that users can bypass the redundant transfer of such large datasets, which are costly and choke up the network.

Many algorithms were developed for federated analytics for both horizontal and vertically partitioned datasets[15,56,57]. For vertically partitioned datasets, secure matrix product algorithms are widely adopted[58–60]. None of these methods used dual optimization to perform interval estimation for vertically partitioned datasets. Dual optimization has been used for support vector machine classifiers[61], but the logistic regression model is the preferred method in genetics. VERTIcal Grid lOgistic regression (VERTIGO) is a distributed algorithm to build a logistic regression model on vertically partitioned datasets using dual optimization[62]. However, VERTIGO provides the only point-estimates, so no confidence interval is provided, and the statistical significance of the estimate in the form of a P-value is not provided. This study is an extension of our previous work, namely VERTIGO, to add standard errors to derive the interval estimates and express the parameter's statistical significance. This paper introduces a novel way of generating and transmitting confidence intervals along with coefficients. We describe our proposed algorithm, provide the mathematical proof, and demonstrate the algorithm performance on both simulated and real datasets.

3.3 METHODS

3.3.1 Synthetic data generation

Synthetic data for 2000 samples and 20 features were created with some distributional assumptions, as follows:

1. Generate two independent matrices, X_1 and X_2, of the dimension 2000

examples × 20 features, using a Uniform[0, 1] distribution

2. Derive a linear combination, $X = 1 + 2X_1 + 3X_2$, of the above two matrices
3. Generate random ground truth parameter vector β with size (20 × 1) using a Uniform[0,1] distribution
4. Apply the sigmoid function to calculate the probabilities for a binary outcome,

$$p = \frac{1}{1 + e^{-X\beta}}$$

5. Generate the binary outcome y with probability p in step 4 using a Bernoulli distribution

Then the generated samples were assigned to mutually exclusive partitions, where the number of partitions, k, was varied from 2 to 4 and each partition represented a client site.

3.3.2 Real data: BURN1000

A synthetic data about a burn study was obtained from R package aplore3. It is included in a companion data archive for the textbook by Hosmer and Lemeeshow[63]. The burn data had eight variables and 1000 samples. The outcome was death, a binary variable of alive or dead. The seven features were age, gender, race, burn facility, total burn surface area, burn involved in inhalation injury, and flame involved in a burn injury.

3.3.3 Real data: PennCath

A real data was downloaded from the Foulkes lab (http://www.stat-gen.org/), and this is the PennCATH cohort data, which arises from a Genome-wide association (GWA) study of coronary artery disease (CAD) and cardiovascular risk factors based at the University of Pennsylvania Medical Center[64]. First of all, the quality control process is performed on the genotype data to check sex discrepancy, minor allele frequency, Hardy-Weinberg equilibrium, and relatedness. In the end, the sample size of the data shrinks from 3850 to 1280. Then the whole dataset was split into two clients, phenotype and genotype. The binary outcome is the disease condition, yes or no. The phenotype data includes age, sex, and additional covariates for each individual, while the genotype data contains 10 principal components for SNPs. Those 10 components, along with phenotype data and one genotype data in 1,000 SNPs, will be put into the VERTIGO-CI algorithm. To evaluate the computation time, we designed studies for 3 batches of trials using 10, 100, and 1000 SNPs.

3.3.4 Model

The logistic model is defined as

$$P(y = \pm 1|X, \beta) = \frac{1}{1 + \exp -y\beta^\top X}$$

where y is a binary outcome, X is the design matrix of sample-by-feature, and β is the model parameter. The goal is to find the estimate for β given observed data X and y. The best estimate for β is the maximizer of the log-likelihood function

$$argmax_\beta l(\beta) = argmax_\beta \log \pi(yX\beta) - \frac{\lambda}{2}\beta^T\beta$$

where π is the sigmoid function and $\frac{\lambda}{2}\beta^T\beta$ is the regularization penalty term to avoid overfitting. Since the above equation cannot be used for a vertically partitioned dataset in its current form, VERTIGO algorithm adopts reparameterization using the dual form of the original optimization equation

$$argmin_\alpha J(\alpha) = argmin_\alpha \frac{1}{2\lambda}\|y\alpha X\|_2^2 - L(\alpha)$$

$$L(\alpha) = -\alpha^T \log(\alpha) - (1-\alpha)^T \log(1-\alpha)$$

This dual form of the maximum likelihood function is generating the same results by optimizing dual parameters with respect to samples rather than features, keeping the information intact[65]. The next step is to update the parameters α using Newton's method[66] by iterating

$$\alpha^{(s+1)} = \alpha^{(s)} - \frac{J'(\alpha^s)}{H(\alpha^s)}$$

where $J'(\alpha)$ and $H(\alpha)$ are, respectively, the first and second derivative of dual object function $J(\alpha)$, defined as

$$J'(\alpha) = \lambda^{-1} y \alpha^T y X X^T + \log \frac{\alpha}{1-\alpha}$$

$$H = \lambda^{-1} \text{diag}(y) X X^T \text{diag}(y) + CI$$

Note that H, Hessian matrix, has been changed in this situation for calculation convenience, and such changes will not harm the convergence as it only changes the step size[67]. C is a positive constant that enables the Hessian matrix to be full rank so its inverse matrix exists. When dual parameters α converges, the desired primal form parameter vector β can be obtained by its relationship to α,

$$\beta = \lambda^{-1} \alpha y^T X$$

This study's novel contribution is producing the standard errors of the point estimates that can be used to report statistical significance by P-Values or confidence intervals. The standard error of the coefficient can be represented as

$$(X'VX)^{-1/2}$$

$$V = \text{diag} \frac{e^{X\hat{\beta}}}{1+e^{X\hat{\beta}}} 1 - \frac{e^{X\hat{\beta}}}{1+e^{X\hat{\beta}}}$$

with the setting of vertically partitioned assumption on X, we have $X = (X_1, X_2, \cdots, X_k)$.

Since V is not separable for its own, the intermediate-term $e^{X_i\hat{\beta}_i}$ can be used to calculate V, by sending each term to clients, so that the final matrix V can be computed. Additionally, V should not be known by the center server because the information of X can be reverse-engineered using previously seen data. So, at this step, the matrix V must be kept secret from the server.

The first connected client to the closed network acts as a lead-client and collects the first intermediate matrix, $e^{X_i\hat{\beta}_i}$, from the other clients. This lead-client generates V and sends it back to all clients. Finally, each client sends the second intermediate matrix, $X_iV^{1/2}$, back to the server. Since the matrix V is hidden to the server, the individual-level data are protected. Since $X'VX$ is separable as follows

$$X'VX = \begin{pmatrix} X_1'VX_1 & X_1'VX_2 & \cdots & X_1'VX_k \\ X_2'VX_1 & X_2'VX_2 & \cdots & X_2'VX_k \\ \vdots & \vdots & \ddots & \vdots \\ X_k'VX_1 & X_k'VX_2 & \cdots & X_k'VX_k \end{pmatrix}$$

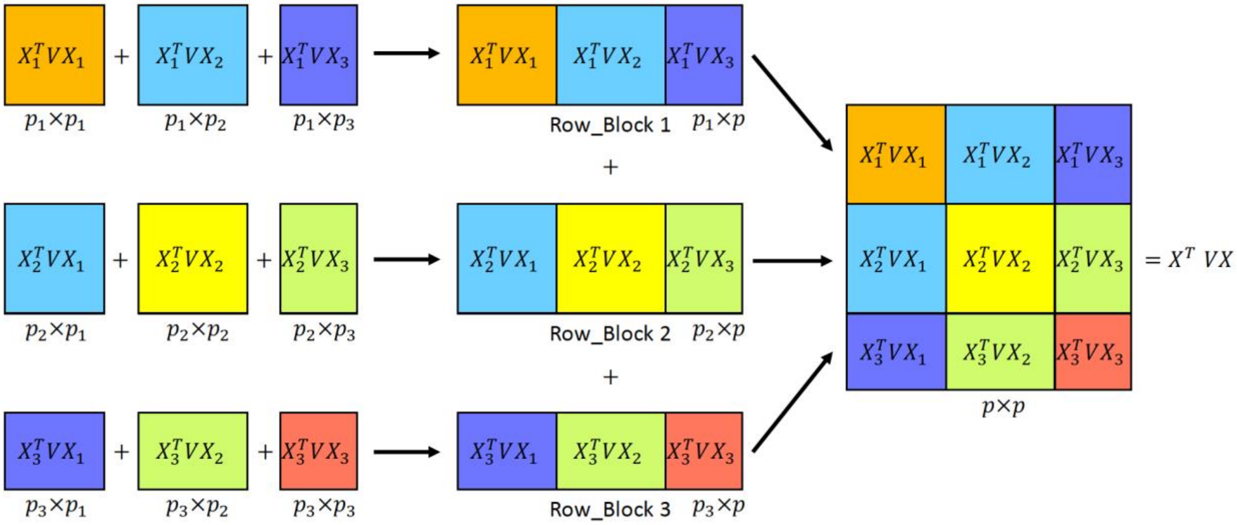
where k is the number of clients, directly interpretable statistics such as the Z score can be calculated as $Z = \beta/\mathrm{diag}((X'VX)^{-1/2})$, and confidence intervals and P-values can be derived. The pseudo-code is presented in Algorithm 3.1. Since $X_i'VX_j$ has a different size, the problem turns into a ‘puzzle solving’ to update the partial block matrices. Thus, putting those matrices in the right places is important. See the matrices-puzzle-solving pseudo-code in Algorithm 3.2. As an example, when $k = 3$, the algorithm will be executed as shown in Figure

Algorithm 3.1: VERTIGO-CI

```
Input: Data matrix of each client $X_i$ ($n$ samples by $p_i$ features), shared outcome $Y$, and penalty parameter λ ($i = 1, \cdots, k$)
Output: Coefficient $\beta^*$, their standard errors and confidence intervals
Procedure:
  1. Each client $i$: sends gram matrix $K_i = X_i X_i^T$ to the server
  2. Server: combines the global gram matrices  to have $K = \sum_i K_i$ 's , initializes dual
     parameters $\alpha^{(0)} = \mathbf{0}$ and broadcasts these parameters back to the clients
  3. Initialize step $s = 0$
  4. Repeat while changes in $\alpha <$ predetermined threshold:
      a. The client  $i$: Computes $E_i^{(s)} = \lambda^{-1} y \alpha^{(s)T} y K_i$ and send the intermediate
         matrix to the server
      b. Server: Combines and calculates
         $E^{(s)} = \sum_i E_i^{(s)}$ and $J'(\alpha^{(s)}) = E^{(s)} + \log \frac{\alpha^{(s)}}{1 - \alpha^{(s)}}$
      c. Server: Computes Hessian matrix $H^{(s)}(\alpha^{(s)}) = \lambda^{-1}\mathrm{diag}(y) K \mathrm{diag}(y) + CI$
         and calculates the inverse matrix$H^{(s)^{-1}}$
      d. Server: Updates the dual parameters using Newton's method
         $\alpha^{(s+1)} = \alpha^{(s)} - J'(\alpha^{(s)}) H^{(s)^{-1}}$,
         then sends the updated $\alpha^{(s+1)}$ back to clients
      e. $s = s + 1$
  5. Set the final alpha as $\alpha^*$, the optimal value of $\alpha$
  6. Each client $i$: Calculates the global optimization $\beta_i^* = \lambda^{-1} \alpha^* y^T X_i$ and sends it to the
     server
  7. Server: Combines the global optimum estimates from each client
     $\beta^* = (\beta_1^{*T}, \cdots, \beta_k^{*T})^T$
  8. Client-to-Client communication:
      a. The client $i$: Calculates $e^{X_i \beta_1^*}$ and sends it to client 1
      b. Client 1: Combines the statistics $e^{X\beta} = \prod_i e^{X_i \beta_i^*}$ and calculates $V$, then sends
         the $V$ back to clients $2, 3, \ldots, k$
  9. Each client $i$: Calculates $X_i V^{1/2}$ and sends to the server
  10. Server: Combines and calculates the standard errors, p-values, and confidence intervals
```

Algorithm 3.2: Matrices-Puzzle-solving

```
Input: Intermediate matrix of each client $X_i' V^{1/2}$, number of clients $k$
Output: The completed intermediate matrix $X'VX$ for calculation of Standard Deviation.
For $i = 1 : k$ do
    For $j = 1 : k$ do
        $RowBlock[i] = [RowBlock[i],\ X_i' V^{1/2} \cdot (X_j' V^{1/2})^T]$
    End for
        $X'VX = [(X'VX)^T,\ RowBlock[i]^T]^T$
End for
```



$X_1^T V X_1$ + $X_1^T V X_2$ + $X_1^T V X_3$
$p_1 \times p_1$
$p_1 \times p_2$
$p_1 \times p_3$
$X_1^T V X_1$ $X_1^T V X_2$ $X_1^T V X_3$
Row_Block 1 $p_1 \times p$
+
$X_2^T V X_1$ + $X_2^T V X_2$ + $X_2^T V X_3$
$p_2 \times p_1$
$p_2 \times p_2$
$p_2 \times p_3$
$X_2^T V X_1$ $X_2^T V X_2$ $X_2^T V X_3$
Row_Block 2 $p_2 \times p$
+
$X_3^T V X_1$ + $X_3^T V X_2$ + $X_3^T V X_3$
$p_3 \times p_1$
$p_3 \times p_2$
$p_3 \times p_3$
$X_3^T V X_1$ $X_3^T V X_2$ $X_3^T V X_3$
Row_Block 3 $p_3 \times p$
$X_1^T V X_1$ $X_1^T V X_2$ $X_1^T V X_3$
$X_2^T V X_1$ $X_2^T V X_2$ $X_2^T V X_3$ $= X^T\ VX$
$X_3^T V X_1$ $X_3^T V X_2$ $X_3^T V X_3$
$p \times p$

Figure 3.1: Example for 3 clients VERTIGO-CI matrices puzzle combination. Here the dimensions of X_1, X_2, X_3, V are $n \times p_1, n \times p_2, n \times p_3$, and $n \times n$ where n is the number of patients and p_i is the number of variables in the client i. And $p = p_1 + p_2 + p_3$ is the total number of variables. 'Row_Block i' is defined as $[X_i V X_1, X_i V X_2, \ldots, X_i V X_k]$ binding k matrices column-wise where k is the number of clients.

Table 3.1: The difference in parameter estimates in synthetic data. The difference was measure in the L_∞ norm, the maximum absolute distance from the ground truth of the 20$estimates. The dataset had 2000 samples, and 20 features were used.

Number of Clients	Difference in Coefficient	Difference in Std Error
2	1.34 x 10^{-6}	5.31 x 10^{-8}
3	1.34 x 10^{-6}	5.22 x 10^{-8}
4	1.34 x 10^{-6}	5.34 x 10^{-8}

3.1. 'Row_Block i' is defined as $[X_iVX_1, X_iVX_2, \ldots, X_iVX_k]$ binding k matrices column-wise where k is the number of clients.

3.3.5 Implementation

We implemented the VERTIGO-CI in Python 3.7, using the numpy, pandas, and scipy modules to perform the mathematical computations. We utilized the asyncio module for network programming to allow asynchronous operations. All testing was performed on Amazon Web Service (AWS) EC2 instance of r5a.2xlarge (64 GB Memory, 8 CPUs) with Ubuntu 18.04 instances in different data centers in five continents (Asia: Seoul, Australia: Sydney, Europe: Dublin, North America: Oregon/Virginia, and South America: Sao Paulo).

3.4 RESULTS

The proposed method's correctness is reported in Table 3.1 using the maximum absolute distance from the ground truth of the 20 estimates for 20 features from the synthetic dataset. All 20 coefficients specified in the simulation model achieved the near-perfect agreement. The runtime of the proposed method increased exponentially with an increase in the sample size and an increase in the number of clients (Figure 3.2). The runtime increased slightly when the number of features was increased. The effect of physical distance among clients and the server

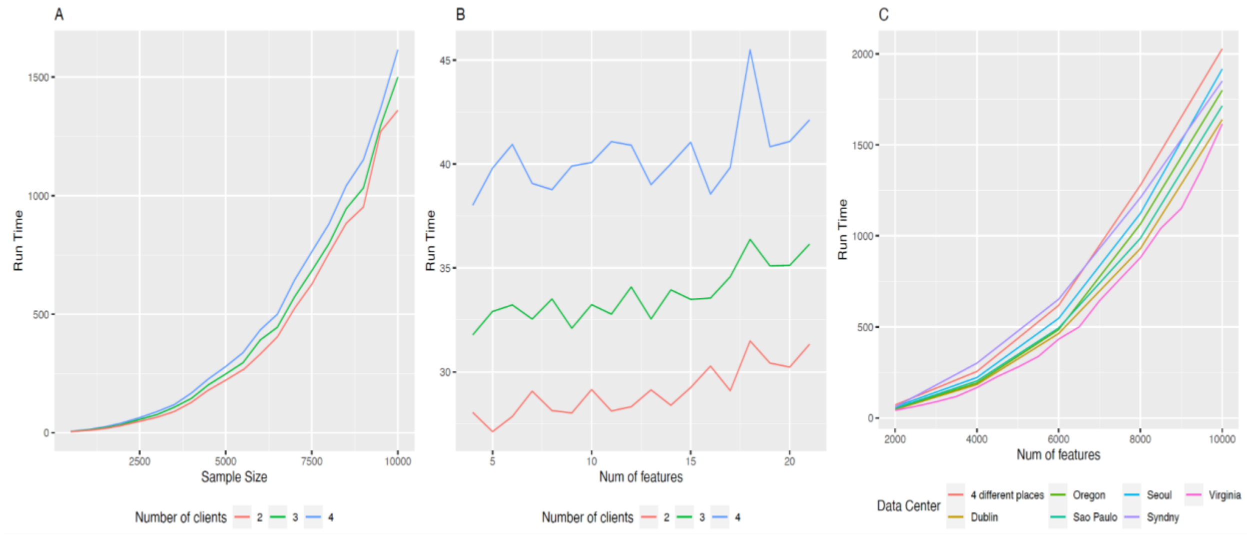


Figure 3.2: Computation time of the synthetic data. The time includes intermediate file transfer in two ways, client-to-client and client-to-server. **A**: Both sample size and number of clients varied under a fixed number of features = 20. **B**: Both feature and client numbers varied under a fixed sample size = 2000. **C**: Runtime by different AWS data centers. The blue line represents the run time of all four clients scattered in four different data centers away from Virginia, where the server is located. The other colors represent the two data centers, one for co-locating all four clients and the other for the server site.

was evaluated using different cloud service providers' data centers. Six different Amazon Web Service (AWS) data centers were selected to co-locate all four clients, while keeping the server in Virginia, US. All four clients were scattered in four different places (blue line), which took the longest execution time. As a baseline (pink), the co-location of all four clients and the central server in one data center (Virginia) achieved the shortest computation time. From Dublin to Sydney, a remote data center was tested to observe the effect of the client data center's physical distance from the server data center, Virginia. Interestingly, trans-US (Oregon - Virginia) took longer run times than trans-Atlantic (Dublin - Virginia) or trans-America (Sao Paulo - Virginia). The reasons may lay on multiple jump boxes along with the connection between Oregon and Virginia, while the submarine cables are connected directly. In BURN1000 (the first real dataset), VERTIGO-CI achieved the near-perfect agreement between the estimates and the ground truth (Table 3.2). Its average runtime varied between 12 and 15 seconds, with the number of clients ranging from 2 to 4 (Table 3.3). In PENNCATH (the second real dataset), the proposed method showed a good agreement between the federated and centralized coefficient estimates (Table 3.4). However, the estimated difference in standard error was the one order of magnitude larger than for the coefficient. The runtime increased linearly with the increase in the number of SNPs, and the mean running time for each trial can be seen in Table 3.5.

Table 3.2: Accuracy of VERTIGO-CI in BURN1000 data. flame: flame involved in burn injury, SE: Standard Error, TBSA: total burn space area in percentage

Variable	**Coefficient**	**Coefficient Difference to the Ground Truth**	**Standard Error**	**Standard Error Difference to the Ground Truth**
Intercept	-3.819841	4.978316e-07	0.296338	-5.805270e-08
Race white	-0.347684	-3.992930e-08	0.153023	-7.573084e-09
Facility	-0.176201	-3.277626e-07	0.139130	-3.553973e-08
Gender male	-0.069838	-2.018457e-08	0.142060	-9.766352e-09
Flame involved	0.291130	-1.952836e-07	0.178000	-2.107444e-08
Inhalation injury	0.439069	7.644087e-08	0.118723	-1.191069e-08
TBSA	1.741145	-5.004354e-08	0.179537	-1.389442e-08
Age	2.075578	3.407076e-08	0.217424	-9.323797e-09

Table 3.3: The runtime in BURN100 data with varied number of clients

Number of Clients	Mean running time (s)
2	12.4515
3	14.1357
4	15.9227

Table 3.4: Accuracy of VERTIGO-CI in PENNCATH data. HDL: High-Density Lipoprotein, LDL: Low-Density Lipoprotein, PC: Principal Component, TG: TriGlyceride, and SE: Standard Error

Variable	Coefficient	Coefficient Difference to the Ground Truth	SE	SE Difference to the Ground Truth
SEX	-1.200262	2.007786e-07	0.005065	1.391678e-01
AGE	-0.032013	1.330223e-06	0.144231	1.393521e-01
HDL	0.015559	2.796433e-07	0.004879	1.855911e-04
TG	0.011913	1.658586e-08	0.001869	7.267426e-04
LDL	0.006471	8.162119e-07	0.001142	7.268230e-04
PC1	1.057632	6.040747e-06	2.405702	3.457518e-05
PC2	-3.234316	1.851210e-05	2.378695	1.801008e-02
PC3	-2.172853	1.293278e-05	2.404689	2.596270e-02
PC4	-1.136879	7.012701e-06	2.392821	1.190317e-02
PC5	1.449743	8.630703e-06	2.408969	1.611641e-02
PC6	0.060668	8.945382e-08	2.401000	8.005713e-03
PC7	2.508987	1.462683e-05	2.424523	2.348583e-02
PC8	-3.037303	1.840741e-05	2.449719	2.515850e-02
PC9	-2.629828	1.576750e-05	2.422779	2.698205e-02
PC10	-0.983910	6.774366e-06	2.396671	2.614376e-02

Table 3.5: The runtime with PENNCATH data with varied number of SNPs

Number of SNPs	Mean running time (s)	Standard Deviation of run time
10	260.2662	51.4977
100	2625.8944	13.6653
1000	26159.9742	0.6426

3.5 DISCUSSION

We proposed a novel method of embedding the client-to-client part to enhance the interpretation of VERTIGO with hypothesis statistics like standard error, Z-score, p-value as well as confidence intervals for each coefficient. Using both synthetic and real datasets, we demonstrated the correctness of VERTIGO-CI by showing that its estimates are identical to those from the logistic regression with acceptable runtime with a small to midsize number of features. Our proposed method's novel contribution is the standard error of the point estimates, which allows statistical decisions using P-value and confidence intervals. As the previous VERTIGO implied, the implementation of a fixed-Hessian matrix on Newton's method can highly reduce the computation complexity. However, the inversion of the fixed-Hessian matrix is still non-trivial.

And another potential problem is the size of the gram-matrix during communication, a gram matrix with 10,000×10,000 size can take up to 60 GB size. We have successfully implemented our VERTIGO-CI on a server in different sites but there is still room for improvement in runtime to handle a very large number of features as in genomics data.

3.6 ACKNOWLEDGEMENTS

This study was supported by National Heart, Lung, and Blood Institute grant R01HL136835 and National Library of Medicine grant R01LM013712.

Chapter 3, in full, in a reprint of the material as it appears in “VERTIcal Grid lOgistic regression with Confidence Intervals (VERTIGO-CI).” Jihoon Kim, Wentao Li, Tyler Bath, Xiaoqian Jiang, Lucila Ohno-Machado. AMIA Jt Summits Transl Sci Proc 2021 May 17;2021:355-364. eCollection 2021, PMID: 34457150. The dissertation author was the primary investigator and the first author of this paper.

Chapter 4 Transforming Cohort Data Collected over the Span of 30-years into a Common Data Model Without External Access to Individual Records: the Strong Heart Study

4.1 ABSTRACT

The Strong Heart Study (SHS) is a prospective cohort study of cardiovascular disease and its risk factors among American Indians. The existing SHS data model is not standardized in a universally accepted way so that researchers require additional efforts of data extraction, harmonization, and analysis.

This study aims at evaluating the feasibility of transforming SHS cohort data source to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), to enable syntactic/semantic interoperability and collaborative research in an efficient manner while protecting individual privacy.

The source database of the SHS was transformed to OMOP CDM. After the source profiling, an extract-transform-load process was designed, implemented, and evaluated in an iterative manner. To evaluate data mapping quality, mapping coverage and conformance were measured; the characteristics of both source and destination databases were compared.

The SHS database was transformed into an OMOP CDM database of 5,930 unique participants, 1,004,527 total inserted rows in 8 tables, and 494 mapped variables. The resulting transformed database had 98.2% mapping coverage and 88.7% conformance rate. A near perfect agreement was observed between source and destination databases. Introducing six novel mapping types increased the coverage.

SHS personnel transformed its longitudinal cohort of American Indians to OMOP CDM, in partnership with an external collaborating team who were not granted access to individual-level data keeping participants' trust. The OMOP-mapped SHS database can promote a

longitudinal study spanning multiple years within SHS and a joint study with external cohorts through syntactic/semantic interoperability.

4.2 INTRODUCTION

The Strong Heart Study (SHS) is a longitudinal study of cardiovascular disease and its risk factors among American Indians in four geographic areas of the United States: Arizona, Oklahoma, North Dakota, and South Dakota since 1989[68]. SHS has become one of the longest running large scale population studies of chronic diseases in any minority groups. It led to discoveries of disproportionate cardiovascular disease (CVD) prevalence, incidence, and risk factors in American Indians, and the existence of previously unrecognized epidemic of CVD among those with type 2 diabetes[69,70]. Recognizing its scientific contributions, the National Heart, Lung, and Blood Institute (NHLBI) at the National Institute of Health (NIH) provided contiguous research funding in multiple study phases that span four to five years, and currently phase 7 study is being conducted. SHS receives data access requests incoming at the rate of at least one per week.

Clinical research is usually driven by hypothesis and its exploratory nature of related research activities tend to focus on the structure and content of analytic datasets[71]. For this reason, outlining all possible research use cases that could be generated from an underlying supporting database gets less interest and is typically beyond the scope of a single investigator. This is a universal problem that applies to any health research database that is continuously funded and repeatedly accessed for frequent data requests over a long period such as SHS. A data model is one solution to store and retrieve data in a cohort database to address multiple anticipated research use cases. A data model is an abstract model that defines which data elements will be stored, how they will be stored together with relationships and constraints, and how the research dataset will be extracted, analyzed, and interpreted through accompanied metadata[71]. However, the existing SHS data model was fragmented within each study phase and did not adopt a common data standard[72], posing challenges for data extraction, harmonization, analysis, interpretation, and collaborative research. Due to phase-specific naming convention and siloed data model within a specific study phase, subsequent studies were required to reprocess and reharmonize the same data, resulting in a waste of time and effort as well as errors. Part of the reason for inefficiency is a funding mechanism, where each study phase has to go through its own grant application process with different project aims and scientific goals in line with technology advancement, change in life-style/environment, and revision of guidelines with newly acquired clinical knowledge.

Data standardization using a common data model (CDM) is a remedy to overcome these limitations by enabling syntactic/semantic interoperability, and collaborative research across fields and fulfilling findability, accessibility, interoperability, and reusability (FAIR) principles[73]. Adoption of CDM-based data standardization also enables analysis methods and codes to be reusable and reduces the inefficient tasks of reprocessing/remapping the same data multiple times[74]. In a recent comparison of four CDMs[75], the Observational Medical Outcome Partnership (OMOP) CDM, developed by the Observational Health Data Sciences and Informatics (OHDSI)[36] Consortium was considered best for a longitudinal registry in reference to the Clinical Data Interchange Standards Consortium Study Data Tabulation Model (CDISC SDTM)[76], the National Patient-Centered Clinical Research Network (PCORnet)[77], and the Food and Drug Administration (FDA) Sentinel programs[78]. OMOP CDM has been widely adopted in large biobanks such as All of Us (AoU) Research Program[7] and Million Veterans Program

(MVP)[8], and multiple institutions in Austria[74], France[79], Germany[80], and the US[81]. Recently, disease registries such as cancer[82] and pulmonary hypertension[13] have been transformed to OMOP CDM. Additionally, previous efforts of data transformation to OMOP CDM were either nation, institution, registry, or disease specific, but few studies exist for minority groups, especially for American Indians. Ours is the first study of transforming a large collection of American Indian health data to the OMOP CDM.

All above mentioned data conversions to OMOP CDM were performed within compute environments with full server access to individual-level data. However, this is not possible with SHS of American Indians that have well-founded cultural concerns about data use that go beyond public concerns and a mistrust of researchers and federal government funded projects[83]. Incidents where informed consent and community understandings have been violated, such as the Havasupai incident[20] and the Barrow Alcohol Study[84], led many American Indian tribes to reject scientific research within their communities. Despite this community concern, the SHS has been very successful in conducting important research and in developing collaborations and interventions by codesigning a research study with tribal communities[85]. SHS acknowledged the advantage of the source database transformation to OMOP CDM to promote higher utilization of a valuable research resource. This need for source data conversion to OMOP CDM under the constraint of protecting individual level data goes beyond SHS to the studies of sensitive data such as addiction and substance abuse, disability, mental health, minor/children, reproductive care, and sexually transmitted disease[86].

In this study, we describe how source SHS data was mapped to OMOP CDM as destination in steps of source profiling, extract-transform-load (ETL), and evaluation, under the condition that the external collaborating team would not have server access to participant level data. We illustrate encountered challenges during mapping and corresponding solutions and evaluation measures of mapping coverage, mapping conformance, and the characteristics of both source and destination databases for mapping agreement evaluation. Beyond new processes, a novel contribution of this paper, we report six different mapping types with their counts and representative examples.

4.3 METHODS

4.3.1 Source data

SHS had longitudinal epidemiologic data of 5,930 total unique participants that started with 3,505 participants in phase 1 (1989-1992) and followed by subsequent studies of 2,793 in phase 2 (1993-1995), 2,484 in phase 3 (1997-1999), 2,773 in phase 4 (2001-2003), 2,387 in phase 5 (2005-2011), and 3,039 in phase 6 (2013-2019)[68]. Written informed consent was obtained from each participant in the SHS cohort.

The source databases were stored in the format of Microsoft Access and SAS version 9.4. Heart outcome data have been collected through Mortality Morbidity Surveillance for more than 30 years, while other data were collected through personal interview survey, physical examination, and lab tests on biosamples collected in each study phase. The survey questions included demography, medical history, life-style, dietary, family health history, leisure-time exercise, and occupation-related physical activities[68]. The physical exam included anthropometric measurement, examination of the heart/lungs/pulses/bruits, blood pressure, 12-lead electrocardiogram, glucose test, lipid panel test from blood, apolipoproteins, and albumin/creatinine. The total number of source variables was 504, with a breakdown of 104 variables from phase 1, 86 from phase 2, 84 from phase 3, 104 from phase 4, 85 from phase 5, and 41 from phase 6. Institutional review boards (IRB) of the respective area's Indian Health Service facility and the University of Oklahoma Health Science Center (OHUSC) have approved the study, in addition to approvals from participating tribal nations and the IRB of the University of California San Diego (UCSD) Health.

4.3.2 OMOP CDM

The OMOP CDM v6.0 had 39 tables in six groups of Clinical Data, Health System, Health Economics, Standardized Derived Elements, Metadata, and Vocabulary[87]. Of these, the Clinical Data group with 7 tables were most actively used to insert the mapped SHS data. Specifically, those tables included <PERSON>, <CONDITON_OCCURRENCE>, <DRUG_EXPOSURE>, <MEASUREMENT>, <OBSERVATION>, <PERSON>, and <VISIT_OCCURRENCE>. The last group, Vocabulary, is a group of 12 tables that contain the concepts, the fundamental unit of meaning to express clinical information. The concepts were derived from 111 vocabularies including Systematized Nomenclature of Medicine (SNOMED) for condition and observation, RxNorm for drug, Logical Observation Identifiers Names and Codes (LOINC) for measurement, and CPT4 for procedure.

4.3.3 Mapping workflow

Figure 4.1 illustrates the workflow of extract-transform-load (ETL) to map source SHS data to destination in OMOP CDM. Each column represents six ETL steps of the workflow that proceeds from left to right. Each column contains key components and stakeholders. Given the source database and metadata files, both OHDSI tool White Rabbit and custom SQL codes were used to profile the source in consultation with the SHS data management team. A list of unique

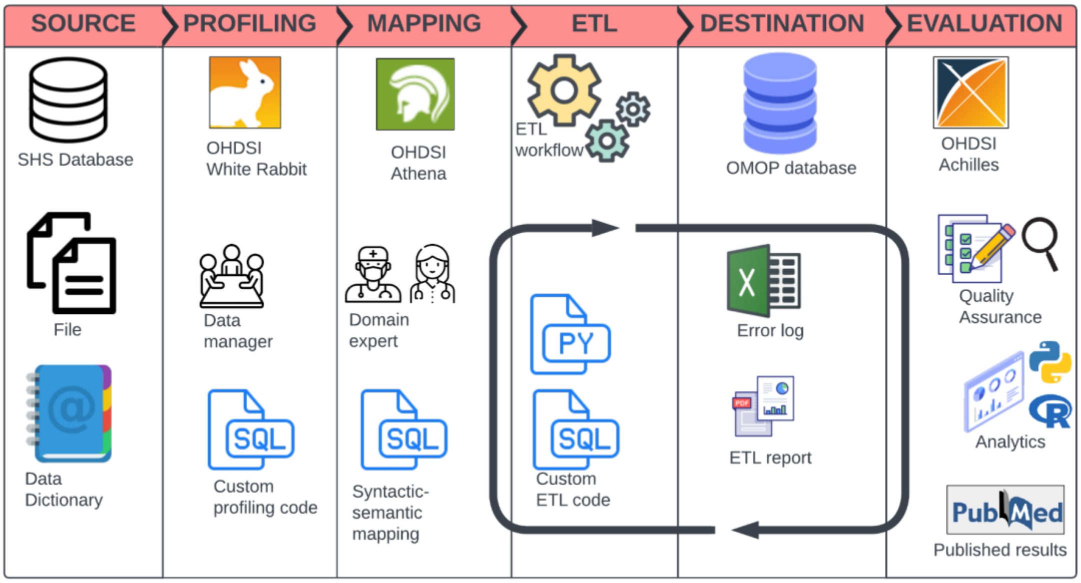


Figure 4.1: Workflow of data mapping from source SHS to destination OMOP databases. ETL: Extract-Transform-Load, OHDSI: Observational Health Data Sciences and Informatics, PY: Python programming language, SQL: Structured Query Language

variable and value pairs were created and reviewed to explore the permissible value, distribution, cardinality, missing rate, and concordance between data dictionary document and actual data in the source. From the scan of the SHS source database, WhiteRabbit generated detailed information on the tables, fields, and values that appeared in a field. WhiteRabbit's output report was used as a reference during the ETL process and SQL code design. Next step was vocabulary mapping, where source concepts were mapped syntactically and semantically to a controlled vocabulary, with help from OHDSI tool Athena and domain experts. For each source terminology, Athena was used to search for relevant OMOP concept identifiers in OMOP vocabularies. Given a SHS source variable 'HbA1c', a commonly used test to diagnose prediabetes and diabetes by measuring average blood sugar levels over the past 3 months, Athena returned the OMOP standard concept identifier '3004410' as a candidate destination variable, with a corresponding concept name 'Hemoglobin A1c/Hemoglobin.total in Blood' that mapped to the LOINC concept code ‘4548-4’, a widely used OMOP concept name. Upon the completion of vocabulary mapping, the data table mapping was followed. The concept identifiers and concept names obtained from vocabulary mapping were included in Python and SQL codes to insert new records into the destination tables. Any error or exception was recorded in the log file and corresponding report during ETL workflow characterization; custom Python and R codes were used for quality assurance. Initial UCSD mapping results were first reviewed by an internal team at the OUHSC and went through the second review independently by the Yonsei University team that did not participate in the first mapping. In case of disagreement between two mapping teams, a joint decision was made through a series of discussions.

4.3.4 Vocabulary mapping type

The required efforts of source-to-destination vocabulary mapping varied by the relationship between source and destination. An ideal case is when the source vocabulary could be matched to the destination vocabulary with 1-to-1 semantic equivalence as in mapping of source variable 'CHF_C' to destination concept 'Congestive heart failure' with corresponding concept identifier '319835'. When such initial attempt failed, the search continued to find the destination vocabulary with 1-1 semantically equivalence as in mapping of source 'WHEN WAS YOUR LAST DRINK, WITHIN WK, MONTH, YR, >1 YR' to destination concept name 'Time since stopped drinking' with corresponding concept identifier 4042875. When both attempts failed, the source vocabulary was classified as unmapped. To better characterize vocabulary mapping and improve the mapping coverage, we extended this approach of three vocabulary mapping types suggested in a previous study[74] and introduced novel vocabulary mapping types. The definitions of six mapping types were data-driven by adding a new mapping type going through unmapped variables and values, further explained with examples in the Results section.

4.3.5 Mapping evaluation

In the final step, software routines were developed to evaluate the integrity and equivalence of the source and the destination databases, and row counts of corresponding tables were compared. SQL queries were developed to gather information on patient, medication, and diagnosis, from the OMOP CDM into temporary tables. The destination tables were then compared record-wise to the corresponding source tables. Descriptive statistics were calculated from the transformed data and compared to original values to determine the success of the ETL process. The whole ETL process was conducted in an iterative manner until a sufficient level of data quality assurance was reached in terms of data completeness, conformance and plausibility[88,89]. Completeness was measured as the frequency of source data attributes present in a data set without reference to data values and OMOP CDM concept identifiers in this study. Conformance was measured as the number of newly created OMOP concept identifiers custom to SHS not existing in OMOP standard concepts. Adding custom concepts that did not exist in OMOP increases the completeness but decreases the conformance. Plausibility refers to the believability or truthfulness of data values after mapping. While conformance and completeness focus on the structure and presence of values, respectively, plausibility focuses on actual values and their distributions[89]. The predefined threshold for acceptable completeness and conformance was 85% or higher. The target threshold for plausibility was a difference of 5% or less in the counts, percentages, and averages between the source and destination for preselected variables: number of participants, gender, death, visit age, hypertension, diabetes, body mass index, and HbA1c.

4.3.6 Privacy-protecting mapping

A windows server version 2016 with (2 CPUs, 8 GB Memory, and 250 GB Storage) was set up at the SHS Coordinating Center (SHSCC), University of Oklahoma. Microsoft Structured Query Language (SQL) Server (version 14.0), OHDSI software tools, White Rabbit and Achilles, Python 3.7 and R 4.2.1 were installed. The source database and flat files were uploaded to SQL Server database with primitive database schema similar to data structure as a source. Then OMOP v6.0 was installed on a separate database schema as a destination. Access to this server and individual-level data therein was limited to a designated programmer at SHSCC who already has been accessing the data for over 10 years. The mapping team external to the SHSCC wrote and sent the SQL and ETL scripts to the SHSCC programmer, who ran the code and returned only the summary statistics such as counts or the names of source variables that had been mapped to OMOP concepts. The aggregate data egress happened only after approval by the SHS domain experts. No server access was granted, other than the SHSCC programmer. And no individual-level data was transferred out of the server.

4.4 RESULTS

4.4.1 Mapping coverage

Of 504 SHS source variables, 494 were mapped to OMOP, achieving 98.2% coverage. The eight tables of the OMOP destination database and their counts are shown in Table 4.1, and the statistics were obtained from 5,930 SHS participants. As expected, the OMOP <measurement> table had the most rows both in total records and records per individual. The numbers of participants were similar across tables except for the <drug_exposure> table. This is because the source SHS stored the history of multiple drug names as a

Table 4.1: Statistics of OMOP-mapped tables. NA: Not Applicable, SD: Standard Deviation

OMOP Table	Rows N	Persons N	Rows per Person MEAN (SD)	Concepts N
care_site	3	NA	NA	1
condition_occurrence	56,955	5,924	9 (6.6)	29
drug_exposure	12,310	3,529	3 (2.5)	2
location	3	NA	NA	1
measurement	495,227	5,927	83 (31.4)	56
observation	417,118	5,930	70 (29.3)	41
person	5,930	5,930	1 (0.0)	NA
visit_occurrence	16,981	5,930	2 (1.1)	NA

single string, which was incompatible to OMOP's data model that stores each itemized medication use in the <drug_exposure> table with corresponding exposure start and end datetimes linked to patient visit records in the <visit_occurrence> table. Each SHS participant had up to six visits with a single visit per study phase. Figure 4.2 shows a heat map, where each entry represents the number of mapped variables from the source domain of SHS in the y-axis to the destination concept classes of OMOP in the x-axis. The largest number of mapped variables was from the source Medical History to the destination Lab test in the last row of the heat map. Clinical Finding and Lab Test were two most populated domain concept classes. The first column in destination, Anatomical Therapeutic Chemical (ATC) Classification System, is a drug classification system and mapped variables come only from a single source, Medical Records Abstraction. The last column represents the variables that were unable to be mapped to existing OMOP concepts so customized concepts had to be created.

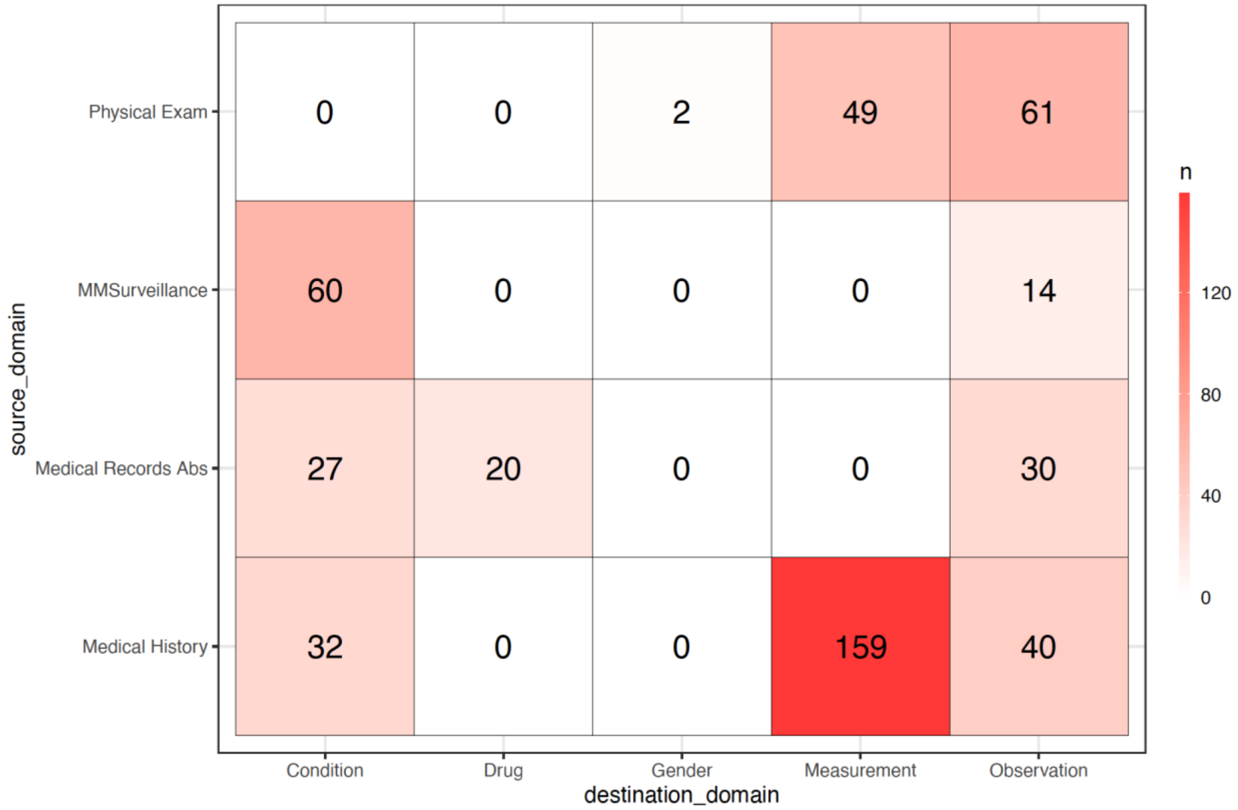


Figure 4.2: Heat map of number of mapped variables from source domains (SHS) to destination (OMOP) concept classes. ATC: Anatomical Therapeutic Chemical classification system of drug, LOINC: Logical Observation Identifiers Names and Codes, SHS custom: newly added concepts for Strong Heart Study

In addition to high coverage, mapped variables well conformed to OMOP standard, resolving various data modeling issues. Table 4.2 illustrates the existing challenges of conforming source data and the corresponding solutions of mapping them to OMOP in the destination. The challenges either stemmed from non-concordance to the data dictionary or presence of missing value within the source database. For example, the source variable ‘S1DMAGE’ representing ‘AGE AT FIRST DIAGNOSIS OF DM DURING STUDY PHASE 1’ did not have any value (100% missing) but it was discovered that the variable name had been changed from ‘MED27’. Also, ‘S1DMAGE’ in phase 1 and ‘S2DMAGE’ in phase 2 were not comparable even though they were measuring the same data element except for the time period.

Table 4.2: Example of challenges encountered during source to destination mapping. In the source database, 'S1DMAGE' represented 'AGE AT FIRST DIAGNOSIS OF DM DURING STUDY PHASE 1'.

Challenge	Example	Solution
Variable names disagree between data dictionary and actual source file	Statement in the Data Dictionary: 'Renamed variable name MED27 to S1DMAGE' Actual data in the source: S1DMAGE variable did not exist	Reported inconsistencies and renamed variables
Cannot compare same concept values longitudinally	Six different variable names existed for the concept 'Age at the first diagnosis of diabetes': S1DMAGE, S2DMAGE, S3DMAGE, S4DMAGE, S5DMAGE, S6DMAGE	Mapped to a common and standard concept identifier, 4307859 ('age at diagnosis')
Multiple values were found when a single value is expected	The source variable 'Age at the first diagnosis of diabetes' is expected to have a single value, but multiple values were found as a participant gave different response when asked at different study phase	Kept the earlies value as a unique value while keeping other values as a record keeping
Missing value exists in a certain study phase	The values for the source variable S5DMAGE were missing	Tracked document to identify and compute the missing values from other variables

4.4.2 Mapping plausibility

Plausibility refers to the believability or truthfulness of data values after mapping[89]. After consultation with SHS subject matter experts, exploratory analysis was conducted to evaluate the agreement between source and destination. In Table 4.3, the numbers of participants were equal, and their breakdown counts by demography, mortality, condition, and measurement values were

Table 4.3: Agreement between source and destination. Selected data items are compared before and after mapping to OMOP CDM. BMI: Body Mass Index (kg/m^2), HbA1C: Hemoglobin A1 C (mg/dL), OMOP: Observational Medical Outcomes Partnership, SD: Standard Deviation, SHS: Strong Heart Study

Item	**Source (SHS)**	**Destination (OMOP)**
Participants, N (%)	5,930 (100.0)	5,930 (100.0)
Gender: female, N (%)	3,460 (58.3)	3,460 (58.3)
Gender: male, N (%)	2,470 (41.7)	2,470 (41.7)
Death, N (%)	2,463 (41.5)	2,463 (41.5)
Visit Age, MEAN (SD)	59.9 (15.8)	59.9 (15.8)
Hypertension**,** N (%)	3,039 (51.2)	3,039 (51.2)
Diabetes, N (%)	2,968 (50.1)	2,968 (50.1)
BMI: with measurements, N (%)	5,919 (99.8)	5,919 (99.8)
BMI: 30 or higher, N (%)	3,675 (62.0)	3, 675 (62.0)
HbA1c: with measurements, N (%)	4,710 (79.4)	4,710 (79.4)
HbA1c: 6.5% or higher, N (%)	1,930 (32.5)	1,930 (32.5)

identical between source and destination. Visit Age was considered instead of chronological age because SHS has been an ongoing longitudinal study since 1989 with new participants coming and old participants passing away. The fraction of diabetic participants with high HbA1c values were in good agreement with the previous study results[90]. To examine study level plausibility, a plot of numbers of records was drawn along time by OMOP tables using OHDSI tool Achilles (Figure 4.3). The number or records are shown using 1M data points from 5,930 participants in five OMOP tables. Six different peaks over time matched with actual six SHS study phases, where each participant made a visit on his/her own date but once per phase.

4.4.3 Mapping type and conformance

Six mapping types were introduced, and their descriptions are shown in Table 4.4 with corresponding counts and examples. The most dominant type (67.8%) was ‘equivalence’, where SHS source variable was mapped to OMOP concept identifier as an exact string match or one-to-one relation such as ‘male gender’. The second most mapping type was ‘added qualifier’ (12.8%), where the source variable was split into a stem and a qualifier part and mapped to multiple OMOP concept identifiers as explained in Figure 4.2. OMOP conversion was not possible with 56 source variables, in such a case, corresponding new variables were created with concept identifier numbers greater than 20,000,000,000, following OMOP recommendation[87]. ‘DEFSTK_C’ is a source variable representing ‘Definite stroke confirmed by chart review’ but could not be mapped to OMOP directly, since OMOP requires actual date time of clinical diagnosis with associated visit record. While customized concept creation allows flexibility and a new mapping, interoperability is decreased as other institutions might not have those custom concepts or might have them, but they are used in different contexts. For example, the same concept identifier ‘20,000,000,001’ could be mapped as a condition in hospital A but a certain

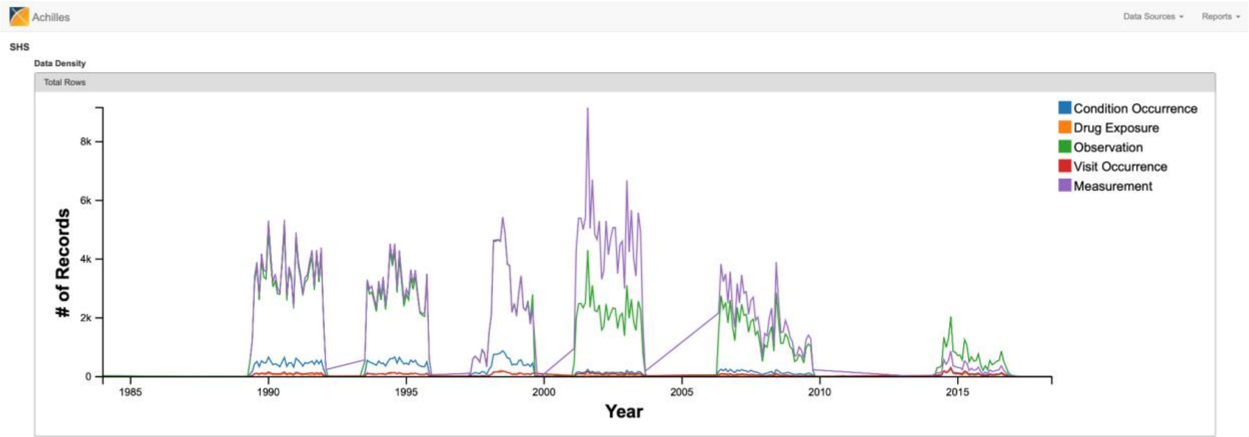


Figure 4.3: Achilles plot of OMOP-mapped records. The numbers of records are shown along time by OMOP tables using OHDSI tool Achilles

Table 4.4: Mapping types of total 494 mapped variables

Type	Description	Mapped Variables N (%)	Representative Example	
			Source (Concept identifier, ‘Concept name’)	**Destination (Concept identifier, ‘Concept name’)**
added qualifier	split source concept into two, stem concept and qualifier concept	63 (12.8)	S2DMAGE, ‘AGE AT FIRST DIAGNOSIS OF DM’	4307859, ‘Age at diagnosis’ 208120, ‘Diabetes mellitus’
equivalence	1-to-1 full equivalence mapping	335 (67.8)	CHF_C, ‘Congestive heart failure’	319835, ‘Congestive heart failure’
introduce custom	unable to map to existing OMOP variables; create a new custom concept	56 (11.3)	DEFSTK_C, ‘Definite stroke confirmed by chart review’	2000000017, ‘Definite stroke confirmed by chart review’
nearest	map to most similar concept	3 (0.6)	S1LDRINK, ‘SHS3 WHEN WAS YOUR LAST DRINK, WITHIN WK, MONTH, YR, >1 YR’	4042875, ‘Time since stopped drinking’
scale conversion	convert log-scaled values to anti-log scale values	5 (1.0)	S2LACR, ‘SHS2 LOG URINARY ALBUMIN / CREATININE Ratio’	3034485, ‘Albumin/Creatinine [Mass Ratio] in Urine’
uphill	map to more general concept	3 (6.5)	S1OTHCVD_C, ‘OTHER CARDIOVASCULA R DISEASE’	1304057, ‘Disorder of cardiovascular system’

medication in hospital B. Our mapping conformance to the standard OMOP was calculated based on the number of 'introduce custom' type mapping, which was 56. Then the mapping conformance rate was 88.7% (=100-100*56/494)[89]. In some cases, a source variable was mapped to a more general concept as no direct matching OMOP concept was found at the same granular level ('uphill' mapping type), the nearest mapping concept was found ('nearest' mapping) or the scale/unit was converted ('scale conversion' mapping type). The representative examples are shown in Table 4.4. There were 10 (2%) unmapped variables. More than half of these were due to study protocol. For example, the remaining unmapped cases were specific qualifiers (e.g, 'HYPERTENSION BY WHO DEFINITION'), composite customs (e.g, 'ANGINA AND ECG' meaning angina pectoris discovered in ECG), specific condition (e.g., 'DEFINITE MI BEFORE EXAM'), or concepts (e.g., 'SHS DIABETES TREATMENT') that are too general.

4.5 DISCUSSION

4.5.1 Principal results

We successfully transformed source SHS to OMOP CDM. We achieved 98.2% completeness and 88.7% conformance rate. Both source and destination databases showed near perfect agreement in the domains of demography, condition, measurement, and observation. We also introduced novel mapping types and applied them to our mapping. We introduced novel mapping types between source and destination. This is an extension of three mapping types of Haberson et al.; (i) one-to-one, (ii) one-to-one but semantically equivalent, and (iii) no equivalent concept exists[74]. The main motivation of this extension was the presence of survey question type data in our source. With availability of large research cohorts like AoU[7] and the MVP[8], both of which are converted to OMOP CDM and have a lot of survey questions, the equivalence mapping to one concept may not be possible anymore and additional types such as ours could be improve the mapping coverage and the usability of mapped data. Harmonizing the data across these resources is important to increase the statistical power of studies.

4.5.2 Implication of data transformation to OMOP CDM

Data standardization enables capturing, classifying, and analyzing patient data using common vocabularies and ontologies through reuse of methods and software tools[87]. Mapping to CDM boosts code reusability and reduces the cost and efforts during a joint analysis of SHS with external cardiovascular studies in diverse populations such as the Jackson Heart Study (JHS) for African Americans[91] or the Multi-Ethnic Study of Atherosclerosis (MESA) for Hispanics[92]. In addition to SQL code reuse, OHDSI provides many software tools for cohort identification, data quality assurance and analysis that can be applied provided that the source database has been transformed to OMOP CDM. In recent COVID-19 research[93], we demonstrated the usefulness of OMOP-CDM in a rapid and efficient way to perform a research analysis across 11 academic hospital databases. Such a fast response to a clinical question about a new pandemic infectious disease like COVID-19 was possible because the vast majority of health databases already mapped to the OMOP CDM. Similarly, in SHS, we anticipate that a clinical question about cardiovascular disease could be answered jointly from diverse population cohorts including SHS, JHS, MESA, All of Us, MVP with similar speed and efficiency.

4.5.3 Comparison with Prior Work

In related works, a longitudinal, community-based health study registry data with over 12,000 active participants was mapped to four different data models[75]. The mapping coverage was 76% with OMOP CDM, 55% with SDTM, 48% with PCORnet and 37% with Sentinel.

When three pulmonary hypertension registry databases were mapped to OMOP CDM[13], the percentage of mapped patients in the source databases ranged between 96% and 99%. The percentage of excluded records during mapping ranged from 7% to 52%. In another study, a regional health claims database of 12,606 patients over age 70 was transformed to OMOP CDM[74]. The mapping coverage was 99.7% for drug codes and 99.2% for diagnosis codes. Unlike previous data conversion to OMOP CDM, our mapping process was conducted in a privacy-preserving manner without server access to individual-level data by investigators external to the data hosting institution. Although not having access to the actual observations slowed down the process of transforming the data, and possibly resulted in fewer findings related to inconsistencies in coding than would be uncovered if all teams had access to the participant data, it was critical that we followed this route in order to honor the commitment with the participants and thus uphold trust across all parties. Our workflow could be applied to OMOP transformation of databases of sensitive data such as sexually transmitted diseases or mental health.

4.5.4 Limitations

Due to the current table schema design of OMOP, we could not add secondary qualifiers during mapping. For example, the concept age at first diagnosis of diabetes according to WHO clinical diagnosis guideline' has nested qualifiers, which cannot be mapped to OMOP CDM directly. Also, behavior survey questions such as smoking and drinking were not mappable to current OMOP CDM as they were often accompanied with different qualifiers such as unit, frequency, elapsed time since start/stop among others. Beyond survey questions, SHS also has electrocardiogram waves and genomic data, but this was outside the scope of present study and OMOP CDM does not support those data yet. Building or extending CDM that accommodate those non-canonical data will be our future work.

4.6 CONCLUSIONS

In this study, we transformed a longitudinal prospective cohort focused on cardiovascular disease in American Indians into the OMOP CDM while preserving individual privacy. The existing data model of SHS previously required repeated and redundant processing of data extraction, harmonization, and analysis, consuming more time and effort than was warranted. The resulting SHS mapped to OMOP CDM will facilitate a large-scale SHS studies spanning multiple phases/years efficiently and will foster a joint study with OMOP-based external cohorts through syntactic/semantic interoperability and ability to reuse code and software tools developed by and shared with a large OHDSI community.

4.7 ACKNOWLEDGEMENTS

This research was supported by NIH grant R01HL136835, and the UCSD collaboration with SHS was initiated with a pilot grant from the iDASH National Center for Biomedical Computing (U54HL108460). The Strong Heart Study has been funded in whole or in part with federal funds from the National Heart, Lung, and Blood Institute, National Institute of Health, Department of Health and Human Services, under contract numbers 75N92019D00027, 75N92019D00028, 75N92019D00029, 75N92019D00030. The study was previously supported by research grants: R01HL109315, R01HL109301, R01HL109284, R01HL109282, and R01HL109319 and by cooperative agreements: U01HL41642, U01HL41652, U01HL41654, U01HL65520, and U01HL65521. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Chapter 4, in full, has been submitted for publication as "Transforming Cohort Data Collected over the Span of 30-years into a Common Data Model Without External Access to Individual Records: the Strong Heart Study." Jihoon Kim, Paulina Paul, Pravina Kota, Tyler Bath, Kai W Post, Yu Rang Park, Ying Zhang, Elisa T Lee, Lucila Ohno-Machado. The dissertation author was the primary investigator and the first author of this paper.

Chapter 5 Patient Perspectives About Decisions to Share Medical Data and Biospecimens for Research

5.1 ABSTRACT

5.1.1 Importance

Patients increasingly demand transparency in, and control of, how their medical records and biospecimens are shared for research. How much they are willing to share and what factors influence their sharing preferences remain under-studied in real settings.

5.1.2 Objective

To determine whether and how various presentations of consent forms result in differences in EHR and biospecimens sharing rates and whether these rates vary according to user interface design, data recipients, data/biospecimen items, and patient characteristics.

5.1.3 Design, Setting, and Participants

A data and biospecimen sharing preference survey was conducted at two academic hospitals after random assignment of patients to 1 of 4 options with different layout and formats of indicating sharing preferences.

5.1.4 Interventions

All participants were presented with a list of data/biospecimen items that could be shared for research within the same healthcare organization or with other non-profit or for-profit institutions. Participating patients were randomly asked to select the items they would like to share (opt-in) or were asked to select items they would not want to share (opt-out). Patients in these 2 groups were further randomized to select only among 18 categories, versus among detailed 59 items (simple versus detailed form layout).

5.1.5 Main Outcomes and Measures

The primary endpoints were the percentages of patients willing to share data and biospecimen categories/items.

5.1.6 Results

Among 1,800 eligible participants, 1,246 (69.2%) who completed their data sharing survey were included in the analysis and 850 (68%) of these responded to the satisfaction survey with mean (standard deviation) age of 51.1 (16.7) years; 59.6% were female and 84% white. The number of participants who declined sharing with the Home Institution, Non-Profit, and For-Profit institutions were 46 (3.7%), 352 (28.3%), and 590 (47.4%), respectively. A total of 836 (67.1%) indicated that they wanted to share all items with researchers from the Home institution. When comparing opt-out to opt-in interfaces, all 59 (100%) variables were associated with the sharing decision. When comparing simple to detailed forms, only 14 (23.7%) variables were associated with the sharing decision.

5.1.7 Conclusion and Relevance

A large percentage of patients were willing to share their data and biospecimens for research.

5.2 INTRODUCTION

Use of personal data without explicit user consent has recently put technology companies such as Facebook in the public spotlight[94–97]. In contrast, it appears that fewer concerns have been raised regarding the use of medical records and biospecimens[98], which are also sensitive, for secondary use purposes such as research. It is unclear whether this is because patients are generally unaware that their "de-identified" records are being made available to researchers[99]; their lack of knowledge that anonymized records can be traced back to individuals[100,101]; or simply because there have not been many widely publicized incidents to date[102].

Current laws and regulations require healthcare institutions to comply with a minimally necessary standard in sharing patient medical records and biospecimens for research.[103] There are legislations[104,105] regulating research reuse of patient medical records and biospecimens, so healthcare institutions can allow "de-identified" data sharing and identified data sharing (as long as proper institutional review board (IRB) approvals have been obtained), unless the patient explicitly declines the use of data/biospecimens for any other purposes than direct patient care. This "all-or-nothing" option is problematic because, alerted by recent high-profile cases, the increasing awareness among the general public regarding inappropriate reuse of personal data without explicit user consent[106] may dramatically change patients' attitudes toward secondary use activities involving their medical records and biospecimens (i.e., patients may start denying research access to all their data and biospecimens, even if they may have no problem sharing most of their data/biospecimens or if they want to share them only with certain types of institutions)[107,108]. The regulatory landscape is also changing (for example, in the USA, as of September 23, 2013, newly enrolled patients who need to sign a Health Insurance Portability and Accountability Act (HIPAA) authorization must opt-in to allow the use of their personal health information (PHI) for optional sub-studies and future secondary use)[109]. In the European Union, the General Data Protection Regulation[110] implemented in 2018 requires that patients consent for clinical data use for research. These issues point to a critical need to better understand patients' sharing preferences.

Surveys using hypothetical scenarios have been conducted[111–113], but there has been a paucity of research studies involving EHR and biospecimens sharing preferences applied in real settings[114]. Tiered-consent (i.e., breaking down the record into smaller units in a consent form and allowing partial use of the EHR) is not routinely available in practice today, limiting patients' rights and participation in how their health data are being shared, while there is increasing evidence that patients want to be asked[99] and what they consider sensitive varies. In California, patient specific permission to share mental health, substance abuse, HIV status and genetic information is required in HIPAA authorization forms, but no other items are specified[115]. In many states, there is no requirement for a patient's specific permission for sharing these types of data items[103].

Our study aimed at understanding patients' preferences towards sharing specific data items in their EHRs and biospecimens with different types of researchers. We hypothesized that there would be different decisions for sharing depending on researchers' affiliations, patient characteristics, and the user interface design format of the consent form in which data sharing preferences were elicited. In this study, we randomly assigned patients to 4 types of preference elicitation forms so we could understand whether the form layout and opting in or out method are associated with patients' sharing preferences. This study followed the American Association for Public Opinion Research (AAPOR) reporting guidelines.

5.3 METHODS

5.3.1 Study Design and Population

Patient participants were recruited from two academic medical centers in Southern California. They were approached either by email invitation, or in person in the waiting area of 10 adult outpatient clinics. Inclusion criteria were (1) age 18 or older; (2) being a patient at either academic medical center; and (3) ability to read English or Spanish. While it was preferred that all research activities be conducted through the research website, the study did provide an option to allow patients who did not have easy access to the internet to participate via paper forms. Preference elicitation and surveys were conducted between May 2017 and September 31, 2018. Study protocol, sharing choice form, informed consent form, survey questionnaires, and health literacy test questions are provided as Supplementary materials. The institutional review boards of the University of California (UC) San Diego and UC Irvine approved this study. The informed consent was obtained from the web portal right after sign-up for online users and via a paper form for other users.

Study participants were invited to select preferences of sharing their data and biospecimens for research use. The preferences for data sharing were honored by the institutions during the study period. Each participant also received periodical reports listing research activities that involved secondary use of their medical records.

The list of data and biospecimens that a participant could choose to share or not share included 59 data/biospecimen types grouped into 18 categories (Box 1). This taxonomy was developed based on a pilot study[114] and 5 focus groups involving 18 patients who also provided input on how to best present the selection options on a computer screen and on paper.

Box 1. List of date elements and categories. Data items and categories included in this study.
* items and categories not previously included in our pilot study

Contact Information*
- Name*
- Home Address*
- Email Address*
- Phone Number*

Demographics
- Age
- Sex
- Race
- Ethnicity
- Sexual orientation*

Socioeconomic Status
- Education
- Marital status
- Insurance Status
- Occupation
- Income

Living Environment and Lifestyle*
- Alcohol Consumption Status
- Recreational Drug Use
- Smoking Status
- Diet*
- Physical Activity/Exercise Level*
- Stress Level*
- Social Isolation*

Sexual Life
Pregnancy History
Adoption History*
Body Measurements
Vital Signs
Allergies*
Current or Previous Disease or Condition
- Substance Abuse Related Disease or Condition
- Mental Health Disease or Condition
- Sexual or Reproductive Disease or Condition
- Other

Family Health History
- Substance Abuse Related Disease or Condition
- Mental Health Disease or Condition
- Sexual or Reproductive Disease or Condition
- Other

Laboratory Test Results
- Genetic Test
- Sexually Transmitted Disease Test

- Drug Screening
- DNA sequencing*
- Other

Biospecimen

- Tissue
- Blood
- Urine*

Imaging Test*

- X-Ray*
- MRI*
- CT Scan*
- Other*

Therapy or Treatment Procedures

- Mental Health Related*
- Genitourinary or Reproductive*
- Cosmetic*
- Bariatric*
- Other*

Medications

- Mental Health Related*
- Other*

Health Care Encounter

- Medical Record Number*
- Visit Dates
- Physician’s Name
- Specialty of the Hospital or Clinic
- Clinic Location

Eleven data categories encompassing 50 data items, 6 data categories (Sexual Life, Pregnancy History, Adoption History, Body Measurements, Vital Signs, Allergies) without detailed data items, plus 3 biospecimen items grouped into one biospecimen category were available for selection, as shown in Box 1. The simple form contained 18 categories and the detailed form contained 53 detailed items plus 6 data categories (i.e., there were 59 sharable items in this detailed form). Combining two interventions (opting method and form layout), each participant was randomized into one of four conditions:

(1) Opt-in Simple,

(2) Opt-in Detailed,

(3) Opt-out Simple, and

(4) Opt-out Detailed.

There was no time limit to complete the sharing preferences, which could be changed over time (preferences as of September 31, 2018 considered in the analysis). Participants indicated their sharing preferences by selecting an item or category that they wanted to share when they received an opt-in form or unselecting what they did not want to share when they received an opt-out form. For the simple forms, when a category was selected, all items that belonged to that category were selected, so we could compare individual items across groups. Participants could assess information about which study used or did not use their data and modify their future sharing choices at any time. The screen shots of our digital consent system are shown in Supplementary Figures B.1 and B.2. Once the intervention period was over, a request to complete a satisfaction survey was submitted to assess participant satisfaction with the study and to obtain information about self-reported socio-demographics. Participants had three months to complete this survey. Monthly reminder emails were sent and participants were compensated with a $10 gift card for the completion of the sharing choice form, and a $10 gift card for completing the satisfaction survey. They were not compensated when they made changes to previous selections.

We implemented the Short Assessment of Health Literacy (SAHL), which is designed to assess health literacy by measuring comprehension of the meaning and relation of 18 sets of keywords[116]. A participant was deemed to have an adequate level of literacy if at least 15 items or 83.3% were answered correctly, otherwise literacy was deemed inadequate according to the SAHL evaluation criteria[116].

5.3.2 Statistical Analysis

The homogeneity of the four randomization groups by variable of interest was assessed with the Chi-square test on baseline characteristics. In a univariate analysis, for each of 59 sharable items, a 2-by-2 table was constructed using shared vs. not shared as response to a binarized exposure variable (i.e., exposure vs. reference). An unadjusted odds ratio and its 95% confidence interval were calculated. Assessed exposure variables included the elicitation form's opting method (opt-out vs. opt-in), form layout (detailed vs. simple), patient's age (>=60 vs. <60), self-reported health status (very good or better vs. worse than very good), health literacy (adequate vs. inadequate), gender (female vs. male), household income (>= US$125k/year vs. < US$125k/year), race (white vs. nonwhite), education (>= 4-year-college vs. < 4-year-college), and site (Site #2 vs. Site #1). A logistic regression was applied for the model-based adjusted odds-ratio after controlling for exposure variables as covariates. Statistical significance was determined by 95% confidence interval of the odds ratio for each sharing choice variable.

5.4 RESULTS

A total of 1,800 patients were eligible for this study. Of these, 1,582 signed a consent form to participate in this study: 1,246 (69.2% of eligible participants) who completed their data sharing preference surveys were included in primary analysis and 850 (68%) of these responded to the survey with mean (standard deviation) age of 51.1 (16.7) years; 507 (59.6%) were female and 677 (79.6%) white. The participant recruitment and randomization processes are summarized in Figure 1.

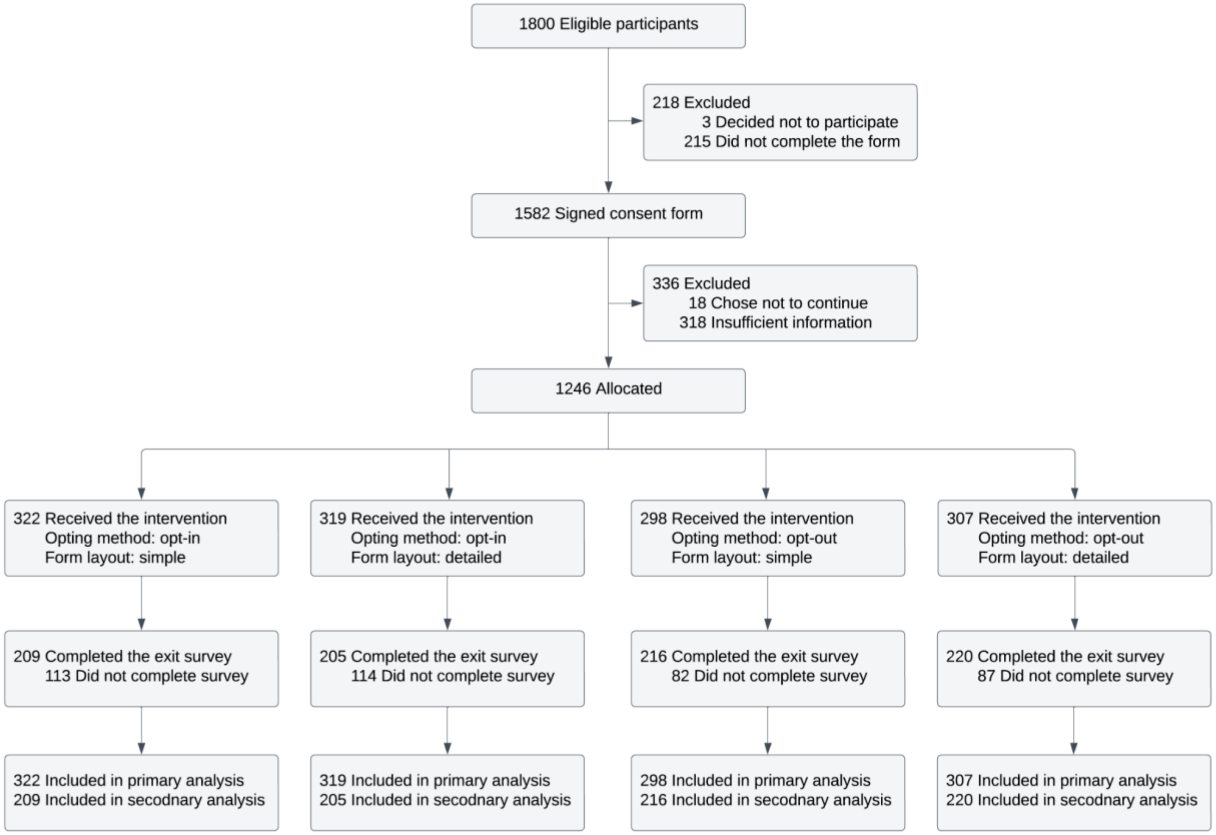

1800 Eligible participants
218 Excluded
3 Decided not to participate
215 Did not complete the form
1582 Signed consent form
336 Excluded
18 Chose not to continue
318 Insufficient information
1246 Allocated
322 Received the intervention
Opting method: opt-in
Form layout: simple
319 Received the intervention
Opting method: opt-in
Form layout: detailed
298 Received the intervention
Opting method: opt-out
Form layout: simple
307 Received the intervention
Opting method: opt-out
Form layout: detailed
209 Completed the exit survey
113 Did not complete survey
205 Completed the exit survey
114 Did not complete survey
216 Completed the exit survey
82 Did not complete survey
220 Completed the exit survey
87 Did not complete survey
322 Included in primary analysis
209 Included in secodnary analysis
319 Included in primary analysis
205 Included in secodnary analysis
298 Included in primary analysis
216 Included in secodnary analysis
307 Included in primary analysis
220 Included in secodnary analysis

Figure 5.1: Study flow diagram.

Randomization assignments, characteristics of the participants who completed sharing preferences and who completed the survey are shown in Table 5.1. A slightly higher number in the Opt-in group reflects the fact that only this option was available for the 40 participants who elected to use paper forms (simple or detailed). Of 12 variables in Table 4.1, none showed a significant difference among 4 randomized groups.

The number of participants who declined sharing with the Home Institution, Non-Profit, and For-Profit institutions were 46 (3.7%), 352 (28.3%), and 590 (47.4%), respectively. There were 291 (23.4%) patients who wanted to share all items with any researcher, while 46 (3.7%)

did not want to share any items. The remaining 909 (72.9%) wanted to share selectively, meaning that they wanted to share at least one item with at least one type of institution, with a general preference towards sharing within the institution in which the patient received care, followed by sharing with researchers from non-profit institutions. The majority of patients (836 or 67.1%) wanted to share all items with researchers from the Home institution.

As explained earlier, for the 4 groups, participants could indicate sharing preferences that could result in eight combinations of three types of researcher's affiliations (i.e., the institution holding their EHRs and biospecimens – home institution, non-profit, and for-profit institutions):

(1) Do not share, regardless of affiliation,

(2) Share with the home institution (H) only,

(3) with non-profit institutions only (NP),

(4) with for-profit institutions only (FP),

(5) with the home institution and non-profit institutions (H+NP),

(6) with the home institution and for-profit institutions (H+FP),

(7) with non-profit institutions and for-profit institutions (NP+FP), and

(8) share with any researcher, regardless of affiliation.

Our analyses were done focusing on (1), (2), (5), and (8), since the other combinations appeared very rarely (4.4%). The participant tendency to decrease sharing as the recipient group got larger was statistically significant, as measured by a chi-square test for trends in proportions ($P <= 4.95E-134$).

Table 5.2 shows the data sharing preferences of all participants. Demographics, allergies, vital signs and body measurements were among the items that the participants were most willing

Table 5.1: Baseline characteristics by intervention group

Variable	Opt-in + Simple (n = 322)	Opt-in + Detailed (n = 319)	Opt-out + Simple (n = 298)	Opt-out + Detailed (n = 307)
Site				
#1	105 (33)	99 (31)	112 (38)	115 (37)
#2	217 (67)	220 (69)	186 (62)	192 (63)
Age				
10-20	1 (0)	1 (0)	0 (0)	0 (0)
20-30	49 (15)	28 (8)	41 (14)	31 (10)
30-40	61 (19)	60 (19)	58 (19)	61 (20)
40-50	46 (14)	54 (17)	41 (14)	41 (13)
50-60	58 (18)	61 (19)	51 (17)	50 (16)
60-70	63 (20)	69 (22)	66 (22)	78 (25)
70-80	35 (11)	39 (12)	36 (12)	34 (11)
80-90	8 (2)	7 (2)	4 (1)	12 (4)
90+	1 (0)	0 (0)	1 (0)	0 (0)
Self-reported health status				
Excellent	31 (10)	27 (8)	31 (10)	36 (12)
Very Good	103 (32)	90 (28)	110 (37)	93 (30)
Good	102 (32)	100 (31)	103 (35)	121 (39)
Fair	53 (16)	61 (19)	46 (15)	51 (17)
Poor	16 (5)	18 (6)	8 (3)	6 (2)
Missing	17 (5)	23 (7)	0 (0)	0 (0)
Health literacy test (SAHL)				
Adequate	271 (84)	272 (85)	267 (90)	285 (93)
Inadequate	34 (11)	24 (8)	31 (10)	22 (7)
Missing	17 (5)	23 (7)	0 (0)	0 (0)

Table 5.1: (continued) Baseline characteristics by intervention group

Variable	Opt-in + Simple (n = 322)	Opt-in + Detailed (n = 319)	Opt-out + Simple (n = 298)	Opt-out + Detailed (n = 307)
Gender				
Female	113 (35)	127 (40)	126 (42)	141 (46)
Male	96 (30)	78 (24)	89 (30)	78 (25)
Other	0 (0)	0 (0)	1 (0)	1 (0)
Non-response in survey	113 (35)	114 (36)	82 (28)	87 (28)
Race				
American Indian/Pacific Islander	3 (1)	2 (1)	2 (1)	2 (1)
Asian	21 (7)	13 (4)	10 (3)	17 (2)
Black	6 (2)	7 (2)	7 (2)	4 (1)
White	157 (49)	165 (52)	181 (61)	174 (57)
Multi-race or other	22 (7)	18 (6)	16 (5)	23 (7)
Non-response in survey	113 (35)	114 (36)	82 (28)	87 (28)
Ethnicity				
Hispanic	24 (7)	25 (8)	17 (6)	26 (8)
Non-Hispanic	185 (57)	180 (56)	199 (67)	194 (63)
Non-response in survey	113 (35)	114 (36)	82 (28)	87 (28)
Education				
High school or less	16 (5)	9 (3)	8 (3)	11 (4)
> High school or < 4-year college	38 (12)	62 (19)	47 (16)	56 (18)
4-year college	75 (23)	50 (16)	64 (21)	64 (21)
Graduate school	80 (25)	84 (26)	97 (33)	89 (29)
Non-response in survey	113 (35)	114 (36)	82 (28)	87 (28)

Table 5.1: (continued) Baseline characteristics by intervention group

Variable	Opt-in + Simple (n = 322)	Opt-in + Detailed (n = 319)	Opt-out + Simple (n = 298)	Opt-out + Detailed (n = 307)
Household Income				
< $25,000	44 (14)	29 (9)	35 (12)	36 (12)
$25,000 - $75,000	40 (12)	68 (21)	55 (18)	58 (19)
$75,000 - $125,000	45 (14)	43 (13)	54 (18)	52 (17)
$125,000 - $200,000	51 (16)	43 (13)	42 (14)	50 (16)
> $200,000	29 (9)	22 (7)	30 (10)	24 (8)
Non-response in survey	113 (35)	114 (36)	82 (28)	87 (28)

Table 5.2: Overall willingness to share with different institutions (n = 1,246). Number of participants (percentage) are shown for. HI: Home Institution, NP: Non-Profit institution, and FP: For-Profit institution.

Variable	Sharing with none	Sharing with HI	Sharing with HI+NP	Sharing with HI+NP+FP
Demographics_Age	78 (6)	321 (26)	259 (21)	572 (46)
Demographics_Sex	81 (7)	317 (25)	257 (21)	575 (46)
VitalSigns	89 (7)	365 (29)	246 (20)	534 (43)
Allergies	90 (7)	353 (28)	250 (20)	539 (43)
Demographics_Race	91 (7)	318 (26)	256 (21)	564 (45)
Demographics_Ethnicity	94 (8)	317 (25)	259 (21)	560 (45)
ImagingTest_X.ray	97 (8)	381 (31)	243 (20)	514 (41)
ImagingTest_CTscan	99 (8)	381 (31)	242 (19)	512 (41)
ImagingTest_MRI	101 (8)	380 (30)	242 (19)	512 (41)
Demographics_SexualOrientation	104 (8)	315 (25)	255 (20)	557 (45)
Lifestyle_Exercise	104 (8)	352 (28)	254 (20)	520 (42)
SES_MaritalStatus	107 (9)	331 (27)	256 (21)	534 (43)
BodyMeasurements	108 (9)	363 (29)	236 (19)	526 (42)
Lifestyle_Diet	111 (9)	346 (28)	252 (20)	522 (42)
Lifestyle_Stress	111 (9)	351 (28)	250 (20)	519 (42)
LabTestResults_GeneticTest	113 (9)	384 (31)	245 (20)	492 (39)
ImagingTest_OtherImagingTest	113 (9)	377 (30)	238 (19)	507 (41)
Medication_OtherMedication	113 (9)	374 (30)	239 (19)	507 (41)
SES_Education	114 (9)	324 (26)	255 (20)	535 (43)
DiseaseCondition_SexualReproductive	114 (9)	370 (30)	247 (20)	501 (40)
LabTestResults_STDtest	114 (9)	394 (32)	235 (19)	490 (39)
LabTestResults_DNAsequencing	114 (9)	390 (31)	242 (19)	488 (39)
SES_Occupation	115 (9)	334 (27)	253 (20)	526 (42)
LabTestResults_DrugScreening	115 (9)	387 (31)	240 (19)	491 (39)

Table 5.2: (continued) Overall willingness to share with different institutions (n = 1,246)

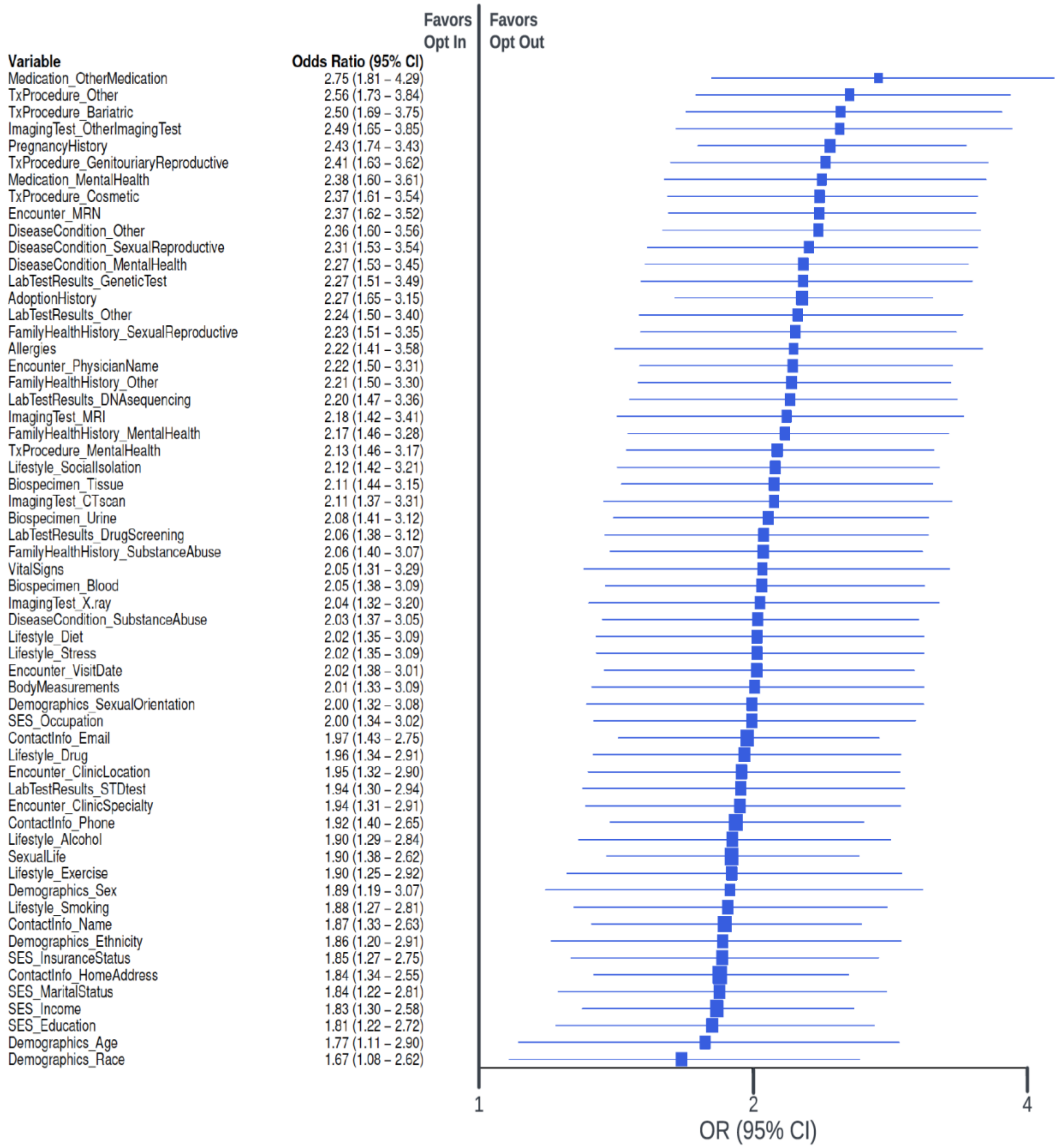
Variable	Sharing with none	Sharing with HI	Sharing with HI+NP	Sharing with HI+NP+FP
Lifestyle_SocialIsolation	117 (9)	350 (28)	249 (20)	515 (41)
LabTestResults_Other	118 (9)	392 (31)	235 (19)	488 (39)
Biospecimen_Blood	118 (9)	388 (31)	231 (19)	495 (40)
DiseaseCondition_MentalHealth	119 (10)	363 (29)	248 (20)	501 (40)
FamilyHealthHistory_MentalHealth	119 (10)	380 (30)	242 (19)	492 (39)
Encounter_ClinicSpecialty	119 (10)	393 (32)	230 (18)	489 (39)
DiseaseCondition_SubstanceAbuse	120 (10)	365 (29)	248 (20)	500 (40)
Lifestyle_Smoking	120 (10)	342 (27)	251 (20)	517 (41)
Lifestyle_Alcohol	121 (10)	345 (28)	254 (20)	512 (41)
Biospecimen_Urine	122 (10)	385 (31)	231 (19)	494 (40)
Medication_MentalHealth	122 (10)	374 (30)	237 (19)	499 (40)
Encounter_ClinicLocation	122 (10)	401 (32)	226 (18)	482 (39)
SES_InsuranceStatus	124 (10)	346 (28)	250 (20)	510 (41)
FamilyHealthHistory_SubstanceAbuse	124 (10)	379 (30)	238 (19)	493 (40)
FamilyHealthHistory_SexualReproductive	124 (10)	380 (30)	240 (19)	490 (39)
DiseaseCondition_Other	125 (10)	362 (29)	244 (20)	501 (40)
Encounter_VisitDate	125 (10)	401 (32)	231 (19)	474 (38)
FamilyHealthHistory_Other	126 (10)	378 (30)	239 (19)	489 (39)
Biospecimen_Tissue	126 (10)	383 (31)	231 (19)	492 (39)
TxProcedure_GenitouriaryReproductive	126 (10)	368 (30)	244 (20)	495 (40)
Lifestyle_Drug	126 (10)	346 (28)	253 (20)	507 (41)
Encounter_PhysicianName	126 (10)	401 (32)	229 (18)	476 (38)
TxProcedure_MentalHealth	129 (10)	371 (30)	243 (20)	491 (39)
TxProcedure_Bariatric	129 (10)	364 (29)	243 (20)	498 (40)

Table 5.2: (continued) Overall willingness to share with different institutions (n = 1,246)

Variable	Sharing with none	Sharing with HI	Sharing with HI+NP	Sharing with HI+NP+FP
TxProcedure_Cosmetic	131 (11)	366 (29)	240 (19)	496 (40)
TxProcedure_Other	131 (11)	366 (29)	244 (20)	494 (40)
Encounter_MRN	134 (11)	413 (33)	228 (18)	456 (37)
SES_Income	163 (13)	318 (26)	247 (20)	497 (40)
ContactInfo_Name	166 (13)	467 (37)	193 (15)	410 (33)
ContactInfo_Email	182 (15)	462 (37)	192 (15)	402 (32)
PregnancyHistory	182 (15)	368 (30)	203 (16)	477 (38)
ContactInfo_HomeAddress	189 (15)	484 (39)	186 (15)	375 (30)
ContactInfo_Phone	193 (15)	485 (39)	187 (15)	373 (30)
SexualLife	194 (16)	375 (30)	204 (16)	462 (37)
AdoptionHistory	195 (16)	354 (28)	201 (16)	481 (39)

to share. Contact information, sexual history, adoption and pregnancy history, and income were the items that the participants were least willing to share.

The sharing preferences were affected by the form's opting method (opt-out vs. opt-in) but not by the layout (detailed vs. simple). Participants were willing to share fewer items when they used the Opt-in form (Supplementary Figures B.1-B.2). Differences according to opting method were significant for all 59 (100%) variables. For form layout (Supplementary Figure B.3), however, only 14 (23.7%) variables had a significant effect on sharing choices. Age >=60 was associated with sharing selections for 56 (95%) variables, and adequate health literacy was associated with sharing selections for all 59 (100%) variables (Supplementary Figures B.4-B.5). The effects of opting method on sharing decision remained significant with one exception (Race); but decreased in magnitude, as shown in Supplementary Figure B.6, after adjusting for participants' characteristics and the form layout. The adjusted odds ratios of sharing in reference to no-sharing for 59 variables were controlled for form layout, age, education, gender, health literacy, household income, self-reported health status, and site in a logistic regression model. For form layout, the number of variables that had significant association with sharing decision decreased from 14 to 9, after adjusting for participants' characteristics and for the opting method.



Variable	Odds Ratio (95% CI)
Medication_OtherMedication	2.75 (1.81 – 4.29)
TxProcedure_Other	2.56 (1.73 – 3.84)
TxProcedure_Bariatric	2.50 (1.69 – 3.75)
ImagingTest_OtherImagingTest	2.49 (1.65 – 3.85)
PregnancyHistory	2.43 (1.74 – 3.43)
TxProcedure_GenitouriaryReproductive	2.41 (1.63 – 3.62)
Medication_MentalHealth	2.38 (1.60 – 3.61)
TxProcedure_Cosmetic	2.37 (1.61 – 3.54)
Encounter_MRN	2.37 (1.62 – 3.52)
DiseaseCondition_Other	2.36 (1.60 – 3.56)
DiseaseCondition_SexualReproductive	2.31 (1.53 – 3.54)
DiseaseCondition_MentalHealth	2.27 (1.53 – 3.45)
LabTestResults_GeneticTest	2.27 (1.51 – 3.49)
AdoptionHistory	2.27 (1.65 – 3.15)
LabTestResults_Other	2.24 (1.50 – 3.40)
FamilyHealthHistory_SexualReproductive	2.23 (1.51 – 3.35)
Allergies	2.22 (1.41 – 3.58)
Encounter_PhysicianName	2.22 (1.50 – 3.31)
FamilyHealthHistory_Other	2.21 (1.50 – 3.30)
LabTestResults_DNAsequencing	2.20 (1.47 – 3.36)
ImagingTest_MRI	2.18 (1.42 – 3.41)
FamilyHealthHistory_MentalHealth	2.17 (1.46 – 3.28)
TxProcedure_MentalHealth	2.13 (1.46 – 3.17)
Lifestyle_SocialIsolation	2.12 (1.42 – 3.21)
Biospecimen_Tissue	2.11 (1.44 – 3.15)
ImagingTest_CTscan	2.11 (1.37 – 3.31)
Biospecimen_Urine	2.08 (1.41 – 3.12)
LabTestResults_DrugScreening	2.06 (1.38 – 3.12)
FamilyHealthHistory_SubstanceAbuse	2.06 (1.40 – 3.07)
VitalSigns	2.05 (1.31 – 3.29)
Biospecimen_Blood	2.05 (1.38 – 3.09)
ImagingTest_X.ray	2.04 (1.32 – 3.20)
DiseaseCondition_SubstanceAbuse	2.03 (1.37 – 3.05)
Lifestyle_Diet	2.02 (1.35 – 3.09)
Lifestyle_Stress	2.02 (1.35 – 3.09)
Encounter_VisitDate	2.02 (1.38 – 3.01)
BodyMeasurements	2.01 (1.33 – 3.09)
Demographics_SexualOrientation	2.00 (1.32 – 3.08)
SES_Occupation	2.00 (1.34 – 3.02)
ContactInfo_Email	1.97 (1.43 – 2.75)
Lifestyle_Drug	1.96 (1.34 – 2.91)
Encounter_ClinicLocation	1.95 (1.32 – 2.90)
LabTestResults_STDtest	1.94 (1.30 – 2.94)
Encounter_ClinicSpecialty	1.94 (1.31 – 2.91)
ContactInfo_Phone	1.92 (1.40 – 2.65)
Lifestyle_Alcohol	1.90 (1.29 – 2.84)
SexualLife	1.90 (1.38 – 2.62)
Lifestyle_Exercise	1.90 (1.25 – 2.92)
Demographics_Sex	1.89 (1.19 – 3.07)
Lifestyle_Smoking	1.88 (1.27 – 2.81)
ContactInfo_Name	1.87 (1.33 – 2.63)
Demographics_Ethnicity	1.86 (1.20 – 2.91)
SES_InsuranceStatus	1.85 (1.27 – 2.75)
ContactInfo_HomeAddress	1.84 (1.34 – 2.55)
SES_MaritalStatus	1.84 (1.22 – 2.81)
SES_Income	1.83 (1.30 – 2.58)
SES_Education	1.81 (1.22 – 2.72)
Demographics_Age	1.77 (1.11 – 2.90)
Demographics_Race	1.67 (1.08 – 2.62)

Figure 5.2: Forest plot of unadjusted odds ratio for opting method (opt-out vs. opt-in). 59 sharable items or categories were sorted by odds ratios and shown with their 95% confidence intervals. For each sharable variable, a 2-by-2 table was constructed using binary outcome (shared vs. not shared) and binary exposure variable, opting method (opt-out vs. opt-in). Then the odds ratio and its 95% confidence interval were calculated. Row labels are expressed in the format of "Category_Item". A category with no item was expressed simply as "Category." Abbreviations: CT, Computerized Tomography; DNA, DeoxyriboNucleic Acid; MRI, Magnetic Resonance Imaging; MRN, Medical Record Number; SES, Social Economic Status; STD, Sexually Transmitted Disease; TxProcedure, Treatment Procedure.

Participants over 60 years old or deemed to have an adequate health literacy level were more willing to share more items than were their counterparts (Supplementary Figures B.4 and B.5). Household income, education level, gender, perceived health status, race and site were not associated with a higher level of sharing for most variables (Supplementary Figures B.8-B.13)

The majority of participants (850, or 68.2%) completed the satisfaction survey. Of these, 815 (95.9%), had no trouble understanding the data/information presented in the forms, while 17 (2.0%) felt that the choices in the forms were inadequate. The vast majority, 837 (98.4%), enjoyed participating in the study. A large number of respondents (517, 60.8%) indicated that having a detailed form layout to make selections had no influence on their sharing decisions, 288 (33.9%) indicated it made them more willing to share, and 27 (3.2%) indicated it made them less willing to share their data and biospecimens. The remaining 2.1% were indifferent (i.e., they did not answer this question). Consistent with previous findings,21 most respondents were highly interested in knowing who would use the data/biospecimens (637, 74.9%) and were also equally willing to share their data/biospecimens for research and healthcare (683, 80.3%).

5.5 DISCUSSION

The finding in this study that most patients are willing to share the majority of their EHRs and biospecimens with researchers is reassuring. Not only biomedical research can benefit from these resources but also a multi-site learning healthcare system[98,117] can continuously advance as a result of data-driven improvements to process and associated outcomes. The finding that 955 (76.6%) participants made sharing choices to select at least one item that they did not want to share with a particular type of researchers is important when considering that this item might lead to a decision to decline sharing of the whole record if only an “all-or-nothing” option is available. This is important because the item to withhold may not be of relevance to a certain study, but the current “all-or-nothing” option, if exercised, would remove that patient’s data from all research studies. The finding that 291 (23.4%) participants wanted to share all items with everyone can help plan for studies based on EHRs and biospecimens that are expected to be shared with a broad range of researchers. The finding that only a small percentage do not want to share any item (46, 3.7%) is also reassuring. Opt-in forms appear to be the most conservative opting method to obtain sharing preferences, resulting in less sharing.

An important finding of this study is that the majority of participants indicated at least one item that should not be shared. There was a clear preference to share the data and biospecimens within the institution in which the patient received care, followed by non-profit institutions. In a system in which people can choose where to receive care it seems plausible that a patient elects to receive care in the most trusted institution, and this trust may more easily transfer to the care of data and biospecimens.

The reluctance to share data and biospecimens with researchers from for-profit institutions needs further investigation, as the category aggregates highly different industries and further refinement might reveal sub-groups that have higher association with decline to share than others. Strategies to convey how data and biospecimens are being used or will be used for research that includes the development of commercial products to improve health outcomes need to be developed and implemented so patients can provide consent that is truly “informed.”

Finally, studies requiring permission to use the whole EHR for research may consider provisions for participants to decline sharing of specific items and for participants to specify the types of researchers who should be authorized to work with their data. This may increase participation and satisfaction.

This study has some limitations. Patients who elect to receive care at academic medical centers may be more familiar with research and more willing to share their data and biospecimens for research than patients who receive care in other types of institutions. Also, health literacy in general was relatively high in our sample, so the results may be optimistic. However, this optimistic figure may be counterbalanced by the fact that patients who participated in our study may be more concerned about data/biospecimen sharing than those who declined participation, so our recruitment may have selected for individuals who would be in general more concerned about the privacy of their data and biospecimens, and thus tended to remove more items that would those who did not want to participate in this study. There could also be geographical factors: both institutions were located in California, where privacy protections for EHRs are higher than in many other states and in which biomedical and data science research are very prominent. However, these limitations do not detract from the fact that we were able to show that it was practical to implement a system that used patient data and biospecimen sharing preferences to guide services that make these resources available for research, and that the majority of patients were willing to share their EHR data and biospecimens for research.

5.6 CONCLUSION

We showed that a tiered-permission system that allowed for specific removal of data items or categories of data could be implemented in practice and showed that it matters to participants with whom the EHRs data and biospecimens will be shared, as there were differences in sharing preferences according to the researchers' affiliations. Participants appreciated being asked about their data and biospecimen sharing preferences. We also showed that the way in which patients' sharing preferences are elicited matters. In this study data and biospecimen sharing preferences were equivalent across institutions but were different according to the opting method (an opt-out version resulted in more sharing than an opt-in version). A simple form layout displaying data categories was associated with sharing preferences that were equivalent to those elicited from a detailed form layout displaying specific data items.

5.7 ACKNOWLEDGEMENTS

This work was supported by grants R01HG008802 and UL1TR001442 from the National Institutes of Health.

Chapter 5, in full, is a reprint of the material as it appears in "Patient Perspectives About Decisions to Share Medical Data and Biospecimens for Research." Jihoon Kim, Hyeoneui Kim, Elizabeth Bell, Tyler Bath, Paulina Paul, Anh Pham, Xiaoqian Jiang, Kai Zheng, Lucila Ohno-Machado. JAMA Network Open. 2019 Aug 2;2(8):e199550. doi: 10.1001/jamanetworkopen.2019.9550. The dissertation author was the primary investigator and the first author of this paper.

Chapter 6 Final Remarks

In the study of horizontally partitioned data described in Chapter 2, we demonstrated that it was practical to answer questions about COVID-19 using EHR data from systems that had different policies and must follow various regulations, without moving individual-level data out of these health systems. We were able to generate descriptive statistics and build a multivariate, iterative regression model without centralizing individual-level data. Our best practice workflow and open-source code for cohort definition, data harmonization, data quality assurance, and federated learning are generalizable to clinical research beyond COVID-19.

With a vertically partitioned data used in Chapter 3 we implemented a federated learning algorithm for logistic regression by reconstructing the covariance matrix in the dual space using a novel ring-structure protocol. We showed the equivalent accuracy and tolerable runtime compared to the centralized version on both synthetic and real datasets. Unlike other federated algorithms that produce only coefficients, our method produced nearly identical results for standard error, Z-score, and p-value as well as confidence interval for each coefficient.

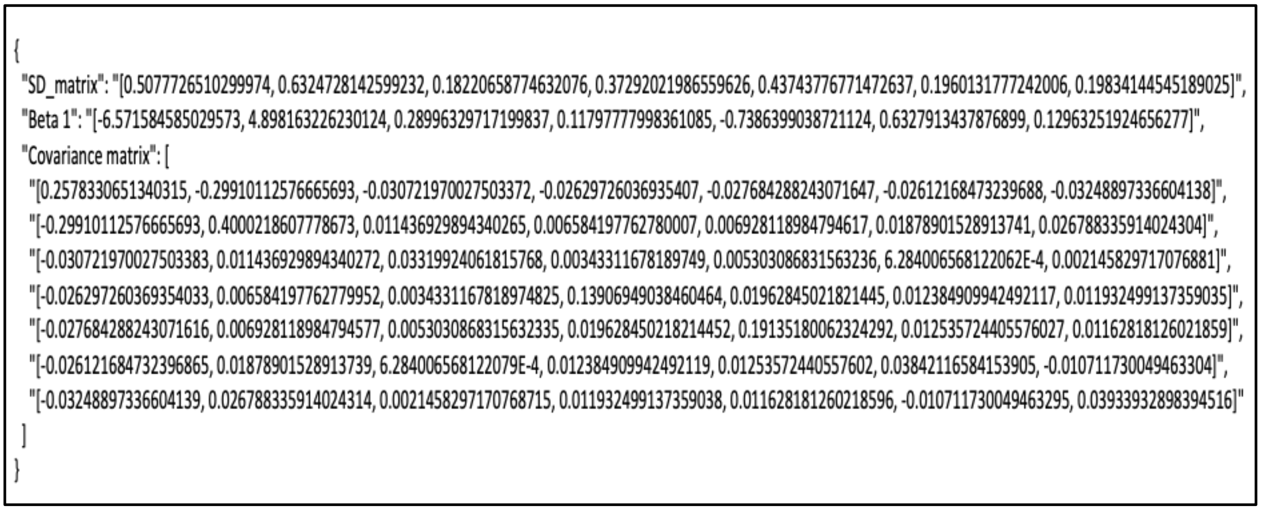
Federated learning algorithms assume that outcome and features variables already have been harmonized, which is still not true with many healthcare research databases. In Chapter 4 we showed that it is feasible for SHS personnel to transform data from their longitudinal cohort of American Indians into a common data model. The novel aspect here was that this was done in partnership with an external collaborating team that was not granted access to individual-level data, thus keeping participants' trust in the data stewards. Our workflow, code set, and open-source scripts created for the American Indian registry could be applied to OMOP transformation of databases containing sensitive patient information such as addiction and substance abuse, disability, mental health, minor/children, reproductive care, and sexual disease or condition.

In Chapter 5 we report on a study that analyzed patient preferences for data sharing. Different sharing rates were associated with the informed consent form's opting method (opt-out vs. opt-in), but not with the form layout (detailed vs. simple) after controlling for several covariates. Participants were willing to share more items when they used the opt-out form. It was reassuring to see that most patients were willing to share data from their EHRs and biospecimens with researchers and only a small percentage of participants was not willing to share any item. We demonstrated that a tiered-permission system that allowed for specific removal of data items or categories of data could be implemented in practice and that it mattered to participants with whom the EHR data and biospecimens would be shared as shown by the differences in sharing preferences according to the researchers' affiliations.

As a future work, we envision to build an end-to-end analysis workflow that starts from honoring patient's data sharing preference through tiered informed consent system, retrieves the records from multiple harmonized databases that use a common data model, performs data quality check and assurance steps, runs federated learning algorithms of prediction/association, and returns the research results back to the patients without allowing access to individual-level data throughout the cycle. We are planning to extend our federated algorithms to be able to handle a very large number of features (e.g., omics data) and large biobank-scale number (e.g., over million) of patients/participants, while incorporating correlations among individuals for clustered and multi-level data. In addition, extension of the existing OMOP CDM to incorporate non-canonical data such as omics, survey, or imaging data is warranted to improve the data coverage. Looking forward, patient data sharing preferences could be more comprehensive if the sharing choice of data items could be stated at a more granular level for each institution, and varied by study, and also could be associated with specific duration of consent. Further studies should include a large number of patients with more diverse backgrounds.

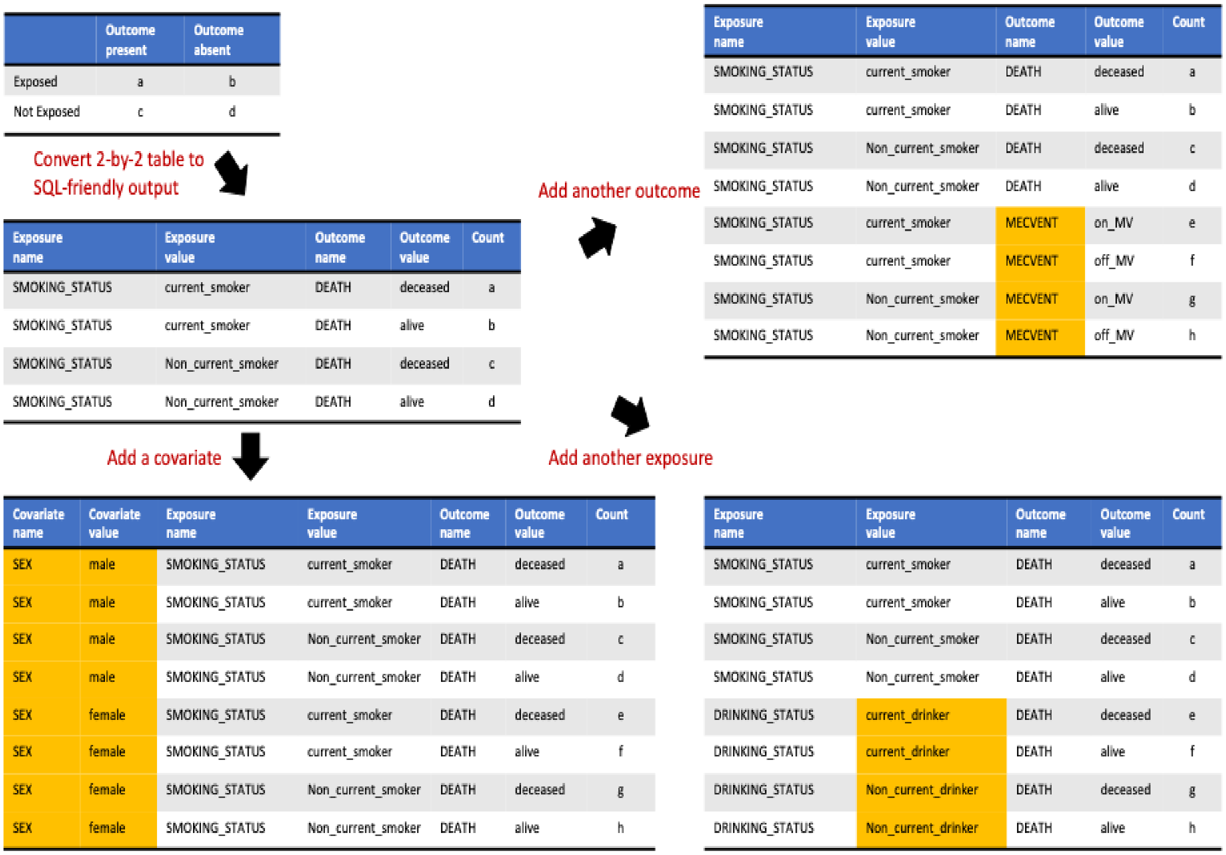
In summary, I developed new algorithms, approaches, and applications of informatics to the real world problems of computing with horizontally and vertically partitioned data, having experts tutor a team to harmonize data without having access to the individual-level data, and eliciting patient data sharing preferences.

Appendix A   Supplemental Material for Chapter 3



```
{
 "SD_matrix": "[0.5077726510299974, 0.6324728142599232, 0.18220658774632076, 0.37292021986559626, 0.43743776771472637, 0.1960131777242006, 0.19834144545189025]",
 "Beta 1": "[-6.571584585029573, 4.898163226230124, 0.28996329717199837, 0.11797777998361085, -0.7386399038721124, 0.6327913437876899, 0.12963251924656277]",
 "Covariance matrix": [
  "[0.2578330651340315, -0.29910112576665693, -0.030721970027503372, -0.02629726036935407, -0.027684288243071647, -0.02612168473239688, -0.03248897336604138]",
  "[-0.29910112576665693, 0.4000218607778673, 0.011436929894340265, 0.006584197762780007, 0.006928118984794617, 0.01878901528913741, 0.026788335914024304]",
  "[-0.030721970027503383, 0.011436929894340272, 0.03319924061815768, 0.00343311678189749, 0.005303086831563236, 6.284006568122062E-4, 0.002145829717076881]",
  "[-0.026297260369354033, 0.006584197762779952, 0.0034331167818974825, 0.13906949038460464, 0.01962845021821445, 0.012384909942492117, 0.011932499137359035]",
  "[-0.027684288243071616, 0.006928118984794577, 0.0053030868315632335, 0.019628450218214452, 0.19135180062324292, 0.012535724405576027, 0.01162818126021859]",
  "[-0.026121684732396865, 0.01878901528913739, 6.284006568122079E-4, 0.012384909942492119, 0.01253572440557602, 0.03842116584153905, -0.010711730049463304]",
  "[-0.03248897336604139, 0.026788335914024314, 0.0021458297170768715, 0.011932499137359038, 0.011628181260218596, -0.010711730049463295, 0.03933932898394516]"
 ]
}
```

Figure A.1: Transferred file in the federated logistic regression. A real example of a transferred file in JSON format among the Consortium Hub and participating sites is shown to illustrate how patient level data are protected. The first row is a vector of standard deviations of the coefficients of the federated logistic regression. The second row is a vector of coefficients. The third row is a covariance matrix between features. At each iteration of the federated logistic regression, this JSON file is being transferred among sites until convergence or the process reaches the predetermined number of iterations. The values in the JSON file above are used in the forest plot of Figure 3.2 in the main manuscript.



	Outcome present	Outcome absent
Exposed	a	b
Not Exposed	c	d

Convert 2-by-2 table to SQL-friendly output

Exposure name	Exposure value	Outcome name	Outcome value	Count
SMOKING_STATUS	current_smoker	DEATH	deceased	a
SMOKING_STATUS	current_smoker	DEATH	alive	b
SMOKING_STATUS	Non_current_smoker	DEATH	deceased	c
SMOKING_STATUS	Non_current_smoker	DEATH	alive	d

Add another outcome

Exposure name	Exposure value	Outcome name	Outcome value	Count
SMOKING_STATUS	current_smoker	DEATH	deceased	a
SMOKING_STATUS	current_smoker	DEATH	alive	b
SMOKING_STATUS	Non_current_smoker	DEATH	deceased	c
SMOKING_STATUS	Non_current_smoker	DEATH	alive	d
SMOKING_STATUS	current_smoker	MECVENT	on_MV	e
SMOKING_STATUS	current_smoker	MECVENT	off_MV	f
SMOKING_STATUS	Non_current_smoker	MECVENT	on_MV	g
SMOKING_STATUS	Non_current_smoker	MECVENT	off_MV	h

Add a covariate

Covariate name	Covariate value	Exposure name	Exposure value	Outcome name	Outcome value	Count
SEX	male	SMOKING_STATUS	current_smoker	DEATH	deceased	a
SEX	male	SMOKING_STATUS	current_smoker	DEATH	alive	b
SEX	male	SMOKING_STATUS	Non_current_smoker	DEATH	deceased	c
SEX	male	SMOKING_STATUS	Non_current_smoker	DEATH	alive	d
SEX	female	SMOKING_STATUS	current_smoker	DEATH	deceased	e
SEX	female	SMOKING_STATUS	current_smoker	DEATH	alive	f
SEX	female	SMOKING_STATUS	Non_current_smoker	DEATH	deceased	g
SEX	female	SMOKING_STATUS	Non_current_smoker	DEATH	alive	h

Add another exposure

Exposure name	Exposure value	Outcome name	Outcome value	Count
SMOKING_STATUS	current_smoker	DEATH	deceased	a
SMOKING_STATUS	current_smoker	DEATH	alive	b
SMOKING_STATUS	Non_current_smoker	DEATH	deceased	c
SMOKING_STATUS	Non_current_smoker	DEATH	alive	d
DRINKING_STATUS	current_drinker	DEATH	deceased	e
DRINKING_STATUS	current_drinker	DEATH	alive	f
DRINKING_STATUS	Non_current_drinker	DEATH	deceased	g
DRINKING_STATUS	Non_current_drinker	DEATH	alive	h

Figure A.2: Extensible output format of site level results. A 2-by-2 table of exposure-outcome association is implemented as a 4-row format in SQL to store the binary exposure and the binary outcome question. The table will expand to add a covariate (Sex), a second outcome (Mechanical Ventilation), or another exposure (drinking status). A name-value format was adopted for clarity and efficiency during data quality check and aggregation.

Table A.1: Pre-coordinated diagnosis codes. Diagnosis codes (OMOP Extension, SNOMED) used to identify patients with COVID-19. At least one occurrence of the diagnosis codes during a hospital encounter with a look back period of 21 days prior to hospitalization captured the patient having a COVID-19 related diagnosis. In contrast to the joint diagnosis codes (ICD-10-CM, SNOMED) mentioned in Figure 3A, there was no further applied logic.

Concept Class Id	Vocabulary Id	Concept Code	Concept Id	Concept Name
Clinical Finding	OMOP Extension	OMOP4873906	756023	Acute bronchitis due to COVID-19
Clinical Finding	OMOP Extension	OMOP4873911	756044	Acute respiratory distress syndrome (ARDS) due to COVID-19
Clinical Finding	OMOP Extension	OMOP4873910	756061	Asymptomatic COVID-19
Clinical Finding	OMOP Extension	OMOP4873909	756031	Bronchitis due to COVID-19
Clinical Finding	SNOMED	1240561000000108	37310284	Encephalopathy caused by 2019 novel coronavirus
Clinical Finding	SNOMED	1240571000000101	37310283	Gastroenteritis caused by 2019 novel coronavirus
Clinical Finding	OMOP Extension	OMOP4873908	756081	Infection of lower respiratory tract due to COVID-19
Clinical Finding	SNOMED	1240541000000107	37310286	Infection of upper respiratory tract caused by 2019 novel coronavirus
Clinical Finding	SNOMED	1240531000000103	37310287	Myocarditis caused by 2019 novel coronavirus
Clinical Finding	SNOMED	1240521000000100	37310254	Otitis media caused by 2019 novel coronavirus
Clinical Finding	SNOMED	1240551000000105	37310285	Pneumonia caused by 2019 novel coronavirus
Clinical Finding	OMOP Extension	OMOP4873907	756039	Respiratory infection due to COVID-19

Table A.2: Excluded diagnosis codes. These ICD-10-CM Codes, mentioned in CDC guideline, were excluded given the high count of false positive COVID-19 patients.

ICD-10-CM Code	**Description**
R05	Cough
R06.02	Shortness of breath
R50.9	Fever, unspecified
Z20.828	Contact with and (suspected) exposure to other viral communicable diseases

Table A.3: Other excluded diagnosis codes. These ICD10CM and SNOMED Concepts were mentioned in N3C - COVID-19 Phenotype Documentation, Version 1.6 (Last updated 6/5/2020) but excluded from our study, as the count of false positives COVID-19 patients was too high.

Concept Class Id	**Vocabulary Id**	**Concept Code**	**Concept Id**	**Concept Name**
Condition	ICD10CM	J96.2	35208101	Acute and chronic respiratory failure
Condition	ICD10CM	J96.22	45581868	Acute and chronic respiratory failure with hypercapnia
Condition	ICD10CM	J96.21	45543283	Acute and chronic respiratory failure with hypoxia
Condition	ICD10CM	J96.20	45596290	Acute and chronic respiratory failure, unspecified whether with hypoxia or hypercapnia
Condition	ICD10CM	J21	1569471	Acute bronchiolitis
Condition	ICD10CM	J21.1	920135	Acute bronchiolitis due to human metapneumovirus
Condition	ICD10CM	J21.8	35207968	Acute bronchiolitis due to other specified organisms
Condition	ICD10CM	J21.0	35207967	Acute bronchiolitis due to respiratory syncytial virus
Condition	ICD10CM	J21.9	35207969	Acute bronchiolitis, unspecified
Condition	ICD10CM	J20	1569470	Acute bronchitis
Condition	ICD10CM	J20.3	35207960	Acute bronchitis due to coxsackievirus
Condition	ICD10CM	J20.7	35207964	Acute bronchitis due to echovirus
Condition	ICD10CM	J20.1	35207958	Acute bronchitis due to Hemophilus influenzae
Condition	ICD10CM	J20.0	35207957	Acute bronchitis due to Mycoplasma pneumoniae
Condition	ICD10CM	J20.8	35207965	Acute bronchitis due to other specified organisms
Condition	ICD10CM	J20.4	35207961	Acute bronchitis due to parainfluenza virus
Condition	ICD10CM	J20.5	35207962	Acute bronchitis due to respiratory syncytial virus
Condition	ICD10CM	J20.6	35207963	Acute bronchitis due to rhinovirus
Condition	ICD10CM	J20.2	35207959	Acute bronchitis due to streptococcus
Condition	ICD10CM	J20.9	35207966	Acute bronchitis, unspecified
Condition	ICD10CM	R06.03	1326788	Acute respiratory distress

Table A.3: (continued) Other excluded diagnosis codes.

Concept Class Id	Vocabulary Id	Concept Code	Concept Id	Concept Name
Condition	ICD10CM	J80	35208069	Acute respiratory distress syndrome
Condition	ICD10CM	J96.0	35208099	Acute respiratory failure
Condition	ICD10CM	J96.02	45596289	Acute respiratory failure with hypercapnia
Condition	ICD10CM	J96.01	45567283	Acute respiratory failure with hypoxia
Condition	ICD10CM	J96.00	45605906	Acute respiratory failure, unspecified whether with hypoxia or hypercapnia
Condition	ICD10CM	J06.9	35207929	Acute upper respiratory infection, unspecified
Condition	ICD10CM	J12.0	35207932	Adenoviral pneumonia
Condition	ICD10CM	R43.0	35211351	Anosmia
Condition	ICD10CM	J40	35208013	Bronchitis, not specified as acute or chronic
Condition	ICD10CM	J18.0	35207952	Bronchopneumonia, unspecified organism
Condition	ICD10CM	R07.1	35211284	Chest pain on breathing
Condition	ICD10CM	R68.83	45577807	Chills (without fever)
Condition	ICD10CM	J96.1	35208100	Chronic respiratory failure
Condition	ICD10CM	J96.12	45572177	Chronic respiratory failure with hypercapnia
Condition	ICD10CM	J96.11	45538489	Chronic respiratory failure with hypoxia
Condition	ICD10CM	J96.10	45548131	Chronic respiratory failure, unspecified whether with hypoxia or hypercapnia
Observation	ICD10CM	Z20.828	45542411	Contact with and (suspected) exposure to other viral communicable diseases
Condition	ICD10CM	B34.2	35205800	Coronavirus infection, unspecified
Condition	ICD10CM	R05	35211275	Cough
Condition	ICD10CM	R50.2	35211385	Drug induced fever
Condition	ICD10CM	R06.0	1572191	Dyspnea

Table A.3: (continued) Other excluded diagnosis codes.

Concept Class Id	Vocabulary Id	Concept Code	Concept Id	Concept Name
Condition	ICD10CM	R06.00	45587496	Dyspnea, unspecified
Condition	ICD10CM	R50.84	45597190	Febrile nonhemolytic transfusion reaction
Condition	ICD10CM	R50	1572254	Fever of other and unknown origin
Condition	ICD10CM	R50.81	45606818	Fever presenting with conditions classified elsewhere
Condition	ICD10CM	R50.9	35211387	Fever, unspecified
Condition	ICD10CM	J12.3	35207935	Human metapneumovirus pneumonia
Condition	ICD10CM	J18.2	35207954	Hypostatic pneumonia, unspecified organism
Condition	ICD10CM	J18.1	35207953	Lobar pneumonia, unspecified organism
Condition	ICD10CM	R06.01	45597165	Orthopnea
Condition	ICD10CM	R06.09	45548944	Other forms of dyspnea
Condition	ICD10CM	J18.8	35207955	Other pneumonia, unspecified organism
Condition	ICD10CM	R50.8	35211386	Other specified fever
Condition	ICD10CM	J98.8	35208108	Other specified respiratory disorders
Condition	ICD10CM	J12.8	35207936	Other viral pneumonia
Condition	ICD10CM	R43.2	35211353	Parageusia
Condition	ICD10CM	J12.2	35207934	Parainfluenza virus pneumonia
Condition	ICD10CM	J12.81	45567260	Pneumonia due to SARS-associated coronavirus
Condition	ICD10CM	J18.9	35207956	Pneumonia, unspecified organism
Condition	ICD10CM	J18	1569469	Pneumonia, unspecified organism
Condition	ICD10CM	R50.82	45597189	Postprocedural fever
Condition	ICD10CM	R50.83	45592424	Postvaccination fever

Table A.3: (continued) Other excluded diagnosis codes.

Concept Class Id	Vocabulary Id	Concept Code	Concept Id	Concept Name
Condition	ICD10CM	J96	1569515	Respiratory failure, not elsewhere classified
Condition	ICD10CM	J96.9	35208102	Respiratory failure, unspecified
Condition	ICD10CM	J96.92	45533563	Respiratory failure, unspecified with hypercapnia
Condition	ICD10CM	J96.91	45605907	Respiratory failure, unspecified with hypoxia
Condition	ICD10CM	J96.90	45567284	Respiratory failure, unspecified, unspecified whether with hypoxia or hypercapnia
Condition	ICD10CM	J12.1	35207933	Respiratory syncytial virus pneumonia
Condition	ICD10CM	R06.02	45534422	Shortness of breath
Condition	ICD10CM	J12	1569465	Viral pneumonia, not elsewhere classified
Condition	ICD10CM	J12.9	35207937	Viral pneumonia, unspecified
Condition	SNOMED	75483001	442555	Breathing painful
Condition	SNOMED	161940008	4059022	Breathless - mild exertion
Condition	SNOMED	161939006	4059021	Breathless - moderate exertion
Condition	SNOMED	161855003	4059003	C/O shivering
Condition	SNOMED	274664007	4168213	Chest pain on breathing
Condition	SNOMED	43724002	434490	Chill
Condition	SNOMED	49727002	254761	Cough
Condition	SNOMED	135883003	4048098	Cough with fever
Condition	SNOMED	11833005	4038519	Dry cough
Condition	SNOMED	267036007	312437	Dyspnea
Condition	SNOMED	161941007	4060052	Dyspnea at rest
Condition	SNOMED	60845006	4263848	Dyspnea on exertion
Observation	SNOMED	840546002	37311059	Exposure to 2019 novel coronavirus
Condition	SNOMED	103001002	4011766	Feeling feverish

Table A.3: (continued) Other excluded diagnosis codes.

Concept Class Id	Vocabulary Id	Concept Code	Concept Id	Concept Name
Condition	SNOMED	386661006	437663	Fever
Measurement	SNOMED	426000000	4141062	Fever greater than 100.4 Fahrenheit
Condition	SNOMED	274640006	4164645	Fever with chills
Condition	SNOMED	23141003	4047610	Gasping for breath
Condition	SNOMED	409702008	4260205	Hyperpyrexia
Condition	SNOMED	44169009	4185711	Loss of sense of smell
Condition	SNOMED	36955009	4289517	Loss of taste
Condition	SNOMED	426976009	4140453	Pain provoked by breathing
Condition	SNOMED	247410004	4090569	Painful cough
Condition	SNOMED	284523002	4109381	Persistent cough
Condition	SNOMED	2237002	4330445	Pleuritic pain
Condition	SNOMED	28743005	4102774	Productive cough

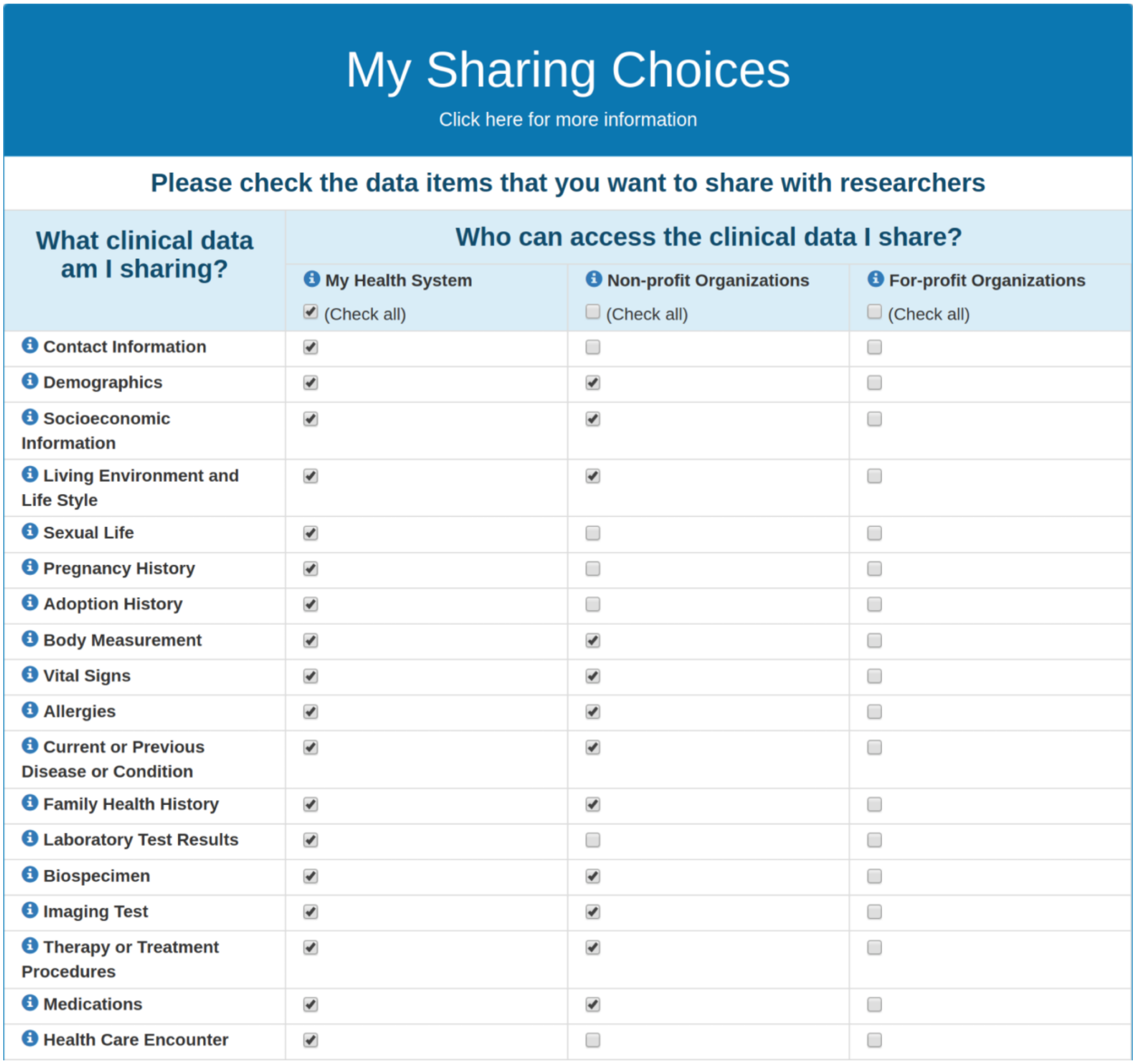
Text A.1: Comparison to other consortia.

The R2D2 consortium is similar to N3C and OHDSI in that OMOP is used as a common data model. R2D2 and 4CE are similar in that both are distributed networks, do not disclose patient level data, and provide the time trend of COVID-19 related metrics on the public website. The R2D2 differs from other consortia (e.g., 4CE, N3C, and OHDSI) in five main points: (1) R2D2 allows the general public to ask questions, (2) Patient privacy is protected by sharing only aggregate level data and adoption of privacy-preserving federated regression method such as GLORE, and (3) our iterative workflow processes with a focus on decentralization and data quality checks lead to an amended data harmonization and high sensitivity and specificity in query results and increased site-level capacity building and independence, with support from the whole consortium. Similar to OHDSI, but unlike N3C and 4CE, (1) the use of EHR data empowers our network to answer questions which include both COVID-19 patient and non-COVID-19 or pre-COVID-19 patients (e.g., 'For the previous 24 months, what are the monthly counts of encounters for breast cancer screening and are there disparities in the patterns as a result of the COVID-19 pandemic?') and (2) full transparency is granted by sharing the finalized SQL codes on public web pages hosted through GitHub and their related results on our webpage. Additionally, the main difference with N3C is architectural: in R2D2 sites do not need to transmit data to a central repository. Similar to OHDSI and 4CE, we utilize a distributed approach in order to attend to the regulations at some institutions. While OHDSI is better suited to perform in-depth analyses for a certain number of questions over longer periods of time, research questions are developed inside the OHDSI consortia, R2D2 provides more shallow information by responding to a larger number of questions requested by the public.

Unlike other important initiatives such as ACT (used by several CTSAs) that also intend to respond to a larger number of questions in a short time, the questions are expressed in natural language, making the approach more flexible although also more manual.

Despite these differences, multiple consortia could work together towards the common goal. First, sharing concept sets and SQL code through a public code repository would be one good starting place. Each consortium would simply download the script generated by another consortium and run it to reproduce and validate the early findings. Second, the documentation of data quality improvement as a knowledge base would be another incentive for different consortia to work together. In our experience, harmonization of measurement values like D-dimer and Vitamin-D took a lengthy process of lab test review, running SQL codes in multiple versions, investigation to the EHR system, updates to the ETL scripts, and brainstorming to understand different site-specific workflows. If each consortium could contribute to provide their best practice to the common knowledgebase, this would save time and efforts of other consortiums and sites. For the same reason, in this rapidly generated study, we did not provide the validation results of our findings against those of other consortia. Instead, we shared the concept sets, SQL code, and aggregate results for others to validate their results on ours. Next step is working with other networks and making concerted efforts to develop codes and validate results together.

Appendix B Supplemental Material for Chapter 5



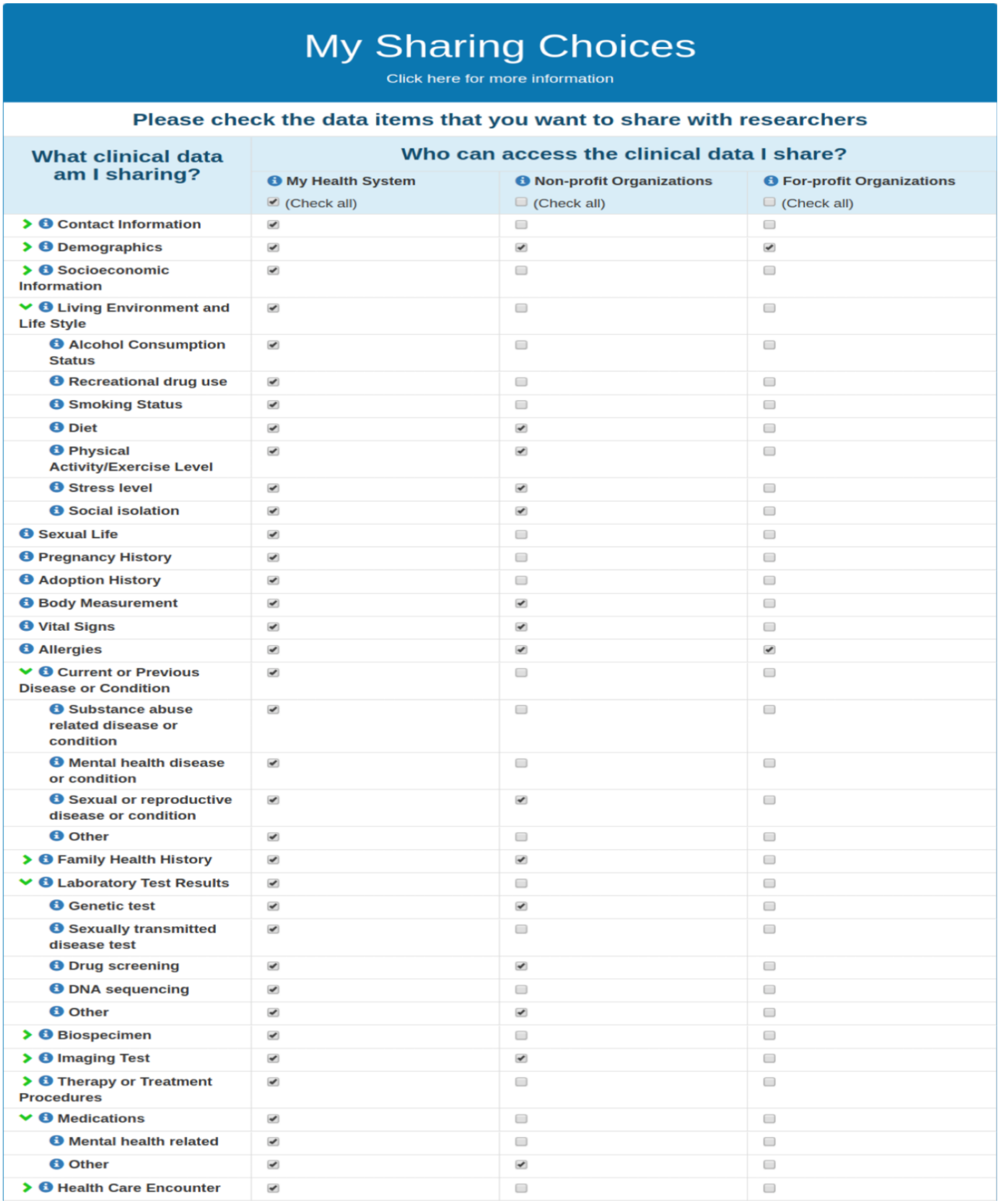
My Sharing Choices

Click here for more information

Please check the data items that you want to share with researchers

What clinical data am I sharing?	Who can access the clinical data I share?		
	ℹ My Health System ☑ (Check all)	ℹ Non-profit Organizations □ (Check all)	ℹ For-profit Organizations □ (Check all)
ℹ Contact Information	☑	□	□
ℹ Demographics	☑	☑	□
ℹ Socioeconomic Information	☑	☑	□
ℹ Living Environment and Life Style	☑	☑	□
ℹ Sexual Life	☑	□	□
ℹ Pregnancy History	☑	□	□
ℹ Adoption History	☑	□	□
ℹ Body Measurement	☑	☑	□
ℹ Vital Signs	☑	☑	□
ℹ Allergies	☑	☑	□
ℹ Current or Previous Disease or Condition	☑	☑	□
ℹ Family Health History	☑	☑	□
ℹ Laboratory Test Results	☑	□	□
ℹ Biospecimen	☑	☑	□
ℹ Imaging Test	☑	☑	□
ℹ Therapy or Treatment Procedures	☑	☑	□
ℹ Medications	☑	☑	□
ℹ Health Care Encounter	☑	□	□

Figure B.1: Screenshot of participant sharing selection form (opt-in and simple form layout)



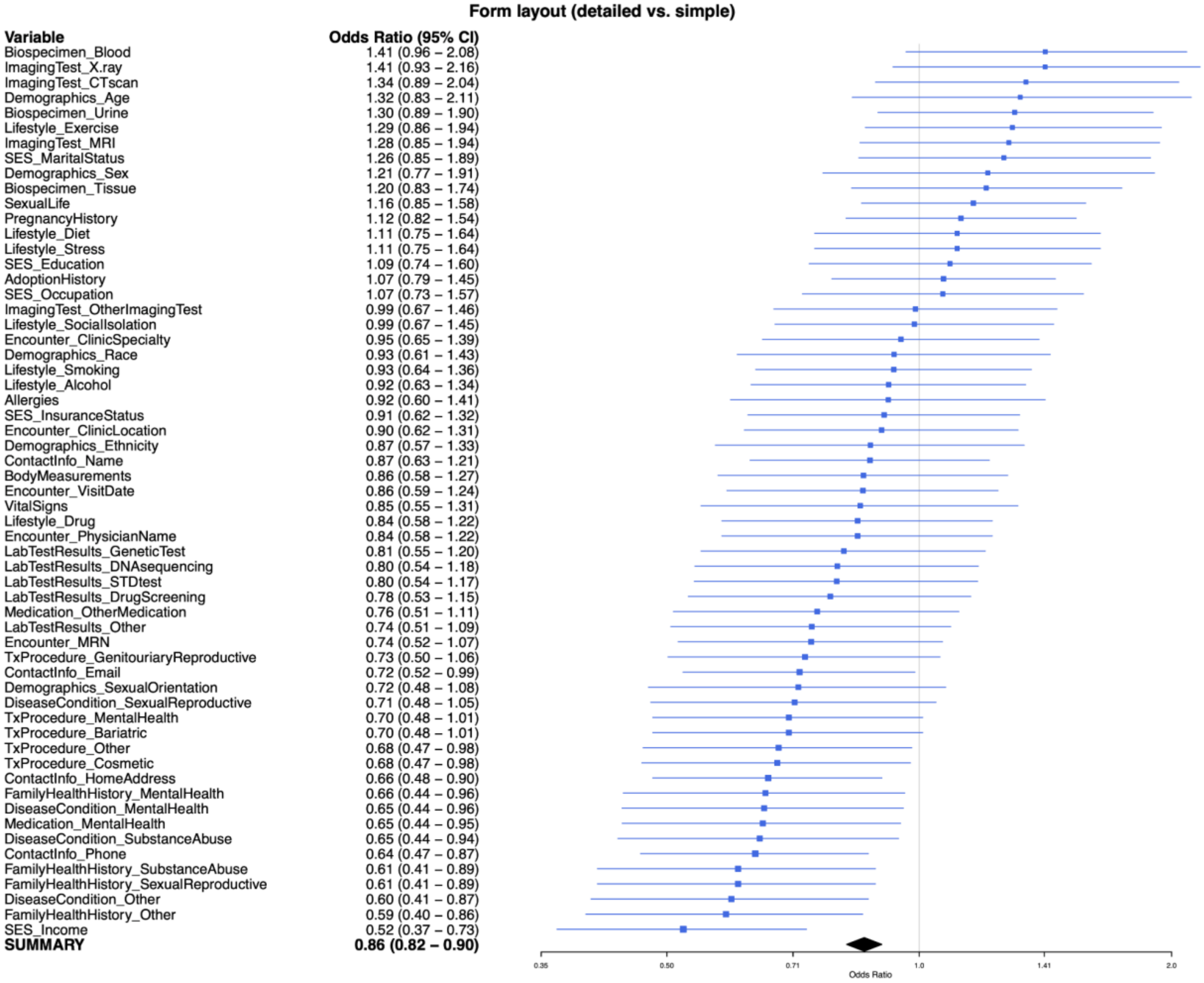
My Sharing Choices

Click here for more information

Please check the data items that you want to share with researchers

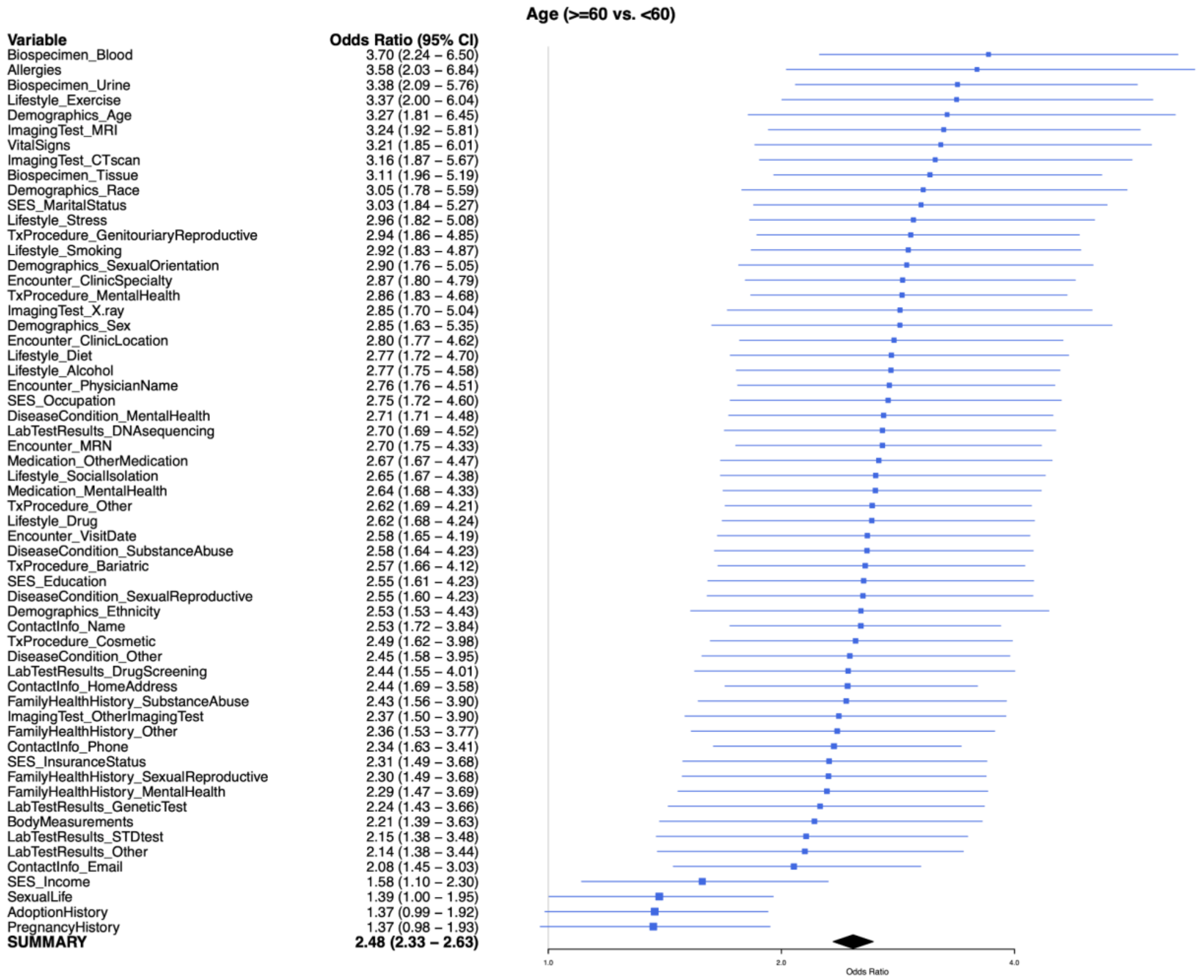
What clinical data am I sharing?	Who can access the clinical data I share?		
	My Health System	Non-profit Organizations	For-profit Organizations
	☑ (Check all)	□ (Check all)	□ (Check all)
> Contact Information	☑	□	□
> Demographics	☑	☑	☑
> Socioeconomic Information	☑	□	□
∨ Living Environment and Life Style	☑	□	□
Alcohol Consumption Status	☑	□	□
Recreational drug use	☑	□	□
Smoking Status	☑	□	□
Diet	☑	☑	□
Physical Activity/Exercise Level	☑	☑	□
Stress level	☑	☑	□
Social isolation	☑	☑	□
Sexual Life	☑	□	□
Pregnancy History	☑	□	□
Adoption History	☑	□	□
Body Measurement	☑	☑	□
Vital Signs	☑	☑	□
Allergies	☑	☑	☑
∨ Current or Previous Disease or Condition	☑	□	□
Substance abuse related disease or condition	☑	□	□
Mental health disease or condition	☑	□	□
Sexual or reproductive disease or condition	☑	☑	□
Other	☑	□	□
> Family Health History	☑	☑	□
∨ Laboratory Test Results	☑	□	□
Genetic test	☑	☑	□
Sexually transmitted disease test	☑	□	□
Drug screening	☑	☑	□
DNA sequencing	☑	□	□
Other	☑	☑	□
> Biospecimen	☑	□	□
> Imaging Test	☑	☑	□
> Therapy or Treatment Procedures	☑	□	□
∨ Medications	☑	□	□
Mental health related	☑	□	□
Other	☑	☑	□
> Health Care Encounter	☑	□	□

Figure B.2: Screenshot of participant sharing selection form (opt-in and detailed form layout)



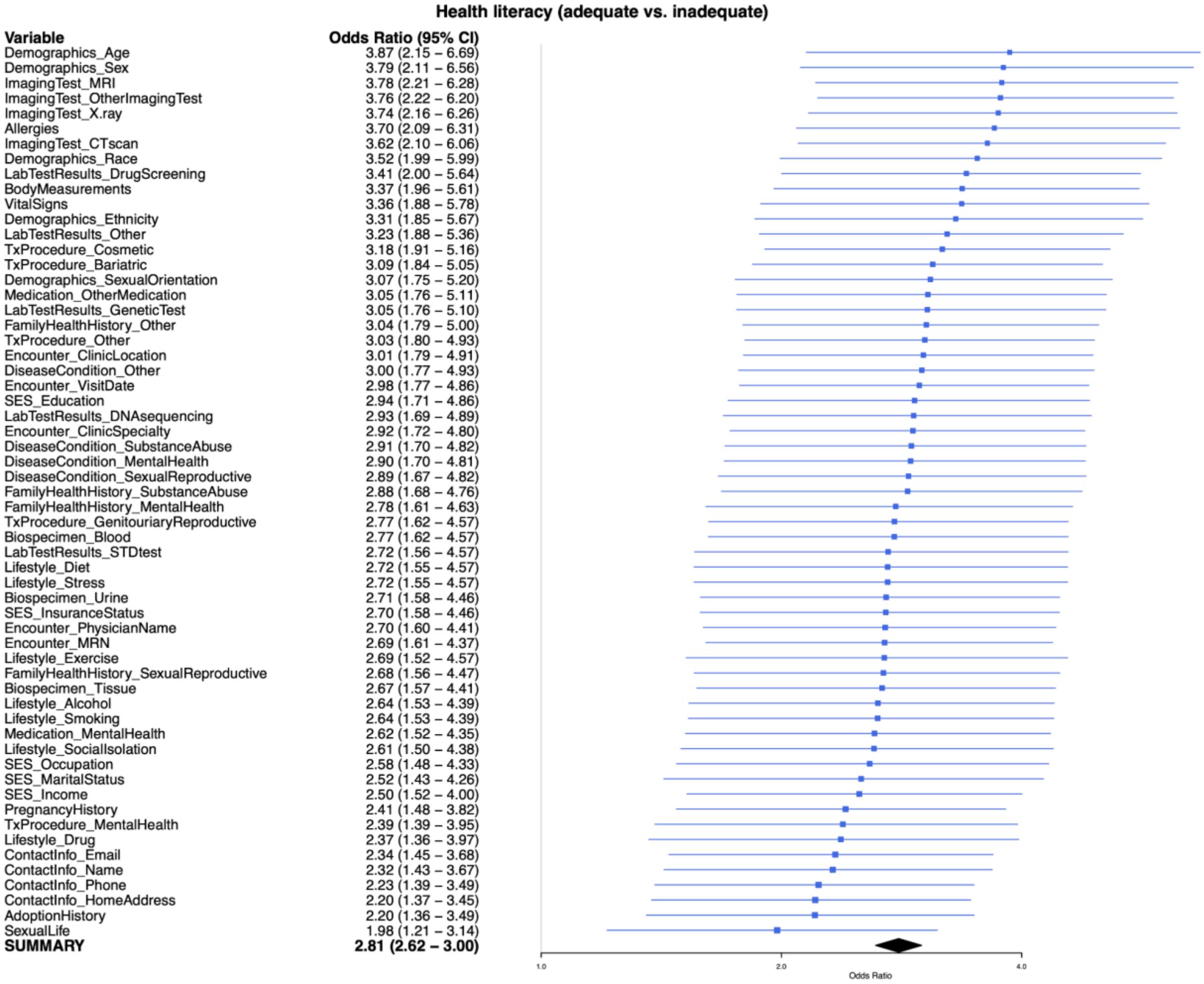
Variable	Odds Ratio (95% CI)
Biospecimen_Blood	1.41 (0.96 – 2.08)
ImagingTest_X.ray	1.41 (0.93 – 2.16)
ImagingTest_CTscan	1.34 (0.89 – 2.04)
Demographics_Age	1.32 (0.83 – 2.11)
Biospecimen_Urine	1.30 (0.89 – 1.90)
Lifestyle_Exercise	1.29 (0.86 – 1.94)
ImagingTest_MRI	1.28 (0.85 – 1.94)
SES_MaritalStatus	1.26 (0.85 – 1.89)
Demographics_Sex	1.21 (0.77 – 1.91)
Biospecimen_Tissue	1.20 (0.83 – 1.74)
SexualLife	1.16 (0.85 – 1.58)
PregnancyHistory	1.12 (0.82 – 1.54)
Lifestyle_Diet	1.11 (0.75 – 1.64)
Lifestyle_Stress	1.11 (0.75 – 1.64)
SES_Education	1.09 (0.74 – 1.60)
AdoptionHistory	1.07 (0.79 – 1.45)
SES_Occupation	1.07 (0.73 – 1.57)
ImagingTest_OtherImagingTest	0.99 (0.67 – 1.46)
Lifestyle_SocialIsolation	0.99 (0.67 – 1.45)
Encounter_ClinicSpecialty	0.95 (0.65 – 1.39)
Demographics_Race	0.93 (0.61 – 1.43)
Lifestyle_Smoking	0.93 (0.64 – 1.36)
Lifestyle_Alcohol	0.92 (0.63 – 1.34)
Allergies	0.92 (0.60 – 1.41)
SES_InsuranceStatus	0.91 (0.62 – 1.32)
Encounter_ClinicLocation	0.90 (0.62 – 1.31)
Demographics_Ethnicity	0.87 (0.57 – 1.33)
ContactInfo_Name	0.87 (0.63 – 1.21)
BodyMeasurements	0.86 (0.58 – 1.27)
Encounter_VisitDate	0.86 (0.59 – 1.24)
VitalSigns	0.85 (0.55 – 1.31)
Lifestyle_Drug	0.84 (0.58 – 1.22)
Encounter_PhysicianName	0.84 (0.58 – 1.22)
LabTestResults_GeneticTest	0.81 (0.55 – 1.20)
LabTestResults_DNAsequencing	0.80 (0.54 – 1.18)
LabTestResults_STDtest	0.80 (0.54 – 1.17)
LabTestResults_DrugScreening	0.78 (0.53 – 1.15)
Medication_OtherMedication	0.76 (0.51 – 1.11)
LabTestResults_Other	0.74 (0.51 – 1.09)
Encounter_MRN	0.74 (0.52 – 1.07)
TxProcedure_GenitouriaryReproductive	0.73 (0.50 – 1.06)
ContactInfo_Email	0.72 (0.52 – 0.99)
Demographics_SexualOrientation	0.72 (0.48 – 1.08)
DiseaseCondition_SexualReproductive	0.71 (0.48 – 1.05)
TxProcedure_MentalHealth	0.70 (0.48 – 1.01)
TxProcedure_Bariatric	0.70 (0.48 – 1.01)
TxProcedure_Other	0.68 (0.47 – 0.98)
TxProcedure_Cosmetic	0.68 (0.47 – 0.98)
ContactInfo_HomeAddress	0.66 (0.48 – 0.90)
FamilyHealthHistory_MentalHealth	0.66 (0.44 – 0.96)
DiseaseCondition_MentalHealth	0.65 (0.44 – 0.96)
Medication_MentalHealth	0.65 (0.44 – 0.95)
DiseaseCondition_SubstanceAbuse	0.65 (0.44 – 0.94)
ContactInfo_Phone	0.64 (0.47 – 0.87)
FamilyHealthHistory_SubstanceAbuse	0.61 (0.41 – 0.89)
FamilyHealthHistory_SexualReproductive	0.61 (0.41 – 0.89)
DiseaseCondition_Other	0.60 (0.41 – 0.87)
FamilyHealthHistory_Other	0.59 (0.40 – 0.86)
SES_Income	0.52 (0.37 – 0.73)
SUMMARY	**0.86 (0.82 – 0.90)**

Figure B.3: Forest plot of unadjusted odds ratio for form layout (detailed vs. simple) The 59 sharable items were sorted by unadjusted odds ratios and shown with their 95% confidence intervals. The pooled odds ratio is displayed at the bottom and labeled SUMMARY. Row labels are expressed in the format of "Category_Item". A category with no item was expressed simply as "Category." Abbreviations: CT, Computerized Tomography; DNA, DeoxyriboNucleic Acid; MRI, Magnetic Resonance Imaging; MRN, Medical Record Number; SES, Social Economic Status; STD, Sexually Transmitted Disease; TxProcedure, Treatment Procedure.



Variable	Odds Ratio (95% CI)
Biospecimen_Blood	3.70 (2.24 – 6.50)
Allergies	3.58 (2.03 – 6.84)
Biospecimen_Urine	3.38 (2.09 – 5.76)
Lifestyle_Exercise	3.37 (2.00 – 6.04)
Demographics_Age	3.27 (1.81 – 6.45)
ImagingTest_MRI	3.24 (1.92 – 5.81)
VitalSigns	3.21 (1.85 – 6.01)
ImagingTest_CTscan	3.16 (1.87 – 5.67)
Biospecimen_Tissue	3.11 (1.96 – 5.19)
Demographics_Race	3.05 (1.78 – 5.59)
SES_MaritalStatus	3.03 (1.84 – 5.27)
Lifestyle_Stress	2.96 (1.82 – 5.08)
TxProcedure_GenitouriaryReproductive	2.94 (1.86 – 4.85)
Lifestyle_Smoking	2.92 (1.83 – 4.87)
Demographics_SexualOrientation	2.90 (1.76 – 5.05)
Encounter_ClinicSpecialty	2.87 (1.80 – 4.79)
TxProcedure_MentalHealth	2.86 (1.83 – 4.68)
ImagingTest_X.ray	2.85 (1.70 – 5.04)
Demographics_Sex	2.85 (1.63 – 5.35)
Encounter_ClinicLocation	2.80 (1.77 – 4.62)
Lifestyle_Diet	2.77 (1.72 – 4.70)
Lifestyle_Alcohol	2.77 (1.75 – 4.58)
Encounter_PhysicianName	2.76 (1.76 – 4.51)
SES_Occupation	2.75 (1.72 – 4.60)
DiseaseCondition_MentalHealth	2.71 (1.71 – 4.48)
LabTestResults_DNAsequencing	2.70 (1.69 – 4.52)
Encounter_MRN	2.70 (1.75 – 4.33)
Medication_OtherMedication	2.67 (1.67 – 4.47)
Lifestyle_SocialIsolation	2.65 (1.67 – 4.38)
Medication_MentalHealth	2.64 (1.68 – 4.33)
TxProcedure_Other	2.62 (1.69 – 4.21)
Lifestyle_Drug	2.62 (1.68 – 4.24)
Encounter_VisitDate	2.58 (1.65 – 4.19)
DiseaseCondition_SubstanceAbuse	2.58 (1.64 – 4.23)
TxProcedure_Bariatric	2.57 (1.66 – 4.12)
SES_Education	2.55 (1.61 – 4.23)
DiseaseCondition_SexualReproductive	2.55 (1.60 – 4.23)
Demographics_Ethnicity	2.53 (1.53 – 4.43)
ContactInfo_Name	2.53 (1.72 – 3.84)
TxProcedure_Cosmetic	2.49 (1.62 – 3.98)
DiseaseCondition_Other	2.45 (1.58 – 3.95)
LabTestResults_DrugScreening	2.44 (1.55 – 4.01)
ContactInfo_HomeAddress	2.44 (1.69 – 3.58)
FamilyHealthHistory_SubstanceAbuse	2.43 (1.56 – 3.90)
ImagingTest_OtherImagingTest	2.37 (1.50 – 3.90)
FamilyHealthHistory_Other	2.36 (1.53 – 3.77)
ContactInfo_Phone	2.34 (1.63 – 3.41)
SES_InsuranceStatus	2.31 (1.49 – 3.68)
FamilyHealthHistory_SexualReproductive	2.30 (1.49 – 3.68)
FamilyHealthHistory_MentalHealth	2.29 (1.47 – 3.69)
LabTestResults_GeneticTest	2.24 (1.43 – 3.66)
BodyMeasurements	2.21 (1.39 – 3.63)
LabTestResults_STDtest	2.15 (1.38 – 3.48)
LabTestResults_Other	2.14 (1.38 – 3.44)
ContactInfo_Email	2.08 (1.45 – 3.03)
SES_Income	1.58 (1.10 – 2.30)
SexualLife	1.39 (1.00 – 1.95)
AdoptionHistory	1.37 (0.99 – 1.92)
PregnancyHistory	1.37 (0.98 – 1.93)
SUMMARY	**2.48 (2.33 – 2.63)**

Figure B.4: Forest plot of unadjusted odds ratio for age (>=60 vs. <60). The 59 sharable items were sorted by unadjusted odds ratios and shown with their 95% confidence intervals. The pooled odds ratio is displayed at the bottom and labeled SUMMARY. Row labels are expressed in the format of "Category_Item". A category with no item was expressed simply as "Category." Abbreviations: CT, Computerized Tomography; DNA, DeoxyriboNucleic Acid; MRI, Magnetic Resonance Imaging; MRN, Medical Record Number; SES, Social Economic Status; STD, Sexually Transmitted Disease; TxProcedure, Treatment Procedure.



Health literacy (adequate vs. inadequate)

Variable	Odds Ratio (95% CI)
Demographics_Age	3.87 (2.15 – 6.69)
Demographics_Sex	3.79 (2.11 – 6.56)
ImagingTest_MRI	3.78 (2.21 – 6.28)
ImagingTest_OtherImagingTest	3.76 (2.22 – 6.20)
ImagingTest_X.ray	3.74 (2.16 – 6.26)
Allergies	3.70 (2.09 – 6.31)
ImagingTest_CTscan	3.62 (2.10 – 6.06)
Demographics_Race	3.52 (1.99 – 5.99)
LabTestResults_DrugScreening	3.41 (2.00 – 5.64)
BodyMeasurements	3.37 (1.96 – 5.61)
VitalSigns	3.36 (1.88 – 5.78)
Demographics_Ethnicity	3.31 (1.85 – 5.67)
LabTestResults_Other	3.23 (1.88 – 5.36)
TxProcedure_Cosmetic	3.18 (1.91 – 5.16)
TxProcedure_Bariatric	3.09 (1.84 – 5.05)
Demographics_SexualOrientation	3.07 (1.75 – 5.20)
Medication_OtherMedication	3.05 (1.76 – 5.11)
LabTestResults_GeneticTest	3.05 (1.76 – 5.10)
FamilyHealthHistory_Other	3.04 (1.79 – 5.00)
TxProcedure_Other	3.03 (1.80 – 4.93)
Encounter_ClinicLocation	3.01 (1.79 – 4.91)
DiseaseCondition_Other	3.00 (1.77 – 4.93)
Encounter_VisitDate	2.98 (1.77 – 4.86)
SES_Education	2.94 (1.71 – 4.86)
LabTestResults_DNAsequencing	2.93 (1.69 – 4.89)
Encounter_ClinicSpecialty	2.92 (1.72 – 4.80)
DiseaseCondition_SubstanceAbuse	2.91 (1.70 – 4.82)
DiseaseCondition_MentalHealth	2.90 (1.70 – 4.81)
DiseaseCondition_SexualReproductive	2.89 (1.67 – 4.82)
FamilyHealthHistory_SubstanceAbuse	2.88 (1.68 – 4.76)
FamilyHealthHistory_MentalHealth	2.78 (1.61 – 4.63)
TxProcedure_GenitouriaryReproductive	2.77 (1.62 – 4.57)
Biospecimen_Blood	2.77 (1.62 – 4.57)
LabTestResults_STDtest	2.72 (1.56 – 4.57)
Lifestyle_Diet	2.72 (1.55 – 4.57)
Lifestyle_Stress	2.72 (1.55 – 4.57)
Biospecimen_Urine	2.71 (1.58 – 4.46)
SES_InsuranceStatus	2.70 (1.58 – 4.46)
Encounter_PhysicianName	2.70 (1.60 – 4.41)
Encounter_MRN	2.69 (1.61 – 4.37)
Lifestyle_Exercise	2.69 (1.52 – 4.57)
FamilyHealthHistory_SexualReproductive	2.68 (1.56 – 4.47)
Biospecimen_Tissue	2.67 (1.57 – 4.41)
Lifestyle_Alcohol	2.64 (1.53 – 4.39)
Lifestyle_Smoking	2.64 (1.53 – 4.39)
Medication_MentalHealth	2.62 (1.52 – 4.35)
Lifestyle_SocialIsolation	2.61 (1.50 – 4.38)
SES_Occupation	2.58 (1.48 – 4.33)
SES_MaritalStatus	2.52 (1.43 – 4.26)
SES_Income	2.50 (1.52 – 4.00)
PregnancyHistory	2.41 (1.48 – 3.82)
TxProcedure_MentalHealth	2.39 (1.39 – 3.95)
Lifestyle_Drug	2.37 (1.36 – 3.97)
ContactInfo_Email	2.34 (1.45 – 3.68)
ContactInfo_Name	2.32 (1.43 – 3.67)
ContactInfo_Phone	2.23 (1.39 – 3.49)
ContactInfo_HomeAddress	2.20 (1.37 – 3.45)
AdoptionHistory	2.20 (1.36 – 3.49)
SexualLife	1.98 (1.21 – 3.14)
SUMMARY	**2.81 (2.62 – 3.00)**

Figure B.5: Forest plot of unadjusted odds ratio for health literacy (adequate vs. inadequate). The 59 sharable items were sorted by unadjusted odds ratios and shown with their 95% confidence intervals. The pooled odds ratio is displayed at the bottom and labeled SUMMARY. Row labels are expressed in the format of “Category_Item”. A category with no item was expressed simply as “Category.” Abbreviations: CT, Computerized Tomography; DNA, DeoxyriboNucleic Acid; MRI, Magnetic Resonance Imaging; MRN, Medical Record Number; SES, Social Economic Status; STD, Sexually Transmitted Disease; TxProcedure, Treatment Procedure.

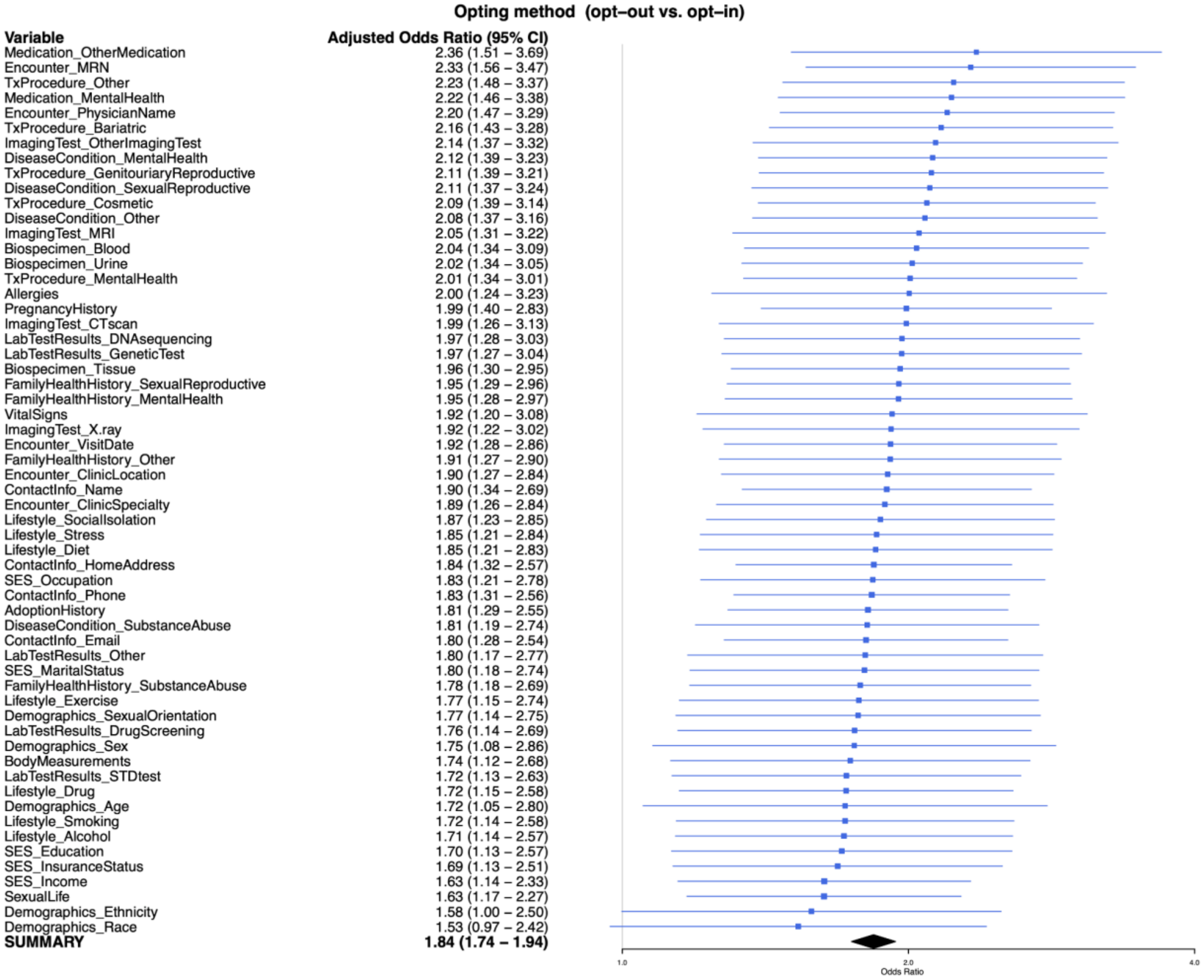


Figure B.6: Forest plot of adjusted odds ratio for opting method (opt-out vs. opt-in). The 59 sharable items were sorted by adjusted odds ratios from the multivariate model (See the Methods for details) and shown with their 95% confidence intervals. The pooled odds ratio is displayed at the bottom and labeled SUMMARY. Row labels are expressed in the format of "Category_Item". A category with no item was expressed simply as "Category." Abbreviations: CT, Computerized Tomography; DNA, DeoxyriboNucleic Acid; MRI, Magnetic Resonance Imaging; MRN, Medical Record Number; SES, Social Economic Status; STD, Sexually Transmitted Disease; TxProcedure, Treatment Procedure.

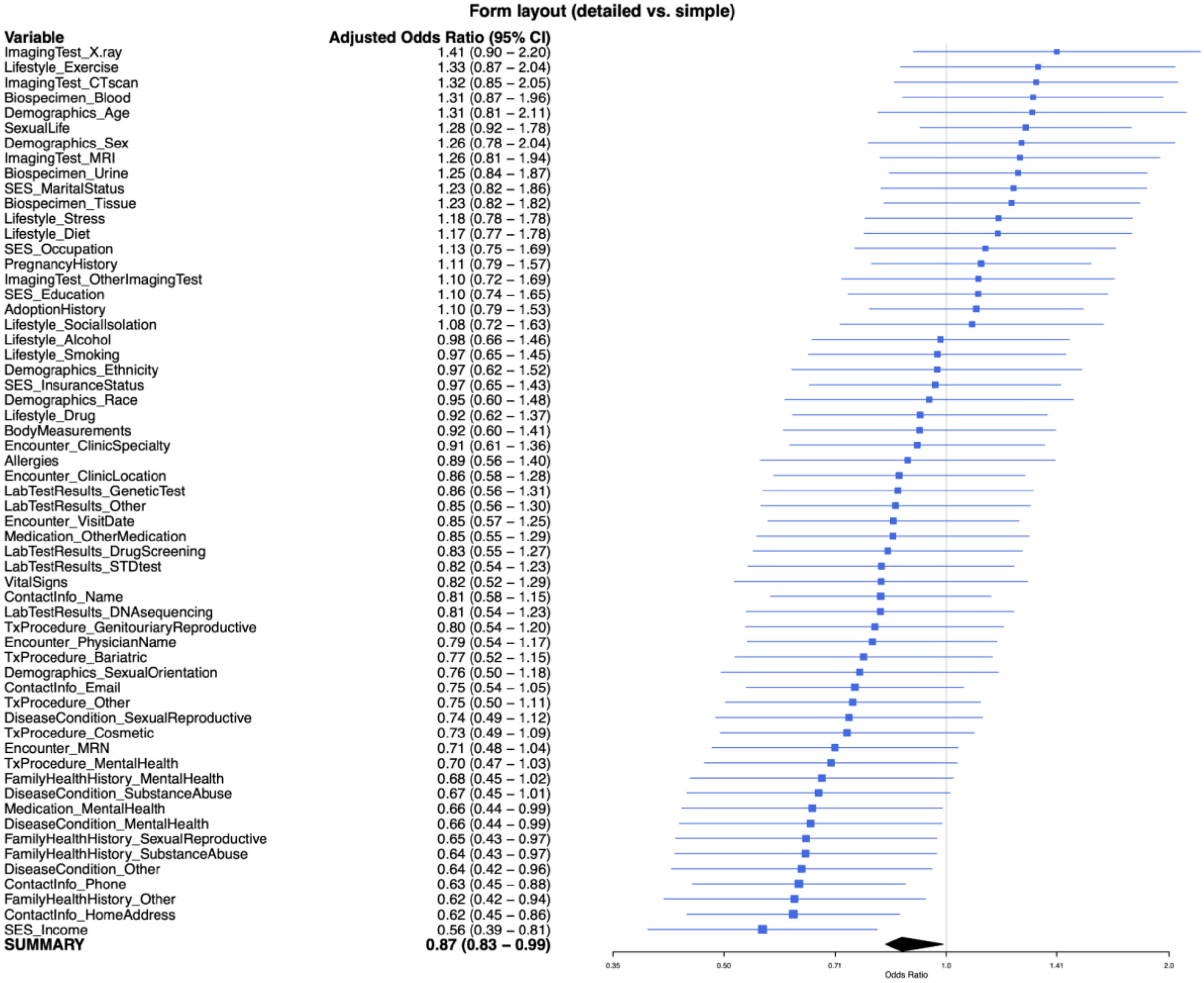
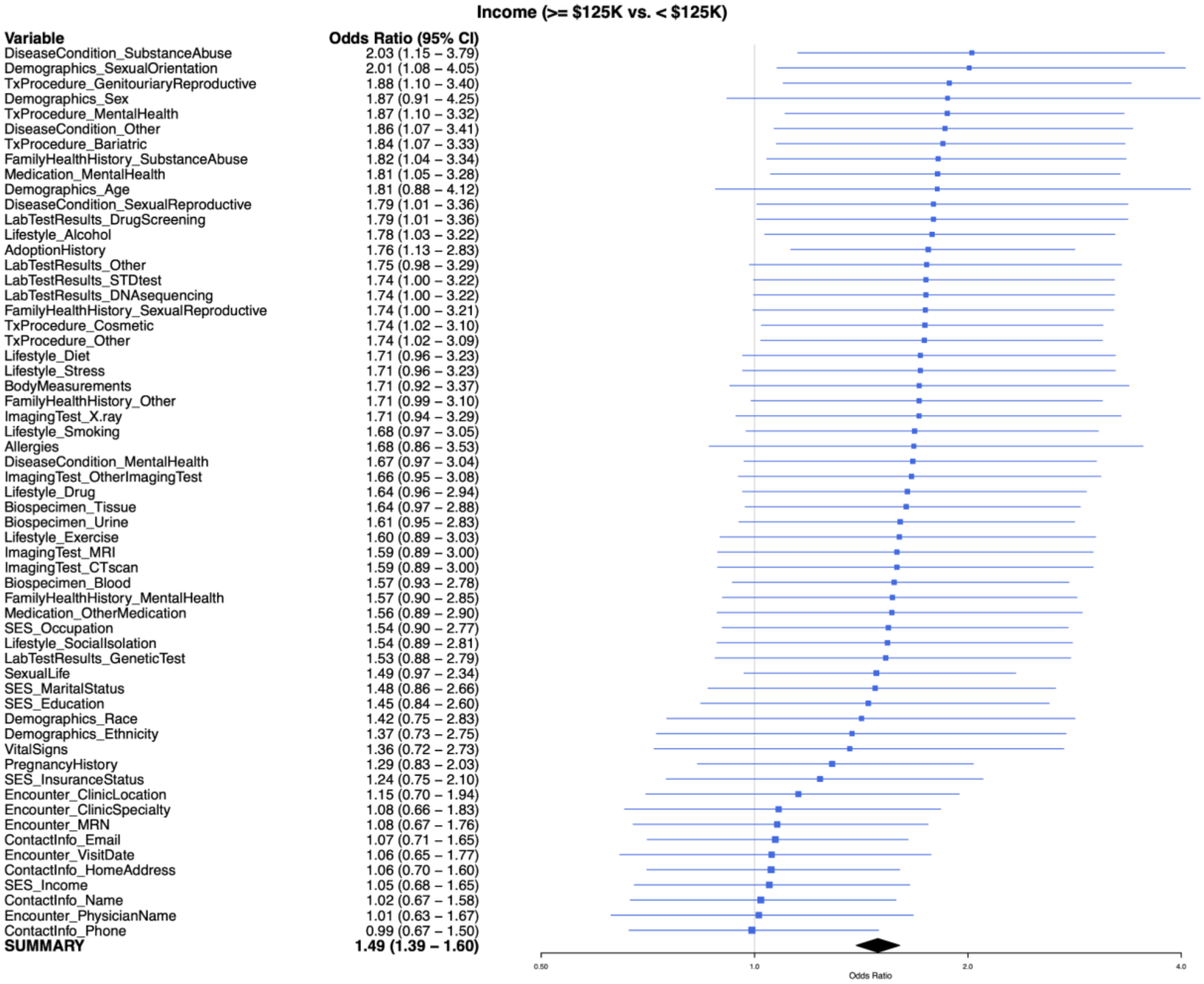


Figure B.7: Forest plot of adjusted odds ratio for form layout (detailed vs. simple). The 59 sharable items were sorted by adjusted odds ratios from the multivariate model (See the Methods for details) and shown with their 95% confidence intervals. The pooled odds ratio is displayed at the bottom and labeled SUMMARY. Row labels are expressed in the format of "Category_Item". A category with no item was expressed simply as "Category." Abbreviations: CT, Computerized Tomography; DNA, DeoxyriboNucleic Acid; MRI, Magnetic Resonance Imaging; MRN, Medical Record Number; SES, Social Economic Status; STD, Sexually Transmitted Disease; TxProcedure, Treatment Procedure.



Variable	Odds Ratio (95% CI)
DiseaseCondition_SubstanceAbuse	2.03 (1.15 – 3.79)
Demographics_SexualOrientation	2.01 (1.08 – 4.05)
TxProcedure_GenitouriaryReproductive	1.88 (1.10 – 3.40)
Demographics_Sex	1.87 (0.91 – 4.25)
TxProcedure_MentalHealth	1.87 (1.10 – 3.32)
DiseaseCondition_Other	1.86 (1.07 – 3.41)
TxProcedure_Bariatric	1.84 (1.07 – 3.33)
FamilyHealthHistory_SubstanceAbuse	1.82 (1.04 – 3.34)
Medication_MentalHealth	1.81 (1.05 – 3.28)
Demographics_Age	1.81 (0.88 – 4.12)
DiseaseCondition_SexualReproductive	1.79 (1.01 – 3.36)
LabTestResults_DrugScreening	1.79 (1.01 – 3.36)
Lifestyle_Alcohol	1.78 (1.03 – 3.22)
AdoptionHistory	1.76 (1.13 – 2.83)
LabTestResults_Other	1.75 (0.98 – 3.29)
LabTestResults_STDtest	1.74 (1.00 – 3.22)
LabTestResults_DNAsequencing	1.74 (1.00 – 3.22)
FamilyHealthHistory_SexualReproductive	1.74 (1.00 – 3.21)
TxProcedure_Cosmetic	1.74 (1.02 – 3.10)
TxProcedure_Other	1.74 (1.02 – 3.09)
Lifestyle_Diet	1.71 (0.96 – 3.23)
Lifestyle_Stress	1.71 (0.96 – 3.23)
BodyMeasurements	1.71 (0.92 – 3.37)
FamilyHealthHistory_Other	1.71 (0.99 – 3.10)
ImagingTest_X.ray	1.71 (0.94 – 3.29)
Lifestyle_Smoking	1.68 (0.97 – 3.05)
Allergies	1.68 (0.86 – 3.53)
DiseaseCondition_MentalHealth	1.67 (0.97 – 3.04)
ImagingTest_OtherImagingTest	1.66 (0.95 – 3.08)
Lifestyle_Drug	1.64 (0.96 – 2.94)
Biospecimen_Tissue	1.64 (0.97 – 2.88)
Biospecimen_Urine	1.61 (0.95 – 2.83)
Lifestyle_Exercise	1.60 (0.89 – 3.03)
ImagingTest_MRI	1.59 (0.89 – 3.00)
ImagingTest_CTscan	1.59 (0.89 – 3.00)
Biospecimen_Blood	1.57 (0.93 – 2.78)
FamilyHealthHistory_MentalHealth	1.57 (0.90 – 2.85)
Medication_OtherMedication	1.56 (0.89 – 2.90)
SES_Occupation	1.54 (0.90 – 2.77)
Lifestyle_SocialIsolation	1.54 (0.89 – 2.81)
LabTestResults_GeneticTest	1.53 (0.88 – 2.79)
SexualLife	1.49 (0.97 – 2.34)
SES_MaritalStatus	1.48 (0.86 – 2.66)
SES_Education	1.45 (0.84 – 2.60)
Demographics_Race	1.42 (0.75 – 2.83)
Demographics_Ethnicity	1.37 (0.73 – 2.75)
VitalSigns	1.36 (0.72 – 2.73)
PregnancyHistory	1.29 (0.83 – 2.03)
SES_InsuranceStatus	1.24 (0.75 – 2.10)
Encounter_ClinicLocation	1.15 (0.70 – 1.94)
Encounter_ClinicSpecialty	1.08 (0.66 – 1.83)
Encounter_MRN	1.08 (0.67 – 1.76)
ContactInfo_Email	1.07 (0.71 – 1.65)
Encounter_VisitDate	1.06 (0.65 – 1.77)
ContactInfo_HomeAddress	1.06 (0.70 – 1.60)
SES_Income	1.05 (0.68 – 1.65)
ContactInfo_Name	1.02 (0.67 – 1.58)
Encounter_PhysicianName	1.01 (0.63 – 1.67)
ContactInfo_Phone	0.99 (0.67 – 1.50)
SUMMARY	**1.49 (1.39 – 1.60)**

Figure B.8: Forest plot of adjusted odds ratio for income (>= $125K vs. < $125K). The 59 sharable items were sorted by adjusted odds ratios from the multivariate model (See the Methods for details) and shown with their 95% confidence intervals. The pooled odds ratio is displayed at the bottom and labeled SUMMARY. Row labels are expressed in the format of “Category_Item”. A category with no item was expressed simply as “Category.” Abbreviations: CT, Computerized Tomography; DNA, DeoxyriboNucleic Acid; MRI, Magnetic Resonance Imaging; MRN, Medical Record Number; SES, Social Economic Status; STD, Sexually Transmitted Disease; TxProcedure, Treatment Procedure.

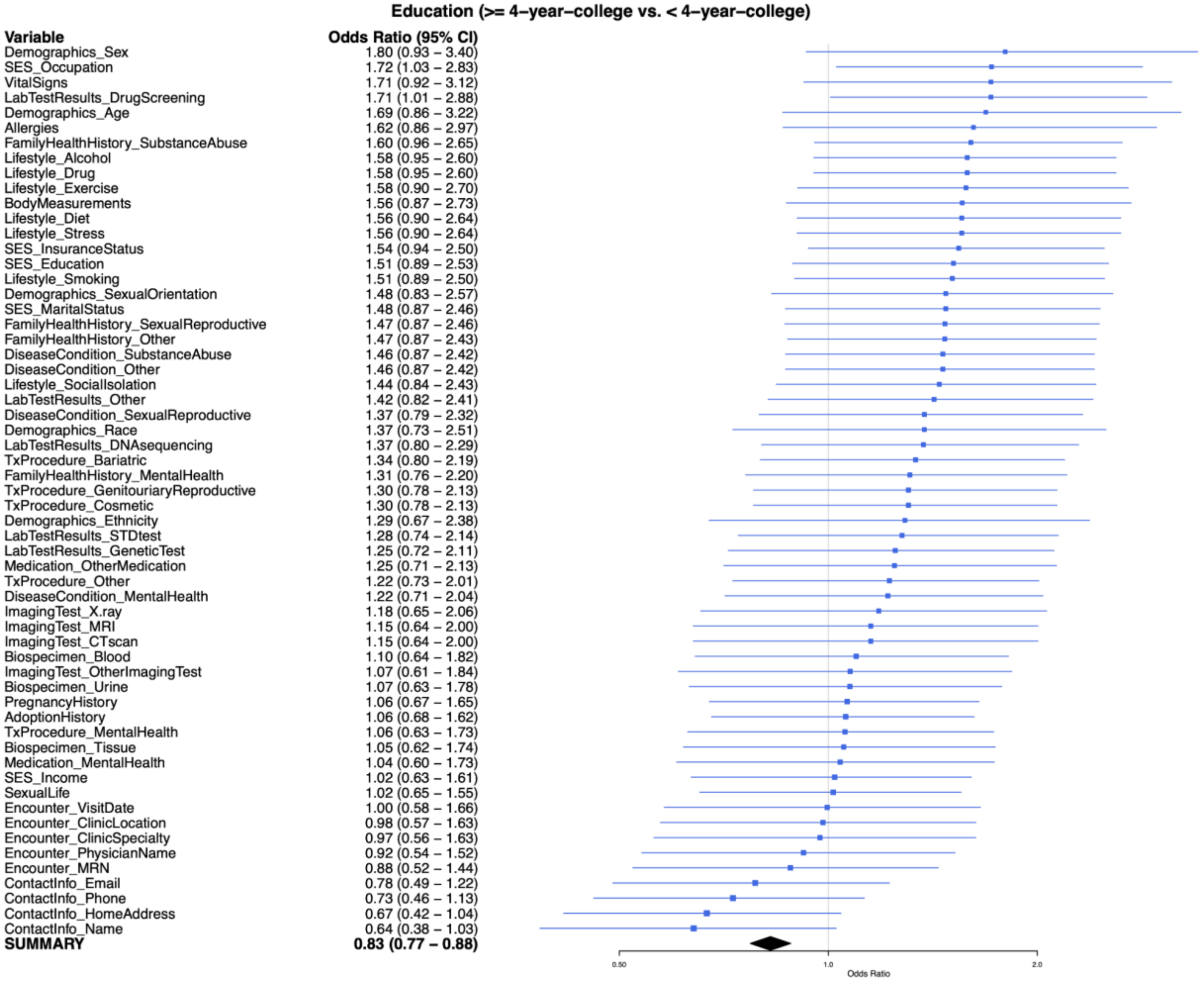
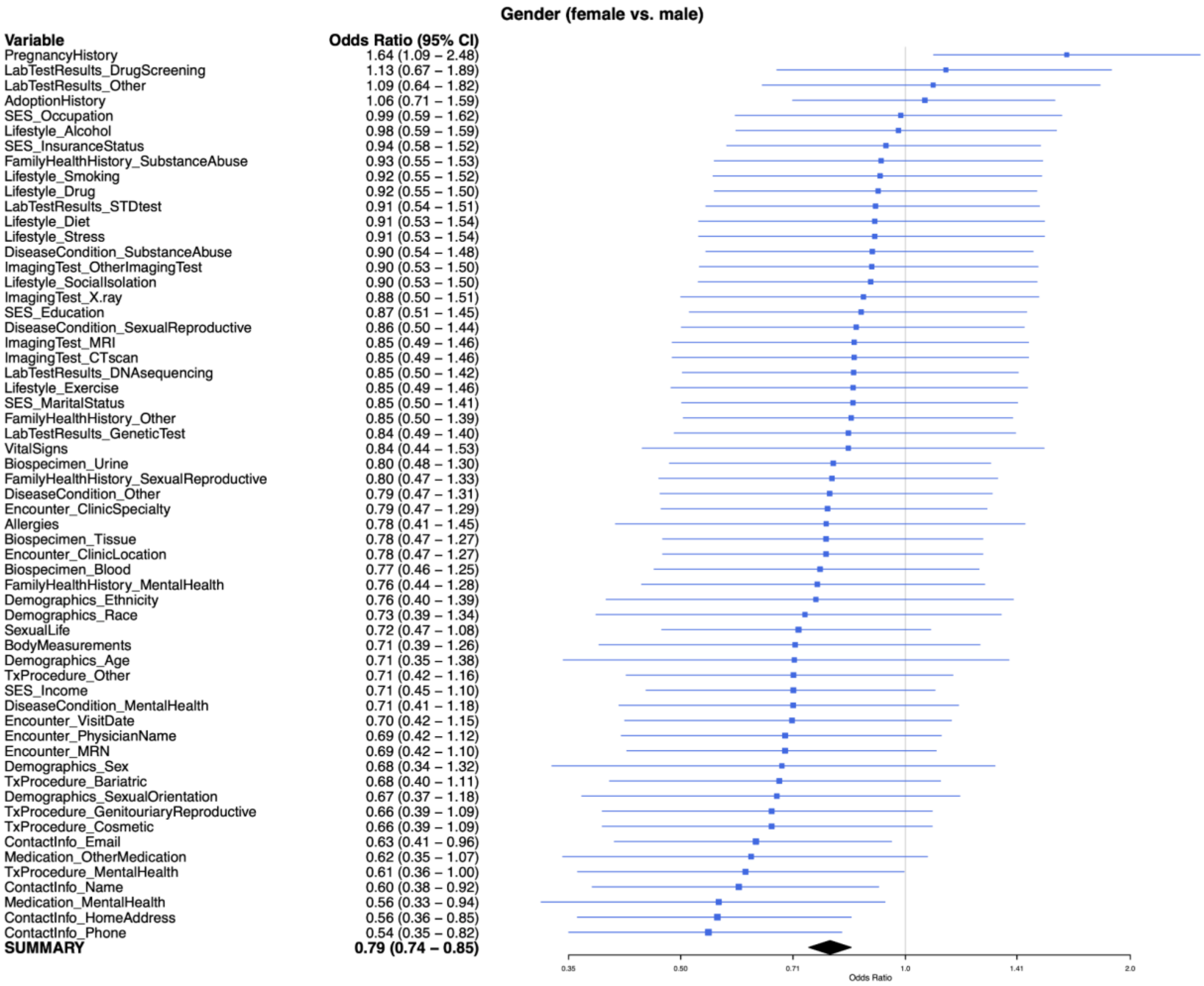


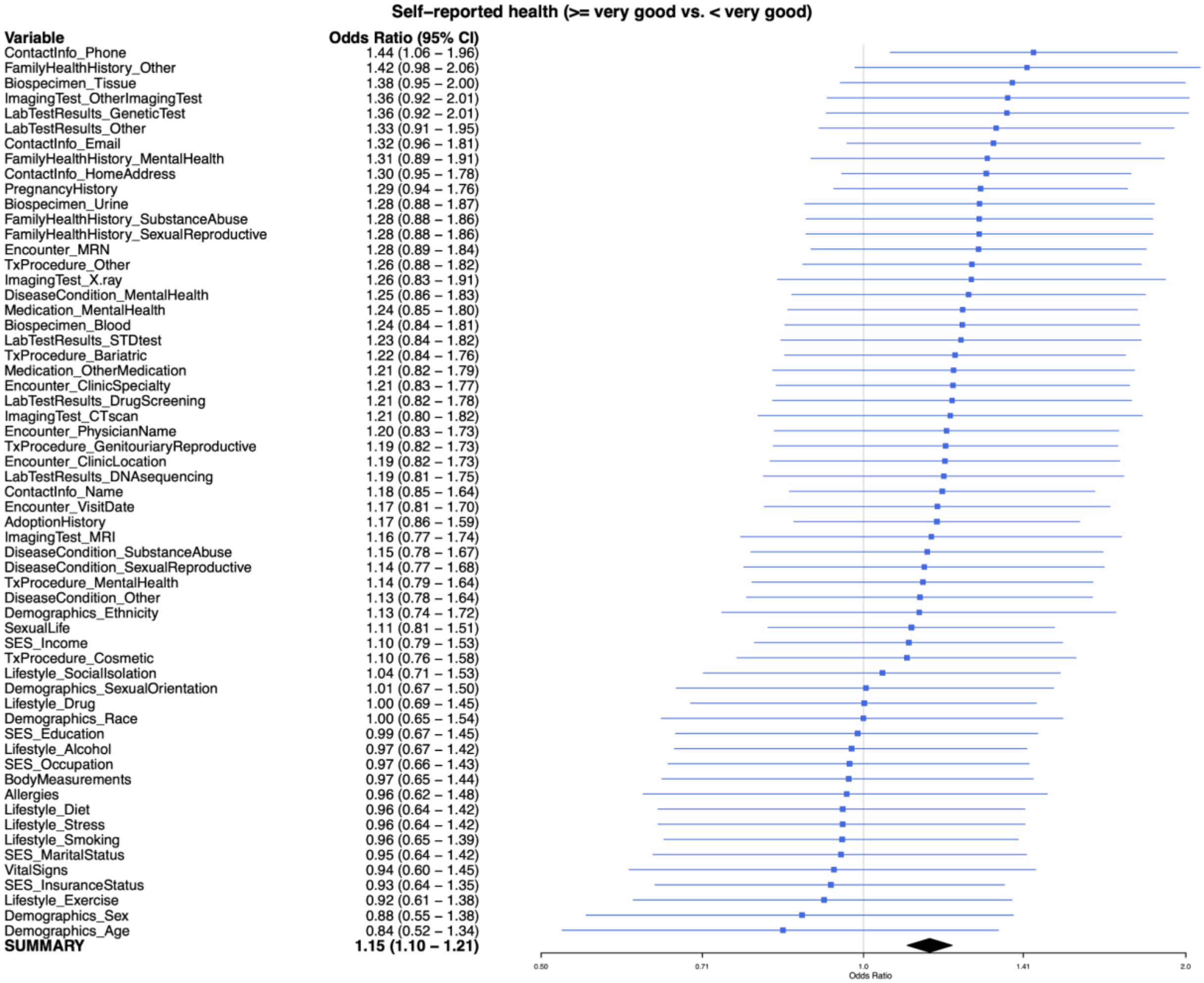
Figure B.9: Forest plot of adjusted odds ratio for education (>= 4-year-college vs. < 4-year-college). The 59 sharable items were sorted by adjusted odds ratios from the multivariate model (See the Methods for details) and shown with their 95% confidence intervals. The pooled odds ratio is displayed at the bottom and labeled SUMMARY. Row labels are expressed in the format of “Category_Item”. A category with no item was expressed simply as “Category.” Abbreviations: CT, Computerized Tomography; DNA, DeoxyriboNucleic Acid; MRI, Magnetic Resonance Imaging; MRN, Medical Record Number; SES, Social Economic Status; STD, Sexually Transmitted Disease; TxProcedure, Treatment Procedure.



Gender (female vs. male)

Variable	Odds Ratio (95% CI)
PregnancyHistory	1.64 (1.09 – 2.48)
LabTestResults_DrugScreening	1.13 (0.67 – 1.89)
LabTestResults_Other	1.09 (0.64 – 1.82)
AdoptionHistory	1.06 (0.71 – 1.59)
SES_Occupation	0.99 (0.59 – 1.62)
Lifestyle_Alcohol	0.98 (0.59 – 1.59)
SES_InsuranceStatus	0.94 (0.58 – 1.52)
FamilyHealthHistory_SubstanceAbuse	0.93 (0.55 – 1.53)
Lifestyle_Smoking	0.92 (0.55 – 1.52)
Lifestyle_Drug	0.92 (0.55 – 1.50)
LabTestResults_STDtest	0.91 (0.54 – 1.51)
Lifestyle_Diet	0.91 (0.53 – 1.54)
Lifestyle_Stress	0.91 (0.53 – 1.54)
DiseaseCondition_SubstanceAbuse	0.90 (0.54 – 1.48)
ImagingTest_OtherImagingTest	0.90 (0.53 – 1.50)
Lifestyle_SocialIsolation	0.90 (0.53 – 1.50)
ImagingTest_X.ray	0.88 (0.50 – 1.51)
SES_Education	0.87 (0.51 – 1.45)
DiseaseCondition_SexualReproductive	0.86 (0.50 – 1.44)
ImagingTest_MRI	0.85 (0.49 – 1.46)
ImagingTest_CTscan	0.85 (0.49 – 1.46)
LabTestResults_DNAsequencing	0.85 (0.50 – 1.42)
Lifestyle_Exercise	0.85 (0.49 – 1.46)
SES_MaritalStatus	0.85 (0.50 – 1.41)
FamilyHealthHistory_Other	0.85 (0.50 – 1.39)
LabTestResults_GeneticTest	0.84 (0.49 – 1.40)
VitalSigns	0.84 (0.44 – 1.53)
Biospecimen_Urine	0.80 (0.48 – 1.30)
FamilyHealthHistory_SexualReproductive	0.80 (0.47 – 1.33)
DiseaseCondition_Other	0.79 (0.47 – 1.31)
Encounter_ClinicSpecialty	0.79 (0.47 – 1.29)
Allergies	0.78 (0.41 – 1.45)
Biospecimen_Tissue	0.78 (0.47 – 1.27)
Encounter_ClinicLocation	0.78 (0.47 – 1.27)
Biospecimen_Blood	0.77 (0.46 – 1.25)
FamilyHealthHistory_MentalHealth	0.76 (0.44 – 1.28)
Demographics_Ethnicity	0.76 (0.40 – 1.39)
Demographics_Race	0.73 (0.39 – 1.34)
SexualLife	0.72 (0.47 – 1.08)
BodyMeasurements	0.71 (0.39 – 1.26)
Demographics_Age	0.71 (0.35 – 1.38)
TxProcedure_Other	0.71 (0.42 – 1.16)
SES_Income	0.71 (0.45 – 1.10)
DiseaseCondition_MentalHealth	0.71 (0.41 – 1.18)
Encounter_VisitDate	0.70 (0.42 – 1.15)
Encounter_PhysicianName	0.69 (0.42 – 1.12)
Encounter_MRN	0.69 (0.42 – 1.10)
Demographics_Sex	0.68 (0.34 – 1.32)
TxProcedure_Bariatric	0.68 (0.40 – 1.11)
Demographics_SexualOrientation	0.67 (0.37 – 1.18)
TxProcedure_GenitouriaryReproductive	0.66 (0.39 – 1.09)
TxProcedure_Cosmetic	0.66 (0.39 – 1.09)
ContactInfo_Email	0.63 (0.41 – 0.96)
Medication_OtherMedication	0.62 (0.35 – 1.07)
TxProcedure_MentalHealth	0.61 (0.36 – 1.00)
ContactInfo_Name	0.60 (0.38 – 0.92)
Medication_MentalHealth	0.56 (0.33 – 0.94)
ContactInfo_HomeAddress	0.56 (0.36 – 0.85)
ContactInfo_Phone	0.54 (0.35 – 0.82)
SUMMARY	**0.79 (0.74 – 0.85)**

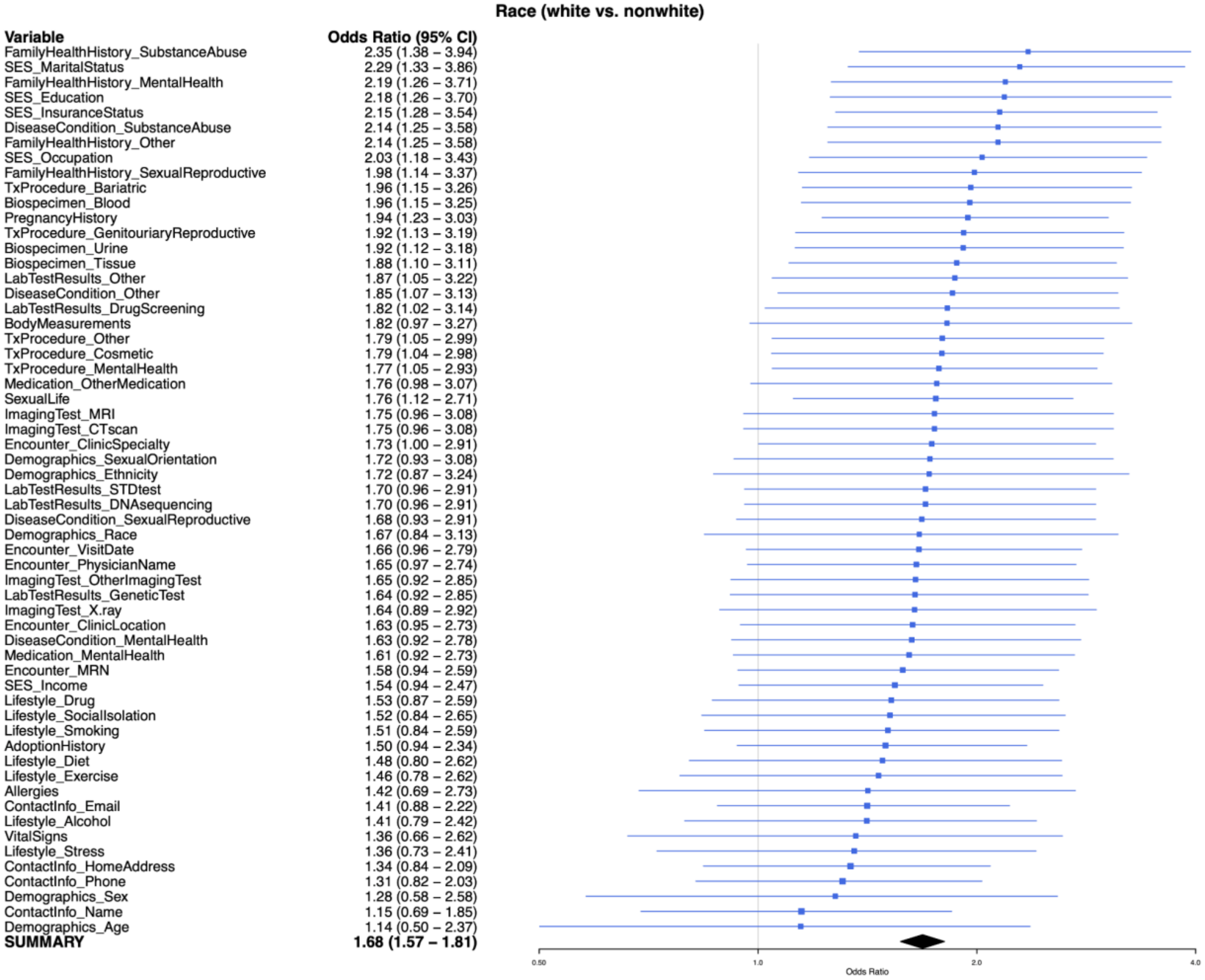
Figure B.10: Forest plot of adjusted odds ratio for gender (female vs. male). The 59 sharable items were sorted by adjusted odds ratios from the multivariate model (See the Methods for details) and shown with their 95% confidence intervals. The pooled odds ratio is displayed at the bottom and labeled SUMMARY. Row labels are expressed in the format of "Category_Item". A category with no item was expressed simply as "Category." Abbreviations: CT, Computerized Tomography; DNA, DeoxyriboNucleic Acid; MRI, Magnetic Resonance Imaging; MRN, Medical Record Number; SES, Social Economic Status; STD, Sexually Transmitted Disease; TxProcedure, Treatment Procedure.



Self–reported health (>= very good vs. < very good)

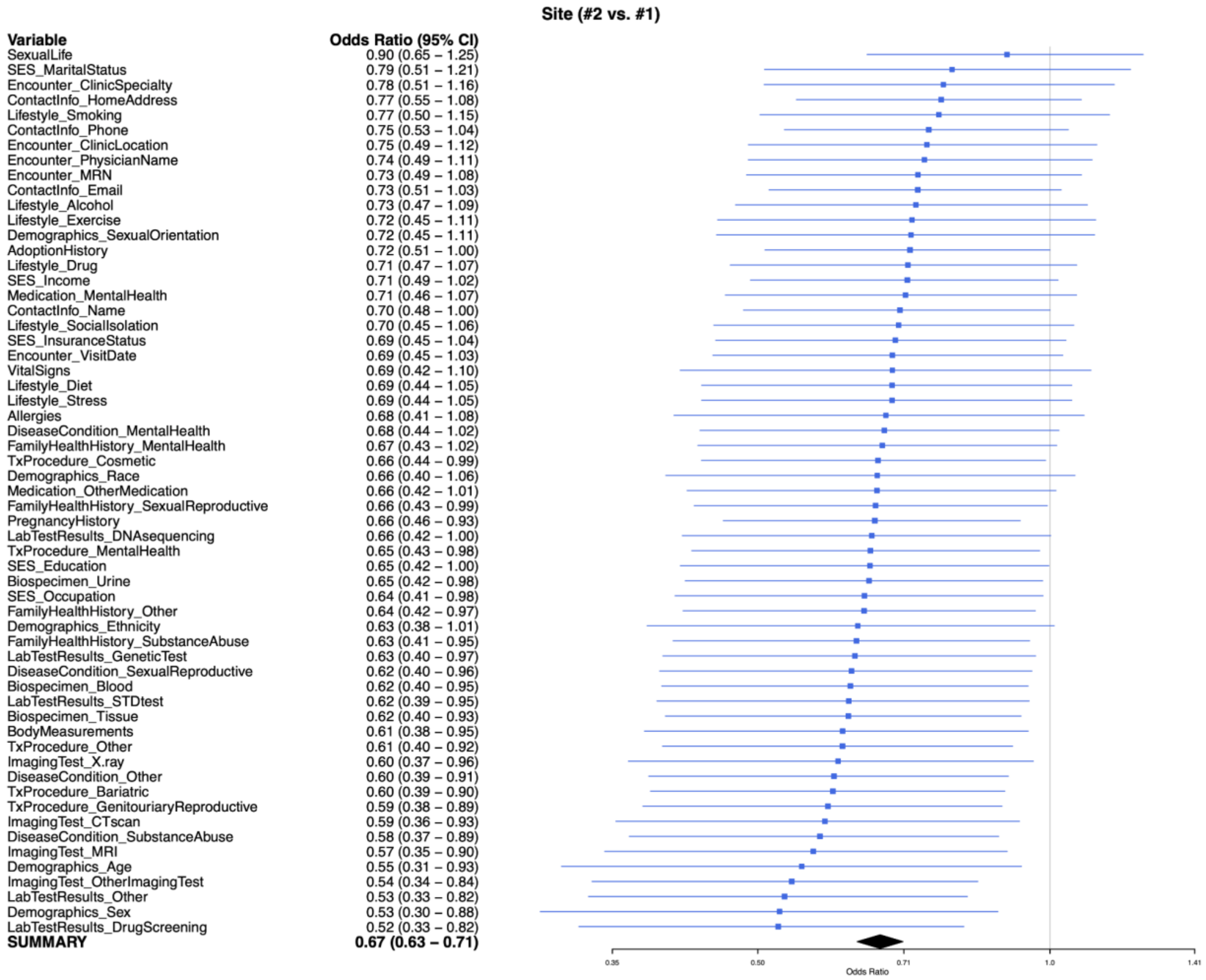
Variable	Odds Ratio (95% CI)
ContactInfo_Phone	1.44 (1.06 – 1.96)
FamilyHealthHistory_Other	1.42 (0.98 – 2.06)
Biospecimen_Tissue	1.38 (0.95 – 2.00)
ImagingTest_OtherImagingTest	1.36 (0.92 – 2.01)
LabTestResults_GeneticTest	1.36 (0.92 – 2.01)
LabTestResults_Other	1.33 (0.91 – 1.95)
ContactInfo_Email	1.32 (0.96 – 1.81)
FamilyHealthHistory_MentalHealth	1.31 (0.89 – 1.91)
ContactInfo_HomeAddress	1.30 (0.95 – 1.78)
PregnancyHistory	1.29 (0.94 – 1.76)
Biospecimen_Urine	1.28 (0.88 – 1.87)
FamilyHealthHistory_SubstanceAbuse	1.28 (0.88 – 1.86)
FamilyHealthHistory_SexualReproductive	1.28 (0.88 – 1.86)
Encounter_MRN	1.28 (0.89 – 1.84)
TxProcedure_Other	1.26 (0.88 – 1.82)
ImagingTest_X.ray	1.26 (0.83 – 1.91)
DiseaseCondition_MentalHealth	1.25 (0.86 – 1.83)
Medication_MentalHealth	1.24 (0.85 – 1.80)
Biospecimen_Blood	1.24 (0.84 – 1.81)
LabTestResults_STDtest	1.23 (0.84 – 1.82)
TxProcedure_Bariatric	1.22 (0.84 – 1.76)
Medication_OtherMedication	1.21 (0.82 – 1.79)
Encounter_ClinicSpecialty	1.21 (0.83 – 1.77)
LabTestResults_DrugScreening	1.21 (0.82 – 1.78)
ImagingTest_CTscan	1.21 (0.80 – 1.82)
Encounter_PhysicianName	1.20 (0.83 – 1.73)
TxProcedure_GenitouriaryReproductive	1.19 (0.82 – 1.73)
Encounter_ClinicLocation	1.19 (0.82 – 1.73)
LabTestResults_DNAsequencing	1.19 (0.81 – 1.75)
ContactInfo_Name	1.18 (0.85 – 1.64)
Encounter_VisitDate	1.17 (0.81 – 1.70)
AdoptionHistory	1.17 (0.86 – 1.59)
ImagingTest_MRI	1.16 (0.77 – 1.74)
DiseaseCondition_SubstanceAbuse	1.15 (0.78 – 1.67)
DiseaseCondition_SexualReproductive	1.14 (0.77 – 1.68)
TxProcedure_MentalHealth	1.14 (0.79 – 1.64)
DiseaseCondition_Other	1.13 (0.78 – 1.64)
Demographics_Ethnicity	1.13 (0.74 – 1.72)
SexualLife	1.11 (0.81 – 1.51)
SES_Income	1.10 (0.79 – 1.53)
TxProcedure_Cosmetic	1.10 (0.76 – 1.58)
Lifestyle_SocialIsolation	1.04 (0.71 – 1.53)
Demographics_SexualOrientation	1.01 (0.67 – 1.50)
Lifestyle_Drug	1.00 (0.69 – 1.45)
Demographics_Race	1.00 (0.65 – 1.54)
SES_Education	0.99 (0.67 – 1.45)
Lifestyle_Alcohol	0.97 (0.67 – 1.42)
SES_Occupation	0.97 (0.66 – 1.43)
BodyMeasurements	0.97 (0.65 – 1.44)
Allergies	0.96 (0.62 – 1.48)
Lifestyle_Diet	0.96 (0.64 – 1.42)
Lifestyle_Stress	0.96 (0.64 – 1.42)
Lifestyle_Smoking	0.96 (0.65 – 1.39)
SES_MaritalStatus	0.95 (0.64 – 1.42)
VitalSigns	0.94 (0.60 – 1.45)
SES_InsuranceStatus	0.93 (0.64 – 1.35)
Lifestyle_Exercise	0.92 (0.61 – 1.38)
Demographics_Sex	0.88 (0.55 – 1.38)
Demographics_Age	0.84 (0.52 – 1.34)
SUMMARY	**1.15 (1.10 – 1.21)**

Figure B.11: Forest plot of adjusted odds ratio for self-reported health (>= very good vs. < very good). The 59 sharable items were sorted by adjusted odds ratios from the multivariate model (See the Methods for details) and shown with their 95% confidence intervals. The pooled odds ratio is displayed at the bottom and labeled SUMMARY. Row labels are expressed in the format of "Category_Item". A category with no item was expressed simply as "Category." Abbreviations: CT, Computerized Tomography; DNA, DeoxyriboNucleic Acid; MRI, Magnetic Resonance Imaging; MRN, Medical Record Number; SES, Social Economic Status; STD, Sexually Transmitted Disease; TxProcedure, Treatment Procedure.



Variable	Odds Ratio (95% CI)
FamilyHealthHistory_SubstanceAbuse	2.35 (1.38 – 3.94)
SES_MaritalStatus	2.29 (1.33 – 3.86)
FamilyHealthHistory_MentalHealth	2.19 (1.26 – 3.71)
SES_Education	2.18 (1.26 – 3.70)
SES_InsuranceStatus	2.15 (1.28 – 3.54)
DiseaseCondition_SubstanceAbuse	2.14 (1.25 – 3.58)
FamilyHealthHistory_Other	2.14 (1.25 – 3.58)
SES_Occupation	2.03 (1.18 – 3.43)
FamilyHealthHistory_SexualReproductive	1.98 (1.14 – 3.37)
TxProcedure_Bariatric	1.96 (1.15 – 3.26)
Biospecimen_Blood	1.96 (1.15 – 3.25)
PregnancyHistory	1.94 (1.23 – 3.03)
TxProcedure_GenitouriaryReproductive	1.92 (1.13 – 3.19)
Biospecimen_Urine	1.92 (1.12 – 3.18)
Biospecimen_Tissue	1.88 (1.10 – 3.11)
LabTestResults_Other	1.87 (1.05 – 3.22)
DiseaseCondition_Other	1.85 (1.07 – 3.13)
LabTestResults_DrugScreening	1.82 (1.02 – 3.14)
BodyMeasurements	1.82 (0.97 – 3.27)
TxProcedure_Other	1.79 (1.05 – 2.99)
TxProcedure_Cosmetic	1.79 (1.04 – 2.98)
TxProcedure_MentalHealth	1.77 (1.05 – 2.93)
Medication_OtherMedication	1.76 (0.98 – 3.07)
SexualLife	1.76 (1.12 – 2.71)
ImagingTest_MRI	1.75 (0.96 – 3.08)
ImagingTest_CTscan	1.75 (0.96 – 3.08)
Encounter_ClinicSpecialty	1.73 (1.00 – 2.91)
Demographics_SexualOrientation	1.72 (0.93 – 3.08)
Demographics_Ethnicity	1.72 (0.87 – 3.24)
LabTestResults_STDtest	1.70 (0.96 – 2.91)
LabTestResults_DNAsequencing	1.70 (0.96 – 2.91)
DiseaseCondition_SexualReproductive	1.68 (0.93 – 2.91)
Demographics_Race	1.67 (0.84 – 3.13)
Encounter_VisitDate	1.66 (0.96 – 2.79)
Encounter_PhysicianName	1.65 (0.97 – 2.74)
ImagingTest_OtherImagingTest	1.65 (0.92 – 2.85)
LabTestResults_GeneticTest	1.64 (0.92 – 2.85)
ImagingTest_X.ray	1.64 (0.89 – 2.92)
Encounter_ClinicLocation	1.63 (0.95 – 2.73)
DiseaseCondition_MentalHealth	1.63 (0.92 – 2.78)
Medication_MentalHealth	1.61 (0.92 – 2.73)
Encounter_MRN	1.58 (0.94 – 2.59)
SES_Income	1.54 (0.94 – 2.47)
Lifestyle_Drug	1.53 (0.87 – 2.59)
Lifestyle_SocialIsolation	1.52 (0.84 – 2.65)
Lifestyle_Smoking	1.51 (0.84 – 2.59)
AdoptionHistory	1.50 (0.94 – 2.34)
Lifestyle_Diet	1.48 (0.80 – 2.62)
Lifestyle_Exercise	1.46 (0.78 – 2.62)
Allergies	1.42 (0.69 – 2.73)
ContactInfo_Email	1.41 (0.88 – 2.22)
Lifestyle_Alcohol	1.41 (0.79 – 2.42)
VitalSigns	1.36 (0.66 – 2.62)
Lifestyle_Stress	1.36 (0.73 – 2.41)
ContactInfo_HomeAddress	1.34 (0.84 – 2.09)
ContactInfo_Phone	1.31 (0.82 – 2.03)
Demographics_Sex	1.28 (0.58 – 2.58)
ContactInfo_Name	1.15 (0.69 – 1.85)
Demographics_Age	1.14 (0.50 – 2.37)
SUMMARY	**1.68 (1.57 – 1.81)**

Figure B.12: Forest plot of adjusted odds ratio for race (white vs. nonwhite). The 59 sharable items were sorted by adjusted odds ratios from the multivariate model (See the Methods for details) and shown with their 95% confidence intervals. The pooled odds ratio is displayed at the bottom and labeled SUMMARY. Row labels are expressed in the format of “Category_Item”. A category with no item was expressed simply as “Category.” Abbreviations: CT, Computerized Tomography; DNA, DeoxyriboNucleic Acid; MRI, Magnetic Resonance Imaging; MRN, Medical Record Number; SES, Social Economic Status; STD, Sexually Transmitted Disease; TxProcedure, Treatment Procedure.



Site (#2 vs. #1)

Variable	Odds Ratio (95% CI)
SexualLife	0.90 (0.65 – 1.25)
SES_MaritalStatus	0.79 (0.51 – 1.21)
Encounter_ClinicSpecialty	0.78 (0.51 – 1.16)
ContactInfo_HomeAddress	0.77 (0.55 – 1.08)
Lifestyle_Smoking	0.77 (0.50 – 1.15)
ContactInfo_Phone	0.75 (0.53 – 1.04)
Encounter_ClinicLocation	0.75 (0.49 – 1.12)
Encounter_PhysicianName	0.74 (0.49 – 1.11)
Encounter_MRN	0.73 (0.49 – 1.08)
ContactInfo_Email	0.73 (0.51 – 1.03)
Lifestyle_Alcohol	0.73 (0.47 – 1.09)
Lifestyle_Exercise	0.72 (0.45 – 1.11)
Demographics_SexualOrientation	0.72 (0.45 – 1.11)
AdoptionHistory	0.72 (0.51 – 1.00)
Lifestyle_Drug	0.71 (0.47 – 1.07)
SES_Income	0.71 (0.49 – 1.02)
Medication_MentalHealth	0.71 (0.46 – 1.07)
ContactInfo_Name	0.70 (0.48 – 1.00)
Lifestyle_SocialIsolation	0.70 (0.45 – 1.06)
SES_InsuranceStatus	0.69 (0.45 – 1.04)
Encounter_VisitDate	0.69 (0.45 – 1.03)
VitalSigns	0.69 (0.42 – 1.10)
Lifestyle_Diet	0.69 (0.44 – 1.05)
Lifestyle_Stress	0.69 (0.44 – 1.05)
Allergies	0.68 (0.41 – 1.08)
DiseaseCondition_MentalHealth	0.68 (0.44 – 1.02)
FamilyHealthHistory_MentalHealth	0.67 (0.43 – 1.02)
TxProcedure_Cosmetic	0.66 (0.44 – 0.99)
Demographics_Race	0.66 (0.40 – 1.06)
Medication_OtherMedication	0.66 (0.42 – 1.01)
FamilyHealthHistory_SexualReproductive	0.66 (0.43 – 0.99)
PregnancyHistory	0.66 (0.46 – 0.93)
LabTestResults_DNAsequencing	0.66 (0.42 – 1.00)
TxProcedure_MentalHealth	0.65 (0.43 – 0.98)
SES_Education	0.65 (0.42 – 1.00)
Biospecimen_Urine	0.65 (0.42 – 0.98)
SES_Occupation	0.64 (0.41 – 0.98)
FamilyHealthHistory_Other	0.64 (0.42 – 0.97)
Demographics_Ethnicity	0.63 (0.38 – 1.01)
FamilyHealthHistory_SubstanceAbuse	0.63 (0.41 – 0.95)
LabTestResults_GeneticTest	0.63 (0.40 – 0.97)
DiseaseCondition_SexualReproductive	0.62 (0.40 – 0.96)
Biospecimen_Blood	0.62 (0.40 – 0.95)
LabTestResults_STDtest	0.62 (0.39 – 0.95)
Biospecimen_Tissue	0.62 (0.40 – 0.93)
BodyMeasurements	0.61 (0.38 – 0.95)
TxProcedure_Other	0.61 (0.40 – 0.92)
ImagingTest_X.ray	0.60 (0.37 – 0.96)
DiseaseCondition_Other	0.60 (0.39 – 0.91)
TxProcedure_Bariatric	0.60 (0.39 – 0.90)
TxProcedure_GenitouriaryReproductive	0.59 (0.38 – 0.89)
ImagingTest_CTscan	0.59 (0.36 – 0.93)
DiseaseCondition_SubstanceAbuse	0.58 (0.37 – 0.89)
ImagingTest_MRI	0.57 (0.35 – 0.90)
Demographics_Age	0.55 (0.31 – 0.93)
ImagingTest_OtherImagingTest	0.54 (0.34 – 0.84)
LabTestResults_Other	0.53 (0.33 – 0.82)
Demographics_Sex	0.53 (0.30 – 0.88)
LabTestResults_DrugScreening	0.52 (0.33 – 0.82)
SUMMARY	**0.67 (0.63 – 0.71)**

Figure B.13: Forest plot of adjusted odds ratio for site (#2 vs. #1). The 59 sharable items were sorted by adjusted odds ratios from the multivariate model (See the Methods for details) and shown with their 95% confidence intervals. The pooled odds ratio is displayed at the bottom and labeled SUMMARY. Row labels are expressed in the format of "Category_Item". A category with no item was expressed simply as "Category." Abbreviations: CT, Computerized Tomography; DNA, DeoxyriboNucleic Acid; MRI, Magnetic Resonance Imaging; MRN, Medical Record Number; SES, Social Economic Status; STD, Sexually Transmitted Disease; TxProcedure, Treatment Procedure.

Text B.2: Study Protocol

Target Recruitment

The required sample size based on a t-test of the sharing scores is 1,010 for two groups with the following assumptions: significance level 0.05, desired power 0.8, and effect size 0.25. We inflate this number by 18% to have 1,200 to accommodate other covariates for fitting a linear model and possible missing values. We will thus recruit a total of 1,200 adult patients from outpatient clinics of UC San Diego and UC Irvine.

Both email and in-person recruitment will be conducted. With email recruitment, detailed information about the study and instruction on participation are provided with the contact email and phone numbers for further details. With in-person recruitment, a study staff approaches patients at the outpatient clinics, explains the study, and provides the instruction on how to participate, if a patient wishes to participate in the study. Participants access the iCONCUR site and create an account and log into the site. Participants see the informed consent form to sign for participating in this study. Once they sign the form, the participants are recorded as enrolled in the study.

After signing the consent form, participants are first asked to complete a brief health literacy test. After the literacy test, the participant are asked to indicate (1) the perceived level of health on a 5-point scale (1: very bad, 5: very good) and (2) whether they have a minor child(ren) or a critically ill/cognitively impaired adult family member for whom they are serving as a surrogate decision maker.

People who want to participate but do not have access to computers or the internet uses paper forms to complete data sharing preferences and presurvey. They also sign a paper informed consent form.

Participant Randomization

Randomization occurs when participants complete the pre-survey and proceed to the sharing preference indication page. iCONCUR will randomly direct the participant to one of four iCONCUR interfaces (opt-in extended, opt-in simple, opt-out extended, and opt-out simple). Paper form users are randomized to either an extended opt-in form or a simple opt-in form.

Survey process in iCONCUR

Participants are asked to indicate sharing preferences for his/her own data first. If the participant identifies him/herself as a surrogate decision-maker, additional questions about data sharing preferences on behalf of family members will be populated automatically. The participants will be asked to indicate their preference for sharing their care dependents' data. After indicating data sharing preferences, the participants are asked whether they want to receive an email notification when their data are used for research. The MyData Use page is populated for every participant whose data are used for research during the study period. Information on the publications or public presentations resulting from the research is also provided in the MyData Use page.

Notification emails are sent to participants who signed up for this service when new information on data use and/or publication becomes available. Those who did not sign up for the email notification service can see the information in iCONCUR anytime they want to.
Toward the end of the survey period, participants receive an email that asks them to revisit iCONCUR and complete a satisfaction survey. Patients receive a $10 gift card for completing the sharing choice indication and a $10 gift card for completing the satisfaction survey.

Statistical analysis plan

The primary research question is the effect of the different tiered interfaces (extended vs. simple) and the form's default state (opt-in vs. opt-out) on the outcome measure of data sharing decision. Our null hypothesis is that the sharing decision will not differ by different form presentations. We will also assess potential associations among covariates. First, univariate statistics will be explored to evaluate the association of various covariates with data sharing scores. Then we will perform multivariate analyses, where standardized coefficients of linear regression will be assessed for their sign and magnitude, p-value, and 95% confidence interval. The R statistical package will be used for data analysis.

Text B.3: iCONCUR – perceived health status and care dependents

Overall how do you rate your health status?

( ) Very poor
( ) Somewhat poor
( ) Fair
( ) Good
( ) Excellent

Do you have anybody who you are legally responsible to take care of?

Name or nickname	**Relationship**
	Children ( ) Parent ( ) Sibling ( ) Partner ( ) Other ( )
	Children ( ) Parent ( ) Sibling ( ) Partner ( ) Other ( )
	Children ( ) Parent ( ) Sibling ( ) Partner ( ) Other ( )
	Children ( ) Parent ( ) Sibling ( ) Partner ( ) Other ( )
	Children ( ) Parent ( ) Sibling ( ) Partner ( ) Other ( )
	Children ( ) Parent ( ) Sibling ( ) Partner ( ) Other ( )

Text B.4: INFORMED CONSENT FOR STUDY PARTICIPATION

University of California, San Diego

Electronic Consent to Act as a Research Subject
iCONCUR ES: informed CONsent for Clinical data and sample Use in Research Extended Study

Who is conducting the study, why you have been asked to participate, how you were selected, and what is the approximate number of participants in the study?
Lucila Ohno-Machado, MD, PhD and Kai Zheng, PhD are conducting a research study. You have been asked to participate in this study because you are a patient at a UC San Diego or UC Irvine health system facility. There will be approximately 1200 participants at either UC San Diego health systems sites or UC Irvine health system sites.

Why is this study being done?
The purpose of this study is to learn patient preferences towards sharing personal medical data for research.

What will happen to you in this study and which procedures are standard of care and which are experimental?
If you agree to be in this study, the following will happen to you:
After agreeing to this consent form, you will be given a link to the online tool called iCONCUR. After creating an account, you will complete a short health literacy test and will be asked about your health status and whether you have a care dependent(s). Once these questions are answered, you can record your preferences towards sharing your medical data for research by using the online iCONCUR tool. Your data sharing preferences will be honored during the study period.
If you indicated care dependent(s), you will be asked to record if you would allow use of their medical data for future research. However, this decision-making about whether to allow use of your care dependent(s) data for research is hypothetical because we do not have the capability at this time to control whether researchers can access your care dependent(s) medical data for research. At the end of the study, you will be asked to fill out a user-satisfaction survey.
You will be assigned by chance to a study group. Your chance of being assigned to each group is 50/50. Neither you nor the researchers will choose which group you are included in. Some of the design and functionality of the online iCONCUR tool may differ depending on which group you are assigned to. The differences do not affect your ability to make sharing choices.
Participation to this study DOES NOT affect the care that you receive from any clinic.

How much time will each study procedure take, what is your total time commitment, and how long will the study last?

This consent process will take 15 minutes or less. You will take around 30 minutes for the initial account set up and health literacy questionnaire. After initial use, you can choose to update your information and preferences as often as you like. Each further use will take around 20 minutes. The user satisfaction survey that will be given at the end of the study will take 20 minutes to complete. This study will last for 18 months after recruitment begins.

What risks are associated with this study?

Participation in this study may involve some added risks or discomforts. These include the following: boredom with the consent or sharing choice process, a potential loss of confidentiality if someone unauthorized gets access to the database where your sharing choices are kept. You may also feel that the study group you are assigned to might not be the group you would prefer to be in. However, please note that you will be assigned to a study group at random (by chance). Your assignment is based on chance rather than a medical decision made by the researchers.

Because this is a research study, there may be some unknown risks that are currently unforeseeable. You will be informed of any significant new findings.

What are the alternatives to participating in this study?

The alternatives to participation in this study are to decline to participate.

What benefits can be reasonably expected?

There may or may not be any direct benefit to you from this study. Participating in this study will give you an opportunity to share your preferences towards having your personal medical data used for future research. And we will honor your preferences during the period of this study. This means that if a researcher requests your medical data for her research while iCONCUR is on-going, we will check your data sharing preferences and won’t provide your medical data if you indicate that you don’t want to share your medical data for research.
Also, the investigator(s) conducting this study may learn more about patient preferences for keeping personal medical data private.

Can you choose to not participate or withdraw from the study without penalty or loss of benefits?

Participation in research is entirely voluntary. You may refuse to participate or withdraw at any time without penalty or loss of benefits to which you are entitled. If you decide that you no longer wish to continue in this study, you will be requested to: call or email research staff to let them know you are withdrawing.

You will be told if any important new information is found during the course of this study that may affect your wanting to continue.

Can you be withdrawn from the study without your consent?
You may be withdrawn from the study for the following reasons: if you do not follow through with using the iCONCUR tool or do not follow the instructions given you by the study personnel.

Will you be compensated for participating in this study?
In compensation for your time, you can receive up to $20 in electronic Amazon gift cards for participating in this research. Ten dollars will be given after you use the iCONCUR tool to record your sharing choices, and an additional $10 will be given if you complete the user satisfaction survey.

Are there any costs associated with participating in this study?
There will be no cost to you for participating in this study.

What about your confidentiality?
Research records will be kept confidential to the extent allowed by law. Your consent form, sharing choices, and any personal information collected during the study will be kept in a secure server. The UC San Diego Institutional Review Board and NIH may review research records.

Who can you call if you have questions?
If you have questions or research-related problems, you may reach research staff at 858-246-1281 or by emailing iconcur@ucsd.edu.

You may call the Human Research Protections Program Office at UC San Diego at (858) 657-5100 to inquire about your rights as a research subject or to report research-related problems.

By clicking "You agree" below you are indicating that you are at least 18 years old, have read this consent form, and agree to participate in this research study. Please print a copy of this page for your records.

__ I agree
__ I do not agree

Electronic signature capture box:

Text B.5: iCONCUR satisfaction survey questions

1. Which gender do you identify as?
   - ( ) Male
   - ( ) Female
   - ( ) Other

2. What race(s) do you identify as (check all that apply)?
   - ( ) American Indian or Alaska Native
   - ( ) Asian
   - ( ) Black or African American
   - ( ) Native Hawaiian or Other Pacific Islander
   - ( ) White
   - ( ) Other

3. Do you identify as Hispanic or Latino/a?
   - ( ) Yes
   - ( ) No

4. What is your highest education level?
   - ( ) High school or less
   - ( ) Beyond high school or <4 years college
   - ( ) 4 year college graduate
   - ( ) Graduate or professional school

5. What is your annual household income?
   - ( ) <$25,000
   - ( ) $25,000-$75,000
   - ( ) $75,000-$125,000
   - ( ) $125,000-$200,000
   - ( ) >$200,000

6. How did you hear about this study?
   - ( ) I got an email about it
   - ( ) I saw a flyer or pamphlet at a clinic
   - ( ) Someone talked to me at a clinic
   - ( ) I don’t remember
   - ( ) A friend or family member told me
   - ( ) Other

7. What were your motivations for participating? (Check all that apply)

( ) I wanted to have control over my data and how it is used
( ) Because of a positive experience with other research studies
( ) To help others
( ) To get compensation
( ) To find out more about medical research in general
( ) I like the focus of this study about controlling access to my data
( ) Someone encouraged me or recommended it to me
( ) I wanted to share my opinion

8. Would you be willing to participate in other studies? We will not use this to contact you in the future but would like to gauge your overall level of willingness.

( ) Yes
( ) No
( ) Depends on the focus of the study
( ) Not sure

9. Have you participated in other research studies in the past?

( ) Yes
( ) No
( ) I don't remember

10. Have you checked the "My Data Use" tab in iCONCUR to see if your data were used by a researcher during your time in this study?

( ) Yes
( ) No

11. Do you think the categories of sharing choices are adequate?

( ) Yes
( ) No
( ) Other (please explain):

12. Is there anything we didn't offer that you think people might want to keep private in their medical record? (Text answer box)

13. Does having these choices make you feel differently about sharing your medical data?

( ) No change
( ) It makes me more willing to share my medical data
( ) It makes me less willing to share my medical data
( ) Other (please explain):

14. Would you feel more comfortable sharing your information if you know who is using it for research?

( ) Yes, I would feel more comfortable
( ) No, I would feel less comfortable
( ) It doesn’t change my comfort level
( ) Other (please explain)

15. Would you like to know about any of the following (Check all the ones that apply)?

( ) What kind of organization the researchers using my data belong to (for example, a profit/non-profit organization, university, healthcare system)
( ) What was the aim of their research
( ) What papers they published using your data
( ) What were the outcomes of their research
( ) Other (please explain):

16. How does your willingness to share your medical data for healthcare (for example, treating patients and improving healthcare environment) compare to your willingness to share for research?

( ) I would be equally willing to share my data for both
( ) I would be more willing to share my data for healthcare than research
( ) I would be more willing to share my data for research than healthcare
( ) I don’t know
( ) Other (please explain):

17. Did you have trouble understanding any of the information presented in the tool?

( ) Yes
( ) No
( ) Other (please explain):

18. Was your experience with iCONCUR satisfactory?

( ) Yes I enjoyed participating
( ) No I was not satisfied with this study: [Text box]

19. Please provide any comments you have on this tool or the study: [Text box]

Text B.6: A Short Assessment of Health Literacy (SAHL) – English Version

Mark the word that is most relevant to the word shown the left-most, bolded column

1 kidney	__urine	__fever	__don’t know
2 occupation	__work	__education	__don’t know
3 medication	__instrument	__treatment	__don’t know
4 nutrition	__healthy	__soda	__don’t know
5 miscarriage	__loss	__marriage	__don’t know
6 infection	__plant	__virus	__don’t know
7 alcoholism	__addiction	__recreation	__don’t know
8 pregnancy	__birth	__childhood	__don’t know
9 seizure	__dizzy	__calm	__don’t know
10 dose	__sleep	__amount	__don’t know
11 hormones	__growth	__harmony	__don’t know
12 abnormal	__different	__similar	__don’t know
13 directed	__instruction	__decision	__don’t know
14 nerves	__bored	__anxiety	__don’t know
15 constipation	__blocked	__loose	__don’t know
16 diagnosis	__evaluation	__recovery	__don’t know
17 hemorrhoids	__veins	__heart	__don’t know
18 syphilis	__contraception	__condom	__don’t know

Source: Lee SY, Stucky BD, Lee JY, Rozier RG, Bender DE. Short Assessment of Health Literacy-Spanish and English: a comparable test of health literacy for Spanish and English speakers. Health Serv Res. 2010 Aug;45(4):1105-20. doi:10.1111/j.1475-6773.2010.01119.x. Epub 2010 May 24. PMID: 20500222; PMCID: PMC2910571.

Bibliography

1. Tsai, C. H., Eghdam, A., Davoody, N., Wright, G., Flowerday, S. & Koch, S. Effects of Electronic Health Record Implementation and Barriers to Adoption and Use: A Scoping Review and Qualitative Analysis of the Content. *Life* **10,** (2020).
2. Seh, A. H., Zarour, M., Alenezi, M., Sarkar, A. K., Agrawal, A., Kumar, R. & Khan, R. A. Healthcare Data Breaches: Insights and Implications. *Healthcare (Basel)* **8,** (2020).
3. Sulmasy, L. S., López, A. M., Horwitch, C. A. & , American College of Physicians Ethics, Professionalism and Human Rights Committee. Ethical Implications of the Electronic Health Record: In the Service of the Patient. *J. Gen. Intern. Med.* **32,** 935–939 (2017).
4. Kumar, A., Guss, Z. D., Courtney, P. T., Nalawade, V., Sheridan, P., Sarkar, R. R., Banegas, M. P., Rose, B. S., Xu, R. & Murphy, J. D. Evaluation of the Use of Cancer Registry Data for Comparative Effectiveness Research. *JAMA Netw Open* **3,** e2011985 (2020).
5. Lee, C. S., Blazes, M., Lorch, A., Pershing, S., Hyman, L., Ho, A. C., Haller, J., Miller, J. W., Chew, E. Y., Lum, F. & Lee, A. Y. American Academy of Ophthalmology Intelligent Research in Sight (IRIS®) Registry and the IRIS Registry Analytic Center Consortium. *Ophthalmol Sci* **2,** 100112 (2022).
6. Ohno-Machado, L., Bafna, V., Boxwala, A. A., Chapman, B. E., Chapman, W. W., Chaudhuri, K., Day, M. E., Farcas, C., Heintzman, N. D., Jiang, X., Kim, H., Kim, J., Matheny, M. E., Resnic, F. S., Vinterbo, S. A. & iDASH team. iDASH: integrating data for analysis, anonymization, and sharing. *J. Am. Med. Inform. Assoc.* **19,** 196–201 (2012).
7. All of Us Research Program Investigators, Denny, J. C., Rutter, J. L., Goldstein, D. B., Philippakis, A., Smoller, J. W., Jenkins, G. & Dishman, E. The ‘All of Us’ Research

Program. *N. Engl. J. Med.* **381,** 668–676 (2019).

8. Gaziano, J. M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen, J., Shannon, C., Humphries, D., Guarino, P., Aslan, M., Anderson, D., LaFleur, R., Hammond, T., Schaa, K., Moser, J., Huang, G., Muralidhar, S., Przygodzki, R. & O'Leary, T. J. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70,** 214–223 (2016).
9. Kroes, J. A., Bansal, A. T., Berret, E., Christian, N., Kremer, A., Alloni, A., Gabetta, M., Marshall, C., Wagers, S., Djukanovic, R., Porsbjerg, C., Hamerlijnck, D., Fulton, O., Ten Brinke, A., Bel, E. H. & Sont, J. K. Blueprint for harmonising unstandardised disease registries to allow federated data analysis: prepare for the future. *ERJ Open Res* **8,** (2022).
10. Yu, Y. W. & Weber, G. M. Balancing Accuracy and Privacy in Federated Queries of Clinical Data Repositories: Algorithm Development and Validation. *J. Med. Internet Res.* **22,** e18735 (2020).
11. Fleurence, R. L., Curtis, L. H., Califf, R. M., Platt, R., Selby, J. V. & Brown, J. S. Launching PCORnet, a national patient-centered clinical research network. *J. Am. Med. Inform. Assoc.* **21,** 578–582 (2014).
12. Visweswaran, S., Becich, M. J., D'Itri, V. S., Sendro, E. R., MacFadden, D., Anderson, N. R., Allen, K. A., Ranganathan, D., Murphy, S. N., Morrato, E. H., Pincus, H. A., Toto, R., Firestein, G. S., Nadler, L. M. & Reis, S. E. Accrual to Clinical Trials (ACT): A Clinical and Translational Science Award Consortium Network. *JAMIA Open* **1,** 147–152 (2018).
13. Biedermann, P., Ong, R., Davydov, A., Orlova, A., Solovyev, P., Sun, H., Wetherill, G., Brand, M. & Didden, E.-M. Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases. *BMC Med. Res.*

Methodol. **21,** 238 (2021).

14. Weber, G. M., Murphy, S. N., McMurry, A. J., Macfadden, D., Nigrin, D. J., Churchill, S. & Kohane, I. S. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. *J. Am. Med. Inform. Assoc.* **16,** 624–630 (2009).

15. Wu, Y., Jiang, X., Kim, J. & Ohno-Machado, L. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. *J. Am. Med. Inform. Assoc.* **19,** 758–764 (2012).

16. Zhang, Z., Mo, L., Huang, C., Xu, P. & written on behalf of AME Big-Data Clinical Trial Collaborative Group. Binary logistic regression modeling with TensorFlow^{TM}. *Ann Transl Med* **7,** 591 (2019).

17. Anthony Reina, G., Gruzdev, A., Foley, P., Perepelkina, O., Sharma, M., Davidyuk, I., Trushkin, I., Radionov, M., Mokrov, A., Agapov, D., Martin, J., Edwards, B., Sheller, M. J., Pati, S., Moorthy, P. N., Wang, S.-H., Shah, P. & Bakas, S. OpenFL: An open-source framework for Federated Learning. *arXiv [cs.LG]* (2021). at <http://arxiv.org/abs/2105.06413>

18. Sadilek, A., Liu, L., Nguyen, D., Kamruzzaman, M., Serghiou, S., Rader, B., Ingerman, A., Mellem, S., Kairouz, P., Nsoesie, E. O., MacFarlane, J., Vullikanti, A., Marathe, M., Eastham, P., Brownstein, J. S., Arcas, B. A. Y., Howell, M. D. & Hernandez, J. Privacy-first health research with federated learning. *NPJ Digit Med* **4,** 132 (2021).

19. Naz, S., Phan, K. T. & Chen, Y. P. A comprehensive review of federated learning for COVID-19 detection. *Int. J. Intell. Syst.* **37,** 2371 (2022).

20. Garrison, N. A. Genomic Justice for Native Americans: Impact of the Havasupai Case on

Genetic Research. *Sci. Technol. Human Values* **38,** 201–223 (2013).

21. Rupp, S. R. Making Room for Patient Autonomy in Health Information Exchange: The Role of Informed Consent Comment. *St. Louis U. L.J.* **56,** 885–916 (2011-2012).

22. Wright, A., Nelson, S., Rubins, D., Schreiber, R. & Sittig, D. F. Clinical decision support malfunctions related to medication routes: a case series. *J. Am. Med. Inform. Assoc.* **29,** 1972–1975 (2022).

23. O'Reilly-Shah, V. N., Gentry, K. R., Van Cleve, W., Kendale, S. M., Jabaley, C. S. & Long, D. R. The COVID-19 Pandemic Highlights Shortcomings in US Health Care Informatics Infrastructure: A Call to Action. *Anesth. Analg.* **131,** 340–344 (2020).

24. Moradian, N., Ochs, H. D., Sedikies, C., Hamblin, M. R., Camargo, C. A., Jr, Martinez, J. A., Biamonte, J. D., Abdollahi, M., Torres, P. J., Nieto, J. J., Ogino, S., Seymour, J. F., Abraham, A., Cauda, V., Gupta, S., Ramakrishna, S., Sellke, F. W., Sorooshian, A., Wallace Hayes, A., Martinez-Urbistondo, M., Gupta, M., Azadbakht, L., Esmaillzadeh, A., Kelishadi, R., Esteghamati, A., Emam-Djomeh, Z., Majdzadeh, R., Palit, P., Badali, H., Rao, I., Saboury, A. A., Jagan Mohan Rao, L., Ahmadieh, H., Montazeri, A., Fadini, G. P., Pauly, D., Thomas, S., Moosavi-Movahed, A. A., Aghamohammadi, A., Behmanesh, M., Rahimi-Movaghar, V., Ghavami, S., Mehran, R., Uddin, L. Q., Von Herrath, M., Mobasher, B. & Rezaei, N. The urgent need for integrated science to fight COVID-19 pandemic and beyond. *J. Transl. Med.* **18,** 205 (2020).

25. Eibensteiner, F., Ritschl, V., Ariceta, G., Jankauskiene, A., Klaus, G., Paglialonga, F., Edefonti, A., Ranchin, B., Schmitt, C. P., Shroff, R., Stefanidis, C. J., Walle, J. V., Verrina, E., Vondrak, K., Zurowska, A., Stamm, T., Aufricht, C. & European Pediatric Dialysis Working Group. Rapid response in the COVID-19 pandemic: a Delphi study from the

European Pediatric Dialysis Working Group. *Pediatr. Nephrol.* **35,** 1669–1678 (2020).

26. Reeves, J. J., Hollandsworth, H. M., Torriani, F. J., Taplitz, R., Abeles, S., Tai-Seale, M., Millen, M., Clay, B. J. & Longhurst, C. A. Rapid response to COVID-19: health informatics support for outbreak management in an academic health system. *J. Am. Med. Inform. Assoc.* **27,** 853–859 (2020).

27. Longhurst, C. A., Harrington, R. A. & Shah, N. H. A ‘green button’ for using aggregate patient data at the point of care. *Health Aff.* **33,** 1229–1235 (2014).

28. Drew, D. A., Nguyen, L. H., Steves, C. J., Menni, C., Freydin, M., Varsavsky, T., Sudre, C. H., Cardoso, M. J., Ourselin, S., Wolf, J., Spector, T. D., Chan, A. T. & COPE Consortium. Rapid implementation of mobile technology for real-time epidemiology of COVID-19. *Science* **368,** 1362–1367 (2020).

29. Chan, A. T., Drew, D. A., Nguyen, L. H., Joshi, A. D., Ma, W., Guo, C.-G., Lo, C.-H., Mehta, R. S., Kwon, S., Sikavi, D. R., Magicheva-Gupta, M. V., Fatehi, Z. S., Flynn, J. J., Leonardo, B. M., Albert, C. M., Andreotti, G., Beane-Freeman, L. E., Balasubramanian, B. A., Brownstein, J. S., Bruinsma, F., Cowan, A. N., Deka, A., Ernst, M. E., Figueiredo, J. C., Franks, P. W., Gardner, C. D., Ghobrial, I. M., Haiman, C. A., Hall, J. E., Deming-Halverson, S. L., Kirpach, B., Lacey, J. V., Jr, Marchand, L. L., Marinac, C. R., Martinez, M. E., Milne, R. L., Murray, A. M., Nash, D., Palmer, J. R., Patel, A. V., Rosenberg, L., Sandler, D. P., Sharma, S. V., Schurman, S. H., Wilkens, L. R., Chavarro, J. E., Eliassen, A. H., Hart, J. E., Kang, J. H., Koenen, K. C., Kubzansky, L. D., Mucci, L. A., Ourselin, S., Rich-Edwards, J. W., Song, M., Stampfer, M. J., Steves, C. J., Willett, W. C., Wolf, J., Spector, T. & COPE Consortium. The COronavirus Pandemic Epidemiology (COPE) Consortium: A Call to Action. *Cancer Epidemiol. Biomarkers Prev.* **29,** 1283–1289 (2020).

30. Terpos, E., Ntanasis-Stathopoulos, I., Elalamy, I., Kastritis, E., Sergentanis, T. N., Politou, M., Psaltopoulou, T., Gerotziafas, G. & Dimopoulos, M. A. Hematological findings and complications of COVID-19. *Am. J. Hematol.* **95,** 834–847 (2020).

31. Meng, X., Deng, Y., Dai, Z. & Meng, Z. COVID-19 and anosmia: A review based on up-to-date knowledge. *Am. J. Otolaryngol.* **41,** 102581 (2020).

32. Richardson, S., Hirsch, J. S., Narasimhan, M., Crawford, J. M., McGinn, T., Davidson, K. W., the Northwell COVID-19 Research Consortium, Barnaby, D. P., Becker, L. B., Chelico, J. D., Cohen, S. L., Cookingham, J., Coppa, K., Diefenbach, M. A., Dominello, A. J., Duer-Hefele, J., Falzon, L., Gitlin, J., Hajizadeh, N., Harvin, T. G., Hirschwerk, D. A., Kim, E. J., Kozel, Z. M., Marrast, L. M., Mogavero, J. N., Osorio, G. A., Qiu, M. & Zanos, T. P. Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. *JAMA* **323,** 2052–2059 (2020).

33. Güemes-Villahoz, N., Burgos-Blasco, B., Arribi-Vilela, A., Arriola-Villalobos, P., Rico-Luna, C. M., Cuiña-Sardiña, R., Delgado-Iribarren, A. & García-Feijoó, J. Detecting SARS-CoV-2 RNA in conjunctival secretions: Is it a valuable diagnostic method of COVID-19? *J. Med. Virol.* **93,** 383–388 (2021).

34. Mehra, M. R., Desai, S. S., Kuy, S., Henry, T. D. & Patel, A. N. Retraction: Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19. N Engl J Med. DOI: 10.1056/NEJMoa2007621. *N. Engl. J. Med.* **382,** 2582 (2020).

35. Dong, X., Li, J., Soysal, E., Bian, J., DuVall, S. L., Hanchrow, E., Liu, H., Lynch, K. E., Matheny, M., Natarajan, K., Ohno-Machado, L., Pakhomov, S., Reeves, R. M., Sitapati, A. M., Abhyankar, S., Cullen, T., Deckard, J., Jiang, X., Murphy, R. & Xu, H. COVID-19 TestNorm: A tool to normalize COVID-19 testing names to LOINC codes. *J. Am. Med.*

Inform. Assoc. **27,** 1437–1442 (2020).

36. Hripcsak, G., Duke, J. D., Shah, N. H., Reich, C. G., Huser, V., Schuemie, M. J., Suchard, M. A., Park, R. W., Wong, I. C. K., Rijnbeek, P. R., van der Lei, J., Pratt, N., Norén, G. N., Li, Y.-C., Stang, P. E., Madigan, D. & Ryan, P. B. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud. Health Technol. Inform.* **216,** 574–578 (2015).
37. Weiskopf, N. G., Bakken, S., Hripcsak, G. & Weng, C. A Data Quality Assessment Guideline for Electronic Health Record Data Reuse. *EGEMS (Wash DC)* **5,** 14 (2017).
38. Wang, X., Zhou, Y., Jiang, N., Zhou, Q. & Ma, W.-L. Persistence of intestinal SARS-CoV-2 infection in patients with COVID-19 leads to re-admission after pneumonia resolved. *Int. J. Infect. Dis.* **95,** 433–435 (2020).
39. Zheng, J., Zhou, R., Chen, F., Tang, G., Wu, K., Li, F., Liu, H., Lu, J., Zhou, J., Yang, Z., Yuan, Y., Lei, C. & Wu, X. Incidence, clinical course and risk factor for recurrent PCR positivity in discharged COVID-19 patients in Guangzhou, China: A prospective cohort study. *PLoS Negl. Trop. Dis.* **14,** e0008648 (2020).
40. Lewnard, J. A., Liu, V. X., Jackson, M. L., Schmidt, M. A., Jewell, B. L., Flores, J. P., Jentz, C., Northrup, G. R., Mahmud, A., Reingold, A. L., Petersen, M., Jewell, N. P., Young, S. & Bellows, J. Incidence, clinical outcomes, and transmission dynamics of severe coronavirus disease 2019 in California and Washington: prospective cohort study. *BMJ* **369,** m1923 (2020).
41. Somani, S. S., Richter, F., Fuster, V., De Freitas, J. K., Naik, N., Sigel, K., Mount Sinai COVID Informatics Center, Bottinger, E. P., Levin, M. A., Fayad, Z., Just, A. C., Charney, A. W., Zhao, S., Glicksberg, B. S., Lala, A. & Nadkarni, G. N. Characterization of Patients

Who Return to Hospital Following Discharge from Hospitalization for COVID-19. *J. Gen. Intern. Med.* **35,** 2838–2844 (2020).

42. Holman, N., Knighton, P., Kar, P., O'Keefe, J., Curley, M., Weaver, A., Barron, E., Bakhai, C., Khunti, K., Wareham, N. J., Sattar, N., Young, B. & Valabhji, J. Risk factors for COVID-19-related mortality in people with type 1 and type 2 diabetes in England: a population-based cohort study. *Lancet Diabetes Endocrinol* **8,** 823–833 (2020).
43. Liu, Y., Du, X., Chen, J., Jin, Y., Peng, L., Wang, H. H. X., Luo, M., Chen, L. & Zhao, Y. Neutrophil-to-lymphocyte ratio as an independent risk factor for mortality in hospitalized patients with COVID-19. *J. Infect.* **81,** e6–e12 (2020).
44. Center for Disease Control and Prevention. ICD-10-CM Official Guidelines for Coding and Reporting FY 2020 (October 1, 2019 - September 30, 2020). *Center for Disease Control and Prevention* Preprint at https://www.cdc.gov/nchs/data/icd/10cmguidelines-FY2020_final.pdf (2020)
45. Haendel, M. A., Chute, C. G., Bennett, T. D., Eichmann, D. A., Guinney, J., Kibbe, W. A., Payne, P. R. O., Pfaff, E. R., Robinson, P. N., Saltz, J. H., Spratt, H., Suver, C., Wilbanks, J., Wilcox, A. B., Williams, A. E., Wu, C., Blacketer, C., Bradford, R. L., Cimino, J. J., Clark, M., Colmenares, E. W., Francis, P. A., Gabriel, D., Graves, A., Hemadri, R., Hong, S. S., Hripscak, G., Jiao, D., Klann, J. G., Kostka, K., Lee, A. M., Lehmann, H. P., Lingrey, L., Miller, R. T., Morris, M., Murphy, S. N., Natarajan, K., Palchuk, M. B., Sheikh, U., Solbrig, H., Visweswaran, S., Walden, A., Walters, K. M., Weber, G. M., Zhang, X. T., Zhu, R. L., Amor, B., Girvin, A. T., Manna, A., Qureshi, N., Kurilla, M. G., Michael, S. G., Portilla, L. M., Rutter, J. L., Austin, C. P., Gersing, K. R. & N3C Consortium. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. *J.*

Am. Med. Inform. Assoc. **28,** 427–443 (2021).

46. Lane, J. C. E., Weaver, J., Kostka, K., Duarte-Salles, T., Abrahao, M. T. F., Alghoul, H., Alser, O., Alshammari, T. M., Biedermann, P., Banda, J. M., Burn, E., Casajust, P., Conover, M. M., Culhane, A. C., Davydov, A., DuVall, S. L., Dymshyts, D., Fernandez-Bertolin, S., Fišter, K., Hardin, J., Hester, L., Hripcsak, G., Kaas-Hansen, B. S., Kent, S., Khosla, S., Kolovos, S., Lambert, C. G., van der Lei, J., Lynch, K. E., Makadia, R., Margulis, A. V., Matheny, M. E., Mehta, P., Morales, D. R., Morgan-Stewart, H., Mosseveld, M., Newby, D., Nyberg, F., Ostropolets, A., Park, R. W., Prats-Uribe, A., Rao, G. A., Reich, C., Reps, J., Rijnbeek, P., Sathappan, S. M. K., Schuemie, M., Seager, S., Sena, A. G., Shoaibi, A., Spotnitz, M., Suchard, M. A., Torre, C. O., Vizcaya, D., Wen, H., de Wilde, M., Xie, J., You, S. C., Zhang, L., Zhuk, O., Ryan, P., Prieto-Alhambra, D. & OHDSI-COVID-19 consortium. Risk of hydroxychloroquine alone and in combination with azithromycin in the treatment of rheumatoid arthritis: a multinational, retrospective study. *Lancet Rheumatol* **2,** e698–e711 (2020).
47. Moore, W. & Frye, S. Review of HIPAA, Part 1: History, Protected Health Information, and Privacy and Security Rules. *J. Nucl. Med. Technol.* **47,** 269–272 (2019).
48. Berkman, B. E., Wendler, D., Sullivan, H. K. & Grady, C. A Proposed Process for Reliably Updating the Common Rule. *Am. J. Bioeth.* **17,** 8–14 (2017).
49. Murphy, J. F. A. The General Data Protection Regulation (GDPR). *Ir. Med. J.* **111,** 747 (2018).
50. California Consumer Privacy Act (CCPA). *State of California - Department of Justice - Office of the Attorney General* (2018). at <https://oag.ca.gov/privacy/ccpa>
51. Pletcher, M. J., Forrest, C. B. & Carton, T. W. PCORnet's Collaborative Research Groups.

Patient Relat. Outcome Meas. **9,** 91–95 (2018).

52. Monsey, L., Best, L. G., Zhu, J., DeCroo, S. & Anderson, M. Z. The association of mannose binding lectin genotype and immune response to Chlamydia pneumoniae: The Strong Heart Study. *PLoS One* **14,** e0210640 (2019).

53. Charbonneau, J., Nicol, D., Chalmers, D., Kato, K., Yamamoto, N., Walshe, J. & Critchley, C. Public reactions to direct-to-consumer genetic health tests: A comparison across the US, UK, Japan and Australia. *Eur. J. Hum. Genet.* **28,** 339–348 (2020).

54. 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A. & Abecasis, G. R. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

55. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* **19,** A68–77 (2015).

56. Chen, F., Wang, S., Jiang, X., Ding, S., Lu, Y., Kim, J., Sahinalp, S. C., Shimizu, C., Burns, J. C., Wright, V. J., Png, E., Hibberd, M. L., Lloyd, D. D., Yang, H., Telenti, A., Bloss, C. S., Fox, D., Lauter, K. & Ohno-Machado, L. PRINCESS: Privacy-protecting Rare disease International Network Collaboration via Encryption through Software guard extensionS. *Bioinformatics* **33,** 871–878 (2017).

57. Lu, C.-L., Wang, S., Ji, Z., Wu, Y., Xiong, L., Jiang, X. & Ohno-Machado, L. WebDISCO: a web service for distributed cox model learning without patient-level data sharing. *J. Am. Med. Inform. Assoc.* **22,** 1212–1219 (2015).

58. Karr, A. F., Lin, X., Sanil, A. P. & Reiter, J. P. Secure Regression on Distributed Databases. *J. Comput. Graph. Stat.* **14,** 263–279 (2005).

59. Slavkovic, A. B., Nardi, Y. & Tibbits, M. M. 'Secure' Logistic Regression of Horizontally

and Vertically Partitioned Distributed Databases. in *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)* 723–728 (2007).

60. Nardi, Y., Fienberg, S. E. & Hall, R. J. Achieving Both Valid and Secure Logistic Regression Analysis on Aggregated Data from Different Private Sources. *JPC J. Planar Chromatogr. - Mod. TLC* **4,** (2012).

61. Yu, H., Vaidya, J. & Jiang, X. Privacy-Preserving SVM Classification on Vertically Partitioned Data. in *Advances in Knowledge Discovery and Data Mining* 647–656 (Springer Berlin Heidelberg, 2006).

62. Li, Y., Jiang, X., Wang, S., Xiong, H. & Ohno-Machado, L. VERTIcal Grid lOgistic regression (VERTIGO). *J. Am. Med. Inform. Assoc.* **23,** 570–579 (2016).

63. Hosmer. *Applied Logistic Regression, 3rd Edition*. (Wiley, 2013).

64. Reilly, M. P., Li, M., He, J., Ferguson, J. F., Stylianou, I. M., Mehta, N. N., Burnett, M. S., Devaney, J. M., Knouff, C. W., Thompson, J. R., Horne, B. D., Stewart, A. F. R., Assimes, T. L., Wild, P. S., Allayee, H., Nitschke, P. L., Patel, R. S., Myocardial Infarction Genetics Consortium, Wellcome Trust Case Control Consortium, Martinelli, N., Girelli, D., Quyyumi, A. A., Anderson, J. L., Erdmann, J., Hall, A. S., Schunkert, H., Quertermous, T., Blankenberg, S., Hazen, S. L., Roberts, R., Kathiresan, S., Samani, N. J., Epstein, S. E. & Rader, D. J. Identification of ADAMTS7 as a novel locus for coronary atherosclerosis and association of ABO with myocardial infarction in the presence of coronary atherosclerosis: two genome-wide association studies. *Lancet* **377,** 383–392 (2011).

65. Minka, T. P. A comparison of numerical optimizers for logistic regression. (2003). at <https://tminka.github.io/papers/logreg/minka-logreg.pdf>

66. Seber, G. A. F. & Lee, A. J. *Linear Regression Analysis, 2nd Edition | Wiley*. (2003).

67. Snyman, J. A. & Wilke, D. N. *Practical Mathematical Optimization*. (Springer International Publishing).

68. Lee, E. T., Welty, T. K., Fabsitz, R., Cowan, L. D., Le, N. A., Oopik, A. J., Cucchiara, A. J., Savage, P. J. & Howard, B. V. The Strong Heart Study. A study of cardiovascular disease in American Indians: design and methods. *Am. J. Epidemiol.* **132,** 1141–1155 (1990).

69. Howard, B. V., Lee, E. T., Cowan, L. D., Devereux, R. B., Galloway, J. M., Go, O. T., Howard, W. J., Rhoades, E. R., Robbins, D. C., Sievers, M. L. & Welty, T. K. Rising tide of cardiovascular disease in American Indians. The Strong Heart Study. *Circulation* **99,** 2389–2395 (1999).

70. Lee, E. T., Howard, B. V., Wang, W., Welty, T. K., Galloway, J. M., Best, L. G., Fabsitz, R. R., Zhang, Y., Yeh, J. & Devereux, R. B. Prediction of coronary heart disease in a population with high prevalence of diabetes and albuminuria: the Strong Heart Study. *Circulation* **113,** 2897–2905 (2006).

71. Kahn, M. G., Batson, D. & Schilling, L. M. Data model considerations for clinical effectiveness researchers. *Med. Care* **50 Suppl,** S60–7 (2012).

72. Jauregui, B., Hudson, L. D., Becnel, L. B., Fitzmartin, R. & Kush, R. The Turning Point for Clinical Research: Global Data Standardization. (2019). at <http://dx.doi.org/>

73. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., ’t Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A.,

Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016).

74. Haberson, A., Rinner, C., Schöberl, A. & Gall, W. Feasibility of Mapping Austrian Health Claims Data to the OMOP Common Data Model. *J. Med. Syst.* **43,** (2019).

75. Garza, M., Del Fiol, G., Tenenbaum, J., Walden, A. & Zozus, M. N. Evaluating common data models for use with a longitudinal community registry. *J. Biomed. Inform.* **64,** 333–341 (2016).

76. Wood, F. & Guinter, T. Evolution and Implementation of the CDISC Study Data Tabulation Model (SDTM). *Pharmaceutical Programming* **1,** 20–27 (2008).

77. Corley, D. A., Feigelson, H. S., Lieu, T. A. & McGlynn, E. A. Building Data Infrastructure to Evaluate and Improve Quality: PCORnet. *J. Oncol. Pract.* **11,** 204–206 (2015).

78. Ball, R., Robb, M., Anderson, S. A. & Dal Pan, G. The FDA's sentinel initiative--A comprehensive approach to medical product surveillance. *Clin. Pharmacol. Ther.* **99,** 265–268 (2016).

79. Lamer, A., Depas, N., Doutreligne, M., Parrot, A., Verloop, D., Defebvre, M.-M., Ficheur, G., Chazard, E. & Beuscart, J.-B. Transforming French Electronic Health Records into the Observational Medical Outcome Partnership's Common Data Model: A Feasibility Study. *Appl. Clin. Inform.* **11,** 13–22 (2020).

80. Maier, C., Lang, L., Storf, H., Vormstein, P., Bieber, R., Bernarding, J., Herrmann, T., Haverkamp, C., Horki, P., Laufer, J., Berger, F., Höning, G., Fritsch, H. W., Schüttler, J., Ganslandt, T., Prokosch, H. U. & Sedlmayr, M. Towards Implementation of OMOP in a

German University Hospital Consortium. *Appl. Clin. Inform.* **9,** 54–61 (2018).

81. Voss, E. A., Makadia, R., Matcho, A., Ma, Q., Knoll, C., Schuemie, M., DeFalco, F. J., Londhe, A., Zhu, V. & Ryan, P. B. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J. Am. Med. Inform. Assoc.* **22,** 553–564 (2015).

82. Carus, J., Nürnberg, S., Ückert, F., Schlüter, C. & Bartels, S. Mapping Cancer Registry Data to the Episode Domain of the Observational Medical Outcomes Partnership Model (OMOP). *NATO Adv. Sci. Inst. Ser. E Appl. Sci.* **12,** 4010 (2022).

83. Pacheco, C. M., Daley, S. M., Brown, T., Filippi, M., Greiner, K. A. & Daley, C. M. Moving forward: breaking the cycle of mistrust between American Indians and researchers. *Am. J. Public Health* **103,** 2152–2159 (2013).

84. Skewes, M. C. & Lewis, J. P. Sobriety and alcohol use among rural Alaska Native elders. *Int. J. Circumpolar Health* **75,** 30476 (2016).

85. Triplett, C., Fletcher, B. J., Taitingfong, R. I., Zhang, Y., Ali, T., Ohno-Machado, L. & Bloss, C. S. Codesigning a community-based participatory research project to assess tribal perspectives on privacy and health data sharing: A report from the Strong Heart Study. *J. Am. Med. Inform. Assoc.* **29,** 1120–1127 (2022).

86. Dyke, S. O., Dove, E. S. & Knoppers, B. M. Sharing health-related data: a privacy test? *NPJ Genom Med* **1,** 160241–160246 (2016).

87. Observational Health Data Sciences. The Book of OHDSI. (2021). at <https://ohdsi.github.io/TheBookOfOhdsi/>

88. Khare, R., Utidjian, L., Ruth, B. J., Kahn, M. G., Burrows, E., Marsolo, K., Patibandla, N., Razzaghi, H., Colvin, R., Ranade, D., Kitzmiller, M., Eckrich, D. & Bailey, L. C. A

longitudinal analysis of data quality in a large pediatric data research network. *J. Am. Med. Inform. Assoc.* **24,** 1072–1079 (2017).

89. Kahn, M. G., Callahan, T. J., Barnard, J., Bauck, A. E., Brown, J., Davidson, B. N., Estiri, H., Goerg, C., Holve, E., Johnson, S. G., Liaw, S.-T., Hamilton-Lopez, M., Meeker, D., Ong, T. C., Ryan, P., Shang, N., Weiskopf, N. G., Weng, C., Zozus, M. N. & Schilling, L. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS (Wash DC)* **4,** 1244 (2016).

90. Lee, E. T., Howard, B. V., Savage, P. J., Cowan, L. D., Fabsitz, R. R., Oopik, A. J., Yeh, J., Go, O., Robbins, D. C. & Welty, T. K. Diabetes and impaired glucose tolerance in three American Indian populations aged 45-74 years. The Strong Heart Study. *Diabetes Care* **18,** 599–610 (1995).

91. Sempos, C. T., Bild, D. E. & Manolio, T. A. Overview of the Jackson Heart Study: a study of cardiovascular diseases in African American men and women. *Am. J. Med. Sci.* **317,** 142–146 (1999).

92. Bild, D. E., Bluemke, D. A., Burke, G. L., Detrano, R., Diez Roux, A. V., Folsom, A. R., Greenland, P., Jacob, D. R., Jr, Kronmal, R., Liu, K., Nelson, J. C., O’Leary, D., Saad, M. F., Shea, S., Szklo, M. & Tracy, R. P. Multi-Ethnic Study of Atherosclerosis: objectives and design. *Am. J. Epidemiol.* **156,** 871–881 (2002).

93. Kim, J., Neumann, L., Paul, P., Day, M. E., Aratow, M., Bell, D. S., Doctor, J. N., Hinske, L. C., Jiang, X., Kim, K. K., Matheny, M. E., Meeker, D., Pletcher, M. J., Schilling, L. M., SooHoo, S., Xu, H., Zheng, K., Ohno-Machado, L. & R2D2 Consortium. Privacy-protecting, reliable response data discovery using COVID-19 patient observations. *J. Am. Med. Inform. Assoc.* **28,** 1765–1776 (2021).

94. Granville, K. Facebook and Cambridge Analytica: What You Need to Know as Fallout Widens. *The New York Times* (2018). at <https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html>

95. What to expect now that Internet providers can collect and sell your Web browser history. *The Washington Post* (2017). at <https://www.washingtonpost.com/news/the-switch/wp/2017/03/29/what-to-expect-now-that-internet-providers-can-collect-and-sell-your-web-browser-history/>

96. MacMillan, D. & McMillan, R. Google Exposed User Data, Feared Repercussions of Disclosing to Public. *WSJ Online* (2018). at <https://www.wsj.com/articles/google-exposed-user-data-feared-repercussions-of-disclosing-to-public-1539017194>

97. Seetharaman, D. Facebook Ignites Debate Over Third-Party Access to User Data. *WSJ Online* (2018). at <https://www.wsj.com/articles/facebook-ignites-debate-over-third-party-access-to-user-data-1521414746>

98. Nair, S., Hsu, D. & Celi, L. A. in *Secondary Analysis of Electronic Health Records* (ed. MIT Critical Data) (Springer, 2016).

99. Rothstein, M. A. Is deidentification sufficient to protect health privacy in research? *Am. J. Bioeth.* **10,** 3–11 (2010).

100. Vaidya, J., Shafiq, B., Jiang, X. & Ohno-Machado, L. Identifying inference attacks against healthcare data repositories. *AMIA Jt Summits Transl Sci Proc* **2013,** 262–266 (2013).

101. El Emam, K. Methods for the de-identification of electronic health records for genomic research. *Genome Med.* **3,** 25 (2011).

102. Ozair, F. F., Jamshed, N., Sharma, A. & Aggarwal, P. Ethical issues in electronic health

records: A general overview. *Perspect. Clin. Res.* **6,** 73–76 (2015).

103. Office for Civil Rights (OCR). Summary of the HIPAA Privacy Rule. *HHS.gov* (2008). at <https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html>

104. Office for Human Research Protections (OHRP). Federal Policy for the Protection of Human Subjects (’Common Rule'). *HHS.gov* (2009). at <https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html>

105. Office for Civil Rights (OCR). Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. *HHS.gov* (2012). at <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html>

106. Mello, M. M., Lieou, V. & Goodman, S. N. Clinical Trial Participants’ Views of the Risks and Benefits of Data Sharing. *N. Engl. J. Med.* **378,** 2202–2211 (2018).

107. Meslin, E. M., Alpert, S. A., Carroll, A. E., Odell, J. D., Tierney, W. M. & Schwartz, P. H. Giving patients granular control of personal health information: using an ethics ‘Points to Consider’ to inform informatics system designers. *Int. J. Med. Inform.* **82,** 1136–1143 (2013).

108. Rowbotham, M. C., Astin, J., Greene, K. & Cummings, S. R. Interactive informed consent: randomized comparison with paper consents. *PLoS One* **8,** e58603 (2013).

109. Health and Human Services Department. Modifications to the HIPAA Privacy, Security, Enforcement, and Breach Notification Rules Under the Health Information Technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act; Other Modifications to the HIPAA Rules. *Federal Register* **78,** 5565–5702 Preprint at

https://www.federalregister.gov/d/2013-01073 (2013)

110. General Data Protection Regulation (GDPR) – Official Legal Text. *General Data Protection Regulation (GDPR)* at <https://gdpr-info.eu/>

111. Powell, J., Fitton, R. & Fitton, C. Sharing electronic health records: the patient view. *Inform. Prim. Care* **14,** 55–57 (2006).

112. Warner, T. D., Weil, C. J., Andry, C., Degenholtz, H. B., Parker, L., Carithers, L. J., Feige, M., Wendler, D. & Pentz, R. D. Broad Consent for Research on Biospecimens: The Views of Actual Donors at Four U.S. Medical Centers. *J. Empir. Res. Hum. Res. Ethics* **13,** 115–124 (2018).

113. Garrison, N. A., Sathe, N. A., Antommaria, A. H. M., Holm, I. A., Sanderson, S. C., Smith, M. E., McPheeters, M. L. & Clayton, E. W. A systematic literature review of individuals' perspectives on broad consent and data sharing in the United States. *Genet. Med.* **18,** 663–671 (2016).

114. Kim, H., Bell, E., Kim, J., Sitapati, A., Ramsdell, J., Farcas, C., Friedman, D., Feupe, S. F. & Ohno-Machado, L. iCONCUR: informed consent for clinical data and bio-sample use for research. *J. Am. Med. Inform. Assoc.* **24,** 380–387 (2017).

115. White, J., Daniel, J. & Posnack, S. Privacy and Security Solutions for Interoperable Health Information Exchange Report on State Law Requirements for Patient Permission to Disclose Health Information. *RTI International report for AHRQ* (2009). at <https://www.healthit.gov/sites/default/files/disclosure-report-1.pdf>

116. Lee, S.-Y. D., Stucky, B. D., Lee, J. Y., Rozier, R. G. & Bender, D. E. Short Assessment of Health Literacy-Spanish and English: a comparable test of health literacy for Spanish and English speakers. *Health Serv. Res.* **45,** 1105–1120 (2010).

117. Cole, A. P., Friedlander, D. F. & Trinh, Q.-D. Secondary data sources for health services research in urologic oncology. *Urol. Oncol.* **36,** 165–173 (2018).