UNIVERSITY OF CALIFORNIA
SANTA CRUZ

Overcoming Data Privacy and Data Gravity Challenges in Bioinformatics Research

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biomolecular Engineering and Bioinformatics

by

James Casaletto

June 2023

The Dissertation of James Casaletto here is approved:

Professor Benedict Paten, Chair

Professor Josh Stuart

Dr. Melissa Cline

Dr. Sylvain Costes

Peter F. Biehl
Vice Provost and Dean of Graduate Studies

Table of contents

List of Figures iv

List of Tables vii

Abstract viii

Dedication ix

Acknowledgements x

1 Introduction 1

2 Federated analysis of BRCA1 and BRCA2 variation in a Japanese cohort 8

3 Modeling the impact of data sharing on variant classification 31

4 Federated analysis for privacy-preserving data sharing 51

5 Analyzing the relationship between gene expression and phenotype in space-flown mice using a causal inference machine learning ensemble 81

6 Conclusion 107

Bibliography 110

A List of publications 127

B Appendix for Modeling the impact of data sharing on variant classification 128

List of Figures

Figure 1. Graphical abstract of data flow 9

Figure 2. Histograms of cumulative log odds for classifying each of 1,000 simulated variants present at a 1e-05 frequency in the population. Classification thresholds are demarcated as vertical hash lines. Benign variants are shown in blue and pathogenic variants are shown in red. 42

Figure 3. Classification trajectories for 20 randomly selected variants at 1e-05 frequency in the population. Classification thresholds are demarcated as horizontal hash lines in the timeline plots. Benign variants are shown in blue and pathogenic variants are shown in red. 43

Figure 4. Classification probabilities over the course of 5 years. The y-axis of these plots is the probability of classifying the variant, converted from the aggregated likelihoods of pathogenicity generated in the simulations. Year 0 constitutes the time just before the sequencing centers share their data and all the variants are unclassified. Year 1 constitutes the moment just after the sequencing centers share their data. As time progresses and more evidence becomes available, some of the variants which were LB get “promoted” to B, and similarly some of the variants which were LP get “promoted” to P. 44

Figure 5. Sensitivity of variant classification to the frequency of observing ACMG/AMP evidence criteria. These “high” and “low” values are taken from the confidence intervals in Table 1. a.) Tornado plot for the sensitivity of Pathogenic and Likely Pathogenic variants. b.) Tornado plot for the sensitivity of Benign and Likely Benign variants. 45

Figure 6. Probabilities of classifying variants at 1e-06 frequency plotted over the course of 5 years. The y-axis of these plots is the probability of classifying the variant, converted from the aggregated likelihoods of pathogenicity generated in the simulations. Year 0 constitutes the time just before the sequencing centers share their data and all the variants are unclassified. Year 1 constitutes the moment just after the sequencing centers share their data. 46

Figure 7. **I**llustrating the logic of “bringing the code to the data”, either because the data is much larger relative to the analytical code f(D) or due to restrictions of exporting the data across organizational boundaries. 53

Figure 8. Legal entities and their roles in a GDPR data sharing agreement. 64

Figure 9. Trust architectures. Figure 4a depicts a centralized trust running external to the federation. Figure 4b depicts a decentralized trust in which any central coordination is distributed across the cluster. Figure 4c depicts a non-clustered environment in which trust is established pairwise between the data consumer and each data owner. 70

Figure 10. A classifier is presented with images of a cow in two different environments, one of which is its typical green pasture and the other of which is a relatively rare beige desert. 84

Figure 11. Box-and-whisker plots for ORO values based on mission. The dashed red line is the arithmetic mean between the two group medians, and the p-values show that the differences between the medians are all significant. 87

Figure 12. Scatter plots of variance versus mean for different transformations used in preprocessing. The vertical and horizontal axes are shown in log scale. 90

Figure 13. PCA plots of GLDS datasets colorized by the different covariates, including a.) library preparation, b.) dataset name, c.) experimental group, d.) mouse strain, and e.) Rodent Research study. 91

Figure 14. Test accuracies for each of the models in the CRISP ensemble, including a dashed red line representing 50% accuracy (null model performance). 94

Figure 15. The top 20 genes and their degree of concordance across the ensemble. The direction of the bar indicates whether higher (up) or lower (down) expression impacts lipid density. 95

Figure S1. Probabilities of classifying variants at 1e-05 frequency plotted over the course of 20 years at 10 small, 5 medium, and 3 large sequencing centers. On the left, all variant classifications and none of the clinical data are shared. On the right, all the clinical data are shared. 129

Figure S2. Histograms of cumulative log odds for classifying each of 1000 simulated variants present at a 1e-06 frequency in the population for one large sequencing center on the left and for all participating sequencing centers on the right over the course of 5 years. Classification thresholds are demarcated as vertical hash lines. Benign variants are in blue and pathogenic variants in red. 129

Figure S3. Classification trajectories over the course of 5 years for 10 randomly selected variants at 1e-06 frequency in the population for one large sequencing center on the left and for all participating sequencing centers on the right. Classification thresholds are demarcated as horizontal hash lines in the timeline plots. Benign variants are in blue and pathogenic variants in red. 130

Figure S4. Histograms of cumulative log odds for classifying each of 1000 simulated variants present at a 1e-06 frequency in the population for one large sequencing center on the left and for all participating sequencing centers on the right. Classification thresholds are demarcated as vertical hash lines. Benign variants are in blue and pathogenic variants in red. 130

Figure S5: Classification trajectories over the course of 20 years for 10 randomly selected variants at 1e-06 frequency in the population for one large sequencing center on the left and for all participating sequencing centers on the right. Classification thresholds are demarcated as horizontal hash lines in the timeline plots. Benign variants are in blue and pathogenic variants in red. 131

Figure S6. Probabilities of classifying variants at 1e-06 frequency plotted over the course of 20 years at 10 small, 7 medium, and 3 large sequencing centers. On the left, all variant classifications and none of the clinical data are shared. On the right, all the clinical data are shared. 131

Figure S7. Histograms of cumulative log odds for classifying each of 1000 simulated variants present at a 1e-05 frequency in the population for one large sequencing center on the left and for all participating sequencing centers on the right. Classification thresholds are demarcated as vertical hash lines. Benign variants are in blue and pathogenic variants in red. 132

Figure S8: Classification trajectories over the course of 5 years for 10 randomly selected variants at 1e-05 frequency in the population for one large sequencing center on the left and for all participating sequencing centers on the right. Classification thresholds are demarcated as horizontal hash lines in the timeline plots. Benign variants are in blue and pathogenic variants in red. 132

Figure S9. Probabilities of classifying variants at 1e-05 frequency plotted over the course of 5 years at 5 small, 3 medium, and 1 large sequencing center. On the left, all variant interpretations and none of the clinical data are shared. On the right, all the clinical data are shared. 133

List of Tables

Table 1. Summary of the variant data. The HGVS terms reflect the NM_007294.3 transcript for BRCA1 and NM_000059.3 for BRCA2. Variants are designated as B (Benign), B/LB (Benign or Likely Benign), LB (Likely Benign), Conflict (Conflicting Interpretations), VUS (Uncertain Significance) or Absent (Not Found). All variants scored against the BayesDel in silico predictor with a score of less than 0.3, within the BP4 scoring range. Additionally, two variants were observed to co-occur with two more more pathogenic variants in the same gene, indicating that at least one of these co-occurrences must be in trans, which meets the standards of BS2 evidence. In BRCA1, we observed co-occurrences of c.4729T>C with c.1518del and c.188T>A, and in BRCA2, we observed co-occurrences of c.964A>C with c.6952C>T, c.5645C>A and c.6244G>T. While these VUS had sufficient evidence for classification on allele frequencies only, these co-occurrences add further support to benign classification. We further observed co-occurrences of BRCA2 c.5660C>T with c.1261C>T and c.4365A>G with c.7480C>T, evidence which could support a benign classification if these variants are observed in co-occurrences with different pathogenic variants in other patient cohorts. 27

Table 2. Key resources table. 28

Table 3. Odds and frequency estimate confidence intervals per ACMG/AMP category per clinical data evidence category. The variable f represents the frequency of the variant itself. 37

Table 4. Odds of pathogenicity per ACMG/AMP classification 37

Table 5. Current number of tests in database and testing rate per center type 38

Table 6. The four experimental groups of mice from the GeneLab datasets. 86

Table 7. Description of and reference for each of the transformations used in the pre-processing of the transcriptomic data. 89

Table 8. One fictitious mouse sample’s gene expression data (truncated to 4 genes) after 5 data transformations, including environment string (“env”) and ORO threshold (“oro_thresh”). 92

Table 9. Mouse genes identified in the CRISP experiment as robustly predictive of the threshold lipid density phenotype. 97

Table 10. Enriched Gene Ontology pathways involving the genes from the CRISP experiment. 97

Table 11. Gene set enrichment analysis using gene ontology gene sets, showing the false discovery rate (FDR) adjusted significance q-value. 98

Table 12. Research articles implicating the changes in expression of the CRISP-identified genes in the spectrum of NAFLD disorders 100

Table 13. Gene result set using EdgeR to find differentially expressed genes between low and high ORO groups. 101

Abstract

Overcoming data privacy and data gravity challenges in bioinformatics research

By

James Casaletto

Next-generation sequencing technologies have generated a massive amount of DNA, RNA, and protein sequences since their inception. However, data privacy policies often restrict sharing such data for the risk of re-identifying individuals from whom the sequences were generated. Even when all the data from a sequencing experiment is available, it is often insufficient for statistical power or training machine learning models. Despite the lack of data, sometimes the data sets are ironically too large to realistically share with researchers. In this thesis, I explore methods to overcome challenges of data privacy and data gravity in bioinformatics research.

In collaboration with QIMR Berghofer and the Riken Center for Integrative Medical Sciences, we used federated methods to analyze genomic data from the BioBank Japan *in situ* to classify variants of uncertain significance while preserving privacy. With the Department of Laboratory Medicine and Pathology at the University of Washington, we developed a statistical model that demonstrates how using responsibly shared clinical evidence alone can classify variants of uncertain significance which occur at the rate of 1 in 100,000 people within just a few years. With researchers from McGill University, we reviewed the state of the art in federated computing technologies and how well they satisfy the privacy restrictions from the General Data Protection Regulation. With researchers from NASA, Amazon, and Intel, we developed a federated learning framework to run between terrestrial and space-borne compute infrastructure, laying the groundwork for subsequent experiments which preclude the need to transfer large datasets across astronomical distances. Finally, at NASA, we used a causal inference machine learning ensemble to infer robust correlation between mouse liver gene expression and a corresponding lipid density phenotype in space-flown mice.

To my loving family, friends, mentors, and colleagues who inspired and motivated this work

Acknowledgements

There are several people to whom I’m very grateful for their inspiration and guidance. I first learned about bioinformatics as a field of study working as a consultant for Celera Genomics. I was so intrigued by this field that I enrolled in the bioinformatics certificate program at the University of Santa Cruz Extension. There I met Professor Sami Khuri who first taught me bioinformatics algorithms, and subsequently introduced me to Professors Josh Stuart and Benedict Paten at the University of California Santa Cruz, and after taking their classes on computational genomics, I joined Professor Paten’s computational genomics lab as a graduate student researcher. There I met Dr. Melissa Cline with whom I had the good fortune of doing most of my research.

Dr. Cline is an amazing researcher and a gifted writer, is very generous with her time and expertise, and an absolute pleasure to work with. Dr. Cline and I developed the idea of using a containerized workflow to find variants of uncertain significance co-occurring *in trans* with known pathogenic variants. Together with then graduate student Charles Markello, Drs Amanda Spurdle and Michael Parsons of QIMR Berghofer, and the research scientists Drs. Yusuke Iwasaki and Yukihide Momozawa from the RIKEN Center for Integrative Medical Sciences, we classified variants of uncertain significance based on the co-occurrence principle using genomic data from BioBank Japan. After developing the container, we collaborated with Beth Sheets from the UCSC Genomics Institute to support Workflow Definition Language (WDL) and mine variant data in the Database of Genotypes and Phenotypes (dbGap) for more co-occurrences. We also used the co-occurrence container to analyze cardiomyopathy-related genes with Dr. Mahsima Shabani of Johns Hopkins University. Last, we worked with Dr. Ian Fore to run the co-occurrence container on the Seven Bridges platform using data from The Cancer Genome Atlas (TCGA).

Next, Dr. Cline and I collaborated with Professor Brian Shirts of the University of Washington medical center and Professor Sean Tavtigian of the University of Utah to develop a model that determines how long it would take to classify variants of uncertain significance using clinical data alone if a set of clinical laboratories would share their data.   My last collaboration with Dr. Cline was in writing a review paper on federated computing. We worked with genomic privacy policy legal researchers Alexander Bernier and Robyn McDougall of the Centre of Genomics and Policy at McGill University. Together we described the state of the art in federated computing (as it pertains to biomedical data) and how well those technologies work to satisfy the constraints of the General Data Protection Regulation.

My dear friend Professor Phil Heller (alumnus of the UCSC BME program) of San Jose State University introduced me to Ryan Scott of NASA Ames who was doing fascinating research on the effects of spaceflight on rodent health. Ryan introduced me to Graham Mackintosh and Drs Sylvain Costes and Lauren Sanders (also an alumnus of the UCSC BME program) who took me in as a graduate student NASA intern. For 18 months, we worked on several projects involving artificial intelligence in life sciences for space biology. In one project, we worked with Patrick Foley and Shashi Jain of Intel Corporation to develop a federated platform that supports building a machine learning model between Earth and the International Space Station. In another project, I worked with Nishan Pantha and Muthukumaran Ramasubramanian of the University of Alabama to build a generative adversarial network which creates synthetic liver transcriptomic data. And my last project during my NASA internship involved collaborating with Hamed Chok of the NASA GeneLab Multi-Omics Analysis Working Group and Professor Adrienne Hoarfrost of the University of Georgia to deploy an ensemble of causal inference machine learning algorithms to identify genes robustly correlated with a phenotype.

I’m very grateful to everyone in Professor Paten’s lab for sharing beers and pizza and giving me their advice on my research, including Drs Adam Novak, Trevor Pesout, Marina Haukness, Charles Markello, Jean Monlong, Ryan Lorig-Roach, Jouni Siren, Andrew Bailey, Colleen Bosworth, and Kishwar Shafin. Also from the lab, I wish to thank fellow graduate students Nick Keener, Robin Rounthwaite, Cecilia Cisar, Melissa Meredith, Mira Mastoras, Jimin Park, Mobin Asri, Prajna Hebbar, Shloka Negi, Matteo Ungaro, Uyen Nguyen, Parsa Eskandar, and Julian Lucas. I wish to extend my gratitude to Dr. Jordan Eizanga who spent hours with me doing math and crunching statistics.

Last, I am very grateful to my wife Holly and daughters Sophia and Aria. Besides their ever-loving support, they agreed to uproot their lives in Silicon Valley so that I could pursue my dream here in Santa Cruz.

Chapter 1: Introduction

Bermuda Principles

In February 1996, leaders of the Human Genome Project (HGP) met in Bermuda where they decided that all human genomic sequence information should be placed in the public domain within 24 hours of being generated. The so-called "Bermuda Principles" were drafted to encourage research and development and to maximize the HGP's benefits to society. These principles redefined an entire industry and established pre-publication data release as the norm in genomics and other research fields. Craig Venter, then CEO of Celera Genomics, declined to be a part of this agreement, arguing that the HGP was over-budget and inefficient. Conversely, members of the HGP questioned Venter's business ethics. For them, pharmaceutical development based on genomic research represented a public good, and that open data and commercial products are mutually beneficial. Indeed, recent financial analyses suggest that genomic sequences in the public domain brought about more commercialization and profitable drug development than did data from the restricted business model from Celera Genomics. [1] [2]

Professor David Haussler of the University of California, Santa Cruz, said of the first draft of the human genome: "There it was, going out into the whole world. Before the Human Genome Project, there had not been a serious discussion about data sharing in biomedical research. The standard was that a successful investigator held onto their own data as long as they could." [3]

AMP vs Myriad Genetics

Patenting genetic data on the *BRCA1* and *BRCA2* genes was an integral part of Myriad Genetics's business model until June 2013 when the United States Supreme Court unanimously decided *Assn. for Molecular Pathology v. Myriad Genetics Inc.*, ruling that isolated naturally occurring sequences of DNA cannot be patented. Opponents argued that these patents would hinder innovation by preventing others from conducting cancer research, would restrict the options available to cancer patients seeking genetic testing, and that the patents are not valid because such sequences are not invented but rather produced by nature. [4]

Since the Supreme Court ruling in 2013, major advances have been made in the field of breast cancer research. In 2013, the four major subtypes of breast cancer (HR+/HER2, HR-/HER2, HR+/HER2+, and HR-/HER2+) were identified. In 2017, the Food and Drug Administration (FDA) approved the first biosimilar drug - trastuzumab-dkst - for breast cancer treatment. In 2019, trastuzumab deruxtecan was approved by the FDA and was very effective in treating metastasized or irremovable HER2+ breast cancer. In 2020, the drug sacituzumab govitecan-hziy was approved by the FDA for treating metastatic triple-negative breast cancer for people who don’t respond to other treatments. And most recently, in October 2021, the FDA approved the drug abemaciclib for patients with HR+/HER2- early breast cancer which constitutes about 70% of all breast cancers.

Fortunately, Myriad has recently reversed its historical company culture by agreeing to submit variants detected by its hereditary cancer risk test, including variants in BRCA1 and BRCA2 genes, to ClinVar starting in the spring of 2023.

HIPAA and GDPR

The Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule protects personally identifiable information (PII) and protected health information (PHI) stored or transmitted by health organizations. The Safe Harbor provision of HIPAA guides researchers by prescribing how to redact data for public sharing. For example, the provision requires removing explicit identifiers (such as name, address, and other personally identifiable information), reporting dates in years, and reducing some or all digits of a postal (or ZIP) code. It is not clear that this is sufficient: perhaps a research participant may still be re-identified even when the data adhere to this standard. In 2014, HIPAA was amended to grant people access to their clinical test results, including genomic data. Putting research data into laypeople's hands leaves many scientists and policy makers uncomfortable, since research data could lack analytic validity, clinical validity, and/or clinical actionability. [5]

In the European Union, the GDPR regulates the use of identifiable personal data: data relating to a person which is identified or identifiable. For data to be considered identifiable, there must be a means by which a person may be re-identified. [6] This is not the case if re-identification is practically impossible on account of the fact that it requires a disproportionate effort in terms of time, cost and man-power, so that the risk of identification appears in reality to be insignificant. If the data controller and proximate third parties do not have a mechanism enabling the re-identification of the concerned individual that is "likely reasonably to be used" [7], then the data are considered to be anonymized and therefore not regulated by the GDPR.

Data Sharing Models

The quality of scientific research is judged, in large part, on its reproducibility which requires that the methods and data be made publicly available. The National Institute of Health (NIH) recently mandated that all the data generated from the research it be made publicly accessible. Through programs such as Big Data to Knowledge (BD2K), Global Alliance for Genomic Health (GA4GH), and The Cancer Genome Atlas (TCGA), the NIH is investing in making its data findable, accessible, interoperable, and reusable (FAIR). But in one sense, data sharing may open a Pandora's box. Defining responsible data sharing is difficult because it's difficult to future-proof. What constitutes "sufficiently private" today may one day become insufficient due to advances in technologies and more data availability. For example, Sweeney et al were able to re-identify 90% of the participants in the Personal Genome Project using publicly available demographic data. Gymrek et al succeeded in correlating variants on the Y chromosome with surnames. Wheeler et al were able to infer predisposition to Alzheimer's disease despite data masking. And Homer et al demonstrated that allele frequency bias could reveal the presence of a target in the case group. Making large amounts of data widely available for a long period of time and re-usable by third-parties involves substantial human and infrastructural resources. Who will support it? And who will guarantee the privacy of patients and research participants? What methods will engender trust? [8]

There are essentially four models that describe how data is shared: public, controlled-access, clique, and upon-request. [9] [10] Public data sharing occurs when data are made available without restrictions, providing the lowest barrier to entry for researchers to re-use the data. The BRCA Exchange Web portal publicly aggregates and shares BRCA variants, including variants expert-classified by members of the Evidence-based Network for the Interpretation of Germline Mutant Alleles (ENIGMA) variant curation expert panel (VCEP), and supports a community for collaborative variant interpretation and curation. Other examples of public data sharing include ArrayExpress (microarray gene expression data) and the Gene Expression Omnibus (GEO). NASA has published several datasets from randomized controlled experiments using model organisms at GeneLab.

Controlled-access sharing occurs when data may be used if certain criteria are met, such as a review of protocols, a commitment to use data only for health-related research, or other elements that affect how one obtains and uses the data. This imposes a modest barrier to entry for reuse efforts and is currently the preferred approach for de-identified genomics data that pose significant reidentification concerns. The UK Biobank and dbGaP are examples of controlled-access data sharing. This allows dataset developers to verify that adequate oversight is in place for research that could potentially lead to reidentification of a study participant.

Clique sharing and sharing upon request occur when researchers form a consortium or make individual agreements to share data, respectively. The researchers within the clique or who own the datasets can select which requesters to whom they grant access. Data made available upon request are not widely shared in practice: the data sharing decisions are left to individual scientists. The data from BIOBANK Japan we used in our first aim follows the clique-sharing model.

Federated Methods

Federated learning is a paradigm with a recent surge in popularity as it holds great promise on learning with fragmented or sensitive data. Instead of centralizing training and testing data into one location, it enables training a shared global model with a central server while keeping the data in local institutions where they originate. Federated learning is promising for healthcare data analytics: the sensitive patient data can stay either in local institutions or with individual consumers without privacy leakage. Each participating site is sent a copy of the model to train on their local data. Once the model has been trained locally over some number of iterations, the sites send only their updated version of the model parameters (aggregated, non-private information) to the central aggregator and keep their individual-level data. The aggregator uses some form of statistical aggregation (e.g. mean or median) on the contributions from all the sites and updates the global model. The updated parameters of the global model are shared again with the other sites, and the process repeats until the global model converges. [11] [12]

The Intel Federated Learning (OpenFL) library is a Python3 platform for federated learning. It enables disparate organizations to train models without sharing privacy-sensitive information or having to move large quantities of data over long distances. The library is composed of two components: the collaborator which learns model parameters based on local data and shares them with the aggregator; and the aggregator which takes the model updates from each of the collaborators and combines them to form the global model. The aggregator is agnostic of the learning framework, and the collaborators can use any deep learning framework such as Tensorflow and PyTorch which supports a periodic callback mechanism. We leverage OpenFL in Chapter 5.

Not all federated methods involve machine learning or multiple nodes. In fact, in Chapter 2, we deploy a single federated container to perform genomic analysis.

Statistical Modeling

All models are wrong, but some are useful. Despite the fact that abstraction implies misrepresentation, it permits defining model parameters and interpreting model results within these parameters. Abstractions are inevitably made for every biology experiment in each laboratory around the world. The use of *in vitro* cell lines and *in vivo* rodent experiments are, for example, by definition, abstractions of human bodies in very complex dynamic environments. These wet lab experiments are similar to their mathematical, dry lab counterparts in that they permit scientists to test hypotheses. The key for any model, be it mathematical, biological, or otherwise, to succeed is to combine data-driven modeling with model-driven experimentation, to the extent possible. The models inform the experimenters about which parameters seem relevant as explanatory of the response. Computer simulations of statistical models that are based on limited data are merely visualizing plausible trajectories forward in time. Predictions could then be made by analyzing multiple potential trajectories using multiple plausible parameter combinations. We trust mathematical models and accept their inherent prediction uncertainties for forecasting weather conditions, for example, because it's a common, well understood phenomenon. Similarly, when using statistical models in other domains like biology, it is mandatory to clearly communicate what models can and cannot do. For clinical purposes, predictive models may not need to accurately describe the complex landscape of genomic variant classification, but to help inform decision making, often upon binary decisions. [13]

Model Organisms

Humans have been traveling to space since the 1960s. We've accumulated years of extensive analysis of the effects of these missions on astronauts. Studies reveal that multiple systems are influenced by the conditions associated with spaceflight, including microgravity, exiting and returning to Earth's atmosphere, confinement in a closed environment, space radiation, changes in gas composition, and an altered diet, all of which may contribute to the observed health effects. Historically, rodent research plays an essential role in understanding of human disease and the ability to evaluate new therapeutics. Similarly, experimentation with rodents in space gives scientists the opportunity to comprehensively evaluate physiologic changes associated with spaceflight, as well as potential mitigation strategies, that exceeds what is possible with human studies. Rodent research has enabled scientists to study the effects of the space environment on health and disease in bone, circadian rhythm, cardiology, ophthalmology, metabolism, the microbiome, and behavior. [14]

After a rodent experiment, the investigator may have a hypothesis about a likely human response to the same conditions, after having considered differences in body weight, exposure, etc. The prediction that the hypothesis entails must then be tested on humans to be verified or falsified in the light of such human data. In essence, the use of animals to generate a hypothesis is not itself predictive. While physics deals with controlled environments upon which reductionism can be practiced, biology does not usually have that luxury. While biological systems are consistent with the laws of physics, biology is not physics. There are properties consequent upon internal organization, epigenetic and epistatic effects, and other factors rooted in evolutionary history which are not found in physics. This means that providing the same intervention to two different biological samples may result in wildly different responses, and humans are no exception. At the very least, using animal models such as rodents informs scientists about where to investigate further, or conversely, where to stop investigating. [15]

Chapter 2: Federated analysis of BRCA1 and BRCA2 variation in a Japanese cohort

Authors: James Casaletto[1], Michael Parsons[2], Charles Markello[1],Yusuke Iwasaki[3], Yukihide Momozawa[3], Amanda B. Spurdle[2], Melissa Cline[1]

Contributions:

In this research, I implemented all the co-occurrence logic in Python, containerized the logic in a Docker container, published the container to GitHub and Dockstore, wrote WDL scripts to launch the container, organized and led all the meetings and email communications with our collaborators and publisher, and wrote the manuscript. I also created a poster which I presented at the GA4GH 9th Plenary.

Affiliations:

1. UC Santa Cruz Genomics Institute, Mail Stop: Genomics, University of California, 1156 High Street, Santa Cruz, CA 95064, USA.
2. QIMR Berghofer Medical Research Institute, 300 Herston Rd, Herston QLD 4006, Australia.
3. Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa 230-0045, Japan

Summary

More than 40% of the germline variants in ClinVar today are variants of uncertain significance (VUS). These variants remain unclassified in part because the patient-level data needed for their interpretation is siloed. Federated analysis can overcome this problem by “bringing the code to the data”: analyzing the sensitive patient-level data computationally within its secure home institution and providing researchers with valuable insights from data that would not otherwise be accessible. We tested this principle with a federated analysis of breast cancer clinical data at RIKEN, derived from the BioBank Japan repository. We were able to analyze these data within RIKEN’s secure computational framework, without the need to transfer the data, gathering evidence for the interpretation of several variants. This exercise represents an approach to help

realize the core charter of the Global Alliance for Genomics and Health (GA4GH): to responsibly share genomic data for the benefit of human health.

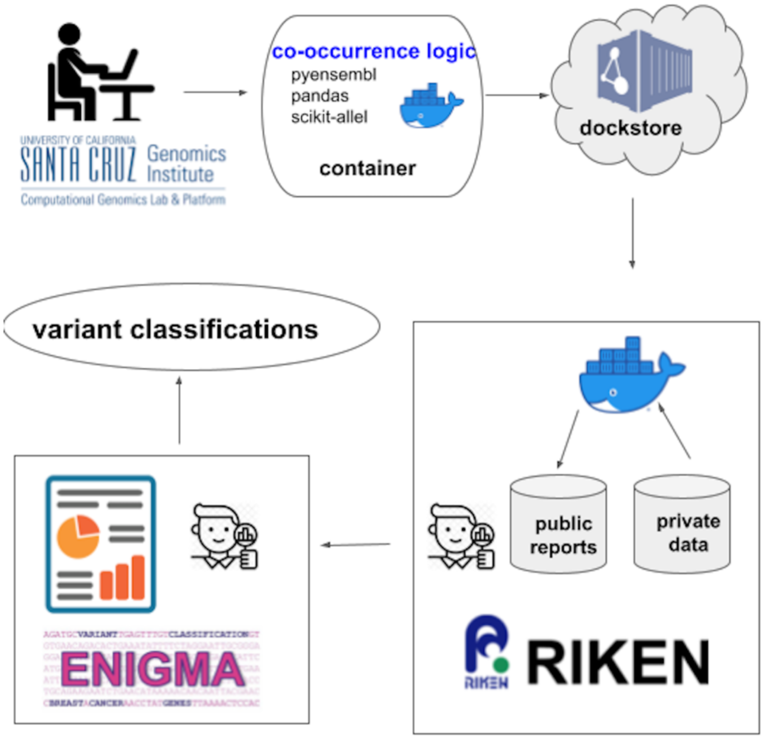

Figure 1. Graphical abstract of data flow

Introduction

One obvious and well-studied example of how genetic variation can impact human health is the risk of cancer presented by pathogenic variation in the *BRCA1* and *BRCA2* genes. Pathogenic *BRCA1/2* variants greatly increase the risk of female breast and ovarian cancer (as reviewed) [16], and also confer significant risk of pancreatic, prostate and male breast cancer (as reviewed) [16]. Genetic testing that identifies a pathogenic variant in these genes enables individuals and their families to better understand their heritable cancer risk, and to manage that risk through strategies such as increased screening, cascade testing of family members, and risk-reducing surgery and medication (as reviewed) [16]. However, these risk-reducing strategies are not available to an individual found to carry a variant of uncertain significance (VUS), a rare variant for which there is insufficient evidence to assess its clinical significance. While individually rare, these VUS are collectively abundant. As of May 2021, ClinVar [17], the world's leading resource on the clinical significance of genetic variants, reports that 8,592/25,028 (34.3%) of *BRCA1/2* variants therein are designated as VUS, while an additional 1,204 (4.8%) have conflicting interpretations. In other words, roughly 40% of *BRCA1/2* unique variants in ClinVar have no clear clinical interpretation. Meanwhile, there are many more variants that have been observed in individuals but are not yet in ClinVar: the Genome Aggregation Database (gnomAD) [18] includes an additional 35,635 *BRCA1/2* variants compiled from genomic sequencing research cohorts. Patients of non-European ancestry are significantly more likely to receive a VUS test report from *BRCA1/2* testing [19], a disparity that stems largely from historical biases in genetic studies [20] [21].

The VUS problem persists in large part because VUS are rare variants; no single institution can readily gather a sufficient set of observations for robust variant classification. Data sharing would seem to be the natural solution, but it faces logistical challenges. Variant interpretation often requires some amount of case-derived information: clinical observations of the variant in patients and their families together with their cancer history. However, case-level data is sensitive and private, and can rarely be shared directly due to regulatory, legal and ethical safeguards [22]. Yet sharing data on rare genetic variants is critical for the advancement of precision medicine, as advocated by organizations including the GA4GH [23], the American College of Molecular Geneticists [24] and the Wellcome Trust [25]. Fortunately, most variant interpretation does not require the case-level data *per se*, but rather variant-level summaries of information derived from those data. The ACMG/AMP Guidelines for variant interpretation [26], which specify forms of evidence for interpreting genetic variants, indicate use of variant-level summary evidence including population frequencies (*BA1, BS1, PM2)*, segregation of the variant and the disorder in patient families (*PP1, BS4),* case-control analysis (*PS4)*, and observations of the VUS *in cis* and *in trans* with known pathogenic variants (*PM3* and *BP2*, depending on the disorder). What is needed is an approach to derive this variant-level evidence from siloed case-level datasets without the need for direct access.

Federated analysis offers such an approach. Rather than an institution sharing its case-level data with external collaborators, those collaborators share an analysis workflow with the institution. The institution runs the workflow on their cohort, generating variant-level data that is less sensitive and can be shared more openly. This can yield valuable evidence for variant interpretation without the sensitive data leaving the home institution [27]. Container technologies support this approach by bundling the software and all its dependencies into a single module for straightforward installation and deployment on a collaborator's system [28]. These technologies include Docker [29], Singularity [30], and Jupyter [31]. Containers and workflows can be shared on the Dockstore platform [32] so that multiple institutions can execute the same software, promoting reproducibility.

We developed analysis workflows to mine tumor pathology, allele frequency, and variant co-occurrence data for *BRCA1* and *BRCA2* from breast cancer patient cohorts at RIKEN, derived from BioBank Japan [33] [34]. This analysis allowed the assessment of new variant interpretation knowledge from a cohort that would not otherwise be accessible. In addition to generating new knowledge on these genetic variants, this yielded new knowledge on the genetics of the Japanese population, which is underrepresented in most genetic knowledge bases. Moreover, we've generalized our container approach to work with any genotype-phenotype combinations of data.

Design

In principle, one could share access to a protected genomics dataset by transferring that data to a trusted third party, such as a secure cloud, but a dataset which contains personally-identifiable information generally cannot or should not be moved from its secure source location. Indeed, the

BioBank Japan data is prohibited from anonymous export. Federated analysis leaves the data securely in place and instead moves the analytic software (which tends to be many orders of magnitude smaller in size than a research cohort) to the data host institution. We designed our federated analysis software to be transparent, modular, and extensible. The analysis software creates multiple reports that capture data quality, associated phenotype, allele frequency, and variant co-occurrence.

Any researcher analyzing a dataset must first ensure that the data values are interpreted correctly; this is especially true when the researcher cannot interact with the data locally. The first report is the data quality report which addresses that need by providing basic statistics (such as minimum, maximum, mean, mode, and median) and reporting any missing or unexpected data values. For this report, we provide a JavaScript Object Notation (JSON) configuration file which defines each of the fields of interest, here as exemplified for the content of the tumor pathology file. The report could be used to check data quality for any delimited file, with or without a header. This data quality report represents a general solution which can be reused for other data sets. The Supplemental Information section includes two full examples with a data quality report.

The second report we generate is the genotype-phenotype report. This report is optional and can only be run when there exists both the variant VCF file as well as a phenotype TSV file. The purpose of this report is the associate a sample's genotype and phenotype directly in the same record. The Supplemental Information section includes two full examples with a genotype-phenotype report.

The third and last report is the variant frequency and co-occurrence report. It was written to summarize the variant counts stratified by patient group (affected vs. control) for estimating allele frequencies; and to report on variants of uncertain significance (VUS) which co-occur *in trans* either with known pathogenic variants in complex heterozygous genotypes, or with themselves as homozygous genotypes. The program takes as input a VCF file and outputs JSON files with the variant counts and the co-occurring variant information. If associated phenotype data is provided, then our software will intersect those phenotype data with the genotype data in the VUS reports. This requires using a tab-separated file with the string 'ID' as the primary key of this table whose values match those in the VCF file. The Supplemental Information section includes three full examples of variant frequency and co-occurrence reports.

To extend on the reporting functionality and generalizability, we provide the ability to integrate and call a custom, domain-specific report which can be leveraged to identify data anomalies in a known domain. This report is optional. In our research, we leveraged this feature to implement a tumor pathology report in which we calculate the number and proportion of triple-negative breast cancers of all breast cancers for which ER, PR, and HER2 test results are available. This pathology report reads a tab-delimited file which is indexed by the sample identifier. Even though these sample identifiers are anonymized, we did not want to risk exposing any identifier in the results. Our tumor pathology report takes as input that same tumor pathology file, and for each pathology feature outputs a summary of the number and proportion of patients stratified by pathogenic variant status, with an odds ratio, confidence interval, and Fisher's exact p-value for the comparison. Additionally, the report includes a comparison of mean age at diagnosis (and entry) for the different patient groups. This can be extended to measure the statistics for any stratification of gene and pathology data. Importantly, this optional custom report can be independently used to validate that the researcher and the collaborator are reading and interpreting the data equivalently. In federated computing, the researcher never has direct access to the data, so any anomalies in the data could be identified if the researcher and collaborating institution agree to independently generate the same report and then compare the results. Indeed, we used this pathology report to validate our federated approach and to verify that there were no data anomalies that would preclude our analysis.

While our research focuses on VUS in *BRCA1* and *BRCA2* genes and associated tumor pathologies, the software was written to work with any genotype-phenotype combinations of data. In the Supplemental Information section, we provide an illustration of how one might assess genetic variation in cardiomyopathy by evaluating VUS in the *MYH7* gene along with associated cardiac phenotype data. All the configuration is passed as command-line options to the program to define such parameters as gene name, whether the data are phased, and which human genome version to use as genomic coordinates. Moreover, all the Python libraries required to run this code are included in the Docker container.

Methods

The dataset

Our analysis revolved around case-control association study data of individuals of Japanese ancestry [33] [34]. These data reside at RIKEN and cannot be accessed outside of that institution. The dataset reports the variants in coding regions of 11 genes associated with hereditary breast, ovarian, and pancreatic cancer syndrome, including *BRCA1* and *BRCA2.* Additionally, the dataset reports the tumor pathology of the breast cancer patients, including ER, PR and HER2 status. The controls within this cohort are individuals who were at least 60 years old when sequenced and who have neither personal nor family history of cancer. The variant data were stored in a Variant Call Format (VCF) file and the associated phenotype (pathology) data were stored in a tab-delimited file. No other files were required for this analysis.

Variant interpretation evidence

We developed Docker containers to collect data for two forms of evidence (ACMG code/s designated in parenthesis): allele frequencies (*BA1, BS1)* and variant co-occurrences *(BS2).* In addition, we estimated *in silico* predictions of variant pathogenicity (*BP4, PP3)* using the BayesDel method for annotation of predicted missense substitutions and insertion-deletion changes [35].

Allele frequencies

By the AMCG/AMP standards, the frequency of a variant in a large, outbred population can offer three different forms of evidence for variant interpretation. First, when the variant is observed at a far greater frequency than expected for the disorder in question, this is such a strong indicator of benign impact (*BA1)* that the variant can be considered benign without any further evidence. Second, when the variant's frequency does not meet the *BA1* threshold but is still greater than expected for the disorder, the frequency represents strong evidence (*BS1)* that can contribute to a benign interpretation. Third, when the variant is absent from controls or reference population datasets, its absence represents moderate evidence (*PM2)* that can contribute to a pathogenic interpretation [26]. While gnomAD is commonly used as source of population frequencies, gnomAD 3.1 contains data from only 2,604 East Asian genomes [18] while gnomAD 2.1 contains data from 9,977 exomes. Similarly, gnomAD 2.1 contained 76 Japanese exomes, while the number of Japanese genomes in gnomAD 3.1 is unknown. Therefore, a Japanese biobank with tens of thousands of samples might plausibly contain additional evidence not available through gnomAD. When considering population frequencies, one must consider the source of the samples and whether individuals affected by the disorder are likely to be present in the dataset [36]. Accordingly, we evaluated the non-cancer subset of gnomAD and the control samples from BioBank Japan. Each ClinGen Variant Curation Expert Panel (VCEP) determines the precise rules for applying the ACMG/AMP standard to the genes and diseases under their purview, including the population frequency thresholds for *BA1* and *BS1* evidence. By the proposed rules of *BRCA* ClinGen Variant Curation Expert Panel (VCEP), the threshold for *BA1* evidence is an allele frequency of greater than 0.001 while the *BS1* frequency threshold is 0.0001.

in silico prediction

By ACMG/AMP standards, if multiple lines of computational evidence predict that a variant will impact either protein function or RNA splicing, that observation can contribute to a pathogenic interpretation (*PP3)*. Conversely, if multiple lines of computation evidence predict that the variant will have no functional impact, that observation can contribute to a benign interpretation (*BP2)*. We estimated the probability that the variant would impact protein function with BayesDel [35], a meta-predictor that has been shown to outperform most others [37]. By the proposed rules of the

BRCA ClinGen VCEP, a BayesDel score of less than 0.3 predicts a benign interpretation while a BayesDel score of greater than 0.3 predicts a pathogenic interpretation.

in trans co-occurrence

In fully penetrant diseases with dominant patterns of inheritance, if one observes a VUS *in trans* (on the opposite copy of the gene) with a known pathogenic variant in the same gene in an individual without the disease phenotype, that observation represents evidence of a benign impact. For *BRCA2* (and more recently *BRCA1*), co-occurrences of two pathogenic variants in the same gene are associated with Fanconi Anemia, a rare debilitating disorder characterized by deficient homologous DNA repair activity, bone marrow failure, early cancer onset and a life expectancy that rarely extends past 40 [38]. Consequently, when an older individual is observed with a *BRCA1* or *BRCA2* VUS as either a homozygous genotype or a compound heterozygous genotype (*in trans* with a pathogenic variant in the same gene), that observation suggests a benign interpretation for the VUS. One caveat is that most clinical sequencing does not report phase; any single co-occurrence of two variants might be *in trans* or *in cis*. However, if a VUS co-occurs with two different pathogenic variants in two different patients, one can assume that at least one of those co-occurrences is *in trans* [39]. Based on these clinical observations, VUS homozygosity or compound heterozygosity with a known pathogenic variant in an individual known or inferred to be without Fanconi Anemia features provides strong evidence against pathogenicity (*BS2*) [37] [38].

Collaboration details

In advance of developing the containers, the authors communicated to determine which data were available and in which format the data were stored. In our research, the variant data were stored in a single VCF (Variant Call Format) file with anonymized sample identifiers, and the pathology data were stored in a single TSV (tab-separated values) file indexed by the same sample identifiers. The data were already prepared in these files in the research that generated the data in the first place, so no additional data preparation steps were required. RIKEN provided a pair of files (one VCF file and one tumor pathology TSV file) with bogus data to preserve privacy but simultaneously allow the UCSC researchers to develop their containers. As previously mentioned, the UCSC team initially developed the container to generate a tumor pathology report. When the UCSC team finished preparing the container for that report, they notified the team at RIKEN to download the container code and run it against the data set. . The instructions for running the container are straightforward and are well-documented in the software repository. After a few iterations and email communications, the reports generated by each team were found to match exactly, thereby validating that accurate analysis could be performed on this data using a federated approach. Subsequently, the UCSC team developed the container to create the co-occurrence and allele frequency report along with the intersection and data quality report. Once those reports were generated, they were sent to the QIMR team to analyze for variant interpretation. In all, the total amount of interaction required to collaborate was minimal, in part because the QIMR team had previously collaborated with the RIKEN team using this same data.

Analysis approach

We created our Docker containers with Python 3.73 code which (a) collects observational statistics on tumor pathology, (b) gathers variant counts for estimating allele frequencies and (c) identifies VUS which either co-occur with a known pathogenic variant in the same gene, or which co-occur with themselves (i.e. homozygous VUS). When reporting co-occurrences, we also reported the age of the patient, to review data against expectations of age at presentation of Fanconi Anemia. To identify VUS, we checked the classifications provided by ClinVar and validated against the ClinGen-approved ENIGMA expert panel in BRCA Exchange [25]. If the clinical significance was ‘Unknown’, or if the variant did not appear in BRCA Exchange, then we labeled the variant a VUS. We applied this container to the BioBank Japan samples. We identified *BRCA1* or *BRCA2* variants which appeared as homozygotes and/or co-occurred with a known pathogenic variant in the same gene. Sequencing data was not phased, but details on the co-occurring variant/s were provided to aid inference of whether a VUS was *in cis* or *in trans*.

Results

We describe here an example of how federated analysis can add information of value for variant interpretation. We analyzed a case-control study of Japanese individuals whose case-level data resides at RIKEN [18,26]. Since these data are not accessible to external researchers, the UC Santa Cruz team developed analysis software, in the form of a Docker container, and shared it with the RIKEN team. The RIKEN team applied the container to analyze this cohort *in situ*, within their secure institutional environment, generating variant-level summary data that contained no personal information and can be shared more openly. The QIMR Berghofer team then applied these data to variant interpretation.

As an initial quality control exercise, we replicated a table from a previous publication on these data [18], using the tumor pathology data. This table contrasts the patients with or without pathogenic variants in terms of factors including family history of seven types of cancer; estrogen, progesterone and herceptin receptor status; and age at diagnosis. We were able to replicate this table precisely, indicating that we were able to process the data accurately. This exercise also demonstrated that our container can be used to generate scientifically meaningful results. While this step was not mandatory for our analysis, we recommend it for the reasons just stated.

Subsequently, we applied the Docker container to analyze the complete patient cohort. We observed 19 *BRCA* variants that have not yet been interpreted by the ClinGen *BRCA1/2* expert panel. For each VUS, we reported its allele frequency in the controls, and any observations of the VUS co-occurring with a known pathogenic variant in the same gene (Table 1). We also annotated variants for single-submitter curations in ClinVar.

Eleven VUS met the standard for stand-alone evidence of benign impact (*BA1*) on the basis of the allele frequencies in the BioBank Japan controls; all of these VUS were predicted bioinformatically to have benign impact (*BP4*). All eleven VUS will meet the standard of benign interpretation on the basis of their frequency evidence from the Japanese cohort. Additionally, two of these variants (*BRCA1* c.4729T>C; *BRCA2* c.964A>C) were observed to co-occur with at least two different pathogenic variants in the same gene, evidence sufficient to apply the *BS2* criterion. Of these eleven VUS, four have single-submitter classifications in ClinVar as Benign or Likely Benign, five have conflicting interpretations, and two are designated by ClinVar as VUS. Based on observations currently in gnomAD [3], seven of these variants would have met the *BA1* criterion, three would have met the *BS1* criterion, and one was absent (meeting the *PM2* criterion). For each of the variants present in gnomAD, East Asian was the continental population with the greatest allele frequency at the 95% confidence level (popmax) [30], a fact that itself adds confidence to the BioBank Japan observations. While seven of the variants could have been interpreted as benign using data in gnomAD, the federated analysis supported the interpretation of four additional variants. This greater sensitivity in the BioBank Japan results reflects the greater cohort size: while gnomAD contains 2,604 East Asian genomes and 9,977 East Asian exomes, the BioBank Japan control group contains 23,731 Japanese individuals.

Five VUS showed strong evidence of benign impact (*BS1*) based on their BioBank Japan allele frequencies, and evidence predictive of benign impact according to BayesDel (*BP4*). These five VUS meet the standard of likely benign interpretation based on their frequency and bioinformatic evidence combined. Additionally, two of these VUS had a single co-occurrence with a pathogenic variant in control individuals; while one should not put too much weight on any single homozygous observation, together with the *BS1* and *BP4* evidence, the data present a consistent picture of benign interpretation supported by multiple lines of evidence. One of these five variants is classified in ClinVar as likely benign, while the other four are classified as VUS. Four of these VUS would reach the *BS1* evidence standard based on their gnomAD population frequencies while a fifth is absent from gnomAD.   The BioBank Japan analysis supports reclassifying five variants, only four of which could be reclassified using data in gnomAD.

Finally, three additional variants were each observed in a single heterozygous co-occurrence and have BayesDel scores predictive of benign impact (*BP4*). With one co-occurrence observation apiece, we cannot predict whether the co-occurrence is *in trans* or *in cis*, so these observations are not themselves sufficient for evidence of benign impact. However, these co-occurrences could contribute to benign evidence when and if the same VUS are observed to co-occur with other pathogenic variant(s) in another cohort. These VUS are rare variants absent from gnomAD and have either conflicting or VUS interpretations in ClinVar.

Discussion

With this demonstration of federated analysis, we analyzed a protected cohort that we would not have been able to access directly, and we gathered knowledge on Japanese genetics to further the interpretation of *BRCA1/2* variants. Of 19 variants currently tagged as VUS by the ClinGen BRCA expert panel, 12 were VUS or conflicting in ClinVar. The suggested interpretations based on bioinformatic and frequency analysis assign a Benign or Likely Benign classification for 16 variants, and highlight the value of extending data capture to a subpopulation not yet well represented in gnomAD. We also demonstrated the federated collection of variant co-occurrences and age at presentation; these data together provided further evidence supporting the Benign and Likely Benign variant interpretations. This analysis would not be feasible with the existing population frequency resources. For example, gnomAD, the resource selected by ClinGen as its standard, does not yet have a large Japanese cohort, and now shares variant co-occurrences but without the patient age information that is needed for ruling out Fanconi Anemia under ENIGMA's variant interpretation rules. These samples had been analyzed previously by the RIKEN and ENIGMA teams [18,26], a fact that explains why an analysis of nearly 30,000 samples revealed only 19 VUS. This federated analysis allowed us to revisit these data with updated classification criteria, as well as collecting new evidence on variant co-occurrences. Further, by developing a tumour pathology report, we provide proof of principle that federated analysis can be designed to capture other clinical features relevant for variant interpretation. These additional data types are generally provided only in summary level data presentations from published cohorts, at best. Additionally, this method can be applied to any other phenotype-genotype relationship that could benefit from otherwise siloed datasets.

We have also demonstrated that there are international sequencing projects that contain valuable information that could be applied today to variant interpretation but are not yet represented in major population data repositories. This is illustrated by the number of Japanese samples analyzed in this study (7,104 cases plus 23,731 controls) versus the size of gnomAD's East Asian cohort (2,604 genomes plus 9,977 exomes). In principle, the gnomAD and the related population genomics resources will grow with time to comprehensively represent all global populations. In practice, due to the high cost of processing external sequence data, gnomAD mostly imports data from cohorts that were sequenced at the Broad, where sequencing data is processed to a common standard (H. Rehm, personal communication, October 4, 2021). For these reasons, capturing global genetic diversity can benefit from gathering evidence from international sources. Because traditional data sharing is blocked by barriers including laws that prohibit exporting genomic sequences, federated analysis can advance data sharing by limiting the scope of data to be shared to the information most needed.

In this instance, the data sharing was simplified by the fact that the RIKEN team had already assembled a case-control dataset on breast cancer, and in doing so, had already reduced the complex phenotypic data to a set of simplified terms. In a typical variant interpretation scenario, the situation is more involved. In genetic testing, the phenotypic data is often absent, or provided in unstructured text fields that must be curated manually prior to any analysis - traditional or federated. Where phenotypic data is available in a structured, electronic form, federated analysis can be viable. The cancer diagnosis (or lack thereof) can be represented through Human Phenotype Ontology (HPO) terms [27], with Disease Ontology [28] terms representing the tumor pathology. For example, if the phenotype file had represented the disease phenotype with HPO terms rather than the simplified representation, one might distinguish between cases and controls in the genotype-phenotype report by recognizing breast cancer cases with the HPO term HP:0003002 (Breast Carcinoma), or potentially the less specific HPO term HP:0100013 (Neoplasm of the Breast). Similarly, if the phenotypic data were associated with cardiomyopathy, one could use the HPO term HP:0001639 to represent hypertrophic cardiomyopathy as a phenotype, or the more general HPO term HPO:0001639 to represent cardiomyopathy. Structured models for phenotypic and genomic data exchange, such as Phenopackets [29], increase the opportunity for federated approaches by improving the data interoperability. With the growth in standards developed by the GA4GH and other organizations, and increasing adoption of electronic data standards worldwide [29], this federated analysis model can be generalized and extended into more areas within genomics. Emerging GA4GH technologies including Beacon V2, Matchmaker Exchange and Data Connect can suggest the presence of samples of interest in remote, siloed cohorts, such as cases with rare monogenic disorders. This federated analysis approach complements such approaches by allowing further analysis of these samples while safeguarding patient privacy.

While gnomAD is a comprehensive source of allele frequency data in genomic research [26], our federated solution does not, *per se*, require using it. Any database deemed more appropriate for a particular use case or cohort may be used as the source of allele frequencies if the data are formatted in a VCF sites file. Similarly, we used ClinVar as our source of ground truth for variant classification, and the ClinVar database may be substituted with another classification database if the data is formatted properly. These data formats are discussed in the Supplementary Information section.

Limitations of the study

Federated computing is being widely adopted, but it does present its own challenges in data privacy and system security. Docker containers are, to an extent, “black boxes”. In order to ascertain whether the analysis is truly both secure and privacy-preserving, an auditor would need to carefully inspect the Dockerfile definition of the container as well as all the software that runs in the container. We mitigated this risk by writing our reports to local text files which could be examined by the RIKEN team before being shared externally. Additionally, we published the software as open source so it may be directly inspected by collaborators. A second, related problem is that one cannot readily determine whether software might damage or compromise the security of the system on which it runs. One promising solution to this problem is certification. Within the emerging field of applications security testing, there are software platforms that can dynamically assess the system accesses of the software under test. While the current platforms are commercial, there will likely be an open-source version in time. Eventually, this may become an element of the GA4GH Cloud Testbed, currently under development. This testbed infrastructure will initially serve as a platform for testing compliance with GA4GH standards and will extend to encompass performance benchmarking. In the future, this platform could potentially report activity that suggests a security risk, such as the details of outgoing network or disk traffic; and publishing these certification results could fit well within the framework of container libraries such as Dockstore. As an immediate solution to this problem, collaborating institutions should run such otherwise unsecured containers in a virtual machine sandbox environment which is completely isolated from their internal network.

Another limitation of our approach is that it requires getting data into the format that our software recognizes, namely tab-separated files and VCF files. In other words, the software is not agnostic of the file format. Moving forward, we will be able to generalize this approach by leveraging the data standards under development by the GA4GH, which will allow methods to compute over generalized data representation models rather than restricting their input to specific file formats. In particular, the standards of the GA4GH Cloud Workstream are already making it easier to leverage software methods across many different computing platforms. Further development will facilitate the streamlined execution of containerized workflows, the representation of phenotypic data, and the sharing of genetic knowledge.

Acknowledgements

We gratefully thank Gunnar Rätsch for instigating this project, and the members of the BRCA Challenge Evidence Gathering Group for discussion on the analytical design. JC is supported by NHGRI grant U54HG007990 and NHLBI grant U01HL137183.

ABS and MTP are supported by funding from the Australian National Health and Medical

Research Council (APP177524). YM is supported by AMED under grant number JP19kk0305010 (to YM). MSC is supported by NCI grant U01CA242954 and BioData Catalyst fellowship OT3 HL147154 from the NHLBI through UNC-CH 5118777.

Author contributions

YI and YM performed the research which generated the variant and pathology data. MSC, ABS and YM planned the analysis of the data. The docker container was developed by JC with input from YM and technical guidance from CM. YI and YM executed the container. MSC, ABS, JC, and MTP analyzed the results, and prepared the manuscript. All authors reviewed the final manuscript.

Declaration of interests

No conflict for MSC, JC, CM, ABS, MTP, YI, and YM.

Gene	BRCA2	BRCA2	BRCA2	BRCA1	BRCA2	BRCA2	BRCA2
Variant (cDNA HGVS)	c.6325G>A	c.7052C>G	c.943T>A	c.4729T>C	c.4365A>G	c.6131G>T	c.964A>C
Variant (Protein HGVS)	p.A2351G	p.A2351G	p.C315S	p.S1577P	p.A2351G	p.G2044V	p.K322Q
ClinVar Classification (May 1, 2021)	B/LB	B/LB	B/LB	B/LB	LB	Conflict	Conflict
gnomAD 2.1.1 Exome Frequency (EAS)	2.55E-03	1.87E-03	5.30E-03	2.65E-04	Absent	4.52E-04	4.31E-04
gnomAD 3.1.1 Genome Frequency (EAS)	2.39E-03	2.02E-03	5.03E-03	2.02E-04	2.01E-03	4.52E-03	2.41E-03
ACMG/AMP Code from gnomAD	BA1	BA1	BA1	BS1	BS1	BA1	BA1
Biobank Japan Frequency (Controls)	1.46E-02	3.16E-03	1.56E-03	1.14E-02	4.64E-04	3.29E-02	2.31E-03
ACMG/AMP Freq from BioBank Japan	BA1	BA1	BA1	BA1	BS1	BA1	BA1
BayesDel Score	-0.61	-0.24	-0.41	0.03	-0.52	-0.16	-0.08
Bioinformatic Code	BP4	BP4	BP4	BP4	BP4	BP4	BP4
ACMG/AMP Class based on Frequency and Bioinformatics	B	B	B	B	LB	B	B
Gene	BRCA1	BRCA1	BRCA2	BRCA2	BRCA2	BRCA2	
Variant (cDNA HGVS)	c.154C>T	c.811G>A	c.5969A>C	c.3395A>G	c.9733T>G	c.5660C>T	
Variant (Protein HGVS)	p.L52F	p.V271M	p.D1990A	p.K1132R	p.S3245A	p.T1887M	

ClinVar Classification (May 1, 2021)	Conflict	Conflict	Conflict	VUS	VUS	VUS	
gnomAD 2.1.1 Exome Frequency (EAS)	1.36E-03	1.32E-03	0	Absent	Absent	1.13E-04	
gnomAD 3.1.1 Genome Frequency (EAS)	4.03E-04	1.21E-03	4.03E-04	0.000201	Absent	Absent	
ACMG/AMP Code from gnomAD	BA1	BA1	BS1	BS1	PM2	BS1	
Biobank Japan Frequency (Controls)	6.78E-03	6.28E-03	2.61E-03	3.75E-03	1.01E-03	1.69E-04	
ACMG/AMP Freq from BioBank Japan	BA1	BA1	BA1	BA1	BA1	BS1	
BayesDel Score	0.14	0.06	-0.08	-0.2	-0.47	-0.29	
Bioinformatic Code	BP4	BP4	BP4	BP4	BP4	BP4	
ACMG/AMP Class based on Frequency and Bioinformatics	B	B	B	B	B	LB	
Gene	BRCA2	BRCA2	BRCA2	BRCA2	BRCA2	BRCA2	
Variant (cDNA HGVS)	c.2672T>A	c.587G>T	c.8040C>G	c.358G>A	c.3983G>A	c.6637T>C	
Variant (Protein HGVS)	p.V891D	p.S196I	p.D2680E	p.V120M	p.S1328N	p.S2213P	
ClinVar Classification (May 1, 2021)	VUS	VUS	VUS	Absent	Conflict	Conflict	
gnomAD 2.1.1 Exome Frequency (EAS)	Absent	1.78E-04	Absent	Absent	0	Absent	

gnomAD 3.1.1 Genome Frequency (EAS)	Absent	Absent	0.000202	Absent	Absent	Absent	
ACMG/AMP Code from gnomAD	PM2	BS1	BS1	PM2	PM2	PM2	
Biobank Japan Frequency (Controls)	9.69E-04	4.64E-04	9.69E-04	0	0	0	
ACMG/AMP Freq from BioBank Japan	BS1	BS1	BS1	PM2	PM2	PM2	
BayesDel Score	-0.05	-0.22	-0.05	-0.48	-0.57	-0.06	
Bioinformatic Code	BP4	BP4	BP4	BP4	BP4	BP4	
ACMG/AMP Class based on Frequency and Bioinformatics	LB	LB	LB	VUS	VUS	VUS	

Table 1. Summary of the variant data. The HGVS terms reflect the NM_007294.3 transcript for *BRCA1* and NM_000059.3 for *BRCA2*. Variants are designated as B (Benign), B/LB (Benign or Likely Benign), LB (Likely Benign), Conflict (Conflicting Interpretations), VUS (Uncertain Significance) or Absent (Not Found). All variants scored against the BayesDel *in silico* predictor with a score of less than 0.3, within the BP4 scoring range. Additionally, two variants were observed to co-occur with two more more pathogenic variants in the same gene, indicating that at least one of these co-occurrences must be *in trans*, which meets the standards of BS2 evidence. In *BRCA1*, we observed co-occurrences of c.4729T>C with c.1518del and c.188T>A, and in *BRCA2,* we observed co-occurrences of c.964A>C with c.6952C>T, c.5645C>A and c.6244G>T. While these VUS had sufficient evidence for classification on allele frequencies only, these co-occurrences add further support to benign classification. We further observed co-occurrences of *BRCA2* c.5660C>T with c.1261C>T and c.4365A>G with c.7480C>T, evidence which could support a benign classification if these variants are observed in co-occurrences with different pathogenic variants in other patient cohorts.

STAR methods

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Sequence and phenotype data	Japanese Genotype-Phenotype Archive	JGAS00000000140

Software and algorithms		
Co-occurrence GitHub repository	This manuscript	https://github.com/BRCAChallenge/federated-analysis
Co-occurrence Dockstore repository	This manuscript	https://dockstore.org/my-workflows/github.com/BRCAChallenge/federated-analysis/cooccurrence
Python 3.7.3	Python Software Foundation	https://www.python.org
Scikit-allel 1.3.1	Miles et al[31]	https://scikit-allel.readthedocs.io/en/stable/
Pandas 1.3.2	Pandas development team[32]	https://pandas.pydata.org/
Bcftools 1.10.2	Danecek et al[33]	https://github.com/samtools/bcftools
Pyensembl 1.8.5	N/A	https://github.com/openvax/pyensembl

Table 2. Key resources table.

Resource availability

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact for this study, James Casaletto (jcasalet@ucsc.edu).

Materials availability

There are no materials that were generated in this study.

Data and code availability

- This paper analyzes existing data from BioBank Japan. The accession number for the dataset is listed in the key resources table.
- All original code has been deposited at GitHub and Dockstore and is publicly available as of the date of publication. URLs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Method details

To run our container, Docker must be installed in the runtime environment at the institution where the data are stored. We tested our container on Docker versions 18.03 and 19.03. The container also requires the appropriate ClinVar VCF file (for GRCh37 or GRCh38) which can be downloaded from their HTTP or FTP site (https://ftp.ncbi.nlm.nih.gov/pub/clinvar/). We used the `bcftools` command to reduce the size of this file to include only the genes of interest. Last, the container requires the gnomAD sites VCF file which can be downloaded from their HTTP site (https://gnomad.broadinstitute.org/downloads). Again, we used the `bcftools` command to reduce the size of this file to include only the genes of interest.

We created a variant co-occurrence and allele frequency report for BRCA1 and BRCA2, but our software has been generalized to find co-occurrences on other genes. Users can specify which version of the human genome (37 or 38), the chromosome and the gene on which to find VUS co-occurring in trans with themselves or with known pathogenic variants. The software runs on both phased and un-phased data, though inferring the genotype phase from un-phased data requires VCEP expertise.

To determine variant classification, users must provide a delimited file with the following fields: `Clinical_significance` and `Genomic_Coordinate_hg37` (or `Genomic_Coordinate_hg38`). Genomic coordinates must have the form of this example variant: “chr13:g.32314514:C>T”, where this represents the variant on chromosome 13, position 32314514 which changes a C nucleotide to a T nucleotide. If the `Clinical_significance` field is defined as “Pathogenic”, “Likely pathogenic”, “Likely_pathogenic”, or “Pathogenic/Likely_pathogenic”, then we interpret that variant as being pathogenic. Similarly, if the `Clinical_significance` field is defined as “Benign”, “Likely benign”, “Likely_benign”, or “Benign/Likely_benign”, then we interpret that variant as being benign. We interpret any other value in the `Clinical_significance` field as being of uncertain significance.

To successfully mine co-occurrence data, our code performs the following steps.

1. Read VCF files

   The genomic variants are defined in a VCF file which our application reads using the `read_vcf()` method of the Python scikit-allel package. We store the variants in a Python dictionary which contains the chromosome, position, reference allele, and alternate allele along with the genotype. The variant classifications are defined in a VCF file which our application reads using the `read_csv()` method of the Python pandas package. We store these classifications in a Python dictionary which contains 3 sets: one for pathogenic variants, one for benign variants, and one for VUS. Last, the allele frequencies are defined in a VCF sites file which our application reads using the `read_csv()` method of the Python pandas package. We store these allele frequencies in the same Python dictionary as the genomic variants.

2. Find variants per sample

   Our application uses multi-threading in Python to parallelize the construction of 3 lists of variants per cohort sample: benign variants, pathogenic variants, and VUS. The classification of variants is determined using the ClinVar VCF file.

3. Intersect variants with phenotype data

   The phenotype data are defined in a tab-delimited file and are read using the `read_csv()` method from the Python pandas package. The `ID` field of the phenotype file is used to match keys in the dictionary of variants per sample. In this way, any phenotypic data can then be associated directly with those variants and not with the samples themselves.

4. Find and annotate co-occurring VUS

   Our application examines the dictionary of variants per sample and the associated genotypes to determine if those VUS co-occur in trans with themselves (homozygous) or with known pathogenic variants (heterozygous). We use the pyensembl Python package to annotate the variants with information such as whether it is exonic or intronic, and whether the variant falls within the known boundary of the gene of interest. Our application then generates the reports which contain the homozygous co-occurring VUS, VUS co-occurring with known pathogenic variants, any associated phenotype data per VUS, and the allele frequency data.

Chapter 3: Modeling the impact of clinical data sharing on variant classification

Authors and affiliations

James Casaletto[1], Melissa Cline[1], Brian Shirts[2]

1. Genomics Institute, University of California, Santa Cruz

2. Department of Laboratory Medicine and Pathology, University of Washington, Seattle

Contributions:

In this research, I implemented all the modeling software in Python, ran all the simulations, generated all the results, published the code to GitHub, organized and led all the meetings and email communications with our collaborator and publisher, and wrote the manuscript.

ABSTRACT

Objective: Many genetic variants are classified, but many more are variants of uncertain significance (VUS). Clinical observations of patients and their families may provide sufficient evidence to classify VUS. Understanding how long it takes to accumulate sufficient patient data to classify VUS can inform decisions in data sharing, disease management, and functional assay development.

Materials and Methods: Our software models accumulation of clinical evidence (and excludes all other types of evidence) to measure their unique impact on variant interpretation. We illustrate the time and probability for VUS classification when laboratories share evidence, when they silo evidence, and when they share only variant interpretations.

Results: Using conservative assumptions for frequencies of observed clinical evidence, our models show the probability of classifying rare pathogenic variants with an allele frequency of 1/100,000 increases from less than 25% with no data sharing to nearly 80% after one year when labs share data, with nearly 100% classification after 5 years. Conversely, our models found that extremely rare (1/1,000,000) variants have a low probability of classification using only clinical data.

Discussion: These results quantify the utility of data sharing and demonstrate the importance of alternative lines of evidence for interpreting rare variants. Understanding variant classification circumstances and timelines provides valuable insight for data owners, patients, and service providers. While our modeling parameters are based on our own assumptions of the rate of accumulation of clinical observations, users may download the software and run simulations with updated parameters.

Conclusion: The modeling software is available at https://github.com/BRCAChallenge/classification-timelines.

OBJECTIVE

Genomic testing is now widely used for patients to determine if their genetics put them at increased risk of heritable disorders and to enable them to manage this risk clinically. For example, a patient with a known pathogenic variant in BRCA1 or BRCA2 should be screened more often for breast, ovarian, and pancreatic cancer. [40] Similarly, asymptomatic patients with familial cardiomyopathy might consider certain lifestyle changes such as losing weight, reducing stress, and sleeping well. [41]

The American College of Medical Genetics (ACMG) and the Association for Molecular Pathology (AMP) define qualitative, evidence-based guidelines for classifying genetic variants. Evidence for variant classification can come from many sources including clinical data, functional assays, and *in silico* predictors. Clinical data, typically derived through genetic testing reports, includes family history, co-segregation, co-occurrence, and *de novo* status. When sufficient evidence is present, a variant curation expert panel (VCEP) may classify the variant as Likely Benign (LB), Benign (B), Likely Pathogenic (LP), or Pathogenic (P) using the ACMG/AMP rules for combining evidence. Variants with little or no evidence to support classification, called Variants of Uncertain

Significance (VUS), create stress for patients and may lead to improper care. Because VUS do not yield medically actionable information, patients with VUS do not benefit from clinical management of their heritable disease risk. Ultimately, the significance of a variant remains uncertain until there is sufficient evidence to classify it. Although computational and functional predictions are helpful, some clinical data linking genotype and phenotype is usually needed to classify most variants. [42] However, there is no centrally available repository of clinical data that can be used for variant classification. Molecular testing laboratories and sequencing centers are the largest source of variant data. Many, but not all, clinical laboratories and sequencing centers actively share variant interpretations through ClinVar; however, they hold most of the clinical data they collect privately, due in large part to patient privacy and regulatory concerns. The shared interpretations for many genetic variants vary or even conflict between laboratories depending on the amount and nature of the evidence provided. [43]

One solution to these problems is for clinical laboratories to develop approaches to centrally share their clinical data associated with specific variants. Widespread sharing of variant pathogenicity evidence would lead to more rapid variant interpretation, greater scientific reproducibility, and novel discoveries. [43] [44] Indeed, the National Institutes of Health recently mandated the sharing of all data for the research which it funds.

While there is no question that data sharing would lead to expedited variant interpretation, and better patient outcomes by extension, under what circumstances is data sharing the most impactful? We have addressed this question by developing open-source software to model the probability of variant classification over time under various forms of data sharing.

The output of our model not only quantifies the value of sharing clinical patient data, the understanding of likely timelines and mechanisms of classification that this modeling illustrates could guide genetics organizations in prioritizing their efforts, inform strategies for functional assay development, improve variant classification guidelines, and enable healthcare providers to develop better strategies for managing specific patients with VUS. Further, the model serves as a platform for testing hypotheses on factors including the rates of gathering clinical evidence on the variant interpretation timeline. While we have informed the model with data according to the scientific literature and our own clinical experience, these factors are modeling parameters that can be modified easily as new evidence emerges, or to test the impact of clinical assumptions on the variant classification rate.

MATERIALS AND METHODS

This section outlines a statistical model that combines clinical information from multiple sequencing centers to create an aggregate, pooled center so that VUS may be classified faster.

Combining multiple forms of variant classification evidence

The evidence that the ACMG/AMP uses to classify variants encompasses several sources of data, including the type of variant (e.g. nonsense or frameshift), *in vitro* functional studies, *in trans* co-occurrence with a pathogenic variant, co-segregation in family members, allele frequency, and *in silico* predictions. They are divided into four levels of strength: "Supporting" (or “Predictive”), "Moderate", "Strong", and "Very Strong". For example, PP1, which represents co-segregation of the disease with multiple family members, is considered “Supporting” evidence for a Pathogenic interpretation. Another form of evidence called BS4 represents the lack of segregation of the disease with the variant in affected family members. The BS4 evidence is considered "Strong" evidence for a Benign interpretation. [45] Tavtigian et al [46] showed that the rule-based ACMG/AMP guidelines can be modeled as a quantitative Bayesian classification framework. Specifically, the ACMG/AMP classification criteria were translated into a naive Bayes classifier, assuming the four levels of evidence and exponentially scaled odds of pathogenicity. While the ACMG/AMP guidelines define rules for the combinations of evidence which lead to variant classifications, the Bayesian framework assigns points to each form of evidence. These points are summed and compared to thresholds to determine the variant’s pathogenicity. We leverage this Bayesian framework to model calculating odds of pathogenicity conditioned on the presence of one or more pieces of evidence for a given variant. For more detail regarding the combination of evidence, see Equations S1-S3 in the supplementary material.

To model the impact of clinical data sharing on variant classification, we exclude all other forms of evidence besides clinical evidence, as described in the following sections. For each variant, our model calculates two odds of pathogenicity: the odds of a VUS being benign and the odds of a VUS being pathogenic, both of which are conditioned on statistically sampled evidence.

Selecting categories of variant evidence for model

Some sources of variant classification evidence are not impacted by data sharing, such as *in silico* prediction scores and functional assay scores. We will not use those categories of evidence in our model so we can specifically quantify the unique contribution of cumulative clinical data to variant interpretation.

Several sources of clinical case and family information will contribute to variant classification over time. As clinical databases grow and data is shared more effectively across institutions, more variants will be classified. Increased clinical information is the major source for variant reclassification as well. [47] We selected the following categories of clinical pathogenic evidence for our model:

- *de novo* variants without paternity and maternity confirmation (PM6)
- co-segregation in family members affected with the disease (PP1)
- *de novo* variants with both paternity and maternity confirmed (PS2)

Similarly, we selected the following categories of benign evidence criteria that relate to clinical information:

- *in trans* co-occurrence with a known pathogenic variant (BP2)
- disease with an alternate molecular basis (BP5)
- lack of segregation in affected family members (BS4)

The more evidence that is gathered over time, the sooner and more likely a VUS will be classified. However, not all the evidence that is gathered over this time will be concordant. [48] Patients who have a pathogenic variant may occasionally present evidence from one or more benign categories, for example, lack of segregation in affected family members due to disease heterogeneity. This presentation of conflicting evidence for a given variant occurs at a low, non-zero frequency. Therefore, we use a combination of pathogenic and benign evidence in the classification of every VUS.

Parameters affecting clinical observations

To model the accumulation of clinical evidence, we defined certain modeling parameters according to the literature and to our own clinical experience. While the values that we have assigned to these parameters constitute well-informed assumptions, these values can be modified to test hypotheses, or as new knowledge emerges over time. In our software, these parameters are encapsulated in a single JSON file, so rerunning the model with revised parameter values requires modifying only one file.

Frequency distribution for evidence

Tavtigian et al calculated the corresponding odds of pathogenicity for each category of evidence and showed that the numerical-based odds are consistent with the rule-based ACMG/AMP guidelines for combining evidence. Those odds are shown in the “Pathogenicity odds of evidence” column of Table 3. Specifically, they determined that, for pathogenic evidence, the odds for "Strong" evidence is 18.7, for "Moderate" is 4.3, and for "Supporting" is 2.08. For benign evidence, the odds for "Strong" evidence is 1/18.7, for "Moderate" it’s 1/4.3, and for "Supporting" it’s 1/18.7. We derived the estimates for the evidence frequencies from the literature. [49] [50] [51] [52]

Evidence category	Estimated benign evidence frequency (low, medium, high)	Estimated pathogenic evidence frequency (low, medium, high)	Pathogenicity odds of evidence

PS2	(0.0001, 0.0015, 0.005)	(0.0006, 0.003, 0.02)	18.7
PM6	(0.0007, 0.0035, 0.01)	(0.0014, 0.007, 0.025)	4.3
PP1	(0.005, 0.01, 0.0625)	(0.05, 0.23, 0.67)	2.08
BS4	(0.025, 0.1, 0.4863)	(0.0001, 0.001, 0.17)	1/18.7
BP5	(0.038, 0.099, 0.36)	(0.00002, 0.0001, 0.00215)	1/2.08
BP2	(1.0 * f, 1.0 * f, 1.0 * f)	(0.001 * f, 0.005 * f, 0.02 * f)	1/2.08

Table 3. Odds and frequency estimate confidence intervals per ACMG/AMP category per clinical data evidence category. The variable f represents the frequency of the variant itself.

Table 3 depicts the odds and estimated frequency confidence intervals for the ACMG/AMP evidence categories that correspond only to clinical evidence. There may be pathogenic evidence observed for benign variants and benign evidence observed for pathogenic variants, though such observations generally occur at a low rate. For example, the frequency of BP2 for pathogenic variants is very unusual, except in tumors or in the case of rare diseases such as Fanconi anemia. Conversely, we assume that the frequency of BP2 evidence for benign variants is quite common and so occurs at the same rate (f) as the variant itself.

Thresholds for odds of pathogenicity

Tavtigian et al defined four threshold ranges for the odds of pathogenicity for each of the four ACMG/AMP variant classifications (Benign, Likely Benign, Likely Pathogenic, Pathogenic).

Classification	**Threshold for odds of pathogenicity**
Benign	(-∞, 0.001)

Likely Benign	[0.001, 1/18.07)
Uncertain significance	[1/18.07, 18.07]
Likely Pathogenic	(18.07, 100]
Pathogenic	(100, +∞)

Table 4. Odds of pathogenicity per ACMG/AMP classification

These thresholds from Table 4 correspond to the values from Table 3 and are consistent with the ACMP/AMP rules for combining evidence. For example, having one piece of strong evidence (e.g. BS4) and one piece of supporting evidence (e.g. BP2) is sufficient to classify a variant as "Likely Benign".

Data from participating sequencing centers

For generating simulated clinical data, we define three categories of sequencing centers: small, medium, and large as shown in Table 5.

Center type	**Current number of patients tested**	**Tests per year**
Small	15,000	3,000
Medium	150,000	30,000
Large	1,000,000	450,000

Table 5. Current number of tests in database and testing rate per center type

We estimated the large database size and testing rate from the online publications of relatively large sequencing labs, [53] [54] and we estimated the small database size and testing rate based on our own experience at the University of Washington Department of Laboratory Medicine (a relatively small laboratory). We estimated the medium database size and testing rate by interpolating between the large and small database values. Our model permits that these estimated sizes be replaced with other hypothetical or real sizes to predict outcomes under different scenarios. In relatively rare circumstances, the same patient may be tested at multiple facilities. In our model, we are assuming that each center has entirely distinct patient populations.

Ascertainment bias

Healthy people from healthy families are underrepresented in many forms of genetic testing. [55] Accordingly, patients with pathogenic variants are observed (or ascertained) more often than those with benign variants, and the forms of evidence that support a pathogenic interpretation accumulate more quickly. How much more likely a person is to present pathogenic evidence than benign evidence is captured in our model as a configurable real-valued constant. We conservatively estimated this term to be 2 based on our experience at the University of Washington Department of Laboratory Medicine.

Prior odds of pathogenicity

The Bayesian prior odds of a variant's pathogenicity represents all other criteria that are not clinical and do not change much, if at all, over time. For this implementation, we sampled a random value from a uniform distribution between 1/18.07 and 18.07 which is the lower and upper bound of the odds of pathogenicity for VUS.

Paradigms for sharing

There are 3 sharing paradigms which we use in our model: sharing nothing, sharing classifications, and sharing evidence. Of the three, we anticipate that sharing nothing will make variant classification most protracted and least probable. The paradigm of sharing classifications is what ClinVar currently enables, and we anticipate that sharing classifications will lead to shorter timelines with higher probabilities of variant classification than sharing nothing. Last, the essence of this research is to model the impact of sharing clinical evidence on the timeline of variant interpretation. We anticipate that sharing clinical data will lead to the shortest timelines with the highest probabilities of classification.

Implementing the simulation

Our statistical model contains one variable: the allele frequency of the VUS of interest. Parameters of the simulation software include the number and types of each of the participating sequencing centers and the number of years for which to run the simulation. Because the variant is of uncertain significance, we gather evidence for both benign and pathogenic classifications simultaneously.

For the first year of our simulation, all the evidence that is assumed to be currently present at each of the individual testing centers is initialized and aggregated. We use the Poisson distribution sampling method when determining how many times the variant is observed, given the VUS frequency. For each year in the simulation, we generate new observations for variants assumed to be benign and assumed to be pathogenic at each sequencing center. We aggregate those observations across participating centers into a single collection to simulate the sharing of data.

We ran simulations as described above 1,000 times to simultaneously generate data points for VUS which occur at the rate of one in every 100,000 people (1e-05), combining data from 10 small centers, 7 medium centers, and 3 large centers generated over 5 years. We then ran simulations for a VUS of frequency 1e-06 (one in every 1,000,000 people) in the same grouping of centers.

We created histograms and scatter plots that show the distribution and progression of the evidence over time. For each year, we plot the probability that each center classifies the variants individually using siloed data or if they collectively pool their data. We calculated the probability of a variant being classified at any sequencing center using the inclusion-exclusion principle in probability [56] assuming all centers would share all variant interpretations. This is a conservative estimate: not all sequencing centers share all their variant interpretations. We performed a sensitivity analysis to show the impact that each of the evidence types has on the probability of being either benign or pathogenic.

RESULTS

In this section, we discuss the results of our simulation with variants over the course of 5 years at 20 participating sequencing centers. We first examine the histograms of the evidence for pathogenic and benign variants after 5 years of observations. Second, we examine the trajectory of evidence over the course of 5 years in scatter plots. Third, we examine the probability scatter plots over the course of 5 years. Fourth, we analyze the sensitivity of our results with respect to each type of evidence. These four sets of results were generated using a variant of 1e-05 frequency. Last, we examine the probability scatter plots over the course of 5 years for a 1e-06 (one-in-a-million) variant.

Histogram plots of variants occurring at 1e-05 frequency

The distribution of evidence gathered individually and combined across all sequencing centers is plotted in Figure 2. As expected, increasing the number of classification data points for the many different variants results in wider Gaussian distributions that increasingly separate from the null assumption of no clinical evidence. More evidence provides more certainty in classifications as evidence exceeds the classification thresholds for an increasing number of variants.

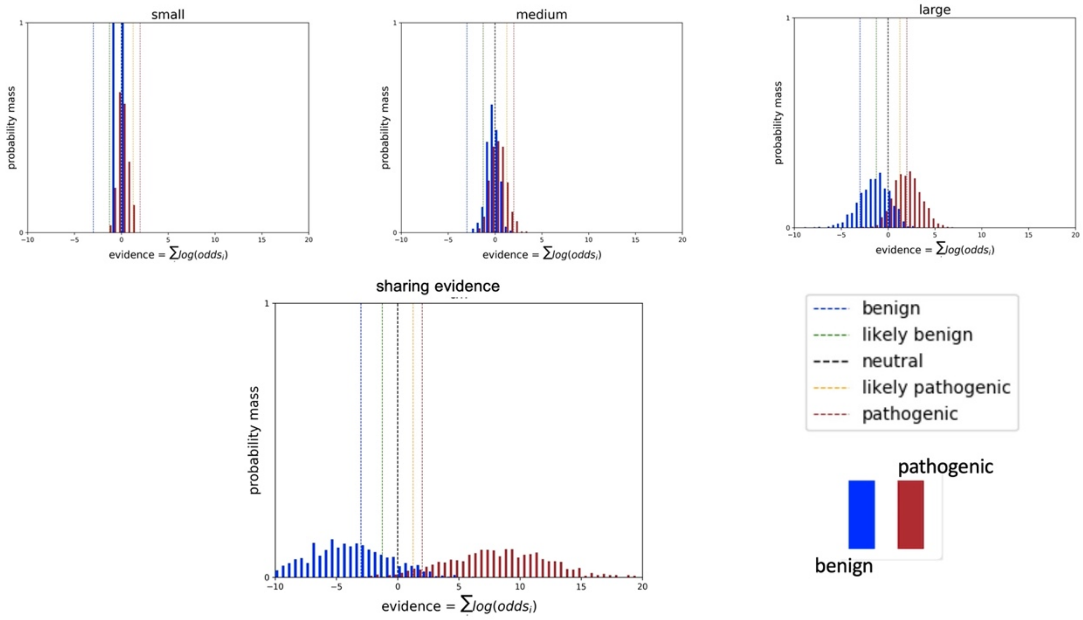

Figure 2 Histograms of cumulative log odds for classifying each of 1,000 simulated variants present at a 1e-05 frequency in the population. Classification thresholds are demarcated as vertical hash lines. Benign variants are shown in blue and pathogenic variants are shown in red.

Trajectories of evidence for variants at 1e-05 frequency

The classification trajectory for individual variants can vary depending on which observations are made and when those are made. Although data accumulation increases the likelihood of classification and the likelihood of correct classification for variants as a group, evidence for individual variants may rise and fall. Figure 2 plots a subset of 20 classification trajectories (10 benign and 10 pathogenic) at a small, medium, and large sequencing center as compared to the combined data across all sequencing centers assumed to be sharing evidence. Trajectories in these scatter plots mimic real-world phenomena: variants may accumulate contradictory evidence; and long time periods may pass with insufficient evidence.

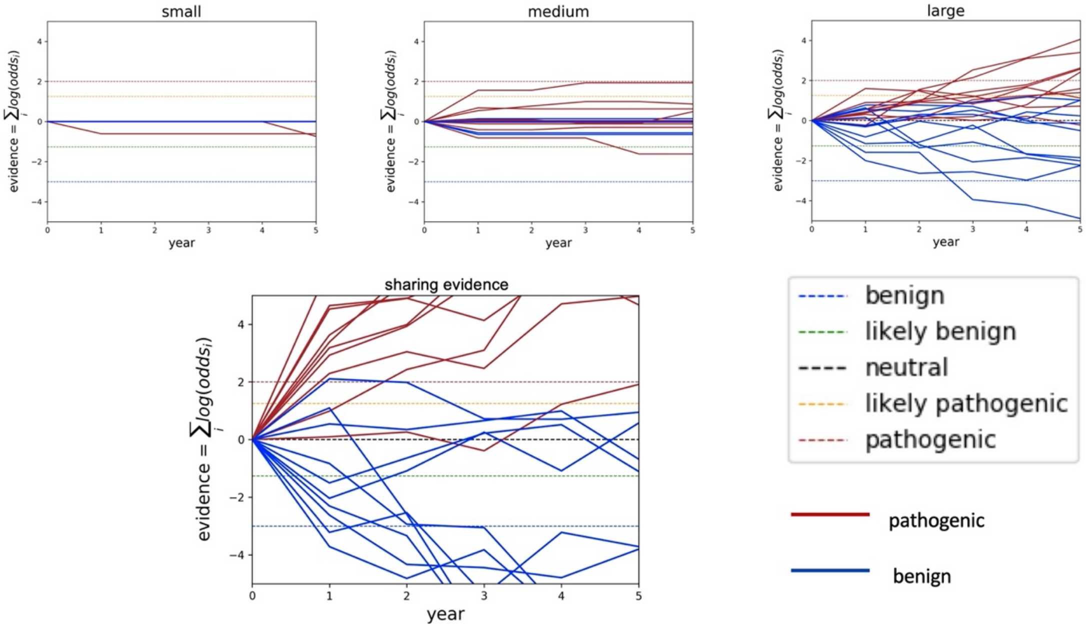

Figure 3 Classification trajectories for 20 randomly selected variants at 1e-05 frequency in the population. Classification thresholds are demarcated as horizontal hash lines in the timeline plots. Benign variants are shown in blue and pathogenic variants are shown in red.

Probabilities of classifying variants at 1e-05 frequency

Figure 3 shows the probability of classifying a variant which occurs at 1e-05 frequency in the population over the course of 5 years under different sharing paradigms. We show a small, medium, and large sequencing center not sharing anything as compared to two forms of sharing: centers sharing their all their variant interpretations but none of their clinical data (labeled "sharing classifications"); and centers sharing all their clinical data (labeled "sharing evidence"). From these graphs, we see that any data sharing increases the likelihood of variant classification. We also see that sharing evidence rather than sharing classifications makes variant interpretation more certain by moving “Likely Benign” variants into the “Benign” classification and similarly moving “Likely Pathogenic” variants into the “Pathogenic” classification. Moreover, sharing evidence rather than sharing classifications reduces the amount of time required to classify variants.

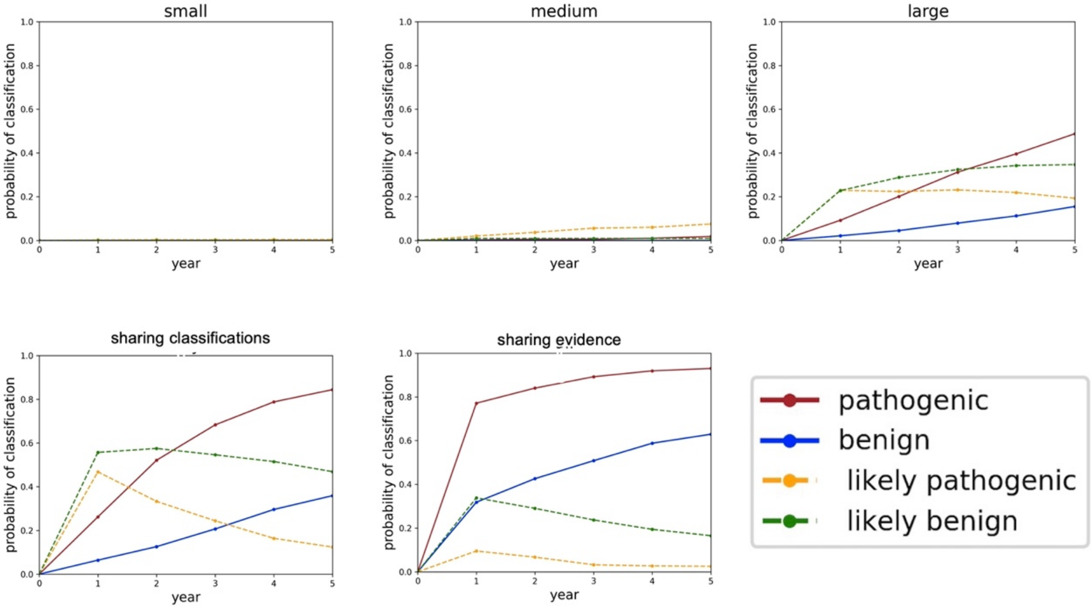

Figure 4 Classification probabilities over the course of 5 years. The y-axis of these plots is the probability of classifying the variant, converted from the aggregated likelihoods of pathogenicity generated in the simulations. Year 0 constitutes the time just before the sequencing centers share their data and all the variants are unclassified. Year 1 constitutes the moment just after the sequencing centers share their data. As time progresses and more evidence becomes available, some of the variants which were LB get "promoted" to B, and similarly some of the variants which were LP get "promoted" to P.

In the supplement, we explore changing the distribution of sequencing centers and the number of years sharing data. In supplementary Figure S1, we see that after 20 years of data sharing, almost all benign and pathogenic variants are classified using clinical data alone. In supplementary Figure S7, we see that reducing the number of participating sequencing centers from 10, 7, and 3 (small, medium, and large) to 5, 3, and 1 significantly reduces the probability of classifying variants using clinical data alone.

Sensitivity analysis for variants at 1e-05 frequency

We estimated conservative confidence intervals around the evidence observation frequencies defined in our model to determine how sensitive the probabilities of classification were to each type of ACMG/AMP evidence. We held all other parameters constant (equal to their expected values) while changing one frequency at a time to the low and high value in their respective interval to determine how sensitive the model is to changes in the frequencies observing different types of clinical data. Based on the assumptions of our experiments, classification of pathogenic variants is most sensitive to BS4 and PP1 evidence criteria (Figure 5a). Classification of benign variants is most sensitive to BS4 and BP5 evidence criteria (Figure 5b). Classifications were not affected by the change in BP2 evidence frequencies for either Benign or Pathogenic variants. Pathogenic variant classification was not affected by changes in BP5 evidence frequency and was therefore dropped from Figure 5a.

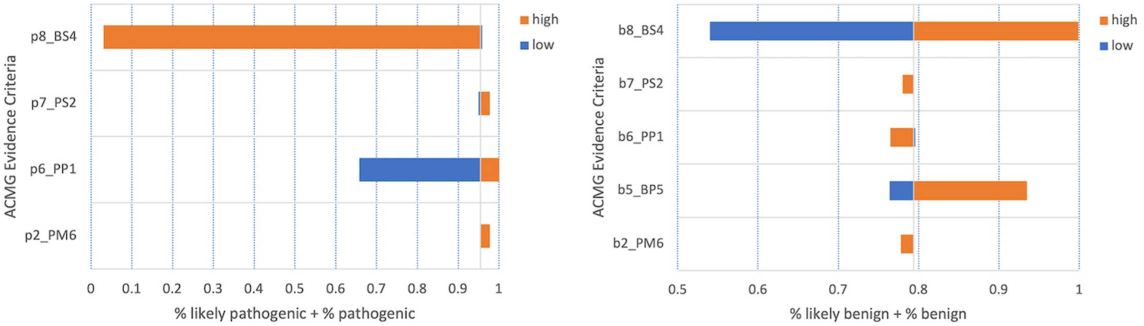

Figure 5 Sensitivity of variant classification to the frequency of observing ACMG/AMP evidence criteria. These "high" and "low" values are taken from the confidence intervals in Table 3. a.) Tornado plot for the sensitivity of Pathogenic and Likely Pathogenic variants. b.) Tornado plot for the sensitivity of Benign and Likely Benign variants.

Probabilities of classifying variants at 1e-06 frequency

For comparison, we evaluated the probability of gathering data for a one-in-a-million variant through data sharing. Figure 6 shows the probability of classifying a 1e-06 variant over the course of 5 years.

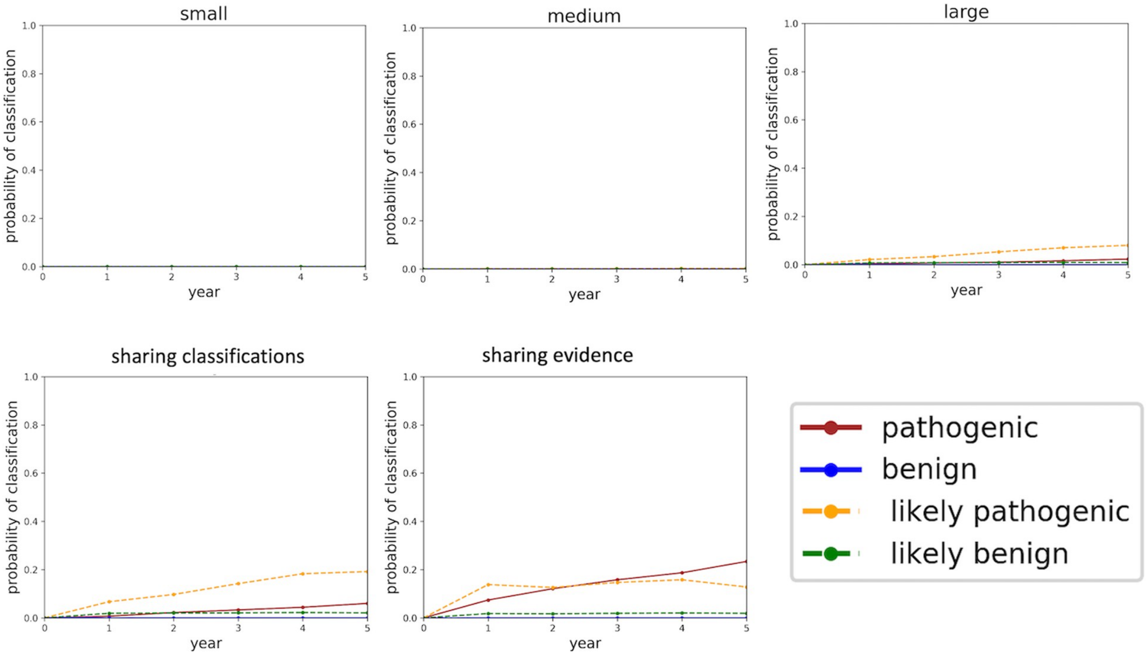

Figure 6 Probabilities of classifying variants at 1e-06 frequency plotted over the course of 5 years. The y-axis of these plots is the probability of classifying the variant, converted from the aggregated likelihoods of pathogenicity generated in the simulations. Year 0 constitutes the time just before the sequencing centers share their data and all the variants are unclassified. Year 1 constitutes the moment just after the sequencing centers share their data.

In addition to these probability plots, we also performed analysis of 1e-06 variants to generate cumulative odds histograms (supplemental Figure S2) and classification trajectories (supplemental Figure S3). These illustrate similar results. To further explore variant classification timelines for 1e-06 variants, we evaluated classification over 20 years of data sharing (supplemental Figures S4-S6). Sharing evidence is predicted to help classify a minority of 1e-06 variants even after 20 years of data sharing.

DISCUSSION

These simulations illustrate that clinical data sharing reduces the time and increases the certainty in classifying VUS. Sharing only variant interpretations rather than clinical data, however, results in longer timelines and lower certainty. For example, the same variant could be interpreted as Likely Pathogenic at one laboratory and as a VUS at a different laboratory based on evidence seen at the two respective laboratories. Similarly, the simulations show that evidence for a given variant can, at times, be contradictory. As defined in the ACMG/AMP classification standards, evidence of pathogenicity may be presented for benign variants (and vice versa), though less frequently than for pathogenic variants. Importantly, our simulations demonstrate that discordant evidence resolves more quickly and with higher certainty when centers share their clinical data rather than only sharing their variant interpretations. These are critical results: mis-classified variants mis-inform healthcare providers and may lead to disastrous patient outcomes. [57] Variants originally classified as Likely Pathogenic or Likely Benign more readily become classified as Pathogenic and Benign, respectively, when data is shared.

Our simulations show that, using clinical evidence alone, classifying pathogenic variants has a higher probability and quicker timeline than for classifying benign. Those ACMG/AMP evidence criteria and classification guidelines that rely on patient clinical data, which we have modeled, require more evidence for benign classification [45] which results in longer timelines. Models indicate that improved guidelines could balance pathogenic or benign evidence categories, or alternatively create a new “lack of pathogenic evidence despite sufficient observations” category of benign evidence.

Additionally, our model can guide functional assay developers as to which variants they should include in their panels. Very rare variants for which we expect insufficient clinical data under any sharing model will need a functional assay to classify it. Functional assays are expensive and require expert interpretation, and this information can maximize the impact of those efforts by identifying variant frequencies and sharing scenarios in which data sharing by itself is insufficient for classification.

We see that highly rare variants (one-in-a-million or less) may be unlikely to be classified by aggregating clinical information alone. Because most variants are highly rare, [58] it's essential that we invest in strategies for the interpretation of highly rare VUS. One strategy is additional investment in cascade testing for highly-rare variants in high-penetrance genes. This is an effective strategy because the variant may be rare in the general population but can still be enriched in the family. [47] Moreover, cascade testing is part of the PP1 and BS4 classification categories, and Figure 5 indicates that both benign and variant classifications are quite sensitive to those categories. Another effective strategy is investment in large-scale functional assays, such as MAVEs (Multiplexed Assays of Variant Effect), which can assay thousands of variants at once. [59]

Most importantly, variant classification timelines will guide prevention, diagnosis, and treatment decisions for patients and their healthcare teams. For example, a patient with a known pathogenic variant in *BRCA1* or *BRCA2* may elect to have a prophylactic mastectomy which, according to the National Cancer Institute, reduces the risk of breast cancer in women who carry a pathogenic *BRCA1* variant by 95%. [26] A patient with a *BRCA1* VUS, on the other hand, may choose to wait if their variant is likely to be classified in the near-term (e.g. within 2 years) but seek alternative options, such as family co-segregation analysis, if that variant will not likely get classified for another 10 years or more. More than half of the variants in the *BRCA1* and *BRCA2* genes are VUS, even though these are two of the most widely studied genes in the human genome. Other Mendelian diseases with highly penetrant alleles have a significantly larger proportion of VUS, so understanding timelines and probabilities of variant classification will have an even higher impact for those genes.

With sufficient clinical data from cooperating sequencing laboratories, these estimates enumerate tangible outcomes that may result from data sharing. It is clear that more variants will be classified and patients will benefit with robust data sharing. This is particularly important over longer time horizons (see supplementary data for 20-year modeling). There are several mature privacy mechanisms that may be leveraged to share data responsibly; differential privacy, secure multi-party computation, homomorphic encryption, blockchain, and federated computing are approaches that have matured and are available today to protect the privacy of those individuals who have shared their data as well as protect the business interests of the institutions which own the data. [60] [61] [62] [63]

CONCLUSION

It is assumed that sharing clinical patient data should improve variant interpretation using the ACMG/AMP variant classification guidelines. Our research provides a framework to explicitly quantify how much and under what circumstances it improves. We have built and made available a model that simulates the generation and sharing of clinical evidence over time. The software provides graphical results to compare sharing clinical data with sharing only interpretations and sharing nothing. Our experiments were based on data estimates from the literature and from our own experience, but readers can define their own values for the frequencies of observations of various ACMG/AMP evidence criteria and experiment with different combinations of centers, different sizes and testing rates, and with different allele frequencies.

FUNDING STATEMENT

J.C. is supported by NHGRI grant U54HG007990 and NHLBI grant U01HL137183. M.C. is supported by NCI grant U01CA242954 and BioData Catalyst fellowship OT3 HL147154 from the NHLBI through UNC-CH 5118777. BS is supported by a grant from the Brotman Baty Institute for Precision Medicine and by the NIH 1OT2OD002748.

COMPETING INTERESTS STATEMENT

The authors do not have any conflict of interest to disclose with respect to this research or manuscript.

CONTRIBUTORSHIP STATEMENT

BS conceived the initial statistical model and wrote a draft in R. JC developed the model software in Python. MC conceived the 'sharing classifications' sharing paradigm. JC wrote the manuscript. JC, MC, and BS reviewed, revised, and approved the final version of the manuscript.

ACKNOWLEDGEMENTS

The authors are very grateful to David Goldgar and Sean Tavtigian for providing their expert insight in the approximation of ACMG/AMP evidence criteria observation frequencies.

DATA AVAILABILITY STATEMENT

There are no external data associated with this manuscript. All the data are generated synthetically in the modeling software, which is available at https://github.com/BRCAChallenge/classification-timelines.

Chapter 4: Federated analysis for privacy-preserving data sharing

Authors:
James Casaletto[1*], jcasalet@ucsc.edu 0000-0002-2339-5362
Alexander Bernier[2], alexander.bernier@mcgill.ca 0000-0001-8615-8375
Robyn McDougall[2], robyn.mcdougall@mcgill.ca 0000-0002-5741-3987
Melissa S. Cline[1], mcline@ucsc.edu, 0000-0002-0148-1956

Author affiliations:

1. UC Santa Cruz Genomics Institute, 1156 High Street, Santa Cruz, CA 95064 U.S.A.
2. Centre of Genomics and Policy, McGill University Faculty of Medicine and Health Sciences,

* Corresponding author

Contributions:

In this paper, I researched all the technology solutions to privacy restrictions, organized and led all the meetings and email communications with our collaborators and publisher, designed and drew all the figures, and wrote the parts of the manuscript which deal with technology solutions to privacy restrictions.

Abstract

Continued advances in precision medicine rely on the widespread sharing of data that relates human genetic variation to disease. However, data sharing is severely limited by legal, regulatory, and ethical restrictions that safeguard patient privacy. Federated analysis addresses this problem by transferring the code to the data: providing the technical and legal capability to analyze the data within its secure home environment rather than transferring the data to another institution for analysis. This allows researchers to gain new insights from data that cannot be moved, while respecting patient privacy and the data stewards' legal obligations. Because federated analysis is a technical solution to the legal challenges inherent in data sharing, the technology and policy implications must be evaluated together. Here we summarize the technical approaches to federated analysis along with a legal analysis of their policy implications.

Introduction

Most diseases have a genetic component [64]. While clinical genetic testing now offers individuals greater opportunities to understand and manage their heritable disease risk, [65] [66] its impact is limited by the many gaps in our understanding of human genetic variation. This is seen in the significant missing heritability of most diseases, with family history predicting disease risk more accurately than genetics. [67] [64] It is also seen in the high rates of "Variant of Uncertain Significance" (VUS) in clinical testing, with recent studies reporting VUS results in roughly 20% of the patients tested with cancer susceptibility gene panel tests. [68] [69] [70] Patients of non-European ethnicities have a significantly higher rate of VUS test results than their European counterparts, resulting in greater mortality from heritable disorders. [71] [72] These problems could be addressed by data sharing, particularly global data sharing. However, sharing human genetic data is limited by a complex network of legal, ethical and regulatory restrictions that aim to protect patient privacy [73] or national [74,75] data sovereignty. [76,77] As a result, most human genetic data remains siloed, and is inaccessible to most researchers and those making clinical inferences.

Yet often, these regulations permit the sharing of aggregated, non-identifiable data, which can advance research. For example, ACMG/AMP variant interpretation guidelines [45] include summary statistics that describe the enrichment or lack of disease in patients with a given genetic variant, but do not directly require individual observations of those patients. If one can share the capabilities to generate these summary statistics, one can share knowledge without transferring the sensitive patient data.

Data federation achieves such sharing through a decentralized architecture, in which a network of data providers maintains full control over the data within a secure computing environment while enabling access to the data by external collaborators. [78] Federated analysis is a form of data federation in which collaborators "bring the code to the data" to analyze data *in situ*, within the data providers' computing environment (see Figure 7).

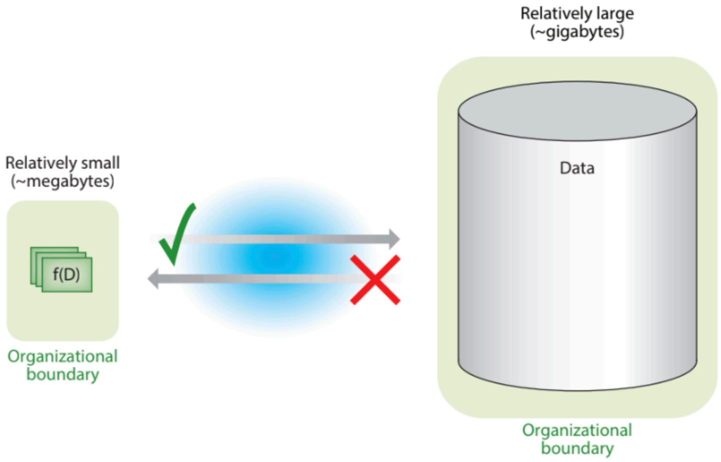

Figure 7. Illustrating the logic of "bringing the code to the data", either because the data is much larger relative to the analytical code f(D) or due to restrictions of exporting the data across organizational boundaries.

When federated analysis generates non-identifiable results from patient-level data, those results can be shared externally, where the original patient-level data cannot. Federated analysis balances the needs of data stewards to restrict access to regulated data with the needs of the scientific and clinical community to gain new insights from data that cannot be transferred for legal, ethical, or technical reasons. Achieving this vision requires technologies that allow researchers to reliably analyze data that they cannot directly access, coupled with privacy safeguards that allow data stewards to assess the risks inherent in such analyses. In short, federated analysis offers technical solutions to legal restrictions on data use and data sharing. As such, to understand its potential and limitations, one must consider the technical and legal aspects of the situation together.

This article begins with a discussion of privacy in genomics, and how that privacy can be compromised. Next, we discuss the central concepts underpinning the General Data Protection Regulation (GDPR), the principal legal framework that regulates the use of personal data in the European Union (EU) and European Economic Area (EEA). The GDPR has been selected instead of other national and international data protection and privacy norms because it is amongst the strictest and most influential data protection laws in the world. The conclusions of the manuscript are nonetheless generalizable to ensure compliance with most other national data protection norms, such as the Health Insurance Portability and Accountability Act (HIPAA) of the United States. We next summarize approaches to federated analysis, presenting examples of successful applications and evaluating both the technical approaches and legal implications. Finally, we review areas for further progress on both the technical and legal fronts, presenting overall recommendations for the technical and organizational implementation of federated data analysis.

Privacy and Privacy Attacks

The sequencing of the human genome enabled unprecedented opportunities for the biomedical research community to make discoveries that are actionable in human health and precision medicine. It used to be that preserving the privacy of participants in biomedical research equated to maintaining the confidentiality of personally identifiable information (PII) and protected health information (PHI) by either publishing aggregated data or by removing PII and PHI. However, over the years, researchers have found ways to uncover protected information from such datasets. In 2008, the Wellcome Trust and the NIH removed access to genomic datasets after it was shown that people could be re-identified from data aggregated from GWAS experiments. For some organizations, removing PII and PHI, or only publishing aggregate data, were not sufficient to protect the privacy of research participants. [79]

Privacy threat model. A privacy threat model defines the most probable attacks on private data, the actors perpetrating the attack, and how the actors would carry out the attack. The actors in a privacy threat model are those entities (people, groups, or organizations) that have some form of access to the data. It’s impossible to protect against every attack, so it is necessary to focus on either the most likely or the most detrimental attacks.

In a federated environment where the upstream data contributors, the data custodians, and the downstream data users belong to different organizations, the threat model becomes more complex because the system could be threatened by participants in any of these organizations. One cannot assume that all participants are 100% reliable. Yet one might assume that some of the participants are reliable, as otherwise there would be no incentive to participate in the system. In short, the safest assumption is that some entity within the system is a potential threat.

A few privacy threat models have been developed and proposed, including the Cloud Security Alliance (CSA), European Network Information and Security Agency (ENISA), and Linkability, Identifiability, Non-repudiation, Detectability, information Disclosure, content Unawareness, and Non-compliance (LINDDUN). [80] Where the CSA does not have specific details for how to preserve privacy, ENISA and LINDDUN, while comprehensive, require a significant investment in time and training to understand and incorporate into a secure, privacy-preserving framework. More purpose-designed approaches such as Cloud Privacy Threat Modeling (CPTM) may be better-suited for designing, implementing, and deploying federated solutions where time and resources are not abundant. [81]

Cyber attacks. Cyber attacks are attempts to read, modify, or delete information through unauthorized access to computer systems. They put the integrity or availability of an otherwise trusted system at risk. For example, a denial-of-service attack is one in which the attacker runs a workload against a service which renders some or all the service either compromised or unavailable. Man-in-the-middle (MITM) attacks involve an unknown third party which can intercept network communications between two otherwise honest entities and impersonate one or both of them. One of the most common and severe cyber attacks is called SQL injection, in which malicious or malformed SQL code is inserted in a Web form that is subsequently unwittingly processed by a SQL service on the backend of the Web form. Such attacks can obtain unrestricted access to databases with sensitive information, resulting in identity theft, loss of information, and fraud. [82]

In the context of federated computing, cyber attacks can be largely mitigated by following common-sense IT security protocols. Federations should require all members to belong to the same virtual private network to block outside traffic. They should use public key cryptography and sufficient authorization requirements within the federation to prevent any leakage of PII and PHI data between federation members. And they should leverage other security measures such as firewalls and multi-factor user authentication across the federation and on a per-host basis.

There are more subtle forms of attack, however, that can be levied within federated environments or in the models that are built and published from federated environments, as described in the following sections.

Re-identification attack. Some privacy attacks require some external data source which contains overlapping data or metadata such that the datasets may be joined to provide a more comprehensive description of an entity such as a person. The ability to join data from multiple sets on some common identifier is called linkability, and the attacks that exploit it are called linkage attacks. Such an external data source might not be available today yet might become available in the future. A re-identification attack occurs when data that have been anonymized or pseudonymized become personally identifiable, an example of which is the re-identification of Massachusetts Governor William Weld in 1997. [83] This re-identification attack required having full ZIP codes, complete birth dates, and gender specified in both health plan data and voter registration data; this linkage enabled the connection between Governor Weld's identity in the voter registration data and his medical records in the health plan data. This attack could be prevented by masking ZIP codes to 2 digits and using only birth year, for example. If those covariate data do not contribute to the utility of the shared dataset, then they should be omitted entirely. This illustrates the principle of data minimization, a fundamental principle of computational data privacy [84].

Reconstruction attack. A reconstruction attack is the ability to partially or entirely reconstruct private data from published aggregate data. In this scenario, a trained model or aggregated dataset was produced from data which contain potentially sensitive information and shared; subsequently, attackers attempt to infer or reconstruct the sensitive information. This broad form of attack includes membership inference and property inference.

A membership inference attack exploits the ability to determine whether a person comes from a source dataset. Methods by which membership attacks may be leveraged are well documented by Shokri and colleagues [85]. One example of membership inference is the attack proposed by Homer et al., in which they demonstrate that it is possible to determine if a person's genomic data were used in the creation of published statistics in a GWAS [86]. Given knowledge of the allele frequencies in the population, the allele frequencies in the GWAS mixture and the genotype information of the person of interest, the attacker calculates how "far" the person of interest is from the reference population and the GWAS mixture using the allele frequencies. The further the person is, the higher the confidence in the membership inference becomes. In another example [87], researchers demonstrated that an attacker who possesses a person's genomic sequence can determine that person's membership in a Beacon, including Beacons that relate to disease, and in this way the Beacon network could leak some of that person's PHI.

A property inference attack attempts to infer some aggregate information about the training set as a whole, such as the environment where the data were produced or the percentage of the data that comes from a particular class (i.e. exploiting skewness). [88] [89] It requires that the attacker have auxiliary datasets that contain some property of interest. With these auxiliary datasets, the attacker can build "shadow models" for each property of interest, and then create a classifier which compares results from the target model against these shadow models to distinguish whether the property in question belongs to the target model. For example, Humbert, et al., demonstrated that certain single nucleotide polymorphisms (SNPs) of family members related to Henrietta Lacks could be inferred using (i) available genomic data, (ii) family relationship structure, (iii) rules of Mendelian inheritance, (iv) minor allele frequencies of the SNPs, and (v) linkage disequilibrium among the SNPs. [90]

Model-poisoning attack. In contrast to the types of attacks discussed previously, model-poisoning attacks may occur within a federation while a model is being built or analysis is being performed on private data. Possible objectives include a denial-of-service attack which simply renders the model ineffective for predictions using out-of-distribution data; or “label flipping”, which targets a subpopulation of the training data such that model predictions involving that subpopulation are erroneous. Yet another, more sophisticated objective of model-poisoning is using that information after the model is built to make inferences about the dataset (e.g. property or membership inference attacks). [91]

The two types of model-poisoning attacks involve either data misconduct or model misconduct. [92] Model misconduct involves changing how the analysis is performed to alter the outcome, while data misconduct requires that the adversary inserts data sufficient to alter the model predictions. For example, if a model that classifies images is trained using images available on the Internet, then an attacker can poison that model by uploading poisoned images to the Internet. The ways to mitigate this risk include limiting the contribution of any single entity, analyzing the nature of the updates to the global model on a per-contributor basis, and performing outlier detection after the model is built. However, models can be poisoned unintentionally as well. For example, the unintentional under-representation of non-European populations in GWAS studies arguably poisons GWAS models against these under-represented populations. [93]

Introducing the General Data Protection Regulation (GDPR)

The GDPR regulates the use of identifiable personal data: data relating to a person which is identified or identifiable. For data to be considered identifiable, there must be a “means likely reasonably to be used either by the controller or by any other person to identify the [concerned individual]”. This is not the case if re-identification is “practically impossible on account of the fact that it requires a disproportionate effort in terms of time, cost and man-power, so that the risk of identification appears in reality to be insignificant”. [6] If the data controller and proximate third parties do not have a mechanism enabling the re-identification of the concerned individual that is “likely reasonably to be used” [7], then the data are considered to be anonymized and therefore not regulated by the GDPR.

There is an apparent tension between the manner in which the Court of Justice of the European Union (CJEU) characterizes this legal test and the manner in which it is articulated in the text of the GDPR. [94] The GDPR calls for a contextual assessment of the reasonable likelihood of re-identification. This suggests that in circumstances in which re-identification is possible, but is improbable or impracticable, the data should be considered non-identifiable and therefore not regulated. [94] The CJEU, on the other hand, appears to characterize data as identifiable unless re-identification is nearly impossible. [6] Nonetheless, both appear to confirm a contextual, risk-based approach to the evaluation of data identifiability. [95]

The more restrained reading of the GDPR identifiability criteria should be preferred, as this interpretation limits the application of the GDPR’s onerous procedural requirements on information that poses a material risk of causing individual re-identification, assessed from the perspective of the data controller. If too many data are considered regulated personal data despite posing a limited risk of re-identification, this could frustrate the functioning of data protection legislation. Furthermore, if identifiability is not assessed from the perspective of the data controller and data processor, but from the perspective of all third parties, it becomes difficult or impossible for regulated parties to determine the boundaries of their legal responsibilities. [6,94,96]

The broad framing of identifiable personal data is a potentially unfortunate public policy choice. GDPR-regulated entities have limited financial, human, and technical resources for ensuring their compliance with the data protection regulation. If data identifiability standards are framed to capture a broad range of data that has a limited risk of re-identification, this framing could prompt regulated actors to scale down their data sharing activities due to the high burden of regulation. It also encourages regulated actors to direct their limited compliance resources to the majority of the data that these actors process rather than structuring their compliance activities in a risk-adjusted manner (i.e., directing their legal compliance resources at the data that has the highest chance of being re-identified). This could lead to actors performing subpar data protection compliance because they lack sufficient resources to ensure appropriate compliance. It could also lead to actors reducing their data sharing due to the potential for GDPR non-compliance, even in circumstances where the data sharing offers large benefits to society or the individual and poses privacy risks that are limited or nonexistent. [97]

If data are identifiable, one must consider the role and the associated legal responsibilities that the GDPR ascribes to the actors that use data, or that determine how data are used. The GDPR uses the concepts of joint controller, controller, and processor to determine the legal obligations of an actor who uses identifiable personal data. Each of these roles bears distinct legal responsibilities. The determination of whether an actor is a joint controller, a controller, or a processor is left to the supervisory authorities (i.e. national regulators) and courts. This means that these roles and responsibilities are not determined by the actor's choice, but rather the manner in which the actor uses data. This determination is left to the national data protection regulators, referred to as 'supervisory authorities,' and is further adjudicated by national courts and the CJEU. [94]

The GDPR defines data processing as "any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means". [94] It goes on to enumerate a non-exhaustive list of examples. In essence, all actions that entail the use or storage of identifiable personal data fall within the definition of data processing.

Controllers determine the purposes and means of personal data processing. Processors perform personal data processing activities at the instruction of controllers. In short, data controllers decide what will be done with personal data while processors implement these decisions. Accordingly, the data controller bears greater legal responsibilities than the data processor.

In some instances, multiple actors will collaborate in determining the purposes and means of personal data processing. This could be the case, for example, if a central organization coordinates and determines the conditions according to which third parties will collect and use personal data for their own purposes. [7,94,94]   In such instances, the law would categorize these multiple actors as joint controllers. The GDPR requires joint controllers to develop contracts to establish their respective and overall responsibilities. If the collective joint controllers are held liable for some data breach, then each joint controller can be held fully responsible for harms that arise from the action of other controllers in the network. The sole ground that enables any single controller to not be held liable for the actions of the others is for this controller to "[prove] that it is not in any way responsible for the event giving rise to the damage". [94]

Data Protection Implications of Federated Computing

Providing guidance on compliance with the substance of the GDPR lies outside the ambit of this article. Rather, we aim to aid health sector data stewards in determining how the structure of their federated data analysis networks, from both a technical and an organizational standpoint, determine the characterization of their activities according to the GDPR—i.e. whether the GDPR would understand them to be controllers, processors, or neither. The characterization first considers whether or not the concerned actor is engaged in the processing of identifiable personal data, and then considers whether this actor could be characterized as a controller (bearing more responsibilities) or a processor (bearing fewer responsibilities).

A broader public policy context animates this analysis. The GDPR has been strongly criticized as an impediment to the research use of data, to the advancement of precision medicine, and to healthcare system functioning in general. [98] Several arguments support this position. Inherent ambiguities in the language of the GDPR create difficulties in determining how to ensure compliance, even when actors behave in good faith. [99] Determinations regarding the appropriate use of data, which are traditionally left to health sector data stewards and research ethics committees, are shifted to generalist data protection regulators, who do not necessarily possess the specialized, domain-specific knowledge required to apply the standards of the GDPR to the health sector. [10] Procedural requirements, such as maintaining data processing reports and documenting self-assessments, can overwhelm the limited resources available to many institutions. [101] Finally, the consequences of non-compliance, which can include poor public perception, administrative fines, and civil liability, can deter health sector institutions from exploring data sharing activities. [74] This perverse incentive is especially strong when one institution's participation would greatly benefit the larger network but participating in the network would not benefit that institution as greatly.

Bearing in mind the foregoing critiques of the GDPR, the critical roles in a federated data analysis network include the following: [78]

First, there are nodes that are responsible for collecting data from individuals and for contributing such data to the network. Second, there are nodes that act as technical data stewards by providing the infrastructure that supports the data storage and processing. Third, there are nodes that act as institutional data stewards, determining which actors can participate in the network as upstream contributors or as downstream recipients of analysis results (and the conditions of such participation). Fourth, there are nodes that act as downstream recipients of analysis results, and that can submit analysis queries to the network and receive responses. Since each node that must perform GDPR compliance activities bears compliance costs in entering the network, the burden of compliance activities scales in proportion to the number of nodes in the network.

One principal advantage of federated data analysis is to enable scalable access to larger quantities of data. There is a tension between the scalable nature of the federated analysis network from a technical perspective, and the inexorable growth of activities required to ensure the network's compliance.   That is, certain legal compliance activities, such as performing Data Protection Impact Assessments (DPIAs), retaining records of data processing activities, or ensuring the alignment of data processing activities with the principles of the GDPR can require intensive and repetitious human effort. This can lead to circumstances in which data processing is cost-effective and scalable, but establishing records of their compliance with select formalistic, procedural elements of the GDPR is prohibitively cost-intensive. It is advantageous to structure networks to reduce the number of regulated data controllers and data processors, to enable more streamlined compliance and ensure that the network remains open to a broad range of prospective nodes, including those that lack the significant resources required to perform burdensome regulatory compliance activities.

In short, nodes in a federated data analysis network should use technical and organizational measures to ensure that the benefits of data analysis are maximized without most network nodes engaging in personal data processing, whether as controllers or processors. [102] If the data analysis is structured such that most participating nodes do not process potentially-identifiable data, using both organizational and technical safeguards, then these nodes will not be required to engage in GDPR compliance activities, and the other nodes are not required to consider these nodes as part of their own GDPR compliance efforts. This ensures that the compliance of the network is cost-effective and simple.

Data Federation and Privacy mechanisms

Most applications run on a single system. A distributed application is written to run across more than one system to leverage the compute, memory, and/or storage resources of multiple systems. For such applications, the programmer must be provided an abstraction that ignores the physical location of the data. [103] Federated computing is a form of distributed computing wherein some or all these data are subject to the stewardship of an entity other than that which provides and/or runs the application. That is, the analytical software is transferred to the location where the data resides rather than the data being exported for analysis. It is particularly well-suited for those datasets which are either too large or too sensitive to move between organizations. [78,104]

There are many forms of federated computing. In federated learning, for example, models are trained over remote datasets in siloed data centers, personal computers, cell phones, and other edge devices while keeping data localized. [104] A federated database, by contrast, is a collection of databases that operate as if they were a single database from a unified portal. [105] The Gene Expression Omnibus (GEO) is a federated database of microarray, next-generation sequencing, and other high throughput functional genomics data. [106] Federated analytics, yet another form of federated computing, distributes predictive or descriptive analytic tasks over one or more systems. [107] In this paper, we will focus on descriptive analytics using genomics data.

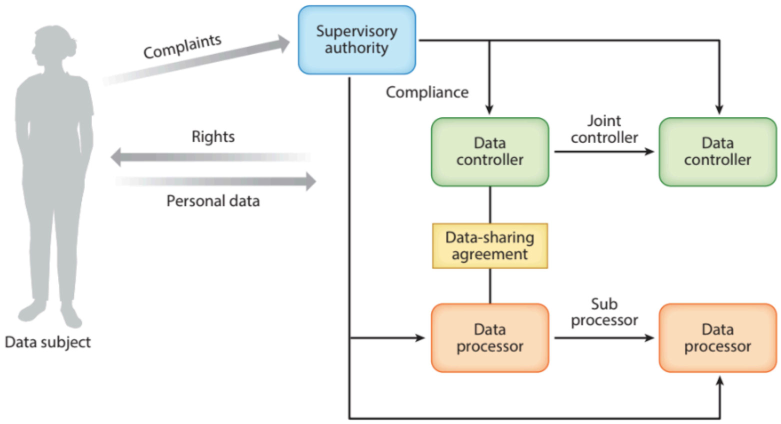

Figure 8. Legal entities and their roles in a GDPR data sharing agreement.

All federated methods involve cooperation between people and organizations and sharing some form of potentially sensitive information. In this section, we discuss different privacy mechanisms that may be used within a federation to eliminate any privacy leakage between federation members.

Secure Multiparty Computation (SMC)

Secure multiparty computation (SMC) or multiparty computation (MPC) was originally formulated as a research question called the "millionaire's problem" in which there are two or more people who are interested in knowing which of them is richer without revealing their actual wealth and without the help of a trusted third party. [108,109] The foundation of SMC entails secret sharing which leverages zero-knowledge proofs, techniques that enable a “prover” to prove a claim to one or more “verifiers” in such a way that they are convinced of its truth without the prover revealing the assertion or any party witnessing the interaction. [110] SMC protocols have a correctness requirement that guarantees that either the output is correct or the protocol terminates early. The number of adversaries (t) that the protocol can tolerate and still be correct (i.e. either terminate the protocol or produce correct output) depends on the type of secret sharing. Using additive secret sharing, the protocol can tolerate all but one honest participant ($t < n$). Using Shamir secret sharing, the protocol can tolerate up to $t < n/2$ passive adversaries and up to $t < n/3$ active adversaries. Examples of SMC in GWAS analyses include Constable, et al., [111] in iDASH 2015, and Cho, et al., [112] on data from the Database of Genotypes and Phenotypes (dbGaP).

SMC is an ideal protocol to leverage in a genomics federation. It is purpose-built for minimizing privacy leakage while maximizing utility among multiple participants operating on local data to construct global results.

Homomorphic encryption (HE)

Encryption is the process of encoding data in such a way that it cannot be interpreted without decoding. The stronger the encryption, the less likely the data can be decoded by brute force. In homomorphic encryption, [113] data are encrypted with a public key and sent to an outside potentially untrusted source to perform computations. That party never decrypts the data, but instead operates on the data in its encrypted format. The party sends the encrypted results back to the originator who, with a private key, can decrypt the results.

Encryption and decryption are notoriously slow, so performing these large-scale genomic analyses on encrypted data has been prohibitive. However, recent work has improved those algorithms through parallelization—a programming technique which divides an application workload into multiple parts, each of which can be run simultaneously on different systems, processors, or cores. In 2018, the winning team of the iDASH Genomic Privacy Challenge implemented a logistic regression approximation for GWAS which was 30x faster than the competing SMC solution. [114] They did so by parallelizing the execution of matrix operations, efficiently encoding the encrypted data, leveraging approximate arithmetic, and optimizing several cryptographic subroutines. These improvements generalize beyond GWAS computation, enabling homomorphic encryption solutions in other domains requiring large-scale statistical analyses on encrypted data. The following year, the winning team of iDash reduced the time necessary to perform imputation on 80,000 SNPs to less than 25 seconds.

Differential privacy (diffP)

Differential privacy is a privacy mechanism which adds noise to a database query result such that the entity submitting the query cannot determine whether any particular individual is a member of that database or not. This addresses membership inference and reidentification attacks. The more noise the mechanism adds, the less likely it is to infer membership, and therefore provides stronger privacy guarantees, but stronger privacy guarantees may render the data less useful.

The concept of epsilon-differential privacy mathematically formulates the privacy guarantee through a parameterized epsilon value which defines an inverse privacy budget—the higher the value, the lower the privacy. [115] A study using differential privacy in a GWAS was published by Uhlerop, et al.. [116] Their solution advocates for the reasonable release of minor allele frequencies for both cases and controls in a way that doesn't compromise privacy and permits sharing of chi-squared statistics and p-values for relevant SNPs.

Controlled access

One form of privacy-preserving technology is a suite of services that permit an end-user to sign in at a portal (authentication) and access different federated resources depending on the privileges assigned to that user (authorization). Users who apply for access to controlled data resources generally must demonstrate a legitimate research purpose and appropriate qualifications. The controlled access mechanism is how most institutions protect their privacy-sensitive genetics and genomics data. The NIH, for example, mandates that all the data from the research it funds be made publicly accessible via controlled access. The Database of Genotypes and Phenotypes (dbGaP) is one such collection of NIH data under controlled access. [117] One way to implement the controlled access approach is to channel data requests through a Data Access Committee (DAC) – a group of individuals who serve as key institutional data stewards to evaluate access requests on a case-by-case basis. It is common for a Data Access Compliance Office (DACO) to coordinate the review of data access requests to enable the streamlined and cost-effective administration thereof. [118] Cheah et al suggest that a DAC should not only protect privacy but should also promote data sharing, motivate data producers, and encourage data re-use with transparent, simple, and clear application procedures[119]. Rahimzadeh et al contend that automated decision support (ADS) for data access requests improves the auditability, consistency, and efficiency of the data access process and ultimately yields fairer outcomes for the research community. [120] The Global Alliance for Genomics and Health Data Use Ontology

(DUO) is their implementation of an ADS system for automating the genomics data access process. [121]

Computational abstractions

Hardware-assisted secure computation: In most operating systems, there exists a user identity called a privileged user which has access to all the data on the system, including data on disk, in main memory, and in the processor caches. If this identity is compromised, then any privacy-sensitive data on that system is at risk of being compromised too. This is referred to as a back-door threat. Intel's Software Guard Extensions (SGX) was first introduced in 2015 with the aim of providing a Trusted Execution Environment (TEE) in which applications can protect critical code and data against malicious privileged system code. In SGX, the code is divided into a trusted part (which processes protected data) and an untrusted part (which does not process protected data). Privileged users do not have access to the trusted part of the application when it is running on the SGX processor, thereby eliminating this back-door threat. [122] The SGX chip has been leveraged to securely run genomics analyses on systems to prevent other applications running on the same system from having access to private data. Two examples include leveraging SGX to protect privacy and simultaneously accelerate computational performance in a GWAS study. [123,124]

Physical machines, virtual machines, and containers: A physical machine, also known as a bare-metal machine, is the collection of hardware components (e.g. disk, CPU, memory) and software components (e.g. kernel, applications) dedicated to and managed by a single operating system. By contrast, a virtual machine is an application which runs on a physical machine's operating system, abstracting the physical components to allow multiple operating systems to run concurrently on a single physical machine. Containers are "light-weight" virtual machines that only abstract those elements of an operating system and application stack that must be provided for a given purpose, excluding components of the operating system that are not needed for that purpose. For federation, both virtual machines and container architectures allow for software to be distributed readily and portably among federation members. [32] While there are known security risks running certain types of container software that expose privileged user access, these risks are being reduced via newer container architectures. [125]

Local versus cloud computing: On-premise (local) computing is a collection of servers, storage devices, networking equipment, power supplies, etc., that operate within the boundaries of an organization. Having dedicated infrastructure in-house entails both the capital expense of purchasing the equipment and the operating expense of managing and running it. Cloud computing is infrastructure that organizations can rent. Cloud consumers still incur the expense of renting time, but they do not incur any expense to purchase or manage the equipment. Cloud computing is especially beneficial for small organizations that do not have resources to own and run their own computing infrastructure, or for any organization which wants to focus its human resources on tasks other than the management of computing infrastructure. Moreover, secure cloud platforms are gaining traction in genomics as a mechanism to share controlled access to data that cannot move for privacy and technical reasons. [126–128]

Federated computing trust architectures

In this review, we consider a federation to be a group of one or more organizations, each with its own privacy-sensitive data sets, forming a single network in which those datasets may be shared in a legally compliant, privacy-preserving manner. Apart from the many technical details that are required to deploy such a solution, at the core of the federation is the trust that participants will comply with the policies set forth to protect the privacy of the individuals who have provided their data and the protocols which enforce the integrity of the federation results. We define three trust architectures of federated computing for genomics data: clustered with centralized trust, clustered with distributed trust, and non-clustered autonomous trust (see Figure 9).

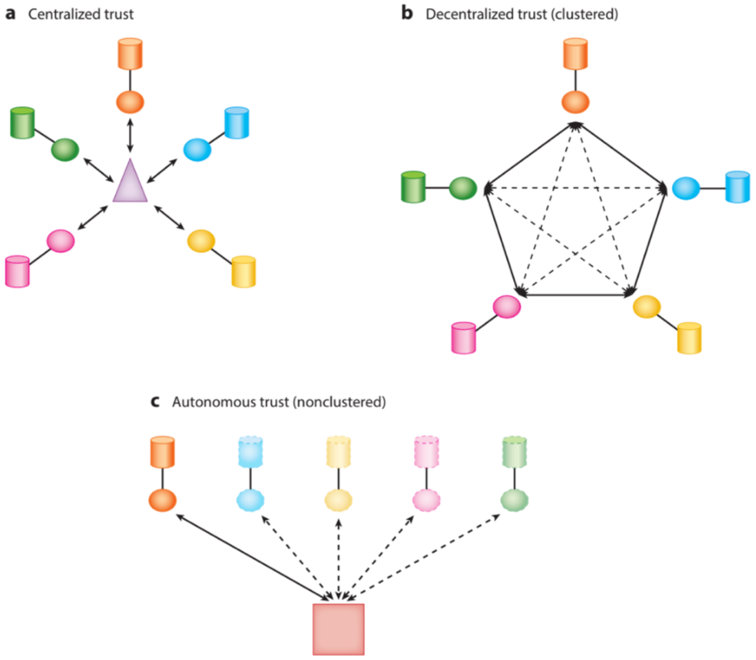
a Centralized trust
b Decentralized trust (clustered)
c Autonomous trust (nonclustered)

Figure 9. Trust architectures. Figure 9a depicts a centralized trust running external to the federation. Figure 9b depicts a decentralized trust in which any central coordination is distributed across the cluster. Figure 9c depicts a non-clustered environment in which trust is established pairwise between the data consumer and each data owner.

Centralized trust

The trusted organization in a centralized trust architecture is an authority that each member of the federation relies upon to establish and maintain overall protocol integrity. In federated learning, for example, updates to the global model are periodically aggregated by the central broker and distributed to each of the federation members. In a controlled access environment, the central authority issues certificates, stores public keys, and provides identity services to authenticate and authorize user access to systems and files. [129] One solution leveraging a centralized trust architecture is the Genomics Research and Innovation Network, which consists of a database of phenotypes and genotypes federated over three participating hospitals, harmonizing the IRB protocols of each participating hospital. [130] This solution requires obtaining the original research participants' consent to use their data in this broader context and to recontact them for further data collection, enrollment in additional studies, and to inform them of potentially medically actionable results.

For GDPR compliance, the creation of a centralized trust raises questions regarding the GDPR role of each organization. It is imperative to categorize each of the organizations acting as network nodes as data controllers, joint data controllers, data processors or non-regulated actors not engaged in the processing of identifiable personal data. For federated analysis, it is optimal to structure a centralized trust as follows. Each node (i.e. organization) that contributes personal data to the federation should be considered a controller of the personal data that it processes. If the identifiable data are processed using third-party hardware or virtual computing resources, the third parties should be considered the data processors. [102] The results that each node contributes to the central analysis node should not contain personal data. Consequently, once the overall federated analysis is synthesized from the organizations acting as network nodes the final output should not contain personal data.

The result of this structure is that each node is engaged in the processing of the personal data at its disposal and must ensure legal compliance for its own personal data processing activities alone. No nodes act as joint controllers. The potential liability of each node is therefore limited to that which results from the analysis of personal data that it processes. Neither the central node nor the recipients of downstream analysis outputs act as data controllers or data processors. [131–133] To achieve this result, each participating node should create a list of the data elements which it processes and determine whether these constitute personal data. Each participating node should also create a list of the data elements that it shares with the central analysis node and confirm that none of these data elements constitute personal data. The central node should confirm that the data elements that it receives do not constitute personal data, alone or in combination with one another. The organizations engaged in the federation should ensure that the final outputs of the analysis are not identifiable. Each node should separately ensure that it does not have at its disposal a “means likely reasonably to be used” of performing the re-identification of the concerned individuals, using the available information. [6,7,96]

Additional measures that can be implemented to further ensure that the information that a node shares with, and receives from, other nodes in the network does not create a risk of re-identification include the following.

Pre-onboarding trust verification mechanisms can be adopted to ensure that nodes participating in a federation can be presumed trustworthy and will not engage in conduct that could create a risk of individual re-identification, for example by ensuring that each node has a bona fide scientific or clinical purpose for engaging in the federation and engages personnel with the necessary technical and scientific training to implement the intended analysis in a manner that reduces the risk of individual re-identification. Second, the use of contracts that bind the institutions participating in the centralized trust to perform their role in compliance with its policies, and to avoid conduct that could lead to individual re-identification. [102] The third critical consideration is ensuring that the final federation outputs do not enable the re-identification of the research participants that contributed data to the analysis. One approach is to release the outputs in a registered access or controlled access database. Another is to add noise or perform other modifications to the data to reduce the risk of individual re-identification.

Decentralized trust

One criticism of the centralized model is that it concentrates power to one single organization. In an inherently distrustful environment (e.g., a federation among industry competitors), this may preclude an organization from joining the federation. Conversely, in an inherently trustful environment, there is no need to select a central authority. In a decentralized trust architecture, there is no central authority. Trust and agreement between members of the federation are arrived at variously by peer-to-peer majority voting, zero knowledge proofs, or some other distributed consensus protocol.

Swarm learning. Warnat-Herresthal et al introduced swarm learning which achieves decentralized trust by exchanging the role of the central federation authority among the federation members. [134] Swarm learning still uses a central server, but that server is elected among the federation members and changes over the lifecycle of model training. It is expected that each of the federation members shall be an aggregation server at some point over the lifecycle of the federation. They leverage blockchain to manage a distributed ledger and smart contract for on-boarding new federation members, electing the federation authority, and for merging model parameters. Warnat-Herresthal et al use their swarm learning approach to train a classifier on transcriptomic data for predicting disease states in COVID-19, tuberculosis, and leukemia. By decentralizing the federation, swarm learning keeps large sensitive data in place, requires no exchange of raw data (encrypted or plaintext), guarantees secure, transparent, and fair onboarding of federation members without a central custodian, allows parameter merging with equal rights for all members, and protects machine learning models from man-in-the-middle attacks.

Incremental learning. Another variation of federated learning that uses decentralized trust is incremental learning. [63] This solution entails a classifier of datasets distributed across multiple organizations. Models are trained at one organization using its local data, after which the model parameters are sent to the next organization which updates the model parameters using its local data. The model is passed through all the organizations participating in the federation and is updated according to their local data. The model may cycle through the organizations for multiple rounds of training until the model converges or a specified number of rounds is reached.

From the perspective of data protection law, similar methods of achieving compliance can be recommended for the decentralized trust as for the centralized trust. The centralized trust is preferred when one of the participating organizations is a logical candidate to securely aggregate local analysis results. In instances in which there is no evident custodial node, a decentralized trust might be beneficial. In a decentralized trust, the following distinctions can be relevant in assessing and mitigating the risk of individual re-identification, and in assigning respective responsibilities among network participants.

The federation members should establish a contractual agreement that defines the commitments, organizational measures, and technological precautions that each node must adopt. This can be challenging to achieve if there is no central node that bears formal organizational responsibilities for data custodianship. That is, often there is no central custodial node that is suitable to bear responsibility for ensuring that each other node respects the contractual commitments applicable to the nodes in the network.

To resolve this challenge, each federation member can bind itself to a multilateral contract between that node and all other participating nodes, establishing the responsibilities of each node. This contract can elaborate the categories of organizations and actors that are authorized to act as network nodes, and the technical and organizational measures that such nodes must implement to ensure that the information shared between nodes remains non-identifiable. [102,132] The measures described should be sufficient to safeguard against re-identification attacks, model poisoning attacks, and other attacks that one or a small number of malicious nodes attempt to perpetrate through their participation in the federation. Because there is no central custodial node that is responsible for performing the verification of compliance on the part of each federation member, it is a best practice to integrate into the multilateral contract a right for each participating node to compel an audit of another specified node for compliance. Alternatively, each node could be subject to independent third-party audits of their compliance with the specified technical, organizational, and contractual terms at pre-specified intervals.

Autonomous trust

In an autonomous trust environment, there is a distinct trust agreement established between the organizations responsible for the stewardship of identifiable personal data, and the downstream organizations that request that analyses be performed on such data. Organizations acting as stewards of identifiable personal data may well be unaware of one another. An example of an autonomous trust federation for processing genomic data is from Casaletto, et al., who developed and shared a container with Biobank Japan to run against a genomic variant dataset from a case-control study of BRCA variants. [135] Due to data protection regulations, the data were not allowed to be shared outside the institution. By running a containerized workflow on the data in situ, they were able to classify genomic variants that were previously unclassified. In a second example, a team of pediatric cancer researchers from the Treehouse platform shared an RNA-Seq analysis container with partner hospitals that were treating pediatric patients with tumors that had proven difficult to treat. While these hospitals were not at liberty to share the RNA-Seq data itself, but they were able to share the gene expression calls estimated by the container; through comparative gene expression analysis against a larger cancer cohort, the research team was able to provide useful new insights for about 70% of these patients. [136] One approach for safeguarding autonomous trusts, to test that such containers produce the agreed-upon output (and do not leak sensitive data), is for the data steward to create a small test dataset, run the container against it, and examine the output.

The autonomous approach may also be used in situations where the data are not necessarily sensitive but rather are too large to transfer. Keeping the data in place and moving analytic code to the data is the central theme of the big data paradigm. The autonomous approach may further be used in situations where the data controller does not have the means to analyze its own data and thus engages a data processor to perform the analyses.

For purposes of compliance with data protection law, each node acting as a data steward would be considered a data controller, while each third-party service provider that provides computational resources to a data steward node would be a processor.

The nodes direct analyses should not be construed as controllers, joint controllers, or data processors, since these nodes do not have access to identifiable personal data and do not determine any “means and purposes” of the personal data processing. To ensure this, contractual agreements should be implemented between each producer node and each user node that provides analysis software, establishing that the local nodes act as controllers and that the requesting users’ role is to receive non-identifiable anonymized data resulting from the local analyses.

The data steward nodes should ensure that none of the outputs shared to analysis recipient nodes constitute identifiable personal data. The data steward nodes should further accept formal responsibility for selecting, implementing, and/or vetting the analyses that the user nodes submit. To ensure that the analysis recipient nodes are not construed as data controllers or data processors, technical mechanisms should be used to limit the queries that the analysis recipient can submit to the data steward nodes. The data steward nodes should have an organizational and/or technical procedure for reviewing and approving the analysis queries that analysis recipient nodes submit prior to their implementation. This helps to ensure that the data steward nodes continue to act as the sole controllers of the concerned personal data, rather than acting as joint controllers in collaboration with the querying analysis recipient nodes. [131,137,138]

Hybrid architectures

Federated solutions for genomics data can mix different trust models in the same architecture; for example, the Canadian Distributed Infrastructure for Genomics (CanDIG) platform. [139] The designers of this federated framework explicitly chose to decentralize authentication and authorization because members of the federation belong to different provinces in Canada. Authentication relies on the identity mechanisms of each of the participating sites. Users log in with their home site credentials rather than with a centralized CanDIG identity. Authorization decisions are made locally at each site, based on the trusted user identity and the nature of the request. In addition to using the decentralized model for authentication and authorization, CanDIG also supports controlled access for registered research users. Controlled access is explicitly granted by data access committees, and researchers with controlled access credentials can access and query datasets. Registered access users sign up and agree to terms of service but with very limited querying ability, and only to those datasets that have opted into such access.

Limitations to adopting federated solutions

Common impediments to broader uptake of federated analysis include the absence of data standards or limited adherence to existing standards, and often irreconcilable interpretations of applicable data protection norms.

A significant rate-limiting factor is that there is not yet clear guidance on best practices for regulatory compliance. National regulators and institutions often interpret data protection norms differently. Furthermore, clinical sites and research institutions frequently share data on a voluntary basis to contribute data to the biomedical data commons and foster its productive downstream use; however, the prospect of such contributions entailing legal liability for the data contributors can deter voluntary data sharing. While best practices are starting to emerge [25,140–142], there is no current standard of practice, so each individual regulator and regulated party determines its own approach to regulatory compliance. This can lead to fragmented data protection compliance emerging in different jurisdictions, and amongst different institutions that produce, use, or share data. Ultimately, what will drive progress is the development of broadly-accepted policy frameworks, promulgated both by regulators and organizations representing regulated parties and civil society. [143]

Other impediments to the adoption of federated solutions involve the complexity of designing, implementing, and deploying federated solutions that are privacy-preserving. In particular, the lack of software development standards and infrastructure deployment best practices for privacy-preserving federated solutions impedes organizations from participating in an inter-organizational federation. Federated analysis requires that the data a priori meet quality and formatting standards, given that the methods developers cannot directly interact with the actual input data. While the type of input data varies somewhat by the method, most methods require some form of genetic variation data and some form of phenotypic data. Genetic variants are commonly represented in HGVS [144] or VCF [145] nomenclatures, which are well-understood but somewhat imprecise. The GA4GH Variant Representation Specification [146] addresses these issues but is not yet widely adopted. For phenotypic or clinical data, the advent of electronic health records has spurred the adoption of the HL7 FHIR standard [147,148], and more recently, the GA4GH Phenopacket standard [149], but where these standards are not yet adopted, mapping unstructured electronic health records to a structured data standard remains a hard problem. [150] Consequently, there is a limited volume of data that meets data standards, with additional data following ad hoc standards or remaining unstructured. Yet this may be a temporary situation. As more software tools emerge that work with data standards, adhering to the standards will ultimately become a cost-saving decision.

Conclusion

The major impediments to sharing genomic data arise from ambiguous regulatory requirements more so than from technological limitations. Federated analysis can in principle overcome these impediments to enable data sharing while respecting data privacy or data sovereignty restrictions. While the exact approaches differ, the principle remains consistent: by keeping the sensitive data under the control of the data controller and sharing analysis software to execute on the data controller's secure system, federated analysis can distill sensitive information down to information that is less sensitive and can therefore be shared more openly, and yet advances knowledge. Nonetheless, the uptake of federated analysis approaches has been hindered by uncertainties around regulatory compliance, of which the GDPR has been the most noteworthy. Few organizations engaged in complex, consortium-level data sharing activities have the appetite to bear significant regulatory risk. These risks can prove considerable for organizations that are categorized as data controllers and data processors, as the interpretive ambiguities inherent in data protection law create a potential for unintentional non-compliance. For organizations categorized as joint controllers, the additional prospect of bearing liability for the activities of other collaborating controllers creates heightened compliance risks. These risks can deter data sharing altogether.

In essence, by ensuring that any data egress from the controller's node is non-identifiable, non-personal data, and that any re-identification of these data is highly unlikely, the practical risks of personal data disclosure, and the related risk of data protection law non-compliance can be minimized. Recent advances in computational data privacy have produced a family of approaches that are robust against many forms of cyber threats, and some even offer a quantifiable level of security; while the computational cost of these methods currently discourages widespread adoption, new hardware developments are making these approaches more tractable. But while technical, organizational, and contractual privacy safeguards can mitigate risk, they cannot eliminate it completely, and the data steward or data controller still bears the largest legal compliance burden. In the future, legislators and regulators must implement both laws and regulatory guidance that diminish both the compliance costs and the prospect of liability for data controllers and data processors that are engaged in prosocial uses of information to ameliorate healthcare and perform research.

Acknowledgements

We wish to thank Robert Cook-Deegan, Angela Page, Tiffany Boughtwood, and Oliver Hofmann for insightful discussions on the scope of this review. Research supported in this publication was supported by the National Institutes of Health: MC was supported by U01CA242954 (NCI), MC and JC were supported by U54HG007990 (BD2K), and JC was supported by R01HG010485 (NHLBI) and U01HL137183 (NHLBI). AB and RM were supported by the CFI Cancer Collaboratory in performing this research.

Chapter 5: Analyzing the relationship between gene expression and phenotype in space-flown mice using a causal inference machine learning ensemble

Authors
James Casaletto[1], Ryan T. Scott[2], Makenna Myrick[3], Graham Mackintosh[4], Hamed Chok[5], Amanda Saravia-Butler[2], Adrienne Hoarfrost[6], Jonathan Galazka[7], Lauren M. Sanders[1], Sylvain V. Costes[7]

[1] Blue Marble Space Institute of Science, NASA Ames
[2] KBR, NASA Ames
[3] University of Florida, Department of Chemistry
[4] Bay Area Environmental Research, NASA Ames
[5] Foresight Science and Technology, Hopkinton, MA
[6] University of George, Department of Marine Science
[7] NASA Ames Research Center, Moffett Field

Contributions:

In this research, I designed and executed all the causal inference experiments, updated the CRISP Python software, organized and led all the meetings and email communications with our collaborators and publisher, and wrote the manuscript.

Abstract

Spaceflight has several detrimental effects on human and rodent health. For example, liver dysfunction is a common phenotype observed in space-flown rodents, and this dysfunction is partially reflected in transcriptomic changes. Studies linking transcriptomics with liver dysfunction rely on tools which exploit correlation, but these tools make no attempt to disambiguate true correlations from spurious ones. In this work, we use a machine learning ensemble of causal inference methods called the Causal Research and Inference Search Platform (CRISP) which was developed to predict causal features of a binary response variable from high-dimensional input. We used CRISP to identify genes truly correlated with a lipid density phenotype using transcriptomic and histological data from the NASA Open Science Data Repositories (OSDR). Our approach identified genes and molecular targets not predicted by previous traditional differential gene expression analyses. These genes are likely to play a pivotal role in the liver

dysfunction observed in space-flown rodents, and this work opens the door to identifying novel countermeasures for space travel.

Introduction

Rodent studies demonstrate that spaceflight negatively impacts liver function [151–153]. Astronaut studies including the seminal NASA Twins Study also reveal a theme of lipid dysregulation [154,155]. Despite these findings, there has been relatively little research studying the impact of microgravity or space radiation on the liver, with more research emphasis on central nervous system effects and carcinogenesis. This is a key knowledge gap considering the disruption of such a critical organ could impact astronaut health and jeopardize the success of future long-term space missions. Identifying the genetic and molecular mechanisms implicated in spaceflight-induced liver dysfunction is required as a first step in precisely mitigating the effects of spaceflight. Traditional statistical methods identify correlations which may or may not be spurious, especially in high-dimensional, high-throughput data analysis [156]. While randomized controlled trials are considered the gold standard for identifying non-spurious, causal relationships between dependent and independent variables [157], such experiments can be very expensive and time consuming, or logistically infeasible, especially in a spaceflight environment where sample sizes are limited. Instead, we turn to new machine learning (ML) approaches to identify genes in transcriptomics data predictive of a lipid metabolic response from spaceflight and ground control rodent liver samples.

Tools which are commonly used to analyze high-dimensional data, and ML algorithms in general, share an intrinsic flaw. They discover those patterns in data which minimize training error, but training data are often flawed by selection bias, label bias, capture bias, and negative set bias [158]. Algorithms which train on biased datasets inherit these data biases. Minimizing training error encourages algorithms to indiscriminately absorb all the correlations found in training data, real or spurious. Spurious correlations resulting from data biases are unrelated to the true underlying signal [159]. Recently, disambiguating true correlations from spurious ones has been studied in the context of causal inference. For this reason, we leverage tools from the causal inference domain to identify genes which are robustly correlated with a phenotype. While such genes are putatively causal, validating true causality is beyond the scope of this research. In this research, we use CRISP – an ensemble machine learning platform developed by the Frontier Development Laboratory (FDL) 2020 Astronaut Health team [160] to enhance biological and medical research with heterogeneous and high-dimensional observational data [161]. The FDL team used CRISP to identify genetic drivers that differentiate two subtypes of colorectal cancer and to implicate operational taxonomic units of the associated microbiome.

The algorithms in the CRISP platform are based on the concept of invariance as a proxy for causal inference. Invariance is a property of a feature which reflects how well a classification algorithm performs using that feature to predict a response invariantly; that is, on data which were generated in different environments, under different circumstances, different conditions, or using different interventions [162,163]. An algorithm based on invariance can identify those features that strongly predict the target label regardless of the background data generating processes that give rise to the dataset. The classic example is a machine learning classifier built to distinguish images of cows from images of camels [164]. A machine learning classifier that is overfit to a particular environment may learn that a cow is an animal that lives in green pastures while a camel is an animal that lives in beige deserts. Given a cow on a sandy beach, this classifier would likely call it a camel (Figure 10). By contrast, a classifier based on invariance would be optimized to ignore the background environment and learn the salient features which truly distinguish a cow from a camel, such as the dimensions of the neck and legs and the shape of the face.

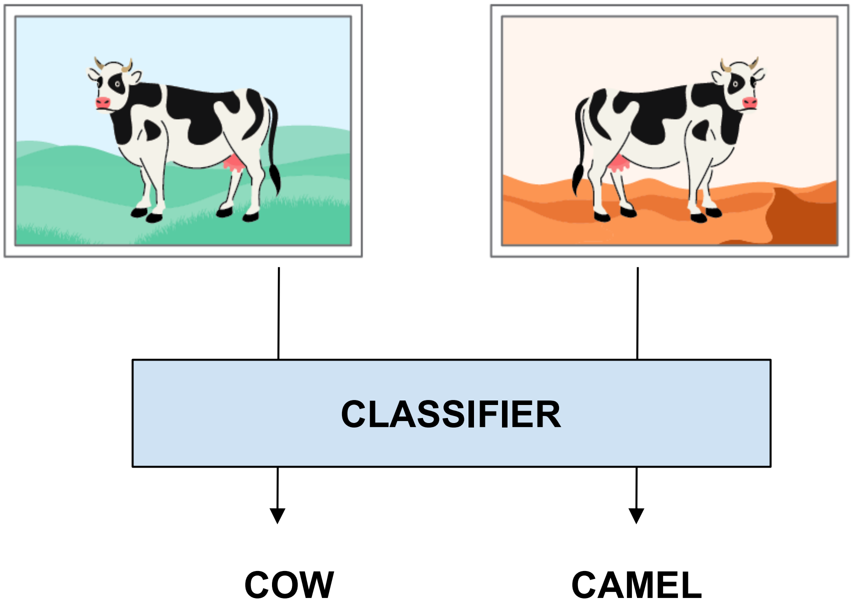

Figure 10. A classifier is presented with images of a cow in two different environments, one of which is its typical green pasture and the other of which is a relatively rare beige desert.

Classification algorithms which exploit invariance promote learning correlations that are stable across training environments, as these are expected to persist on out-of-distribution data (i.e. data generated in environments not seen by the algorithm during training) and therefore be robustly correlated to and more likely causal of, the response variable [165].

To compare our method with traditional differential gene expression tools, we use EdgeR and DESeq2. We also compare our results with those derived from generic machine learning classifiers including random forest and empirical risk minimization. Overall, we find that CRISP identifies a biologically relevant set of genes which are uniquely predictive of a high lipid density response in space-flown mice. Gene set enrichment and pathway analyses reveal that the dysfunctional regulation of the genes identified by CRISP is implicated in the spectrum of diseases caused by non-alcoholic fatty liver disease (NAFLD). The mice in the experiment flight group were only in space for a maximum of 54 days, yet their gene expression profiles were altered significantly enough to manifest markers of NAFLD. Our study provides the first machine learning analysis of gene expression predictive of a disease-related response to spaceflight in the liver.

Materials and methods

The data we used for our experiment include transcriptomic and histology data from the liver tissue of space-flown and ground-control mice.

NASA GeneLab data

The NASA GeneLab repository provides AI-ready datasets allowing rapid deployment of machine learning algorithms for data mining. This is possible because GeneLab is a FAIR database (Findable, Accessible, Interoperable, Reusable) [166,167] with rich metadata providing full context for the data and experiments. This study uses transcriptomic data from four GeneLab datasets: GLDS-47 [168] (version 11), GLDS-48 [169] (version 10), GLDS-137 [170] (version 6), and GLDS-168 [171] (version 10). These datasets were generated from three rodent research (RR) missions: RR-1 CASIS, RR-1 NASA, and RR-3. Two different strains of mice were used: C57 and Balb/C. The RR-1 CASIS experiment was designed to study the effects of microgravity of C57 mice on muscle degeneration due to spaceflight (GLDS-47). The RR-1 NASA mission was designed to validate the experimental hardware and scientific capabilities on the International Space Station (GLDS-48). The RR-3 mission was designed to study countermeasures in Balb/C mice for loss of mass in muscle and bone that have been observed in spaceflight (GLDS-137 and GLDS-168). The GLDS-168 dataset was not based on a separate mission but rather to test the utility of External RNA Control Consortium (ERCC) RNA sequencing controls and therefore constitute technical replicates in our experiments. These rodent research missions were originally designed as randomized controlled experiments, with mice randomly assigned to the groups described in Table 6.

Experimental group	Description
Basal	Housed in standard vivarium cages on Earth, euthanized 1 day after launch.
Vivarium	Housed in standard vivarium cages on Earth, euthanized n days after launch.
Ground	Housed in ISS habitat cages on Earth, euthanized n days after launch.
Flight	Housed in ISS habitat cages on ISS, euthanized n days after launch.

Table 6. The four experimental groups of mice from the GeneLab datasets.

We are re-using these data to explore the relationship between the transcriptomes and a phenotype, constituting the data as observational in our research. Indeed, the causal inference algorithms in the CRISP platform ensemble were designed to run on observational data.

ALSDA liver phenotype data

Liver tissues used for gene expression were also quantified for lipid density using the Oil red O (ORO) staining technique. ORO is a fat soluble, hydrophobic dye that stains lipid molecules red [172]. ORO percent positivity was calculated for each sample from the stained images, providing a scalar value that measures the ORO positivity - higher ORO positivity values directly correspond to higher lipid densities. ORO positivity is the *de facto* histological biomarker for diagnosing the spectrum of disorders in non-alcoholic fatty liver disease (NAFLD) *post mortem*.

This histological data and the ORO percent positivity values are available alongside the transcriptomic data in the NASA OSDR, which integrates the GeneLab omics data with all other spaceflight-relevant data in the Ames Life Sciences Data Archive (ALSDA) [173]. Having all spaceflight-relevant experimental data curated into a FAIR system demonstrates in this work how rapidly one can deploy AI/ML algorithms to gain new knowledge from several datasets, something difficult to achieve with standard systems biology approaches.

Because OSD-168 are comprised of technical replicates, the ORO positivity data are associated with the biological replicates in OSD-47, OSD-48, and OSD-137.

Data preparation

A typical machine learning pipeline includes a data preprocessing step. At the very least, the data must be prepared to satisfy the assumptions and requirements of the algorithms which use the data. In this section, we discuss how we prepared the data prior to running it through the CRISP platform. CRISP requires that the features be real-valued, that the target be binary, and that the environment string be ASCII text, as described in the following sections.

Binarized target

The ORO positivity scalar values in our dataset range from 0.91 to 26.94, but the CRISP platform only permits binary targets (low and high) for classification. We converted the scalar value to a binary value using the median value between flight and ground groups as a binary threshold. We calculated a per-mission threshold as the mean between the flight and non-flight group medians, shown in the box-and-whisker plot of Figure 11 as a horizontal dashed red line.

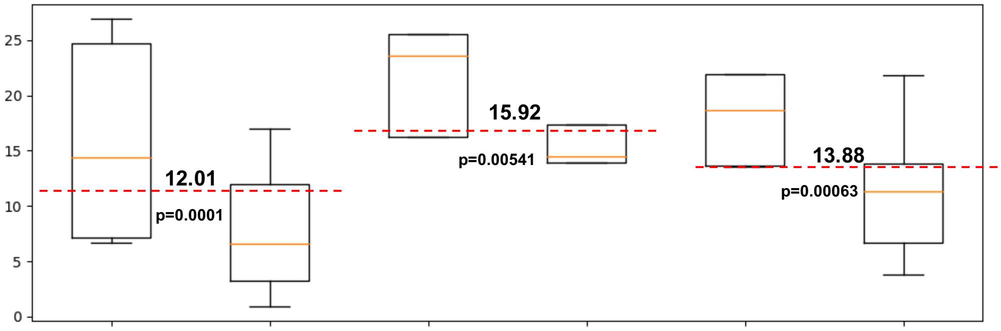

Figure 11. Box-and-whisker plots for ORO values based on mission. The dashed red line is the arithmetic mean between the two group medians, and the p-values show that the differences between the medians are all significant.

Using the thresholds indicated in Figure 11, a sample was assigned a binary target of 0 if its ORO positivity is less than the threshold value and 1 if its ORO positivity is greater than or equal to the threshold value.

Environment string

The environment string is the crux of the invariance approach in the CRISP implementation of causal inference. To force the ensemble of classifiers to find features that truly correlate with the target across environments, as in the example of identifying key features of cow and camel faces in different landscapes, we identify environments within our dataset across which any gene with a true correlation to the target should be invariant. For example, in our study, samples are grouped into environments based on technical, non-biological differences such as how the RNA libraries were prepared or data transformations such as log-scaling. The models are trained using the samples partitioned by environment, and then the models are tested across the other environments to see how well they perform on data from other environments. In this study, we used the following guidelines in our choice of environment string:

1. It shall not leak the target
2. It shall represent known or presumptive interventions or perturbations of the features
3. It should partition the data into subsets that are each sufficiently large for testing
4. It should give the highest accuracy and/or most biologically plausible results

Causal inference algorithms which exploit environment invariance theoretically perform better on out-of-distribution data when the number of environments used to train the model is high. However, if the number of samples in the dataset are small (which is intrinsic to transcriptomic experiments), then the environment string must be selected such that there are enough samples in each partition to test for significance.

Feature transformations

There are many types of transformations, including standardization, normalization, log-scaling, and outlier removal that are commonly performed on data as pre-processing steps in a machine learning pipeline. Some data transformations are more volatile than others. Gonzalez et al [174] and others have shown that while data preprocessing is a necessary step in a machine learning pipeline, there isn’t much agreement about what the best preprocessing technique is. Moreover, a data transformation may change the data so drastically that it destroys some of the underlying signals of interest. This lack of data preprocessing standard exists in transcriptomic data analysis as well [175]. In our research, instead of choosing one pre-processing method, we used several methods, each of which is supported in the literature. Table 7 shows the different types of transformations we leveraged in our preprocessing pipeline, including a literature reference describing its efficacy as a transcriptomic data transformation. We performed each data transformation on the dataset separately, then merged the multiple transformed datasets together. This technique of merging differently transformed data into a single dataset is referred to as *data augmentation* and is a common practice in machine learning [176]. Each transformation was considered a separate environment across which CRISP must find invariant genes. We exploit CRISP's built-in search for invariance across environments and consider each transformation as a perturbation of the data akin to a causal intervention.

Transformation	**Description**	**Reference using transformation for transcriptomic data**
Log scale	Scales values to their log base 2	Quinn et al [27]
Z-score	Scales values to their number of standard deviations from the mean	Zwiener et al [28]
Square root	Scales values to their square roots	Zhang et al [29]
Median of ratios	Scales values to account for sequencing depth, gene length, and outliers	Robinson et al [30]
Centered log ratio	Transforms data to eliminate over-dispersion	Anders et al [31]
Box-Cox	Transforms non-normal data into a normal shape	Sun et al [32]

Table 7. Description of and reference for each of the transformations used in the pre-processing of the transcriptomic data.

Figure 12 shows the original data distribution and the data after having been transformed and plotted as (variance vs mean) coordinates in log scale.

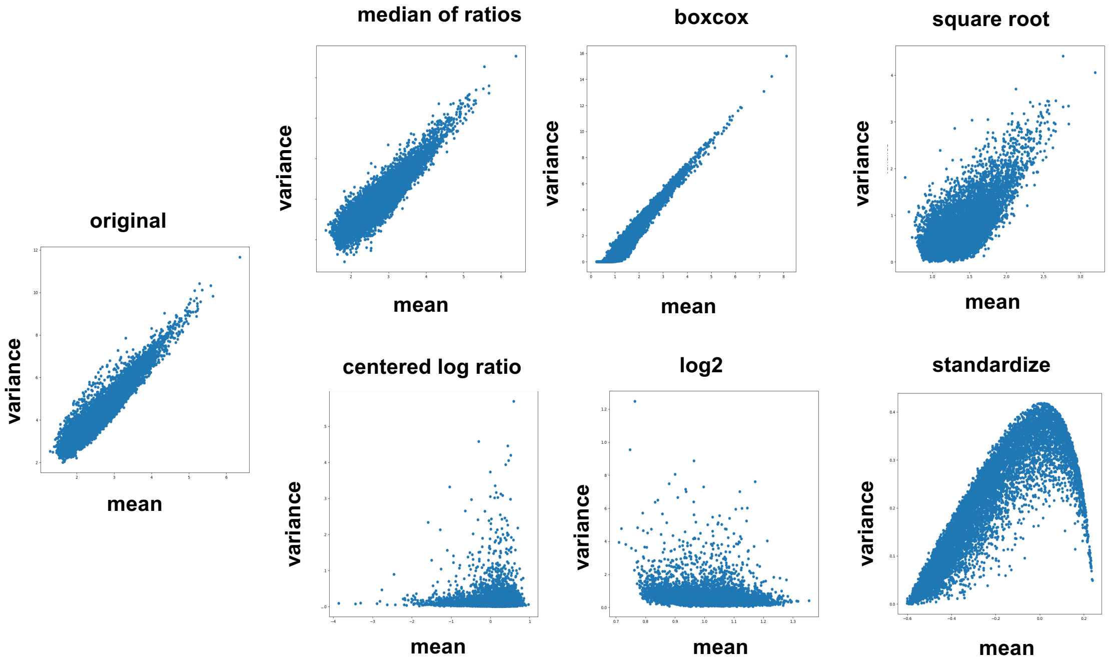

Figure 12. Scatter plots of variance versus mean for different transformations used in preprocessing. The vertical and horizontal axes are shown in log scale.

The mean-variance plot of the differently transformed data reveals that certain transformations change the data significantly while other transformations are relatively mild in effect, compared to the original raw data. It also reveals the fact that the variance of the raw data is proportional to the means of the counts – a condition known as heteroscedasticity. This condition results in non-robust error estimates in linear regression models because the error estimates are less accurate where the variance is higher. In other words, linear regression models built on heteroscedastic data are not uniformly reliable across the distribution of the independent variable. Because CRISP does not use linear regression for any of the models in the ensemble We used multiple transformations which have been demonstrated to specifically address transcriptomic data heteroscedasticity in the literature (Table 7).

In addition to these transformations, we applied some basic filtering on the input to remove transcripts with no ENSEMBL identifiers, don't code for a protein, or have mostly low or zero counts. This filtering produced the final set of 11,854 genes which we subsequently used in all downstream analysis.

Technical batch effects

We examined the dataset for batch effects and found that the type of library preparation of the RNA-seq experiments underlying the transcriptomic data clearly separate the samples, as shown in Figure 13a.

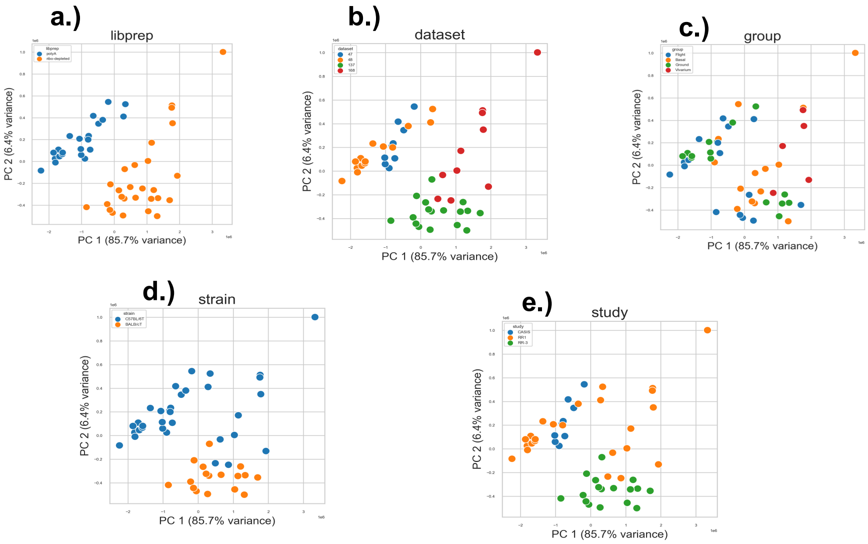

Figure 13. PCA plots of GLDS datasets colorized by the different covariates, including a.) library preparation, b.) dataset name, c.) experimental group, d.) mouse strain, and e.) Rodent Research study.

To account for this batch effect, we include the library preparation in the environment string. We also include the data transformation name in the environment string.

Constructing the environment string

Because the environment is used by CRISP to partition the data into subsets for training and testing, the higher the number of values in the environment, the higher the number of partitions and therefore the fewer samples used for training and testing. We already have a limited number of data points, and having small training and testing sets leads to less reliable results. Therefore, we restricted our choice of environment string to include only known perturbations of the data – i.e. transformation and library preparation – and excluded perturbations of unknown effect such as mission and strain.

sample	Scd1	Apoa1	C3	Apoe	env	Oro Thresh
Mmus_C57-FLT_R1F1_boxcox	151.32	18.21	1.49	4282.92	polyA:boxcox	1
Mmus_C57-FLT_R1F1_clr	0.78	-0.23	0.82	0.94	polyA:clr	1
Mmus_C57-FLT_R1F1_log	18.25	16.98	18.37	13.66	polyA:log	1
Mmus_C57-FLT_R1F1_mor	362.52	139.69	32.83	4470.34	polyA:mor	1
Mmus_C57-FLT_R1F1_zscore	0.08	-0.03	0.01	0.21	polyA:zscore	1
Mmus_C57-FLT_R1F1_sqrt	602.39	373.71	57.95	66.8.63	polyA:sqrt	1

Table 8. One fictitious mouse sample's gene expression data (truncated to 4 genes) after 6 data transformations, including environment string ("env") and ORO threshold ("oro_thresh").

Table 8 shows a snippet of one sample's fictitious input data after having performed the 6 data transformations. The environment string (here, called "env") is a concatenation of the library preparation (here, "polyA") and the transformation name (e.g. "boxcox"). Only the binary ORO threshold (called "oro_thresh") for the sample remains unchanged across the same sample as well as for its respective technical replicates of OSD-168, if they exist.

Running CRISP experiments

With the data in place, we turn now to how we run the CRISP *in silico* experiments. CRISP experiments are configured with several parameters in a JSON configuration file. The `test_val_split` parameter defines how much data to leave out for testing and validation. By default, 10% is dedicated for testing and 10% is dedicated for validation, leaving 80% for training. The `max_features` parameter defines the number of features each model in the ensemble should find as most predictive of the target. The default value is 20, and that's the value we used in our CRISP experiment. The `data_options` define the file location containing the dataset, which columns in that dataset are to be used as predictors, which column is the environment variable, and which column is the target variable. There are several other parameters that may be configured in this JSON configuration file which get consumed by the machine learning algorithms configured in the ensemble.   The JSON configuration we used in our experiment is shown in Supplementary Figure 1.

CRISP ensemble voting

CRISP is an ensemble of machine learning algorithms which perform binary classification. Ensembles are used to combine a set of multiple "weak" learners into a single "strong" learner to minimize training errors [183]. Each model in the ensemble is trained on the dataset to identify the features most predictive of the target. After training, each model selects the features (20 by default) which it found as most predictive of the target. For the linear models -- linear invariant risk minimization (linear IRM) and linear invariant causal prediction (linear ICP) -- the features which are most predictive are those coefficients of the linear model with the highest absolute values. For non-linear models -- non-linear invariant risk minimization (non-linear IRM) and non-linear invariant causal prediction (non-linear ICP) -- the most predictive features are selected through sensitivity analysis.   After each model has selected its top-most predictive features, the ensemble votes to elect a single set of features to present as the final result of the experiment. CRISP attributes the highest weight to the feature that the highest accuracy model gives the largest coefficient. Conversely, the lowest ranking feature from the worst performing model will be attributed the lowest weight.   Furthermore, the higher degree of concordance across the ensemble (i.e. how many models found the feature in their top 20 list), the higher the weighted coefficient of that feature.   In this way, CRISP identifies those features that are most predictive of the target from the highest number of best performing models.

CRISP updates

Accompanying this paper we publish a CRISPv1.1 code release, with the following updates. First, the weights of the features each model finds need to be on the same scale in order to compare them. To this end, we used the `MinMaxScaler()` class from the `scikit-learn`

Python package. Second, we updated the linear IRM code to output continuous feature weights such that the coefficients are now on the unit interval [0, 1]. Third, we changed the default batch size from 128 to 8 due to our small sample size. Fourth and perhaps most importantly, we changed the feature reduction mode of non-linear IRM to use 3 hidden linear layers instead of a single hidden linear layer to further address the issue of heteroscedasticity.

Results

In this section, we show the results of the CRISP experiment. We validate our findings using pathway analysis, gene set enrichment analysis, adverse outcome effect, and a search of relevant literature. We show the results of the gene expression analysis using DESeq2 and EdgeR to compare the CRISP results with commonly used bioinformatic tools.

CRISP results

Each of the models in the ensemble trains on the same data set to identify the features which best predict the target across all environments. Figure 14 depicts the accuracy of each model in the ensemble, and Figure 15 shows the top 20 genes most predictive of lipid density.

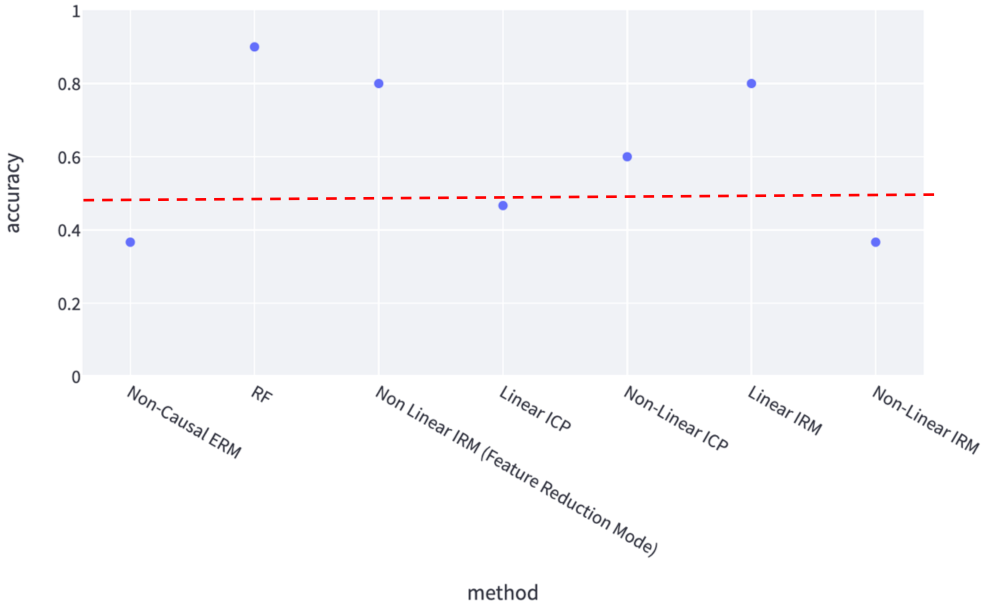

Figure 14. Test accuracies for each of the models in the CRISP ensemble, including a dashed red line representing 50% accuracy (null model performance).

Empirical risk minimization and random forest are not causal predictors and do not participate in CRISP's selection of the features. They are included in the experiment only as a basis of comparison to the causal predictors (non-linear IRM, linear ICP, non-linear ICP, and linear IRM). Figure 14 shows that the random forest (RF) accuracy was the highest (about 90%). We will discuss later in this section that the 20 genes which RF found as most predictive of lipid density are most likely spuriously correlated. Linear IRM has about 80% accuracy and therefore contributes the most to the ensemble results. Conversely, linear ICP performed the worst across the ensemble (about 50% accuracy) and therefore contributes the least to the ensemble results.

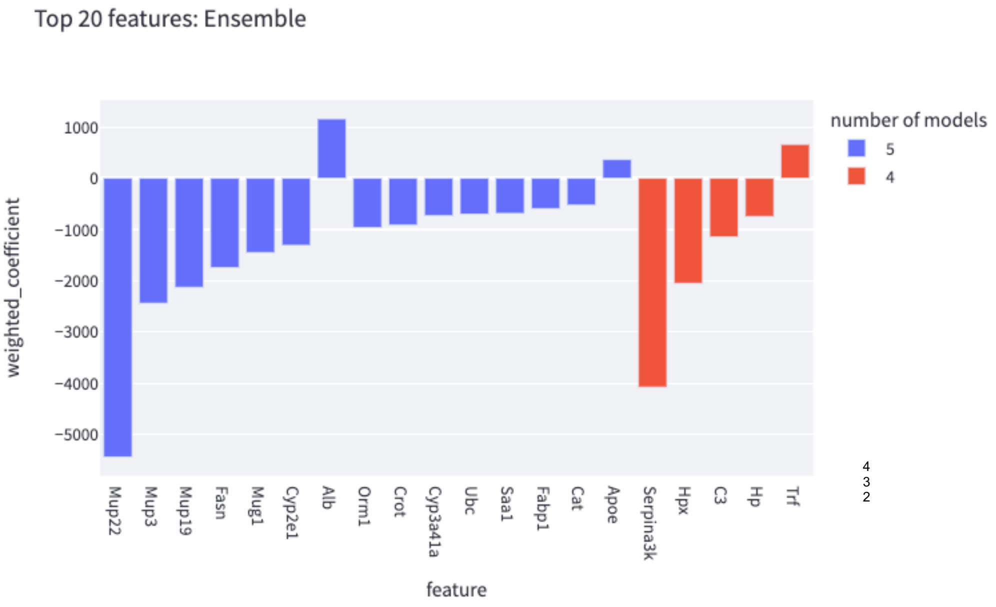

Figure 15. The top 20 genes and their degree of concordance across the ensemble. The direction of the bar indicates whether higher (up) or lower (down) expression impacts lipid density.

The diagram in Figure 15 shows the degree to which each gene across the ensemble is predictive of the lipid density response variable. Based on Figure 15, we see that the *Mup22* gene was found in the top 20 genes most predictive of lipid density of all 5 models. By contrast, the *Trf* gene was found in the top 20 genes most predictive of lipid density in 4 of the models. The biological functions of the 20 CRISP genes are shown in Table 9. The functions were derived from the information provided in the NCBI Gene tool at https://www.ncbi.nlm.nih.gov/gene/.

gene	**name**	**function of protein**
Alb	Albumin	Abundant plasma protein essential for maintaining oncotic pressure that functions as a carrier protein for various molecules such as steriods and fatty acids in blood
Apoe	Apolipoprotein E	Involved in the transport of lipoproteins in the blood
C3	complement component 3	C3 plays a central role in the classical, alternative and lectin activation pathways of the complement system.
Cat	catalase	Enables aminoacylase activity and catalase activity.
Crot	carnitine O octanoyltransferase	Plays a role in lipid metabolism and fatty acid beta-oxidation.
Cyp2e1	cytochrome P450, family 2, subfamily e, polypeptide 1	Enables monooxygenase activity. Implicated in several diseases, including fatty liver disease.
Cyp3a41a	cytochrome P450, family 3, subfamily a, polypeptide 41A	Predicted to enable several functions, including caffeine oxidase activity; iron ion binding activity; and monooxygenase activity.
Fabp1	fatty acid binding protein 1, liver	Predicted to be involved in positive regulation of fatty acid beta-oxidation.
Fasn	fatty acid synthase	Enables fatty acid synthase activity. Involved in lipid biosynthetic process.
Hp	haptoglobin	Plasma glycoprotein that binds free hemoglobin
Hpx	hemopexin	Enables heme binding activity.
Mug1	murinoglobulin 1	Predicted to enable endopeptidase inhibitor activity and protease binding activity.
Mup19	major urinary protein 19	The MUP family proteins bind to, concentrate, and stabilize many volatile scent substances (e.g. pheromones), thereby controlling both pheromone transport in circulation and pheromone release into the air from urine scent marks
Mup22	major urinary protein 22	
Mup3	Major urinary protein 3	

Orm1	orosomucoid 1	Regulates inflammation and metabolism.
Saa1	serum amyloid A 1	Acts upstream of or within cholesterol metabolic process and response to bacterium.
Serpina3k	serine (or cysteine) peptidase inhibitor, clade A, member 3K	Acts upstream of or within response to cytokine and response to peptide hormone.
Trf	transferrin	Predicted to enable iron chaperone activity; iron ion binding activity; and transferrin receptor binding activity.
Ubc	ubiquitin C	Predicted to enable protease binding activity; protein tag; and ubiquitin protein ligase binding activity.

Table 9. Mouse genes identified in the CRISP experiment as robustly predictive of the thresholded lipid density phenotype.

We will discuss later in this section that some of these 20 genes which CRISP found are not only involved in lipid metabolism but also have been implicated in NAFLD.

Validation using pathway analysis

We submitted the 20 genes resulting from our CRISP experiment, and as background all 11,160 genes which were used as features in our CRISP experiment, to the ShinyGO pathway analysis tool (http://bioinformatics.sdstate.edu/go/). This tool finds the Gene Ontology pathways which overlap with the query gene set as compared to the background gene set. Table 10 shows the 4 enriched pathways that ShinyGO identified.

Pathways	**nGenes**	**Pathway genes**	**Fold enrichment**	**Enrichment FDR**
Alcoholic liver disease	4	139	21.2	1.6E-03

Table 10. Enriched Gene Ontology pathways involving the genes from the CRISP experiment.

Using the enrichment false discovery rate metric of significance, the only significant pathway relating to our gene set is alcoholic liver disease.

Validation using gene set enrichment analysis

We used the 20 genes from the CRISP experiment to perform Gene Set Enrichment Analysis (GSEA, http://www.gsea-msigdb.org/gsea/msigdb/annotate.jsp) using the Human Molecular Signatures Database (MSigDB) ontology gene sets (C5) collection. This tool finds those gene ontologies from all ontology gene sets (GO: Gene Ontology and HPO: Human Phenotype

Ontology) that have a significant overlap with the query gene set in the mouse genome. Table 11 shows the top 10 gene sets in the C5 collection that significantly overlap with these 20 mouse genes.

Gene set name	Description	FDR q-value
GOCC BLOOD MICROPARTICLE	A phospholipid microvessicle that is derived from any of several cell types, such as platelets, blood cells, endothelial cells, or others.	1.73E-14
GOCC VESICLE LUMEN	The volume enclosed by the membrane or protein that forms a vesicle.	2.17E-7
GOBP MONOCARBOXYLIC ACID METABOLIC PRO PROCESS	The chemical reactions and pathways involving monocarboxylic acids.	3.06E-8
GOBP TRIGLYCERIDE METABOLIC PROCESS	The chemical reactions and pathways involving triglyceride.	3.06E-8
GOBP LIPID METABOLIC PROCESS	The chemical reactions and pathways involving lipids, compounds soluble in an organic solvent but not, or sparingly, in an aqueous solvent.	3.51E-8
GOBP CELLULAR LIPID METABOLIC PROCESS	The chemical reactions and pathways involving lipids	3.51E-8
GOBP NEUTRAL LIPID METABOLIC PROCESS	The chemical reactions and pathways involving neutral lipids.	3.52E-7
GOBP DEFENSE RESPONSE	Reactions, triggered in response to the presence of a foreign body or occurrence of an injury.	3.52E-7
GOBP SMALL MOLECULE METABOLIC PROCESS	The chemical reactions and pathways involving small molecules	3.84E-7
GOBP ORGANIC ACID METABOLIC PROCESS	The chemical reactions and pathways involving fatty acids.	1.11-7

Table 11. Gene set enrichment analysis using gene ontology gene sets, showing the false discovery rate (FDR) adjusted significance q-value.

Most of these gene sets are directly involved in lipid metabolism.

Validation from scientific literature

Table 12 shows one research article that implicates each of the 20 CRISP genes in NAFLD.

CRISP genes	Research implicating gene expression changes in NAFLD

Alb	Liver-Specific Expression of Transcriptionally Active SREBP-1c is Associated with Fatty Liver and Increased Visceral Fat Mass (PMID: 22363740)
Apoe	Empagliflozin Attenuates Non-Alcoholic Fatty Liver Disease (NAFLD) in High Fat Diet Fed ApoE Mice by Activating Autophagy and Reducing ER Stress and Apoptosis (PMID: **33467546**)
C3	Association between complement C3 and prevalence of fatty liver disease in an adult population: a cross-sectional study from the Tianjin Chronic Low-Grade Systemic Inflammation and Health (TCLSIHealth) cohort study (PMID:25856141)
Cat	Catalase and nonalcoholic fatty liver disease (PMID: 30120555)
Crot	Osthol attenuates hepatic steatosis via decreased triglyceride synthesis, not by insulin resistance (PMID: 25206279)
Cyp2e1	Relevance of CYP2E1 to non-alcoholic fatty liver disease (PMID: 23400921)
Cyp3a41a	CYP3A Activity and Expression in Non-alcoholic Fatty Liver Disease (PMID: 26231377)
Fabp1	The human liver fatty acid binding protein (FABP1) gene is activated by FOXA1 and PPAR-alpha; and repressed by C/EBP-alpha: Implications in FABP1 down-regulation in non-alcoholic fatty liver disease (PMID: 23318274)
Fasn	Expression of fatty acid synthase in non-alcoholic fatty liver disease (PMID:20606731)
Hp	Haptoglobin 2-2 Genotype is Associated with More Advanced Disease in Subjects with Non-Alcoholic Steatohepatitis: A Retrospective Study (PMID: 30820874)
Hpx	Non-acoholic fatty liver diease and livler secretome (PMID: 364441472)
Mug1	Proteome Dynamics Reveals Pro-Inflammatory Remodeling of Plasma Proteome in a Mouse Model of NAFLD (PMID: 27439437)
Mup19	Comparison of hepatic gene expression profiles between three mouse models of Non-alcoholic Fatty Liver Disease (PMID: 35005119)
Mup22	Multi-Omics Characterizes the Effects and Mechanisms of CD1d in Nonalcoholic Fatty Liver Disease Development (PMID: 35465315)
Mup3	Comparison of hepatic gene expression profiles between three mouse models of Non-alcoholic Fatty Liver Disease (PMID: 35005119)
Orm1	Orosomucoid in liver diseases (PMID: 34963738)
Saa1	Hepatocytes derived increased SAA1 promotes intrahepatic platelet aggregation and aggravates liver inflammation in NAFLD (PMID: 33813276)
Serpina3k	PNPLA3 and SERPINA1 Variants Are Associated with Severity of Fatty Liver Disease at First Referral to a Tertiary Center (PMID:33804385)
Trf	Iron depletion attenuates steatosis in a mouse model of non-alcoholic fatty liver disease: Role of iron-dependent pathways (PMID: 33839281)
Ubc	Nutrigenomics of High Fat Diet Induced Obesity in Mice Suggests Relationships between Susceptibility to Fatty Liver Disease and the Proteasome (PMID: 24324835)

Table 12. Research articles implicating the changes in expression of the CRISP-identified genes in the spectrum of NAFLD disorders

Quite notably, every one of the 20 genes that CRISP identified has previously been implicated in the spectrum of diseases associated with NAFLD, suggesting that they may be either potential biomarkers or molecular targets for NAFLD.

Comparing CRISP results to other analyses

In this section, we compare the results of our CRISP experiments with other tools which associate features with targets including DESeq2, random forest, and empirical risk minimization.

Comparing results from CRISP and DESeq2

We used DESeq2 version 1.34.0 to perform differential gene expression analysis (DGEA) on the data on the same set of genes as we used in CRISP to identify which genes are significantly differentially expressed between the high and low ORO groups. DESeq2 performs its own filtering and normalization steps, so we did not transform the data. Because of the distinct batch effect due to library preparation, we added the library preparation covariate to the column data and included it in the DESeq2 design formula along with the ORO threshold value. DESeq2 did not find any genes that were significantly differentially expressed when setting the FDR-adjusted p-value cutoff to 0.05.

Comparing results from CRISP and EdgeR

We used EdgeR version 3.36.0 to perform DGEA on the data on the same set of genes as we used in CRISP to identify which genes are significantly differentially expressed between the high and low ORO groups. EdgeR performs its own filtering and normalization steps, so we did not transform the data. Because of the distinct batch effect due to library preparation, we added the library preparation covariate to the column data and included it in the EdgeR design formula along with the ORO threshold value. Table 13 shows the 6 differentially expressed genes from the EdgeR experiment when setting the FDR-adjusted p-value cutoff to 0.05.

gene	name	function
Atp1a2	ATPase Na+/K+	Integral membrane protein responsible for establishing and

	transporting subunit alpha 2	maintaining the electrochemical gradients of Na and K ions across the plasma membrane
Des	Desmin	This gene encodes a muscle-specific class III intermediate filament. Homopolymers of this protein form a stable intracytoplasmic filamentous network connecting myofibrils to each other and to the plasma membrane and are essential for maintaining the strength and integrity of skeletal, cardiac and smooth muscle fibers.
Hsd3b1	hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 1	Predicted to be involved in several processes, including C21-steroid hormone metabolic process; hippocampus development; and response to corticosterone.
Tpm2	tropomyosin 2	This gene encodes beta-tropomyosin, a member of the actin filament binding protein family, and mainly expressed in slow, type 1 muscle fibers.
Star	steroidogenic acute regulatory protein	Predicted to enable cholesterol binding activity. Acts upstream of or within cellular lipid metabolic process; glucocorticoid metabolic process; and regulation of steroid biosynthetic process.
Akr1b7	aldo-keto reductase family 1, member B7	Predicted to act upstream of or within cellular lipid metabolic process.

Table 13. Gene result set using EdgeR to find differentially expressed genes between low and high ORO groups.

Submitting these genes to the GSEA tool, we see that these genes are implicated in hyperplasia, abnormal myocardium morphology, reduced systolic function, weakness of facial musculature, and areflexia. The ShinyGO tool finds several pathways overlapping these 6 genes, from cardiomyopathy to ovarian steroidogenesis, and various metabolic pathways including galactose, cholesterol, fructose, and lipid metabolism. Given such a wide range of pathways and gene sets that affect so many different organs and pathologies, it is not clear what conclusions could be drawn from the EdgeR gene result set.

Analyzing results from the random forest classifier

CRISP includes non-causal algorithms in its ensemble to compare results with causal algorithms. While random forest lays no claim to identify features causal of a target, the algorithm is one of the best-performing and highly used ML classification algorithms [187]. Similar to its causal counterparts, the random forest algorithm calculates a feature importance metric which can be used to define confidence in the results. Among the 20 genes that random forest identified as predictive of the lipid density response, none of them are involved in the lipid metabolic process according to the NCBI Gene tool. Neither the ShinyGO nor the GSEA enrichment tools found any overlapping pathways or gene ontologies using these 20 genes as input. We conclude that the random forest classifier found spurious correlations between the expression of those 20 genes and the lipid density response.

Analyzing results from empirical risk minimization

Following the same analysis as with random forest, here we analyze the results from empirical risk minimization (ERM). ERM uses the canonical minimization of the sum of the squares of the residuals as its unconditional objective function. It does not partition the data into environments like IRM. As such, it is perhaps the closest comparison to the linear IRM and non-linear IRM results from CRISP. Among the 20 genes that ERM identified as predictive of the lipid density response, none of them are involved in the lipid metabolic process according to NCBI. Neither the ShinyGO nor the GSEA enrichment tools found any overlapping pathways or gene ontologies using these 20 genes as input. We conclude that the ERM classifier found spurious correlations between the expression of those 20 genes and the lipid density response.

Discussion

NAFLD represents a spectrum of metabolic disorders from fatty liver alone to steatohepatitis, steatonecrosis, and nonalcoholic steatohepatitis (NASH). In 2020, it was estimated to have a prevalence of 25.24% globally [188]. NAFLD is considered a “first-world” metabolic disorder; it results from a sedentary lifestyle which increases insulin resistance and promotes lipid accumulation. Hepatocytes swell, primarily with triglycerides, inducing inflammation and later fibrosis. Insulin-resistant adipose tissue with ectopic fat deposits initiate lipotoxicity, the primary driver of hepatocyte injury that ultimately manifests as hepatic, pancreatic, and cardiac dysfunction and disease. Diagnosis of NAFLD *in vivo* is a negative test that excludes other possible causes such as excessive alcohol consumption, drug exposure, and genetic predisposition. Astronauts flying on long space missions are necessarily subject to a sedentary lifestyle in addition to the myriad risks that spaceflight imposes on liver health, yet we still lack a specific biological marker that could precisely characterize the condition.

Using a causal inference machine learning ensemble, we've identified a set of genes which are robustly correlated to lipid density. These genes are consistent with gene set enrichment and pathway analysis tools as well as an abundance of research performed which found each of these genes implicated in NAFLD. On the other hand, machine learning methods such as empirical risk minimization and random forest do not distinguish between spurious and non-spurious correlations. In our experiment, their top 20 genes are not correlated with lipid metabolism, despite the random forest classifier having the highest accuracy (90%) of all the models in predicting lipid density. Our results show that the traditional DESeq2 package did not find any genes significantly correlated to our phenotype. The EdgeR package found 6 genes whose expression was significantly correlated to the lipid density response variable. However, gene set enrichment and pathway analysis tools found a wide variety of conditions and processes associated with these 6 genes, leaving the results difficult to interpret. We attribute these quantitative and qualitative differences in results to the environment invariance modeling approach that the CRISP ML algorithms use.

The NAFLD pathway was the only pathway significantly enriched by *C3*, *Fasn*, *Cyp2e1*, and *Fabp1* genes from the CRISP gene result set. We explored the Bradford-Hill criteria (strength of association, consistency, specificity, plausibility, etc) to further establish the causality of these genes in the CRISP result set. For example, the strength of association between these 4 genes and the NAFLD pathway is 0.0016 (adjusted for multiple tests) and is well below the 0.05 threshold of significance. As noted in Table 8, the literature is consistent in finding these 4 genes implicated in experiments on NAFLD. As opposed to the non-specific pathways enriched by the EdgeR genes, the CRISP genes uniquely enrich the NAFLD pathway. And given that these 4 genes are directly involved in lipid metabolism, it is certainly plausible that they would be causal of excessive lipid density. However, our goal in this research is not to establish causality.

Instead, our approach is to use causal inference methods to identify stable predictors robustly correlated to a response. Causation implies correlation, and the extent to which the genes which CRISP found are causal of the lipid density phenotype, they are certainly robustly correlated.

In this work, we leveraged transcriptomic and histological data from the NASA OSDR. The findings presented here are in keeping with previous work from the GeneLab Analysis Working Groups (AWGs), hundreds of volunteer scientists who set OSDR data standards and perform large meta-analyses with OSDR data. Previous AWG publications identified multi-omic lipid dysregulation in space-flown mouse liver [152] and disruption of mitochondrial function in space-flown mouse and human tissues [155]. These studies and the current study demonstrate the power of data re-use from publicly available, standardized datasets.

In our study, we tried to mitigate the small sample size (a typical issue for spaceflight experiments) by augmenting the data with different transformations. However, having more biological replicates to enrich the underlying gene expression signals would be ideal. Additionally, because of the small sample size, we did not create a validation set for validating the model selection. We therefore recommend future spaceflight experiments to increase the cohort size and make the liver a standard tissue to collect, process, and analyze. Further, we acknowledge one caveat of this study is that the OSD-137 study used Balb/c mouse strain, while the other three studies used the C57BL6 strain. These two mouse strains are known to have differing responses to spaceflight. [152] Thus, a future study would benefit from further investigating the responses of the different mouse strains in larger cohorts. We include in Supplemental Figures 4 and 5 the model accuracies and CRISP ensemble gene set results when using strain in the environment variable rather than library preparation. We observe that while the CRISP experiments have 50% concordance in their gene sets, using library preparation as the environment variable yields more biologically specific and plausible results. The gene ontologies and pathways enriched by the genes that CRISP found when using strain in the environment variable are not uniformly related to lipid metabolism and include pathways such as cardiomyopathy and thyroid hormone synthesis. Because the mean lipid density is significantly higher between flight and non-flight groups, we expected the accuracy to be higher in the CRISP models predicting the lipid density response when using the condition (flight vs non-flight) in the environment variable rather than library preparation. Indeed, the linear IRM method achieves nearly 100% accuracy as shown Supplemental Figure 2. We conclude that this improved accuracy is due to the fact that using the condition in the environment variable leaks the target, and such a model would therefore not perform well on out-of-distribution data.

To further validate our findings, one could explore other data associated with these same rodents including proteomic and methylation data, as well as leverage existing molecular tests of liver function. A future analysis of the gene expression of ground models of NAFLD may help validate the genes identified here by CRISP. Future work could focus on randomized controlled experiments specifically designed to manipulate the function of each putative gene to verify they do, in fact, cause high lipid density in liver tissue. For example, reverse transcriptase PCR or quantitative immunohistochemistry studies on liver samples from spaceflown rodents or from ground-based rodent or cell culture studies could be key validation studies. Adverse outcome pathways (AOPs) related to fatty liver, steatosis, and NAFLD could be explored to identify the molecular pathways responsible for the lipid density phenotype and to gain a more complete purview of the putative dysfunction. Additionally, to further elucidate a link to human manifestation of NAFLD or similar liver dysfunction, it will be important to address the species differences between mouse and human.

In this research, we provide a novel approach to identify robust, non-spurious correlations of gene transcripts associated with liver dysfunction during spaceflight. These gene transcripts constitute potential biomarkers of NAFLD for targeted monitoring or therapeutics development in the future that would otherwise be more time consuming or impossible to identify with traditional statistical or experimental approaches. Given the expense of randomized controlled experiments, having a targeted set of genes putatively causal of the response is invaluable.

Conclusion

Our results demonstrate that using a causal inference framework based on environment invariance has the potential to find features which are robustly correlated with a target. Furthermore, our results show that traditional statistical and machine learning approaches which do not attempt to disambiguate spurious correlations from non-spurious ones may fail to provide meaningful results in high-dimensional, low-sample data sets. While the results of our research robustly correlate gene expression to a lipid dysregulation phenotype in liver tissue, our approach can be generalized to other tissues, phenotypes, and even other -omics data. For NASA to embark on longer and more frequent space missions, understanding the impact of spaceflight on biological function is paramount. To this end, there are many data sets in GeneLab left to explore, and more are continuing to be published.

Acknowledgments

This research was supported by the NASA Human Research Program (HRP), Science Mission Directorate, and Biological and Physical Sciences (BPS) Division. We thank the HRP Radiation Element Translational Radiation Research and Countermeasures team and the GeneLab team for collaboration. Resources supporting this work were provided by the NASA High-End Computing (HEC) Program through the NASA Center for Climate Simulation (NCCS) at Goddard Space Flight Center.

Data Availability

All the mouse liver transcriptomic data and histological lipid data are available at https://osdr.nasa.gov.

Chapter 6: Conclusion

Data federation solves the problems many biomedical scientists encounter in which the data they wish to research is either too sensitive or too large to export. Rather than bringing the data to where the analytical code resides, data federation decentralizes the solution and leaves the data secure in its home institution. Collaborators can then bring the code to the data and analyze it *in situ*. This is the recurring theme throughout my thesis.

In Chapter 2, I collaborated with Dr. Cline from UCSC, scientists at the Riken Institute in Japan, and a breast cancer gene variant curation expert panel in New Zealand. I containerized co-occurring variant logic to analyze *BRCA* variation stored at BIOBANK Japan. Our collaborators at the Riken Institute downloaded our container and ran it against the BIOBANK data to generate a list of variants of uncertain significance which co-occur with known pathogenic variants. Our efforts demonstrated that aggregated results from privacy-sensitive, patient-level data can be used to classify previously unclassified variants. We also ran this container on the Seven Bridges platform using TCGA data, on the Tera platform using dbGaP data, and on local servers using variant data from cardiomyopathy studies at Johns Hopkins, thereby demonstrating the generalizability of our approach. There are many other biobanks throughout the world with variant data from different genes involved in cancer and other diseases to explore. We published our research in Cell Genomics.

In Chapter 3, I collaborated with Dr. Cline from UCSC and Prof. Brian Shirts from University of Washington to ask the question: how long would it take and with what probability would variants of uncertain significance get classified if a number of variously sized genomic sequencing centers responsibly shared their clinical data? I modeled the accumulation of clinical evidence from these sequencing centers sampling from a binomial distribution and used Tavtigian's Bayesian approach for variant classification to simulate classifying variants using this evidence. We found that sharing clinical data as opposed to sharing classifications leads to faster and more likely variant classifications for rare variants (i.e. 1 per 100,000 people). For very rare variants (i.e. 1

per 1,000,000 people), clinical data alone is not sufficient, even after 20 years of collecting and sharing it among 3 large sequencing centers. This modeling approach not only quantitatively demonstrates the value of clinical data sharing, it also provides a usable tool for geneticists to predict how long it will take for a VUS to get classified. My co-authors plan to use this model on real, un-simulated data from large sequencing centers including Ambry Genetics and Invitae for making such predictions. We published our research in the Journal of American Medical Informatics Association.

In Chapter 4, I collaborated with Dr. Cline from UCSC and genomic legal privacy scholars from McGill University to ask the question: how well do federated approaches to genomic data analysis satisfy strict privacy regulations such as those in the GDPR? We discussed how various technological approaches such as homomorphic encryption, secure multiparty computation, differential privacy, and federated computing may both provide sufficient data for biomedical research and preserve the privacy of those who contribute their data. This paper can serve as a guide for technological and legal experts who are planning to have multiple institutions collaborate to build models or perform other analyses which generate aggregated results from privacy-sensitive, patient-level data. Our review will be published in the August 2023 edition of the Annual Review of Genomics and Human Genetics.

In Chapter 5, I collaborated with scientists from NASA to ask the question: which genes are most robustly correlated to the high lipid density observed in space-flown mice? Previous NASA rodent research missions observed that the livers in space-flown mice were visibly larger than their ground control counterparts. I leveraged a causal inference machine learning ensemble to analyze gene expression data associated with a lipid density phenotype. The ensemble found 20 genes, each of which is directly involved in lipid metabolism, that, taken together, significantly enrich the non-alcoholic fatty liver disease pathway from the KEGG database. I also ran traditional tools (DESeq2 and EdgeR) and standard machine learning classifiers such as ERM and random forest as a basis of comparison. The traditional tools either didn’t find any genes correlated to the lipid density phenotype (DESeq2), or they found several genes which do not have anything to do with lipid metabolism (ERM and random forest). EdgeR found 6 genes significantly differentially expressed between flight and ground control mice, but those genes enrich a wide variety of KEGG pathways, including cardiomyopathy, cholesterol, and fructose metabolism. This ensemble platform has been enabled for federated learning, and we have already developed and tested the communication protocol to run federated experiments between Earth and the International Space Station to determine the effect of spaceflight on female reproductive health. Our research is in review at Nature Portfolio Journal (Microgravity) at the time of this writing.

I look forward to working at NASA and continuing my research into federated methods for building models and performing other analyses. NASA's Artemis mission is currently underway to build a lunar colony with the longer-range vision of sending humans to Mars. It is critical to the success of these missions that we understand and mitigate the negative effects of spaceflight on the health of living organisms.

Bibliography

1. Jones, M., Kathryn, R. A. A. & Cook-Deegan, R. The Bermuda Triangle: The Pragmatics, Policies, and Principles for Data Sharing in the History of the Human Genome Project. *Journal of the History of Biology* **51**, 693–805 (2018).
2. How Diplomacy Helped to End the Race to Sequence the Human Genome. *Nature* **582**, 460–460 (2020).
3. Powell, K. The broken promise that undermines human genome research. *Nature* **590**, 198–201 (2021).
4. Bakshi, A. M. Gene Patents at the Supreme Court: Association for Molecular Pathology v. Myriad Genetics. *Journal of Law and the Biosciences* **1**, 183–89 (2014).
5. Evans, B. J. HIPAA's Individual Right of Access to Genomic Data: Reconciling Safety and Civil Rights. *The American Journal of Human Genetics* **102**, 5–10 (2018).
6. v, B. *Bundesrepublik Deutschland, Case C-582/14. ECLI:EU:C:2016:779*. (Court of Justice of the European Union, 2016).
7. Borgesius, F. Z. The Breyer case of the court of justice of the European Union: IP addresses and the personal data definition. *Eur. Data Prot. L. Rev* **3**, (2017).
8. Reardon, J. *et al.* Hallam Stevens, and The Genomic Open workshop group. *GigaScience* **5**, (2016).
9. Scheibner, J. *et al.* Revolutionizing Medical Data Sharing Using Advanced Privacy-Enhancing Technologies: Technical, Legal, and Ethical Synthesis. *Journal of Medical Internet Research* **23**, (2021).
10. Byrd, J. B., Greene, A. C., Prasad, D. V., Jiang, X. & Greene, C. S. Responsible, Practical Genomic Data Sharing That Accelerates Research. *Nature Reviews Genetics* **21**, 615–29 (2020).
11. Rieke, N. *et al.* The Future of Digital Health with Federated Learning. *Npj Digital Medicine* **3**, 119 (2020).

12. Xu, J. *et al.* Federated Learning for Healthcare Informatics. *Journal of Healthcare Informatics Research* **5**, 1–19 (2021).

13. Enderling, H. & Wolkenhauer, O. Are All Models Wrong? *Computational and Systems Oncology* **1**, (2021).

14. N.A.S.A. Rodent Gene Expression Data Added to Data Repository”. (2021).

15. Saelens, W., Cannoodt, R. & Saeys, Y. A Comprehensive Evaluation of Module Detection Methods for Gene Expression Data. *Nature Communications* **9**, 1090 (2018).

16. Collins, J. M. & Isaacs, C. Management of Breast Cancer Risk in BRCA1/2 Mutation Carriers Who Are Unaffected with Cancer. *The Breast Journal* **26**, 1520–27 (2020).

17. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research* **46**, 1062–1067 (2018).

18. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

19. Kurian, A. W. BRCA1 and BRCA2 mutations across race and ethnicity: distribution and clinical implications. *Current Opinion in Obstetrics & Gynecology* **22**, 72–78 (2010).

20. Landry, L. G., Ali, N., Williams, D. R., Rehm, H. L. & Bonham, V. L. Lack Of Diversity In Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice. *Health Affairs* **37**, 780–785 (2018).

21. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).

22. Harris, T. L. & Wyndham, J. M. Data Rights and Responsibilities: A Human Rights Perspective on Data Sharing. *Journal of Empirical Research on Human Research Ethics* **10**, 334–337 (2015).

23. Siu, L. L. *et al.* Facilitating a culture of responsible and effective sharing of cancer genome data. *Nat Med* **22**, 464–471 (2016).

24. Directors, A. C. M. G. B. Laboratory and clinical genomic data sharing is crucial to improving genetic health care: a position statement of the American College of Medical Genetics and Genomics. *Genet Med* **19**, 721–722 (2017).

25. Wright, C. F. *et al.* Genomic variant sharing: a position statement. *Wellcome Open Res* **4**, 22 (2019).

26. Li, M. M. *et al.* Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer. *The Journal of Molecular Diagnostics* **19**, 4–23 (2017).

27. Suver, C., Thorogood, A., Doerr, M., Wilbanks, J. & Knoppers, B. Bringing Code to Data: Do Not Forget Governance. *J Med Internet Res* **22**, 18087 (2020).

28. Schulz, W., Durant, T., Siddon, A. & Torres, R. Use of application containers and workflows for genomic data analysis. *J Pathol Inform* **7**, 53 (2016).

29. Matthias, K. & Kane, S. P. *Docker: up and running*. (O'Reilly Media, Inc, 2015).

30. Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: Scientific containers for mobility of compute. *PLoS ONE* **12**, 0177459 (2017).

31. Toomey, D. *Jupyter for Data Science: exploratory analysis, statistical modeling, machine learning, and data visualization with Jupyter*. (Packt, 2017).

32. O'Connor, B. D., Yuen, D., Chung, V., Duncan, A. G. & Liu, X. K. The Dockstore: enabling modular, community-focused sharing of Docker-based genomics tools and workflows. vol. 6 (2017).

33. Momozawa, Y. *et al.* Germline pathogenic variants of 11 breast cancer genes in 7,051 Japanese patients and 11,241 controls. *Nat Commun* **9**, 4083 (2018).

34. Robinson, P. N. *et al.* The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *The American Journal of Human Genetics* **83**, 610–615 (2008).

35. Feng, B.-J. PERCH: A Unified Framework for Disease Gene Prioritization: HUMAN MUTATION. *Human Mutation* **38**, 243–251 (2017).

36. Harrison, S. M., Biesecker, L. G. & Rehm, H. L. Overview of Specifications to the ACMG/AMP Variant Interpretation Guidelines. *Current Protocols in Human Genetics* **103**, (2019).

37. Tian, Y. *et al.* REVEL and BayesDel Outperform Other in Silico Meta-Predictors for Clinical Variant Classification. *Scientific Reports* **9**, 12752 (2019).

38. Auerbach, A. D. Fanconi anemia and its diagnosis. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **668**, 4–10 (2009).

39. Tavtigian, S. V. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *Journal of Medical Genetics* **43**, 295–305 (2005).

40. Couch, F. J., Nathanson, K. L. & Offit, K. Two Decades After *BRCA:* Setting Paradigms in Personalized Cancer Care and Prevention. *Science* **343**, 1466–1470 (2014).

41. Wexler, R. K., Elton, T., Pleister, A. & Feldman, D. Cardiomyopathy: an overview. *Am Fam Physician* **79**, 778–784 (2009).

42. Berg, J. S. Exploring the importance of case-level clinical information for variant interpretation. *Genetics in Medicine* **19**, 3–5 (2017).

43. Harrison, S. M. *et al.* Clinical laboratories collaborate to resolve differences in variant interpretations submitted to ClinVar. *Genetics in Medicine* **19**, 1096–1104 (2017).

44. Cline, M. S. *et al.* BRCA Challenge: BRCA Exchange as a global resource for variants in BRCA1 and BRCA2. *PLoS Genet* **14**, e1007752 (2018).

45. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17**, 405–424 (2015).

46. Tavtigian, S. V. *et al.* Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genetics in Medicine* **20**, 1054–1060 (2018).

47. Tsai, G. J. *et al.* Outcomes of 92 patient-driven family studies for reclassification of variants of uncertain significance. *Genetics in Medicine* **21**, 1435–1442 (2019).

48. Balmaña, J. *et al.* Conflicting Interpretation of Genetic Variants and Cancer Risk by Commercial Laboratories as Assessed by the Prospective Registry of Multiplex Testing. *JCO* **34**, 4071–4078 (2016).

49. Hampel, H. *et al.* Assessment of Tumor Sequencing as a Replacement for Lynch Syndrome Screening and Current Molecular Tests for Patients With Colorectal Cancer. *JAMA Oncol* **4**, 806 (2018).

50. Susswein, L. R. *et al.* Pathogenic and likely pathogenic variant prevalence among the first 10,000 patients referred for next-generation cancer panel testing. *Genetics in Medicine* **18**, 823–832 (2016).

51. Mannan, A. U. *et al.* Detection of high frequency of mutations in a breast and/or ovarian cancer cohort: implications of embracing a multi-gene panel in molecular diagnosis in India. *J Hum Genet* **61**, 515–522 (2016).

52. Santos, C. *et al.* Pathogenicity Evaluation of BRCA1 and BRCA2 Unclassified Variants Identified in Portuguese Breast/Ovarian Cancer Families. *The Journal of Molecular Diagnostics* **16**, 324–334 (2014).

53. Genetic Testing For Clinicians | Ambry Genetics. https://www.ambrygen.com/providers.

54. Real world data insights | Invitae. https://www.invitae.com/en/partners/data-insights.

55. Ranola, J. M. O., Tsai, G. J. & Shirts, B. H. Exploring the effect of ascertainment bias on genetic studies that use clinical pedigrees. *Eur J Hum Genet* **27**, 1800–1807 (2019).

56. Kahn, J., Linial, N. & Samorodnitsky, A. Inclusion-exclusion: Exact and approximate. *Combinatorica* **16**, 465–477 (1996).

57. Manrai, A. K. *et al.* Genetic Misdiagnoses and the Potential for Health Disparities. *N Engl J Med* **375**, 655–665 (2016).

58. Shirts, B. H., Jacobson, A., Jarvik, G. P. & Browning, B. L. Large numbers of individuals are required to classify and define risk for rare variants in known cancer risk genes. *Genetics in Medicine* **16**, 529–534 (2014).

59. Starita, L. M. *et al.* Variant Interpretation: Functional Assays to the Rescue. *The American Journal of Human Genetics* **101**, 315–325 (2017).

60. Dwork, C. Differential Privacy: A Survey of Results. in *Theory and Applications of Models of Computation* (eds. Agrawal, M., Du, D., Duan, Z. & Li, A.) vol. 4978 1–19 (Springer Berlin Heidelberg, 2008).

61. Acar, A., Aksu, H., Uluagac, A. S. & Conti, M. A Survey on Homomorphic Encryption Schemes: Theory and Implementation. *ACM Comput. Surv.* **51**, 1–35 (2019).

62. Ozercan, H. I., Ileri, A. M., Ayday, E. & Alkan, C. Realizing the potential of blockchain technologies in genomics. *Genome Res.* **28**, 1255–1263 (2018).

63. Sheller, M. J. *et al.* Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci Rep* **10**, 12598 (2020).

64. Kessler, C. Genomics and Precision Medicine: Implications for Critical Care. *AACN Adv. Crit. Care* **29**, 28–35 (2018).

65. Hegde, M. *et al.* Development and Validation of Clinical Whole-Exome and Whole-Genome Sequencing for Detection of Germline Variants in Inherited Disease. *Arch. Pathol. Lab. Med* **141**, 798–805 (2017).

66. Stranneheim, H. & Wedell, A. Exome and genome sequencing: a revolution for the discovery and diagnosis of monogenic disorders. *J. Intern. Med* **279**, 3–15 (2016).

67. Wainschtein, P., Jain, D. & Zheng, Z. TOPMed Anthropometry Working Group, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, et al. *Nat. Genet* **54**, 263–73 (2022).

68. Hermel, D. J., McKinnon, W. C., Wood, M. E. & Greenblatt, M. S. Multi-gene panel testing for hereditary cancer susceptibility in a rural Familial Cancer Program. *Fam. Cancer* **16**, 159–66 (2017).

69. Cragun, D. *et al.* Panel-based testing for inherited colorectal cancer: a descriptive study of clinical testing performed by a US laboratory. *Clin. Genet* **86**, 510–20 (2014).

70. Selkirk, C. G., Vogel, K. J., Newlin, A. C., Weissman, S. M. & Weiss, S. M. Cancer genetic testing panels for inherited cancer susceptibility: the clinical experience of a large adult genetics practice. *Fam. Cancer* **13**, 527–36 (2014).

71. Roberts, M. E., Susswein, L. R., Janice Cheng, W., Carter, N. J. & Carter, A. C. Ancestry-specific hereditary cancer panel yields: Moving toward more personalized risk assessment. *J. Genet. Couns* **29**, 598–606 (2020).

72. Caswell-Jin, J. L., Gupta, T., Hall, E., Petrovchich, I. M. & Mills, M. A. Racial/ethnic differences in multiple-gene sequencing results for hereditary cancer risk. *Genet. Med* **20**, 234–39 (2018).

73. Bernier, A., Molnár-Gábor, F. & Knoppers, B. M. The international data governance landscape. *J Law Biosci* **9**, (2022).

74. Mahesh, K. P. *Genomic Sovereignty in South Africa: Ethico-Legal issues*. (Faculty of Health Sciences, University of the Witwatersrand, 2014).

75. Schwartz-Marín, E. & Méndez, A. A. The law of genomic sovereignty and the protection of "Mexican genetic patrimony. *Med. Law* **31**, 283–94 (2012).

76. Tsosie, K. S., Yracheta, J. M., Kolopenuk, J. A. & Geary, J. We Have "Gifted" Enough: Indigenous Genomic Data Sovereignty in Precision Medicine. *Am. J. Bioeth* **21**, 72–75 (2021).

77. Hudson, M., Garrison, N. A., Sterling, R., Caron, N. R. & Fox, K. Rights, interests and expectations: Indigenous perspectives on unrestricted access to genomic data. *Nat. Rev. Genet* **21**, 377–84 (2020).

78. Thorogood, A., Rehm, H. L., Goodhand, P., Page, A. J. H. & Joly, Y. International federation of genomic medicine databases using GA4GH standards. *Cell Genom* **1**, (2021).

79. Heeney, C., Hawkins, N., de Vries, J., Boddington, P. & Kaye, J. Assessing the Privacy Risks of Data Sharing in Genomics. *Public Health Genomics* **14**, 17–25 (2011).

80. Wuyts, K., Sion, L. & Joosen, W. LINDDUN GO: A Lightweight Approach to Privacy Threat Modeling. in *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)* 302–309 (IEEE, 2020). doi:10.1109/EuroSPW51379.2020.00047.

81. Gholami, A. *et al.* Privacy Threat Modeling for Emerging BiobankClouds. *Procedia Computer Science* **37**, 489–496 (2014).

82. Halfond, W. G., Viegas, J., Orso, A., & Others. A classification of SQL-injection attacks and countermeasures. in *Proceedings of the IEEE international symposium on secure software engineering* vol. 1 13–15 (IEEE, 2006).

83. Sweeney, L., Abu, A. & Winn, J. Identifying Participants in the Personal Genome Project by Name (A Re-identification Experiment. (2013).

84. Senarath, A. & Arachchilage, N. A. G. A data minimization model for embedding privacy into software systems. *Computers & Security* **87**, 101605 (2019).

85. Shokri, R., Stronati, M., Song, C. & Shmatikov, V. Membership Inference Attacks against Machine Learning Models. Preprint at http://arxiv.org/abs/1610.05820 (2017).

86. Homer, N. *et al.* Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. *PLoS Genet* **4**, e1000167 (2008).

87. Shringarpure, S. S. & Bustamante, C. D. Privacy Risks from Genomic Data-Sharing Beacons. *Am. J. Hum. Genet* **97**, 631–46 (2015).

88. Ateniese, G. *et al.* Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers. *International Journal of Security and Networks* **10**, 137–50 (2015).

89. Ganju, K., Wang, Q., Yang, W., Gunter, C. A. & Borisov, N. Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations. in *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security* 619–33 (ACM, 2018).

90. Humbert, M., Ayday, E., Hubaux, J.-P. & Telenti, A. Addressing the concerns of the lacks family: quantification of kin genomic privacy. in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security - CCS '13* 1141–1152 (ACM Press, 2013). doi:10.1145/2508859.2516707.

91. Jere, M. S., Farnan, T. & Koushanfar, F. A Taxonomy of Attacks on Federated Learning. *IEEE Secur. Privacy* **19**, 20–28 (2021).

92. Kuo, T.-T. & Pham, A. Detecting model misconducts in decentralized healthcare federated learning. *International Journal of Medical Informatics* **158**, 104658 (2022).

93. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–64 (2016).

94. Commission, E. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation. (2016).

95. Finck, M. & Pallas, F. They who must not be identified—distinguishing personal from non-personal data under the GDPR. *International Data Privacy Law* **10**, 11–36 (2020).

96. Purtova, N. The law of everything. Broad concept of personal data and future of EU data protection law. *Law, Innovation and Technology* **10**, 40–81 (2018).

97. Beskow, L. M., Hammack-Aviran, C. M. & Brelsford, K. M. Thought Leader Comparisons of Risks in Precision Medicine Research. *Ethics Hum Res* **42**, 35–40 (2020).

98. Gourd, E. GDPR obstructs cancer research data sharing. *Lancet Oncol* **22**, (2021).

99. McLennan, S., Celi, L. A. & Buyx, A. COVID-19: Putting the General Data Protection Regulation to the Test. *JMIR Public Health Surveill* **6**, (2020).

100. Vlahou, A., Hallinan, D., Apweiler, R., Argiles, A. & Beige, J. Data Sharing Under the General Data Protection Regulation. *Time to Harmonize Law and Research Ethics? Hypertension* **77**, 1029–35 (2021).

101. Mascalzoni, D., Bentzen, H. B., Budin-Ljøsne, I., Bygrave, L. A. & Bell, J. Are Requirements to Deposit Data in Research Repositories Compatible With the European Union's General Data Protection Regulation? Ann. *Intern. Med* **170**, 332–34 (2019).

102. Hintze, M. Data Controllers, Data Processors, and the Growing Use of Connected Products in the Enterprise: Managing Risks, Understanding Benefits, and Complying with the GDPR. *SSRN Journal* (2018) doi:10.2139/ssrn.3192721.

103. Lamport, L. & Lynch, N. Distributed Computing: Models and Methods. in *Formal Models and Semantics* 1157–1199 (Elsevier, 1990). doi:10.1016/B978-0-444-88074-1.50023-8.

104. Chaterji, S. *et al.* Federation in genomics pipelines: techniques and challenges. *Briefings in Bioinformatics* **20**, 235–244 (2019).

105. Sheth, A. P. & Larson, J. A. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Comput. Surv.* **22**, 183–236 (1990).

106. Clough, E. & Barrett, T. The Gene Expression Omnibus Database. in *Statistical Genomics* (eds. Mathé, E. & Davis, S.) vol. 1418 93–110 (Springer New York, 2016).

107. Wang, D., Shi, S., Zhu, Y. & Han, Z. Federated Analytics: Opportunities and Challenges. *IEEE Network* **36**, 151–158 (2022).

108. Yao, A. C. Protocols for secure computations. in *23rd annual symposium on foundations of computer science* 160–64 (IEEE, 1982).

109. Micali, S., Goldreich, O. & Wigderson, A. How to play any mental game. in *Proceedings of the Nineteenth ACM Symp. on Theory of Computing, STOC* 218–29 (ACM, 1987).

110. Goldwasser, S., Micali, S. & Rackoff, C. The knowledge complexity of interactive proof-systems. in *Providing Sound Foundations for Cryptography: On the Work of Shafi Goldwasser and Silvio Micali* 203–25 (2019).

111. Constable, S. D., Tang, Y., Wang, S., Jiang, X. & Chapin, S. Privacy-preserving GWAS analysis on federated genomic datasets. *BMC Med Inform Decis Mak* **15**, S2 (2015).

112. Cho, H., Wu, D. J. & Berger, B. Secure genome-wide association analysis using multiparty computation. *Nat Biotechnol* **36**, 547–551 (2018).

113. Gentry, C. *A fully homomorphic encryption scheme*. (Stanford University, 2009).

114. Blatt, M., Gusev, A., Polyakov, Y. & Goldwasser, S. Secure large-scale genome-wide association studies using homomorphic encryption. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 11608–11613 (2020).

115. Dwork, C. & Roth, A. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* **9**, 211–407 (2014).

116. Uhlerop, C., Slavković, A. & Fienberg, S. E. Privacy-Preserving Data Sharing for Genome-Wide Association Studies. *J Priv Confid* **5**, 137–166 (2013).

117. Mailman, M. D., Feolo, M., Jin, Y., Kimura, M. & Tryka, K. The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet* **39**, 1181–86 (2007).

118. Joly, Y., Dove, E. S., Knoppers, B. M., Bobrow, M. & Chalmers, D. Data sharing in the post-genomic world: the experience of the International Cancer Genome Consortium (ICGC) Data Access Compliance Office (DACO. *PLoS Comput. Biol* **8**, (2012).

119. Cheah, P. Y. & Piasecki, J. Data Access Committees. *BMC Med. Ethics* **21**, (2020).

120. Rahimzadeh, V., Lawson, J., Rushton, G. & Dove, E. S. Leveraging Algorithms to Improve Decision-Making Workflows for Genomic Data Access and Management. *Biopreserv. Biobank* **20**, 429–35 (2022).

121. Lawson, J., Cabili, M. N., Kerry, G., Boughtwood, T. & Thorogood, A. The Data Use Ontology to streamline responsible access to human biomedical datasets. *Cell Genom* **1**, (2021).

122. Costan, V. & Devadas, S. *Intel SGX Explained*. (Cryptology ePrint Archive, 2016).

123. Carpov, S. & Tortech, T. Secure top most significant genome variants search: iDASH 2017 competition. *BMC Med. Genomics* **11**, (2018).

124. Sadat, M. N. *et al.* SAFETY: Secure gwAs in Federated Environment through a hYbrid Solution. *IEEE/ACM Trans. Comput. Biol. and Bioinf.* **16**, 93–102 (2019).

125. Mitra-Behura, S., Fiolka, R. P. & Daetwyler, S. Singularity Containers Improve Reproducibility and Ease of Use in Computational Image Analysis Workflows. *Front Bioinform* **1**, (2021).

126. Schatz, M. C., Philippakis, A. A., Afgan, E., Banks, E. & Carey, V. J. Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space. *Cell Genom* **2**, (2022).

127. Turnbull, C. Introducing whole-genome sequencing into routine cancer care: the Genomics England 100 000 Genomes Project. *Ann. Oncol* **29**, 784–87 (2018).

128. Turro, E., Astle, W. J., Megy, K., Gräf, S. & Greene, D. Whole-genome sequencing of patients with rare diseases in a national health system. *Nature* **583**, 96–102 (2020).

129. Afgan, E., Jalili, V., Goonasekera, N., Taylor, J. & Goecks, J. Federated Galaxy: Biomedical Computing at the Frontier. *IEEE Int Conf Cloud Comput* **2018**, (2018).

130. Mandl, K. D. *et al.* The Genomics Research and Innovation Network: creating an interoperable, federated, genomics learning system. *Genetics in Medicine* **22**, 371–380 (2020).

131. European Union, C. C-40/17 Fashion ID GmbH & Co. KG v Verbraucherzentrale NRW eV. (2018).

132. Board, E. D. P. Guidelines 07/2020 on the concepts of controller and processor in the GDPR. (2021).

133. Wrigley, S. When People Just Click": Addressing the Difficulties of Controller/Processor Agreements Online. in *Legal Tech, Smart Contracts and Blockchain* 221–52 (Springer, 2019).

134. Warnat-Herresthal, S. *et al.* Swarm Learning for decentralized and confidential clinical machine learning. *Nature* **594**, 265–270 (2021).

135. Casaletto, J. *et al.* Federated analysis of BRCA1 and BRCA2 variation in a Japanese cohort. *Cell Genomics* **2**, 100109 (2022).

136. Vaske, O. M. *et al.* Comparative Tumor RNA Sequencing Analysis for Difficult-to-Treat Pediatric and Young Adult Patients With Cancer. *JAMA Netw Open* **2**, e1913968 (2019).

137. European Union, C. C-210/16 Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein v Wirtschaftsakademie Schleswig-Holstein GmbH. (2018).

138. European Union, C. Case C-25/17 Tietosuojavaltuutettu v Jehovan todistajat — uskonnollinen yhdyskunta. (2018).

139. Dursi, L. J. *et al.* CanDIG: Federated network across Canada for multi-omic and health data discovery and analysis. *Cell Genomics* **1**, 100033 (2021).

140. Azzariti, D. R., Riggs, E. R., Niehaus, A., Rodriguez, L. L. & Ramos, E. M. *Points to consider for sharing variant-level information from clinical genetic testing with ClinVar*. vol. 4 (Cold Spring Harb Mol Case Stud, 2018).

141. Dyke, S. O. M., Knoppers, B. M., Hamosh, A., Firth, H. V. & Hurles, M. Matching" consent to purpose: The example of the Matchmaker Exchange. *Human Mutation* **38**, 1281–85 (2017).

142. Shabani, M., Dyke, S. O. M., Marelli, L. & Borry, P. Variant data sharing by clinical laboratories through public databases: consent, privacy and further contact for research policies. *Genet. Med* **21**, 1031–37 (2019).

143. Rehm, H. L., Page, A. J. H., Smith, L., Adams, J. B. & Alterovitz, G. GA4GH: International policies and standards for data sharing across genomic research and healthcare. *Cell Genom* **1**, (2021).

144. JT, D. Describing Sequence Variants Using HGVS Nomenclature. *Methods Mol. Biol* **1492**, 243–51 (2017).

145. Danecek, P., Auton, A., Abecasis, G., Albers, C. A. & Banks, E. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–58 (2011).

146. Wagner, A. H., Babb, L., Alterovitz, G., Baudis, M. & Brush, M. The GA4GH Variation Representation Specification: A computational framework for variation representation and federated identification. *Cell Genom* **1**, (2021).

147. Ayaz, M., Pasha, M. F., Alzahrani, M. Y., Budiarto, R. & Stiawan, D. The Fast Health Interoperability Resources (FHIR) Standard. *Systematic Literature Review of Implementations, Applications, Challenges and Opportunities. JMIR Med Inform* **9**, (2021).

148. Bender, D. & Sartipi, K. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. in *Proceedings of the 26th IEEE international symposium on computer-based medical systems* 326–31 (IEEE, 2013).

149. Jacobsen, J. O. B., Baudis, M., Baynam, G. S., Beckmann, J. S. & Beltran, S. The GA4GH Phenopacket schema defines a computable representation of clinical data. *Nat. Biotechnol* **40**, 817–20 (2022).

150. Hong, N., Wen, A., Shen, F., Sohn, S. & Wang, C. Developing a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data. *JAMIA Open* **2**, 570–79 (2019).

151. Jonscher, K. R. *et al.* Spaceflight Activates Lipotoxic Pathways in Mouse Liver. *PLoS ONE* **11**, e0152877 (2016).

152. Beheshti, A. *et al.* Multi-omics analysis of multiple missions to space reveal a theme of lipid dysregulation in mouse liver. *Sci Rep* **9**, 19195 (2019).

153. Vinken, M. Hepatology in space: Effects of spaceflight and simulated microgravity on the liver. *Liver International* liv.15444 (2022) doi:10.1111/liv.15444.

154. Garrett-Bakelman, F. E. *et al.* The NASA Twins Study: A multidimensional analysis of a year-long human spaceflight. *Science* **364**, eaau8650 (2019).

155. da Silveira, W. A. *et al.* Comprehensive Multi-omics Analysis Reveals Mitochondrial Stress as a Central Biological Hub for Spaceflight Impact. *Cell* **183**, 1185-1201.e20 (2020).

156. Squair, J. W. *et al.* Confronting false discoveries in single-cell differential expression. *Nat Commun* **12**, 5692 (2021).

157. Hariton, E. & Locascio, J. J. Randomised controlled trials - the gold standard for effectiveness research: Study design: randomised controlled trials. *BJOG: Int J Obstet Gy* **125**, 1716–1716 (2018).

158. Torralba, A. & Efros, A. A. Unbiased look at dataset bias. in *CVPR 2011* 1521–1528 (IEEE, 2011). doi:10.1109/CVPR.2011.5995347.

159. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building Machines That Learn and Think Like People. (2016) doi:10.48550/ARXIV.1604.00289.

160. Ganju, S. *et al.* Learnings from Frontier Development Lab and SpaceML -- AI Accelerators for NASA and ESA. (2020) doi:10.48550/ARXIV.2011.04776.

161. Budd, S. *et al.* Prototyping CRISP: A Causal Relation and Inference Search Platform applied to Colorectal Cancer Data. in *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)* 517–521 (IEEE, 2021). doi:10.1109/LifeTech52111.2021.9391819.

162. Arjovsky, M., Bottou, L., Gulrajani, I. & Lopez-Paz, D. Invariant Risk Minimization. Preprint at http://arxiv.org/abs/1907.02893 (2020).

163. Peters, J., Bühlmann, P. & Meinshausen, N. Causal inference using invariant prediction: identification and confidence intervals. Preprint at http://arxiv.org/abs/1501.01332 (2015).

164. Beery, S., van Horn, G. & Perona, P. Recognition in Terra Incognita. (2018) doi:10.48550/ARXIV.1807.04975.

165. Bühlmann, P. Invariance, Causality and Robustness. *Statist. Sci.* **35**, (2020).

166. Berrios, D. C., Beheshti, A. & Costes, S. V. FAIRness and Usability for Open-access Omics Data Systems. *AMIA Annu Symp Proc* **2018**, 232–241 (2018).

167. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).

168. Globus, R., Cadena, S. & Galazka, J. Rodent Research-1 (RR1) National Lab Validation Flight: Mouse liver transcriptomic, proteomic, and epigenomic data. (2015) doi:10.26030/K5C1-JD05.

169. Globus, R. *et al.* Rodent Research-1 (RR1) NASA Validation Flight: Mouse liver transcriptomic, proteomic, and epigenomic data. (2015) doi:10.26030/JQ04-0N51.

170. Globus, R., Galazka, J., Smith, R. & Cramer, M. Rodent Research-3-CASIS: Mouse liver transcriptomic, proteomic, and epigenomic data. (2017) doi:10.26030/9K6W-4C28.

171. Galazka, J. RR-1 and RR-3 mouse liver transcriptomics with and without ERCC control RNA spike-ins. (2020) doi:10.26030/RWYP-9325.

172. Levene, A. P., Kudo, H., Thursz, M. R., Anstee, Q. M. & Goldin, R. D. Is oil red-O staining and digital image analysis the gold standard for quantifying steatosis in the liver? *Hepatology* **51**, 1859–1859 (2010).

173. Scott, R. T. *et al.* Advancing the Integration of Biosciences Data Sharing to Further Enable Space Exploration. *Cell Reports* **33**, 108441 (2020).

174. Gonzalez Zelaya, C. V. Towards Explaining the Effects of Data Preprocessing on Machine Learning. in *2019 IEEE 35th International Conference on Data Engineering (ICDE)* 2086–2090 (IEEE, 2019). doi:10.1109/ICDE.2019.00245.

175. Lause, J., Berens, P. & Kobak, D. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. *Genome Biol* **22**, 258 (2021).

176. Wong, S. C., Gatt, A., Stamatescu, V. & McDonnell, M. D. Understanding Data Augmentation for Classification: When to Warp? in *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)* 1–6 (IEEE, 2016). doi:10.1109/DICTA.2016.7797091.

177. Quinn, T. P., Crowley, T. M. & Richardson, M. F. Benchmarking differential expression analysis tools for RNA-Seq: normalization-based vs. log-ratio transformation-based methods. *BMC Bioinformatics* **19**, 274 (2018).

178. Zwiener, I., Frisch, B. & Binder, H. Transforming RNA-Seq Data to Improve the Performance of Prognostic Gene Signatures. *PLoS ONE* **9**, e85150 (2014).

179. Zhang, Z. *et al.* Novel Data Transformations for RNA-seq Differential Expression Analysis. *Sci Rep* **9**, 4820 (2019).

180. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**, R25 (2010).

181. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106 (2010).

182. Sun, Z. & Zhu, Y. Systematic comparison of RNA-Seq normalization methods using measurement error models. *Bioinformatics* **28**, 2584–2591 (2012).

183. Schapire, R. E. The strength of weak learnability. *Mach Learn* **5**, 197–227 (1990).

184. Vilariño, D. L., Cabaleiro, J. C., Martínez, J., Rivera, F. F. & Pena, T. F. Graph-based approach for airborne light detection and ranging segmentation. *J. Appl. Remote Sens* **11**, 015020 (2017).

185. Mitra, S., De, A. & Chowdhury, A. Epidemiology of non-alcoholic and alcoholic fatty liver diseases. *Transl Gastroenterol Hepatol* **5**, 16–16 (2020).

Appendix A: List of publications

The following is a list of publications for which I was first author during my time as a graduate student.

Casaletto, J. *et al.* Federated analysis of BRCA1 and BRCA2 variation in a Japanese cohort. *Cell Genomics* **2**, 100109 (2022).

Casaletto, J., Cline, M. & Shirts, B. Modeling the impact of data sharing on variant classification. *Journal of the American Medical Informatics Association* **30**, 466–474 (2023).

Appendix B: Supplement for *Modeling the impact of data sharing on variant classification*

In this section, we provide more detail for how we combine evidence in the Bayesian framework. We also provide the graphs for additional experiments run with different configurations, as described in the main text.

Combining evidence

A single piece of evidence is represented as an odds of pathogenicity. Clinical evidence observed for the same variant from unrelated patients is modeled as independent, so the odds from multiple observations may be combined multiplicatively. The odds of a variant V_i being pathogenic (belonging to the class P) given all the evidence X_j is the product of all the evidence, expressed as odds, as shown in Equation S1.

$$\mathrm{odds}(\mathrm{Vi} \ \in \ \mathrm{P} \,|\, \mathrm{Xj}) = \prod_{j=1}^{n} \mathrm{Xj} \qquad \textbf{Equation S1}$$

We convert the odds of pathogenicity to a log scale as shown in Equation S2.

$$\log(\mathrm{odds}(\mathrm{Vi} \ \in \ \mathrm{P} \,|\, \mathrm{Xj}) = \sum_{j=1}^{n} \log(\mathrm{Xj}) \qquad \textbf{Equation S2}$$

For a single variant, we compare this sum to the thresholds for Benign, Likely Benign, Likely Pathogenic, and Pathogenic in log scale. The same logic is applied to calculate the odds that the variant is Benign (belonging to the class B), as shown in Equation S3.

$$\log(\mathrm{odds}(\mathrm{Vi} \ \in \ \mathrm{B} \,|\, \mathrm{Xj}) = \sum_{j=1}^{n} \log(\mathrm{Xj}) \qquad \textbf{Equation S3}$$

10 small, 7 medium, and 3 large centers for 1e-05 over 20 years

The following plots show the same collection of sequencing centers we used in the main text for an allele frequency of 1e-05 but for a timespan of 20 years instead of 5 years. These plots show that after 20 years of sharing either data or classifications, all the pathogenic variants and almost all benign variants get classified. The difference between the sharing models is how quickly variants are classified and “promoted” from “Likely Benign” and “Likely Pathogenic” to “Benign” and “Pathogenic”, respectively.

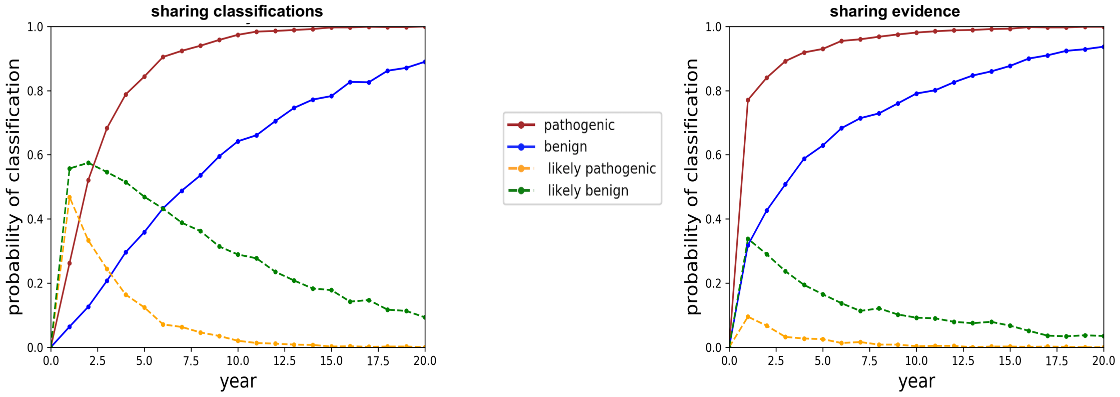

Figure S1. Probabilities of classifying variants at 1e-05 frequency plotted over the course of 20 years at 10 small, 5 medium, and 3 large sequencing centers. On the left, all variant classifications and none of the clinical data are shared. On the right, all the clinical data are shared.

10 small, 7 medium, and 3 large centers for 1e-06 over 5 years

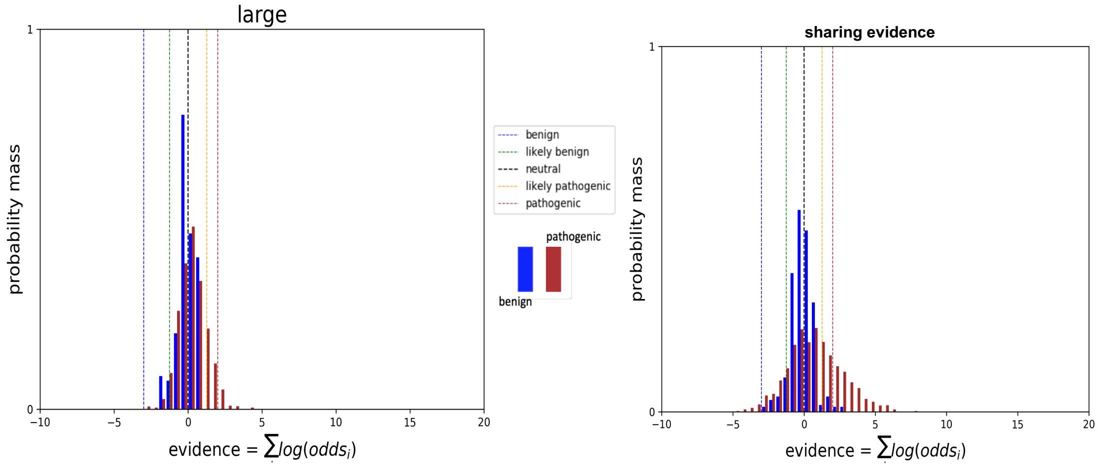

Figure S2. Histograms of cumulative log odds for classifying each of 1000 simulated variants present at a 1e-06 frequency in the population for one large sequencing center on the left and for all participating sequencing centers on the right over the course of 5 years. Classification thresholds are demarcated as vertical hash lines. Benign variants are in blue and pathogenic variants in red.

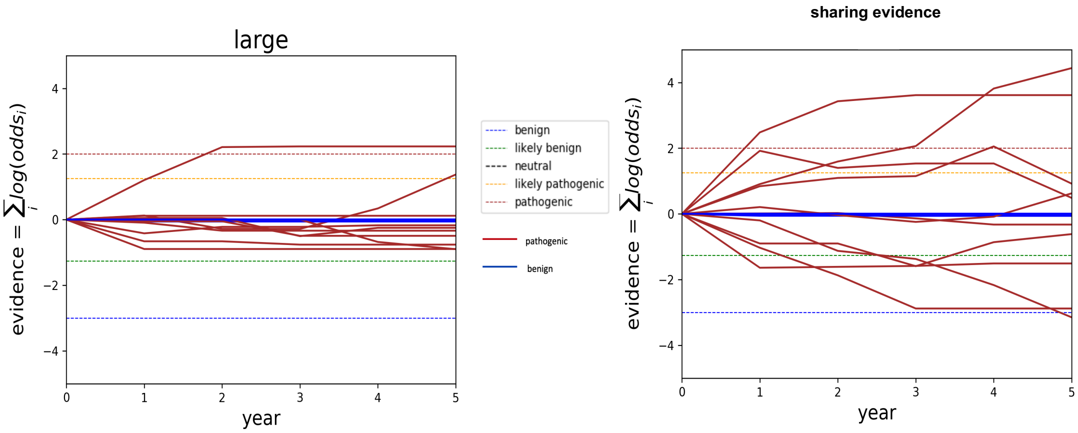

Figure S3. Classification trajectories over the course of 5 years for 10 randomly selected variants at 1e-06 frequency in the population for one large sequencing center on the left and for all participating sequencing centers on the right. Classification thresholds are demarcated as horizontal hash lines in the timeline plots. Benign variants are in blue and pathogenic variants in red.

10 small, 7 medium, and 3 large centers for 1e-06 over 20 years

The following plots show the same collection of sequencing centers we used in the main text for an allele frequency of 1e-06 but for a timespan of 20 years instead of 5 years. These plots show that even after 20 years of sharing data, there remains insufficient clinical data to classify variants which occur at 1e-06 frequency in the population.

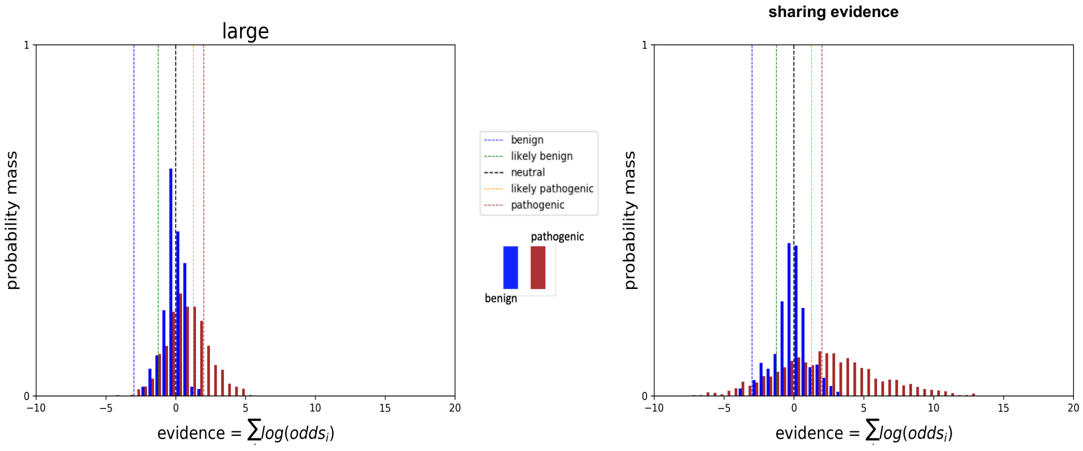

Figure S4. Histograms of cumulative log odds for classifying each of 1000 simulated variants present at a 1e-06 frequency in the population for one large sequencing center on the left and for all participating sequencing centers on the right. Classification thresholds are demarcated as vertical hash lines. Benign variants are in blue and pathogenic variants in red.

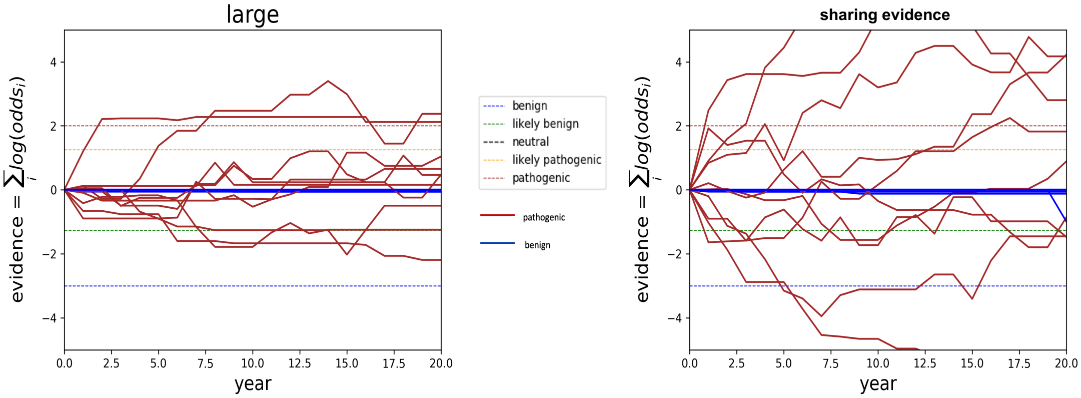

Figure S5: Classification trajectories over the course of 20 years for 10 randomly selected variants at 1e-06 frequency in the population for one large sequencing center on the left and for all participating sequencing centers on the right. Classification thresholds are demarcated as horizontal hash lines in the timeline plots. Benign variants are in blue and pathogenic variants in red.

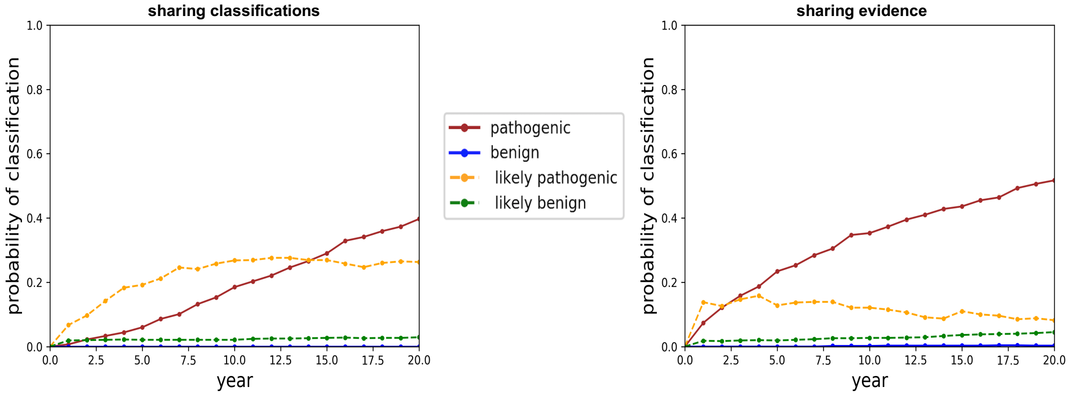

Figure S6. Probabilities of classifying variants at 1e-06 frequency plotted over the course of 20 years at 10 small, 7 medium, and 3 large sequencing centers. On the left, all variant classifications and none of the clinical data are shared. On the right, all the clinical data are shared.

5 small, 3 medium, and 1 large centers for 1e-5 over 5 years

The following plots show the same allele frequency and timespan we used in the main text but with a smaller collection of participating centers. These plots show that sharing clinical data with few centers significantly reduces the probability of classifying variants at the same frequency. The probability of classifying variants as pathogenic reduces from about 95% to about 60%, and the probability of classifying variants as benign reduces from about 60% to about 20%.

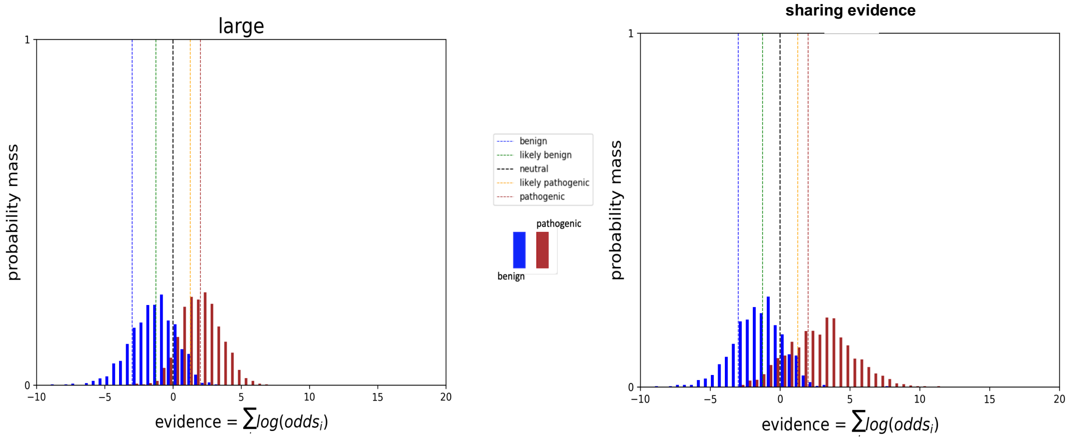

Figure S7. Histograms of cumulative log odds for classifying each of 1000 simulated variants present at a 1e-05 frequency in the population for one large sequencing center on the left and for all participating sequencing centers on the right. Classification thresholds are demarcated as vertical hash lines. Benign variants are in blue and pathogenic variants in red.

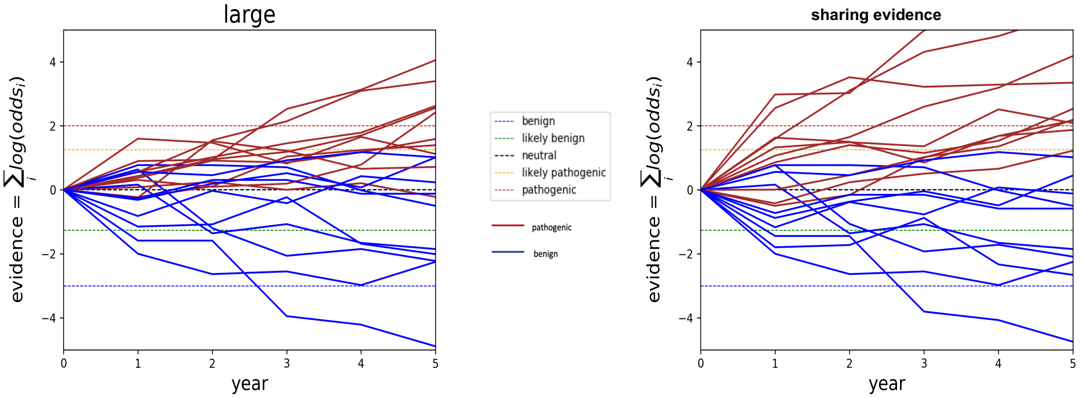

Figure S8: Classification trajectories over the course of 5 years for 10 randomly selected variants at 1e-05 frequency in the population for one large sequencing center on the left and for all participating sequencing centers on the right. Classification thresholds are demarcated as horizontal hash lines in the timeline plots. Benign variants are in blue and pathogenic variants in red.

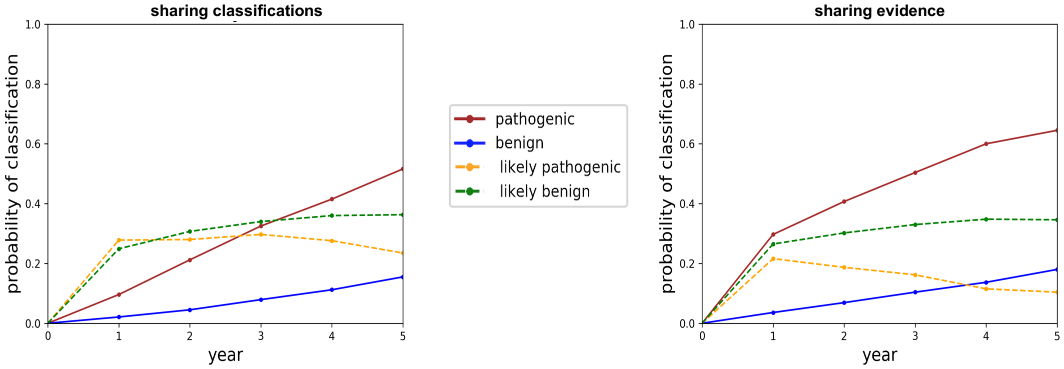

Figure S9. Probabilities of classifying variants at 1e-05 frequency plotted over the course of 5 years at 5 small, 3 medium, and 1 large sequencing center. On the left, all variant interpretations and none of the clinical data are shared. On the right, all the clinical data are shared.