Deriving a sociotechnical model for discovery in genomics-enabled learning health systems

Kathleen Diveny Ferar

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

David R. Crosslin, Chair

Peter Tarczy-Hornoch

Gail P. Jarvik

Annie T. Chen

Program Authorized to Offer Degree:

Department of Biomedical Informatics and Medical Education

University of Washington

Abstract

Deriving a sociotechnical model for discovery

in genomics-enabled learning health systems

Kathleen Diveny Ferar

Chair of the Supervisory Committee:

David R. Crosslin

Department of Biomedical Informatics and Medical Education

Recent advances in genetic sequencing technologies and analysis tools have made genomic data widely available for medical research. Despite the expectation that genomic data will revolutionize medicine, there exist major evidence gaps in demonstrating the utility of clinical genomics for improving patient outcomes and increasing healthcare efficiency. One promising avenue for reducing this evidence gap and accelerating the pace of clinically relevant discoveries is to foster environments in which genomic research and clinical care exist symbiotically. However, the technical and sociocultural requirements for conducting genomic research in clinical environments are not well-defined. The learning health system (LHS) framework is one lens through which the barriers and enablers of clinical genomic discovery can be identified and organized. Furthermore, drawing on experiences from clinical research consortia like the Clinical

Sequence Evidence-Generating Research (CSER) Consortium and the Electronic Medical Records and Genomics (eMERGE) Network can help identify requirements that are unique to genomic research initiatives that straddle the research-clinical boundary. In this work, we sought to derive a sociotechnical model for clinical genomic discovery in genomics-enabled learning health systems (GLHSs). We first identified data coordination challenges, strategies, and recommendations from the clinical genomics research data integration process in the CSER Consortium and found that the social processes involved in data coordination are tantamount to the informatics tools used to facilitate data coordination (**Aim 1**). We then explored medical geneticist perspectives on clinical genomic discovery by interviewing 20 board-certified medical geneticists in CSER, eMERGE, and the University of Washington medical system (**Aim 2**). Using constructivist grounded theory methods, we developed a preliminary model of GLHS discovery that utilizes the concepts of representation, responsibility, risks and benefits, relationships, and resources ("5R") to capture the negotiations and constraints involved in clinical-research integration in genomics. To demonstrate the utility of merging electronic health record (EHR) data with genomic data for discovery, we then conducted a logistic regression-based genome-wide association study for *C. diff.* infection (CDI) using merged genetic and EHR data from 12 clinical sites in the eMERGE Network and found a strong gene-disease association in the HLA-DRB locus (P=8.06 x 10^{-14}) that predisposed carriers to CDI (**Aim 3**). Finally, we conducted a systematic literature review of proposed enablers of clinical genomic discovery and synthesized the qualitative results from the literature review and recommendations from Aim 1 with the *a priori* framework developed in Aim 2 using best-fit framework synthesis (BFFS) (**Aim 4**). We found that the vast majority of themes identified in the literature were accommodated by the *a priori* framework, suggesting that the 5R model of GLHS discovery is

an adequate representation of processes involved in learning health research. Using additional qualitative evidence identified during BFFS, we developed an enhanced 5R sociotechnical model to demonstrate how iterative, multidirectional negotiation and tool development can facilitate virtuous cycles of learning in clinical genomics research.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

"If I have seen further, it is by standing on the shoulders of giants." – Sir Isaac Newton

The list of people to thank for getting to this point could easily be longer than this dissertation, but in brief, thank you to my dissertation committee (Dr. David Crosslin, Dr. Peter Tarczy-Hornoch, Dr. Gail Jarvik, Dr. Annie Chen, and Dr. Debby Tsuang) for their sage advice over the years. A very special thank you to David Crosslin for being a patient mentor, trusted colleague, and friend.

Thank you to all members of the Clinical Sequence Evidence Generating Research (CSER) Consortium, especially those in the Data Wranglers and Project Managers working groups, for teaching me how to herd cats. Thank you to all my predecessors in the Electronic Medical Records and Genomics (eMERGE) Network for doing the heavy lifting of data collection and cleaning before I even arrived at UW. Thank you to the clinicians who participated in my interview study despite their ever-crowding schedules. It was an honor to work with so many brilliant and passionate scientists.

Thank you to all my friends and colleagues at UW who helped foster such a beautiful and collaborative work environment. I could not have asked for a kinder or more supportive group of people to go through this experience with. And finally, thank you to my wonderful husband Keenan, mom Kelly, dad Chuck, brother Steven, sister Elizabeth, and Grandmother Joyce for their boundless love and support. I love all of you dearly.

DEDICATION

For my late grandfather, Wallace (“Wally Pop”) Wayne Neeley.

For always asking, “Kathleen, did you excel today?” in his signature Alabama drawl.

I sure tried, Wally Pop!

CHAPTER 1: INTRODUCTION

1.1 Background

Since the completion of the Human Genome Project in 2003, scientists and clinicians alike have expected genomics to revolutionize human healthcare. Indeed, genomics has led to advancements in medicine that would not have been possible without the novel insights into health, disease, and basic biology that our DNA reveals in such extraordinary detail. Treatments such as gene therapy and cancer immunotherapy provide hope to patients and families where hope never seemed like an option, and diagnostic techniques for rare inherited disorders help to end diagnostic odysseys and pave the way for new treatment options. However, translating genomics research more broadly into clinical practice remains a challenge despite the already delivered and expected promises of genomic medicine [1].

One mechanism that has been proposed for realizing the full potential of genomic research and medicine is to formalize the integration of genomic research and clinical care. It is well-recognized that large research evidence bases are required to demonstrate the clinical utility and actionability of genomic variants in different populations and clinical environments [2]. However, a lack of evidence for the economic and clinical utility of using newer genomic discoveries to guide clinical care has led to low adoption of cutting-edge genomic medicine among healthcare organizations and a lack of support from healthcare payers and policymakers to advance genomics-informed clinical care [3]. Integrating genomic research with medical practice can contribute to an enhanced evidence base for the validity and clinical utility of genomic findings [4]. Genomic research that is conducted within healthcare organizations is

naturally benefited by its close proximity to rich clinical data that can be used to identify disease associations in large, diverse patient cohorts, and by its proximity to clinical outcomes data that can be used to monitor patients over time [5]. Nonetheless, a host of technical, social, cultural, and ethical questions remain regarding how best to conduct genomic discovery in clinical settings [6].

The learning health system (LHS) framework is a useful lens for investigating the challenges and enablers of conducting genomic discovery in a clinical context. Proposed by the Institute of Medicine (IOM, now known as the National Academy of Medicine) in 2007, the LHS framework champions a healthcare system in which data is generated as a by-product of routine care, data is transformed into knowledge through research, and new knowledge is iteratively used to improve the quality of care and improve healthcare efficiency [7]. The core elements of this model include a robust data infrastructure, care improvement through clinical decision support, and rewards for high-value care and transparency. LHS systems that are adapted to accommodate clinically-generated genomic data have previously been referred to as genomics-enabled learning health systems (GLHSs), and have been strongly supported by a variety of clinical, research, and policy stakeholders [8]. For example, the 2020 National Human Genome Research Institute (NHGRI) strategic vision report identified the development and implementation of GLHSs as an important new research frontier in the field of genomics [9]. However, the basic LHS model must be enhanced to accommodate challenges that are amplified in genomics, such as the large size and complexity of the raw data, a lack of data standards, disparities in access to genomic testing, difficulties in implementing effective clinical decision support tools, and a lack of insurance coverage for genetic testing [8]. The “data to knowledge”

process in a GLHS is particularly susceptible to these challenges, given its historical exclusivity to research environments and the novelty of genomic data relative to other forms of clinical data. While previous studies have identified barriers for incorporating genomic data into an LHS, none have specifically addressed the challenges of genomic discovery [10–14].

Multi-site clinical genomics research projects are ideal environments for developing and evaluating genomic data integration and discovery techniques in a clinical context. The Clinical Sequence Evidence-Generating Research (CSER) Consortium, for example, is an ideal environment for studying clinical research data integration techniques and evaluating these techniques in an integrated research-clinical context [15]. In addition, the Electronic Medical Records and Genomics (eMERGE) Network conducts discovery and translational work in genomics using the combined powers of genomic and EHR data [16]. These projects experience many of the same challenges that are identified in the LHS model, including data integration issues, questions of privacy and patient consent, and funding challenges [17]. Clinically embedded genomic research projects are uniquely positioned at the interface of research and clinical care and can thus inform strategies for harnessing genomic data for clinical use [18]. They are also ideal environments for generating gene-disease association discoveries using merged clinical and genetic data and demonstrating the utility of harnessing multiple data types for medical genomics research. One such disease of interest is *Clostridioides difficile* (*C. diff.*) infection (CDI), formerly known as *Clostridium difficile* infection, which presents a large epidemiological and economic burden to the US healthcare system and may be impacted by host genetic risk factors [19].

Given the complex technical and sociopolitical landscape of clinical research integration in genomics, in this dissertation we aim to develop a comprehensive sociotechnical model for GHLS discovery that examines the relationships between individuals and the complex social, political, and technical environments in which they operate. The development of such a model warrants the triangulation of multiple research methods, perspectives, and data sources to reveal and synthesize different aspects of reality [20,21]. We therefore approach the topic of clinical research integration from four different angles: 1. Researcher perspectives on data integration for clinical genomic research (**Aim 1**); 2. Clinician perspectives on genomic knowledge generation in clinical environments (**Aim 2**); 3. An applied example of knowledge generation using clinical data (**Aim 3**); and 4. Systematic literature review and qualitative evidence synthesis (**Aim 4**). The integrative conceptual model that results from this work can be used to facilitate the design and development of GLHS discovery research programs by "harnessing the natural properties which emerge (often spontaneously) at the interface between the socio (human behavioural) and technical components of complex systems" (Braithwaite et al. 2008, p. 37) [22,23]. In this way, the actual and expected contributions of genomic research to human health can begin to converge.

1.2 Dissertation Aims

1.2.1 Aim 1: Researcher perspectives on clinical and genomic data coordination

In this aim, we identify 14 lessons learned and 11 broad recommendations for survey, phenotype, and sequence data coordination through retrospective analysis of digital artifacts generated as a by-product of the data coordination process in the CSER Consortium. While these

recommendations are grounded in the experiences of a large, NIH-funded research program, they are thematically interoperable with data coordination initiatives in general, and have practical implications in the areas of planning, communication, informatics, analytics, and data governance.

1.2.2 Aim 2: Medical geneticist perspectives on clinically embedded genomic discovery

In this aim, we explore the perspectives of board-certified medical geneticists on integrating genomic discovery research with clinical care. Using constructivist grounded theory methods [24,25], we identify perceived drivers and barriers for GLHS discovery, and offer an *a priori* theoretical framework for understanding the technical, social, and ethical forces that influence the shifting boundaries between research and clinical care in genomics.

1.2.3 Aim 3: Discovery of genetic risk factors for *C. diff.* infection using merged clinical and genomic data

In this aim, we use merged genomic and clinical data from the eMERGE Network to conduct a logistic regression based GWAS of CDI cases and controls to identify common genetic variants associated with higher risk of developing CDI. We also demonstrate the utility of using clinical data for gene-disease association studies and provide a practical example of clinical genomic discovery in action.

1.2.4 Aim 4: Development of an integrative sociotechnical model for genomics-enabled learning health system discovery

The objectives of this aim are twofold. First, we conduct a systematic literature review of studies that have identified enabling factors of genomic discovery and validation research in the LHS model and develop a theory of change model to describe the current landscape of this body of literature. Second, we use best-fit framework synthesis (BFFS) to compare the *a priori* model from Aim 2 with qualitative evidence identified in Aim 1 and the systematic literature review to create an integrative sociotechnical model for GLHS discovery.

CHAPTER 2: RELATED WORK

2.1 Data coordination in clinical research

Sharing rich clinical and genomic datasets within and between institutions is essential for advancing medical genomics research, and ultimately for achieving the LHS vision [26,27]. However, there are many known technical, ethical, and political challenges with sharing clinical and genomic data, such as insufficient or nonexistent data harmonization infrastructures, identifiability concerns, and a lack of trust between the public and healthcare institutions [28]. Several genomics research consortia have identified strategies for improving data coordination in clinically-embedded environments [16,29,30], but standards and expectations for clinical and genomic data coordination have yet to be established. In this work, we seek to contribute to the development of best practices and standards for clinical genomic data coordination, which can be applied to both LHS environments and multi-site research projects in general.

2.2 Genomic discovery in genomics-enabled learning health systems

While the LHS model has received considerable attention since it was first proposed by the IOM in 2007, the concept of a genomics-enabled LHS has not been well-defined [8]. The original LHS model outlines the technical, social, and political requirements for conducting rapid learning using clinical data, but the novelty and complexity of genomic data necessitate the development of an enhanced conceptual GLHS model [31]. The line between research and clinical care in genomics has historically been blurred–perhaps more so than in other medical disciplines–due to the rapid evolution of technologies in the field and the direct diagnostic implications of many discoveries, but significant ethical and legal conflicts of interest have

arisen as genomic research has naturally shifted into clinical spaces, and vice versa [6]. Systematically integrating genomic research into clinical environments therefore has the potential to exacerbate existing tensions between research and clinical priorities in genomics. In this work, we seek to identify and relate the components of a novel GLHS model from the perspectives of medical geneticists, who work at the bleeding edge of research and clinical care in genomics.

2.3 Electronic health record and genomic data integration for discovery

Leveraging both participant-level clinical data and genomic data is essential for making clinically relevant genomic discoveries [32]. The eMERGE Network has been a leader in this area of research, and has successfully developed harmonized clinical phenotypes across a network of EHRs, which have been used to conduct genome-wide association studies (GWAS) for diseases such as herpes zoster [33], peripheral arterial disease [34], and dementia [35]. Additional GWAS that leverage rich clinical data are needed to identify opportunities for new clinical interventions and potential therapeutic targets for diseases that present a significant burden to patients and health systems [36]. For example, *C. diff.* infection (CDI) is a leading infectious cause of nosocomial diarrhea in North America and is associated with a high global burden of disease [37]. A previous GWAS of 16,464 patients (1,160 CDI cases; 15,304 controls) from the Geisinger MyCode cohort [38] was conducted using a CDI phenotyping algorithm developed by the eMERGE Network, and several variants in the human leukocyte antigen (HLA) region were suggestive of increased CDI risk. In this work, we conduct an additional GWAS in a cohort of 99,000 eMERGE participants using the eMERGE CDI phenotyping algorithm to identify genetic risk factors that are significantly associated with CDI.

2.4 Qualitative evidence synthesis in learning health systems research

Systematically integrating evidence from the literature with a guiding conceptual model can facilitate a broader understanding of complex research landscapes, such as the LHS research landscape [39]. While previous systematic literature reviews and scoping reviews have assessed enabling and inhibiting factors of systems that conform to the original LHS model, none have assessed the literature surrounding genomics-enabled implementations of the LHS model [40–44]. Similarly, Enticott et al. (2021) [45] developed an integrative LHS framework for the Australian healthcare system using evidence synthesis, but additional work is needed to develop an integrative GLHS model that is tailored to the US healthcare system. In this work, we seek to systematically synthesize themes from the body of literature that addresses the GLHS concept, and to leverage qualitative evidence synthesis methods that facilitate the development of an integrative sociotechnical model for GLHS discovery.

CHAPTER 3: LESSONS LEARNED FROM MULTI-INSTITUTIONAL CLINICAL RESEARCH DATA INTEGRATION (AIM 1)

3.1 Introduction

Data coordination is foundational to data-driven discovery work. While this process is more commonly referred to as "data management" [46–48], we use the term "coordination" to emphasize the communicative and collaborative aspects of managing research data. Significant collaboration between institutions, clinicians, researchers, policymakers, and patient-participants is required to yield datasets that advance biomedical research [49]. Few organizations are more acutely aware of the challenges of data coordination than multi-institutional clinical research programs, which experience conflicting research and clinical priorities across multiple institutions. The CSER Consortium [15] was one such multi-site program that consisted of seven clinically embedded genomic medicine research projects. While all projects shared a common goal of investigating the utility of integrating genomic sequencing into clinical care, their specific research aims, methods, patient populations, and clinical environments varied widely. Over a period of three years, the consortium worked with an internal Data Coordinating Center (DCC) to harmonize seven distinct survey, phenotype, and sequencing datasets from the second phase of CSER into a single resource. Neither the first phase nor the second phase of CSER was originally designed for genomic discovery research. However, investigators from the first phase of CSER challenged the viability of the traditional research-clinical dichotomy in the rapidly evolving field of genomics [18]. As one CSER site noted,

> We believe [the CSER studies] are intrinsically both [research and clinical care]. Given the nature of the data generation and analysis process and the regular rates of change in

> genome interpretation, each family is in a very real sense a research project. However, the consequences of the results are often of substantial and direct clinical impact, and thereby these efforts are also clinical care. [Site 4] [18].

Given the fluidity between genomic testing, research, and clinical care in CSER, we argue that the data coordination experiences of clinical research consortia can reveal challenges that might be faced by clinically embedded genomic discovery programs and offer potential solutions.

In this aim, we identify 14 lessons learned and 11 broad recommendations for survey, phenotype, and sequencing data coordination through retrospective analysis of digital artifacts generated as a by-product of the coordination process. While these recommendations are grounded in the experiences of a large, NIH-funded research program, they are thematically interoperable with data coordination initiatives in general, and have practical implications in the areas of planning, communication, informatics, analytics, and data governance. The content of this chapter is largely derived from a paper by Muenzen et al. (2022) titled, "Lessons learned and recommendations for data coordination in collaborative research: The CSER consortium experience" [50].

3.2 Related work

3.2.1 The promises and perils of data coordination

It is widely recognized that sharing clinical and research data within and between institutions is essential for advancing medical research and precision medicine [26,27]. Harnessing the ever-

growing troves of Next-Generation Sequencing data will help elucidate the complex interactions between genetics, environment, and human health and disease [51]. Combining genomic data with clinical data is necessary for identifying genetic variants that drive both rare and common disease, and for characterizing the range of clinical presentations associated with each [52,53]. While there has been significant progress in understanding monogenic disease since the advent of exome and genome sequencing, the impacts of non-coding, multigenic, and multi-allelic variation on phenotype are poorly understood [51]. To identify relationships between complex genetic factors and human health and disease, sufficient genomic and clinical evidence must be accumulated [53,54]. The "digitalization of medicine" (Auffray et al. 2016, p. 1) [52] through EHRs has contributed to a collective pool of clinical data that could be used to facilitate genomic discovery research, but both genomic and EHR data are largely siloed in different testing centers, research databases, and medical institutions [28]. Improving clinical data integration strategies within and between healthcare institutions is therefore an important precursor to discovery, but the standards and expectations for clinical and genomic data coordination are not well established.

There are known technical, legal, ethical, financial, political, and cultural barriers to sharing and aggregating health-related data for research purposes. Data collected across heterogeneous environments are inherently difficult to harmonize because they are likely collected, structured, and stored using different standards [55]. Across healthcare institutions, incompatible EHR platforms–or the lack of an EHR altogether–make automated data integration difficult [56]. Even if data can be shared between EHRs, health data is largely unstructured and sophisticated Natural Language Processing (NLP) tools are required to convert clinical text into a format that is useful

for large-scale research. Data quality is also a major concern when using EHR data for research, since clinical data are notoriously incomplete, inconsistent, and inaccurate [57,58]. While technical challenges are most commonly reported in the literature [55,59], many policymakers, biomedical researchers, and ethicists have argued that the legal and ethical challenges of sharing clinical data are the most problematic [59,60]. Clinical records contain highly sensitive Protected Health Information (PHI), which is protected by the federal Health Insurance Portability and Accountability Act (HIPAA) of 1996. While data containing PHI can be securely transferred between departments and institutions, the HIPAA Privacy and Security rules significantly restrict access to health data for research purposes. Patients may object to sharing some or all of their health data with individuals other than their healthcare providers and may want to be re-consented for every new use of their data [61]. There are additional privacy risks when sharing genomic data, which can potentially be used to re-identify individuals [62]. In this way, the scientific imperative of sharing rich clinical and genomic data across institutions and country borders conflicts with the moral imperative of protecting individual privacy [63–65]. The ethical conundrum of sharing clinical and genomic data is heightened in underrepresented minority communities, where patient trust in the medical and research enterprises are low due to historical wrongs committed by both enterprises [66]. Additionally, the policy landscape that governs clinical and research data sharing is fragmented at best, and "despite its abundance, has not resulted in a cohesive system of incentives able to reconcile the interests and expectations of different stakeholders" (Blasimme et al. 2018, p. 706) [26]. Finally, the scientific and medical communities have not yet achieved a "culture" of data sharing, in which trust and reciprocity between researchers, clinicians, patients, and research participants are central to the mission of sharing data for research purposes [67]. The number of stakeholders involved in coordinating

clinical and genomic data is large, and mutual understanding of the roles, responsibilities, and capabilities between stakeholders is rare [52].

3.2.2 Current solutions to data coordination challenges

Although many barriers to sharing clinical data have been identified, developing solutions to mitigate these barriers is challenging. To address the technical challenges of sharing clinical data, previous research has suggested that standardized metadata models should be developed to harmonize heterogeneous datasets retrospectively [17,68,69]. Although standardized capture of electronic health data is preferred, this is a major bottleneck in biomedical data sharing and is largely driven by EHR vendors [28]. Others have suggested that both data capture and metadata standards be harmonized internationally, but this solution has its own extensive set of barriers that all require complex solutions [68,70]. Data anonymization and more sophisticated cryptology approaches like blockchain have been proposed to alleviate security and privacy issues of health data sharing [28,68,71]. However, it is well-known that as the privacy of data increases, the utility of data decreases [72]. This tension is somewhat alleviated by de-identified public and controlled-access genomic databases like the 1000 Genomes Project [73], the UK Biobank [74], the National Center for Biotechnology Information (NCBI) Database of Genotypes and Phenotypes (dbGaP) [75], the NIH All of Us Research Hub [76], and the NHGRI Analysis Visualization and Informatics Lab-space (AnVIL) [77]. However, these broad data sharing mechanisms do not eliminate participant privacy issues [64], and do not often satisfy the need for more detailed clinical information. To address issues of consent, new digital consent technologies and models like dynamic consent have been proposed [26,71]. These same approaches might be useful for engaging minority communities in conversations about health

data and consent for research use and building trust [78,79]. To address issues of policy fragmentation, studies have suggested that a unified, international policy for health data sharing be developed that addresses multiple data types, encompasses a broad set of policy themes, and balances competing values of different data sharing stakeholders [26,71,80,81]. However, the extreme variability in healthcare networks and policy landscapes across the globe make this solution difficult. Finally, to increase scientific and healthcare community engagement in data sharing, some have suggested that academic and healthcare leadership take an active role in identifying and encouraging best practices in data sharing, maintaining the necessary infrastructure, and contributing to policy and guideline development [82]. For this approach to be effective, however, best practices in data coordination and guideline development must first be identified.

3.2.3 Data coordination in clinical genomics research projects

Clinical genomics research consortia face many of the same data coordination challenges that are encountered when sharing clinical data for research because they operate at the interface of research and clinical care. Examining the experiences and approaches of multi-site clinical genomics consortia is therefore an important precursor to defining best practices for heterogeneous clinical data coordination. Additionally, the experiences of Coordinating Centers (CCs) and Data Coordinating Centers (DCCs) within these consortia are valuable to document, since they are the entities that develop and orchestrate protocols for coordinating clinical research data [83]. For example, the eMERGE Network CC used centralization storage and data harmonization, network-wide Data use Agreements (DUAs), and standardized privacy and security policies to coordinate clinical and genetic data across 18 sites over the history of the

network [16]. The Li-Fraumeni Exploration (LiFE) Consortium experienced challenges coordinating communication between international members and harmonizing variant and clinical data across 8 research sites [29]. The LiFE DCC developed standardized data dictionaries, data transfer agreements, DUAs, and QA/QC measures to address technical and communication challenges. The Global Enteric Multicenter Study (GEMS) experienced challenges coordinating data across 8 different countries and across sites that had "diverse cultural, social, and technological backgrounds" (Biswas et al. 2012, p. S260) [30]. The GEMS DCC found it useful to implement a standardized data management software to collect clinical case reports but found that requiring all sites to use an electronic data capture system was not culturally appropriate. Although the reported experiences and data coordination strategies of past clinical research consortia are informative for clinical data integration strategies at a high level, a more detailed and nuanced look at specific data coordination tools, methods, and motivations used by clinical research consortia is necessary for building a comprehensive understanding of both effective and ineffective data coordination strategies.

The second phase of the CSER Consortium has been well-documented in the literature since its inception in 2018, especially with regards to its position at the research-clinical interface and to its experiences with harmonizing outcomes measures. Although CSER's initial goal was to investigate the clinical implementation of genomic sequencing in diverse populations, consortium members recognized that the genetic data collected during the study could be used for discovery purposes [15]. Efforts to make CSER genomic, clinical, and outcomes data available for future research align with the more generalizable goal of using electronic clinical data for secondary research, making CSER an excellent case study for post-hoc data

harmonization. The consortium also previously experienced challenges developing and implementing consensus outcomes measures across diverse clinical sites and research projects and identified the importance of team science approaches to data harmonization [84]. The current study builds on this previous body of work by contextualizing both harmonization issues and questions of the research-clinical interface within the organizing framework of multi-site data coordination in CSER.

3.3 Methods

In this study, we used Fleming's proposed artifact study model [85] to characterize the culture of the CSER Consortium through the lens of data coordination, and to ultimately identify cross-cutting lessons learned, recommendations, and themes in clinical research data coordination. Fleming proposed this model in 1974 as a method for characterizing human cultures through the analysis of human-made objects. While artifact analysis has historically been used to study how physical artifacts like decorative art or hand-made tools reflect the cultures in which they were developed, digital artifacts such as email exchanges and audio recordings are frequently generated as a result of computer-based work and are similarly indicative of modern work culture [86]. For example, Fang et al. (2022) [87] identified digital artifacts as an essential part of knowledge coordination in distributed teams, where the "technology practices" of team members are "embedded in digital artefacts" (Fang et al. 2022, p. 537). We therefore used artifact analysis to systematically uncover the practices and perspectives of those involved in CSER data coordination.

3.3.1 Artifact collection

Digital artifacts were identified by the primary investigator (K.F.), who was involved in development and maintenance of all technical and communicative aspects of the CSER DCC. To guide artifact collection, we identified five aspects of CSER data coordination that warranted examination and collected relevant digital artifacts: 1. Consortium structure and communication; 2. Data coordination timeline; 3. Informatics architecture; 4. Survey data harmonization; and 5. Sequence data collection. For each of the five categories, the primary investigator identified digital artifacts that were relevant to data coordination, including: 1. Email exchanges between primary investigator and stakeholders, and official consortium emails; 2. Material from the CSER Consortium private and public-facing website; 3. Papers previously published by CSER Consortium members; 4. GitHub code repositories for digital tools; 5. CSER REDCap databases; 6. Documents generated and distributed by the CSER DCC to consortium members; and 7. Official documents, such as Funding Opportunity Announcements (FOAs) and NIH policy descriptions. **Table 3.1** shows which artifact types were collected for each data coordination component.

Data Coordination Component	Email	CSER Website	CSER Papers	GitHub Code	REDCap Database	DCC Docs	Official Docs
1. Consortium structure and communication	✓	✓	✓				✓
2. Data coordination timeline	✓	✓			✓	✓	
3. Informatics architecture	✓			✓	✓	✓	
4. Survey data harmonization	✓	✓	✓	✓	✓	✓	
5. Sequence data collection	✓			✓		✓	✓

Table 3.1. Digital artifacts used for artifact analysis in five areas of CSER data coordination.

3.3.2 Artifact analysis

The five stages of a traditional artifact analysis include identification, evaluation, cultural analysis, and interpretation [85]. While artifacts were initially evaluated individually based on their history, form, construction, and function, they were ultimately described in combination with one another to facilitate the identification of cross-cutting themes in data coordination. The cultural analysis involved identifying tensions and relationships between the technical, sociocultural, and political aspects of the data integration process. Identification of those relationships was facilitated by visually mapping relationships between entities using contextual design techniques developed by Beyer & Holtzblatt [88]. Finally, a core set of emergent themes in data integration was identified and interpreted in the context of the core pillars identified in the

value-creating LHS framework proposed by Menear et al. (2019) [89] to relate the findings to the LHS model.

3.4 Results

3.4.1 Consortium structure and communication

CSER consisted of a Steering Committee and eight main working groups with members from the following contact institutions and CSER projects: 1. Baylor College of Medicine (KidsCanSeq); 2. Kaiser Permanente Northwest (CHARM); 3. University of North Carolina at Chapel Hill (NCGENES 2); 4. Icahn School of Medicine at Mount Sinai (NYCKidSeq); 5. University of California, San Francisco (P^3EGS); 6. HudsonAlpha Institute for Biotechnology (SouthSeq); and 7. The National Human Genome Research Institute (ClinSeq). Consortium activities were facilitated by a Coordinating Center based at the University of Washington and were guided by an external committee, the CSER Advisory Panel, consisting of six experts in genomic medicine and a community advocate. While all CSER sites shared a common goal of investigating the applications and outcomes of genomic sequencing in clinical care, the patient populations, specific research aims, and study protocols differed widely between sites **(Figure S3.1)**. Detailed descriptions of CSER working groups, study populations, and sequencing methodologies are described in Amendola et al. (2018) [15] and Goddard et al. (2020) [84].

Consortium communication was facilitated through monthly working group video calls, biweekly Coordinating Center calls, monthly Steering Committee calls, and tri-annual consortium-wide meetings. At the start of the COVID-19 pandemic in early 2020,

communications became entirely virtual. However, this did not significantly change communication between working groups, as communication had been largely virtual to begin with. The DCC interacted extensively with the Data Wranglers Working Group (established by the DCC in Fall 2019) and the Project Managers Working Group (established in Spring 2019). Interactions largely consisted of monthly video calls and ad-hoc calls with individual site analysts and project managers.

The DCC collaborated with several external organizations that helped maintain the technical infrastructure that the consortium used to securely manage its aggregated survey and sequence data. The Institute of Translational Health Sciences (ITHS) at the University of Washington managed the Research Electronic Data Capture (REDCap) database [90,91] that the DCC used for centralized CSER data storage, and maintained a secure web server that hosted the consortium's R Shiny [92] data management tool. The DCC also collaborated extensively with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-Space (AnVIL) consortium, which was responsible for hosting shared CSER genomic, clinical, survey, and phenotypic data in the AnVIL cloud computing ecosystem [77].

3.4.2 Timeline of CSER data harmonization, collection, and analysis activities

The second phase of CSER began in August 2017. Harmonized measures were developed throughout 2018, and sites adopted the harmonized measures in late 2018. As described in Goddard et al. 2020 [84], sites designed most of their data collection instruments independently and began recruitment and/or survey administration up to 18 months after the consortium start date. By the time the consortium had finalized the harmonized measures in late 2018, several

sites had already begun administering surveys and were tasked with administering some harmonized items that they had not previously implemented. The DCC developed the initial harmonized database and custom data collection platform throughout the Fall and Winter of 2019-2020. The DCC began coordinating the centralized intake of common survey measure responses in early 2020 and continued to collect this data until the end of the recruitment and follow-up periods at each site. Initial requests for—and preliminary analysis of—harmonized survey data began in Fall 2020, and the first submissions of genome and exome data to the AnVIL cloud platform began in Spring 2021, shortly after the AnVIL platform was designated as an official NIH data repository [93]. A timeline of major consortium-wide activities related to data harmonization, collection, and analysis is shown in **Figure 3.1.**

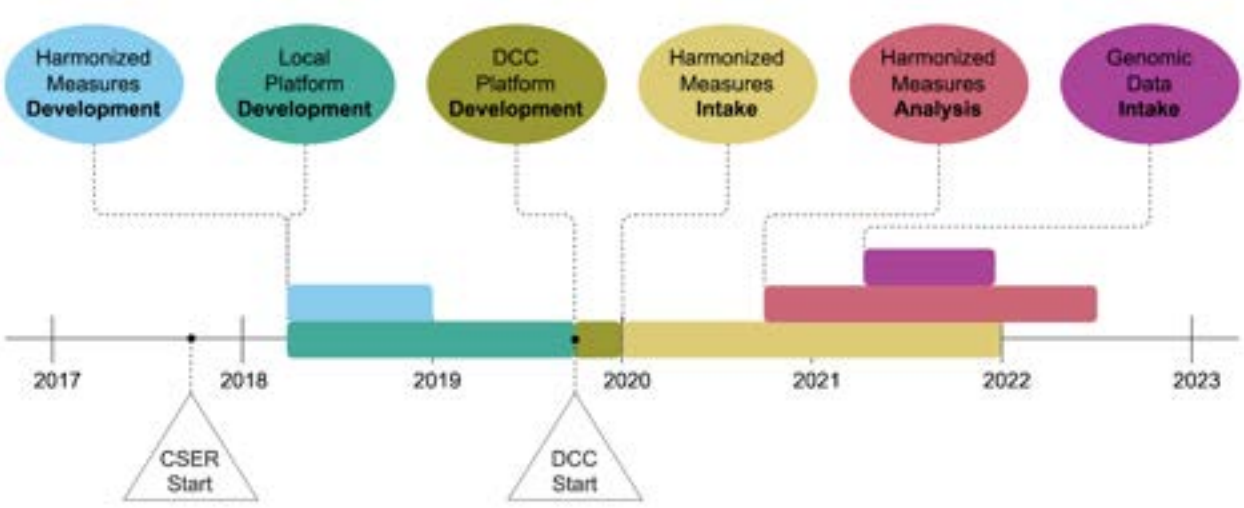


Figure 3.1. CSER Phase 2 data coordination and analysis timeline.

3.4.3 Informatics architecture

The DCC utilized a suite of informatics tools and platforms to securely store and share consortium data. These platforms included:

Local site servers and data capture tools. Data was collected and stored locally by each CSER site before it reached the DCC. Sites collected survey data using platforms including REDCap, SurveyMonkey, and custom-developed web applications. Some measures (like participant ages) were pulled directly from the EHR by sites if they were not collected through harmonized surveys. Methods for survey data storage also varied by site, with some sites using REDCap databases or similar platforms designed for clinical research, and others using relational or non-relational database management systems for optimized storage and querying of large datasets. The vast majority of survey data quality assurance (QA) and quality control (QC) was performed at CSER sites prior to DCC submission. These QA/QC measures included, but were not limited to, checks for missing data, range value checks, and outlier analyses. Genomic data was stored on servers with high disk capacity at each site or using secure cloud storage services like Amazon S3 or Microsoft Azure.

REDCap database. A secure instance of REDCap was hosted and maintained by the University of Washington ITHS and populated by CSER sites using data submission tools maintained by the DCC. All harmonized survey measures, case-level sequencing results, and participant-level sequencing metrics (e.g., aggregated case-level results) were centrally stored in REDCap and were linked at the participant level using a unique identifier called a “CSER ID.”

CSER Data Hub. The DCC used a custom R Shiny web interface called the "Data Hub" to securely exchange harmonized survey data, case and participant-level sequencing metrics, and documentation within the consortium. See "Informatics" and "Data De-Identification and Security" for more details on the architecture and security features of the Data Hub.

AnVIL storage and compute platform. The NIH-funded AnVIL consortium develops and maintains the AnVIL cloud ecosystem, which was built using Google Cloud storage and compute resources. The AnVIL is a component of the emerging federated data ecosystem paradigm in genomics [94], which is meant to improve genomic data sharing and interoperability without compromising data security or privacy. The AnVIL is authorized to share both open access (unrestricted) and controlled access (restricted) data derived from human samples [93]. Permission to access and use controlled-access data is granted on a case-by-case basis by a relevant NIH Data Access committee and is moderated through the database of Genotypes and Phenotypes (dbGaP) Authorized Access System [95]. CSER sites were required to submit their genomic Binary Alignment Map (BAM) and Variant Call Format (VCF) files, sequence, and sample metadata (e.g., reference genome build, sample source), and phenotypic data (e.g., disease codes, sex, race/ethnicity) to the AnVIL platform. Data stored in the AnVIL could then be analyzed in Terra [96], a cloud platform developed by the Broad Institute of MIT and Harvard to facilitate biomedical research data sharing and analysis.

3.4.4 Collection and aggregation of harmonized survey measures

To collect common survey measures administered at each site, the DCC developed a REDCap database using the harmonized survey measures developed by the consortium in 2018 [97] , and worked with the Data Wranglers Working Group to map site-specific data models to a harmonized data model using a three-phase approach:

Phase 1: Model. To facilitate mapping between site datasets and the DCC harmonized database, the DCC developed tab-delimited import templates and accompanying data dictionaries for six harmonized survey types **(Figure S3.2)**. All patient surveys were divided into two distinct variable sets to distinguish between surveys administered to a parent or guardian proxy of a pediatric participant and those administered to an adult participant. The DCC also developed standardized import templates and data dictionaries for participant-level and case-level genetic sequencing metrics **(Figure S3.3)**. All templates and data dictionaries were distributed as downloadable zip files on the Data Hub.

Phase 2: Map. Site analysts developed semi-automated variable mapping pipelines using the data handling software(s) of their choice (e.g., Excel, R, Python, Stata, SAS), and used these pipelines to generate harmonized datasets from the harmonized data model developed in Phase 1.

Phase 3: Upload. Staff at each site shared their harmonized datasets through a custom data upload interface on the Data Hub, which ensured that the datasets met the specifications of the models developed during Phase 1, and automatically transferred data to the DCC REDCap

database using the redcapAPI R package [98]. Initial submissions for each of the harmonized survey types and sequencing metrics occurred in 2-3-month intervals throughout 2020 and 2021. All sites repeated Phases 2 and 3 on a quarterly basis until the end of follow-up to update existing participant records, and to create records for newly recruited participants.

3.4.5 Genomic sequence data collection in the NHGRI Analysis Visualization and Informatics Lab-space

The CSER DCC facilitated the transfer of genome and exome data and metadata from site platforms to the AnVIL platform. The DCC developed harmonized metadata models in collaboration with members of the AnVIL team and other CSER members, using standards previously developed by dbGaP and The Cancer Genome Atlas (TCGA) Program as references. To facilitate the transfer of sequence data and metadata to the AnVIL platform, the DCC developed sample scripts for securely transferring data to Google Cloud buckets and made these scripts available for download on an SFTP server hosted by the University of Washington Genome Sciences department. The DCC also provided step-by-step instructions for preparing data, submitting required data ingest forms, and using sample scripts for batch sequence data transfers.

3.5 Lessons learned

Throughout 2020 and 2021, the DCC worked to meet the evolving data coordination needs of the CSER Consortium as it actively collected sequence and survey data from study participants. The following section describes the approaches that the CSER Consortium used to navigate the complexities of multi-site data sharing and offers a set of lessons learned from its data

coordination experiences **(Table 3.2)**. Lessons learned are referenced in the text using numbered identifiers (e.g., Lesson Learned 1a, Lesson Learned 1b) to exemplify connections between experiences and lessons learned.

Category	Lessons Learned
Communication	1a. Identify primary points of contact for addressing different data coordination requirements (e.g., technical infrastructure, data mapping, consortium policy) using existing communication patterns among working groups and sites 1b. Define the unique roles of different working groups in the data coordination process, and use those roles to guide inter-group communication 1c. Send periodic update emails with consolidated information (progress, resources, action items) to key data coordination stakeholders
Harmonization	2a. Provide data managers with standardized data collection instruments (templates) and specifications for mapping variables to those instruments (data dictionaries) 2b. Deploy rigorous version control methods for data coordination resources that change over time, and ensure that data managers are informed of changes 2c. Implement standardized protocols and timelines for making changes to data collection instruments 2d. Engage a multidisciplinary group of consortium members to develop and approve standardized data models
Informatics	3a. Consolidate informatics tools and resources within a secure, centralized platform 3b. Utilize available IT expertise and resources at participating institutions 3c. Prioritize security of informatics tools and disseminate security information to consortium members
Compliance	4a. Engage a multidisciplinary group of consortium members to develop a harmonized set of data sharing consent categories 4b. Use multiple data sharing specifications (e.g., institutional certifications, informed consents, data use letters) to map site-level consent groups to consortium-level consent categories
Analytics	5a. Document data quality issues and unique aspects of the harmonized dataset, and plan to distribute documentation to both current and future data users 5b. Facilitate access to onboarding resources for users of shared data analysis platforms like the AnVIL

Table 3.2. Data coordination lessons learned in the CSER Consortium.

3.5.1 Communication

As the DCC integrated with the consortium throughout 2020, additional communication channels beyond monthly Data Wranglers Working Group calls were formed to fully support the consortium's data coordination requirements. While the Data Wranglers primarily served the role of handling site-level survey and sequence data and developing computational pipelines to convert data into a harmonized format, the Project Managers provided the necessary project-level guidance to ensure that data was being shared securely and responsibly, such as tracking regulatory documents, overseeing data collection, and developing data QA/QC measures. Together, the two working groups contributed to the development of feasible and efficient DCC harmonized data upload requests and data dictionaries, assisted in coordinating responses to new data requests (including site-specific data), assisted in troubleshooting challenging data elements (e.g., consent categories), responded to requests for project-specific information, and kept track of data submission timelines **(Lesson Learned 1a)**. The DCC, Data Wranglers, and Project Managers communicated through an iterative, multi-directional feedback loop throughout the project period to ensure that all groups were equipped to fulfill their respective data coordination responsibilities **(Lesson Learned 1b)**.

Multiple working groups requested that the DCC share important data coordination updates with the rest of the consortium. To increase transparency of ongoing work and maintain an organized list of action items, the DCC sent update emails to the Data Wranglers Working Group, Project Managers Working Group, Sequence Analysis and Diagnostic Yield Working Group, and Principal Investigators (PIs) first on a biweekly and eventually on a monthly basis to

communicate important DCC activities, inform consortium members of key resources, and track new data coordination requirements. To communicate DCC activities and goals with the broader consortium, the DCC also gave regular progress updates during biweekly and monthly Coordinating Center and Steering Committee calls, respectively. These updates helped other working groups and consortium stakeholders anticipate availability of shared data, and allowed consortium members outside of the Project Managers, Data Wranglers, and Sequence Analysis and Diagnostic Yield Working Groups to regularly provide feedback and ask questions about current and planned DCC initiatives **(Lesson Learned 1c)**.

Interactions between the DCC and groups external to the consortium were largely facilitated by weekly or biweekly standing meetings, including those with AnVIL project managers and the University of Washington ITHS staff. These meetings helped the DCC receive timely assistance and feedback from technical support teams, and to communicate questions and concerns raised by CSER members. **Figure 3.2** shows the different groups involved in CSER data coordination, their responsibilities, and the types of communication that took place between different stakeholders.

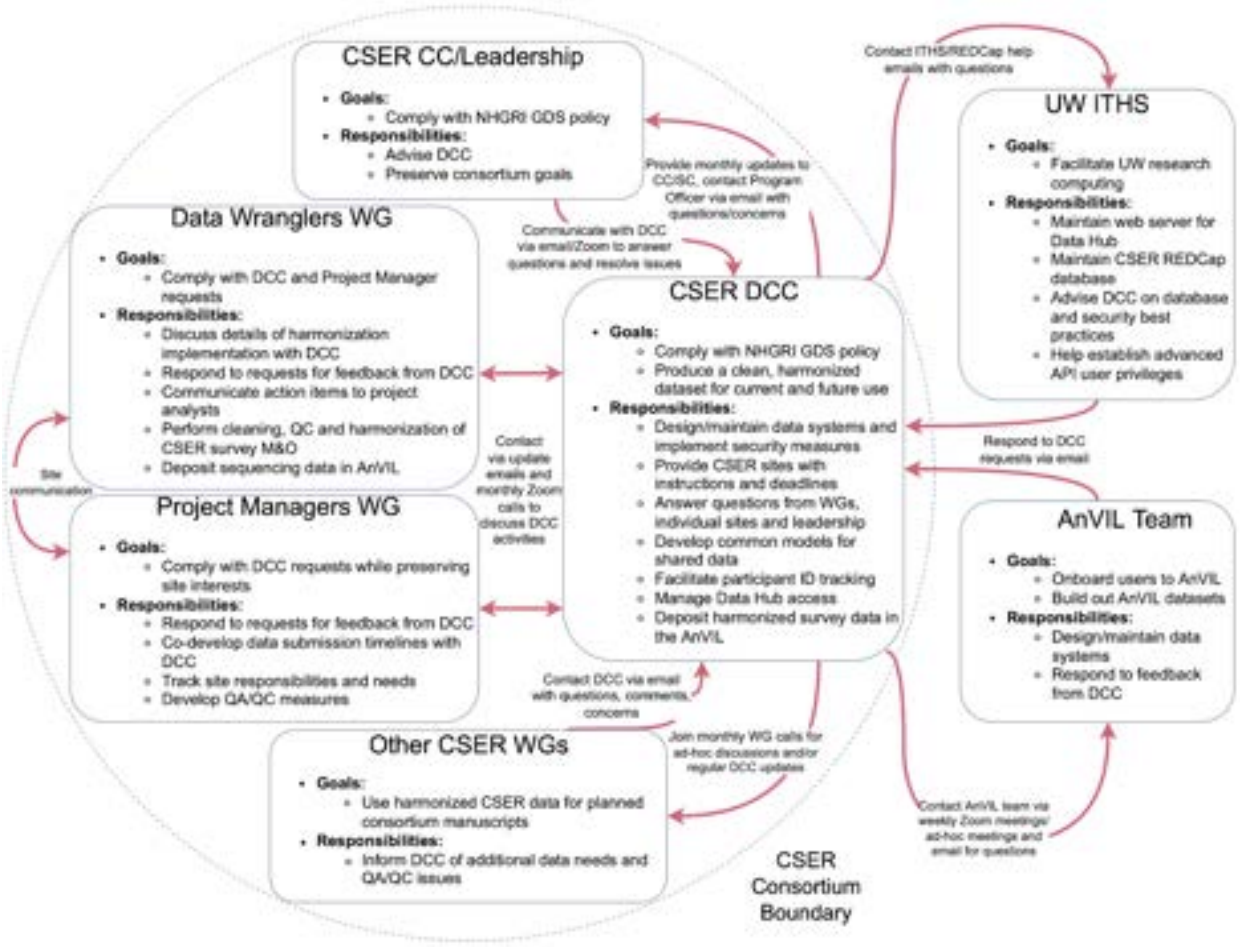


Figure 3.2. Methods of communication between groups involved in CSER data coordination.

3.5.2 Harmonization

3.5.2.a Survey data harmonization

Throughout 2020 and 2021, the DCC developed a variety of strategies to facilitate the harmonization and intake of common survey measures. As described in Goddard et al. 2020 [84], the CSER Measures and Outcomes Working Group previously led the consortium through identifying 31 survey domains across CSER projects that captured measures related to the common research aim of evaluating the personal and clinical utility of genome and exome

sequencing, while accommodating natural heterogeneity in study designs and patient populations. Common survey measures were presented to research participants in a wide variety of study environments, altered to meet the needs of individual sites, and collected and stored using different data modeling strategies. As a result, measures were harmonized across many factors, including question wording, response scales, and variable naming. While measure harmonization was important for achieving cross-site interoperability of research findings, it was also a time-consuming effort that required careful planning and use of limited resources.

In CSER's experience, achieving and sharing semantically interoperable data was far more complex than simply sharing data. As described in "Consortium structure and communication," the seven CSER projects served different patient populations, investigated unique research questions, and used different clinical sequencing interventions **(Figure S3.1)**. Furthermore, sites developed their own data collection tools before a clear set of centralized data sharing expectations was established. To reconcile differences between site-specific implementations of common survey measures, the DCC developed standardized data import templates and data dictionaries to guide harmonized survey mapping, as described in "Collection and aggregation of harmonized survey measures" **(Lesson Learned 2a)**. The complexity of this process is illustrated in **Figure 3.3**, which depicts the mapping process for a single variable in the Communication Satisfaction measure from the first Patient Post-Return of Results (RoR) survey. By the end of the survey mapping phase for all six harmonized surveys and two sequencing metric reports, sites had implemented mapping logic for over 1100 variables.

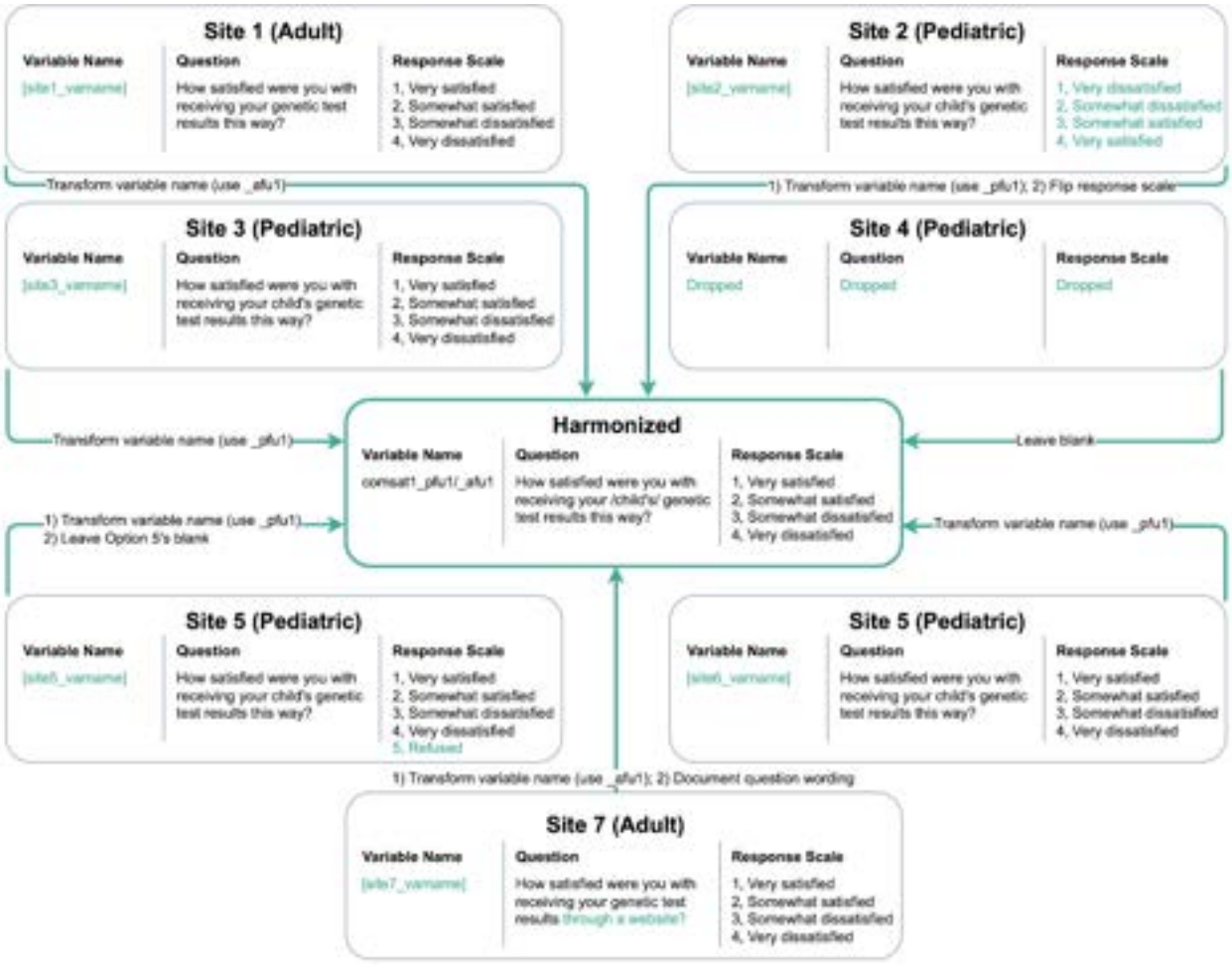


Figure 3.3. Sample harmonization process for one variable in the Communication Satisfaction measure, across all seven CSER projects. To map participant responses to the Participant Post-Return of Results (RoR) Follow-Up #1 harmonized import template, each site created a local mapping between the site-level variable name and the harmonized variable name (comsat1_pfu1 for pediatric surveys, comsat1_afu1 for adult surveys) and documented any differences in question wording. Some sites were also required to map alternate response encodings to the harmonized response scale. For example, Site 2 administered the question with a reversed response scale (where 1 =' Very satisfied' on the harmonized scale, and 4 = 'Very satisfied' on the site scale), and modified harmonized responses accordingly (1 = 4, 2 = 3, 3 = 2, 4 = 1). Similarly, Site 5 administered the question with an additional response option, and was instructed to map these responses to blank values (5 = ' ').

The primary goal of the survey mapping phase (Phase 2) was for each site to develop a semi-automated pipeline that could be used to quickly update harmonized datasets with new or modified data on a quarterly basis. However, the pipeline development process was complex and

time-intensive for each site and involved frequent updates to mapping logic. Updates included relatively simple changes like variable name modifications and harmonized response scale adjustments, but also included more complex updates like the addition of new variables that were deemed necessary for accurate, reliable, and secure downstream analysis of harmonized data **(Table S3.1)**. For example, the elapsed time since RoR variable was first proposed during a Data Wranglers Working Group meeting in July 2020, when it was discovered that not all participant or provider follow-up surveys could be administered or collected within the harmonized time frames specified **(Figure S3.2)**, and that having more granular elapsed time data could improve the accuracy of downstream analyses. A placeholder variable was developed and then iteratively refined before seeking Steering Committee and IRB approval. The finalized variable required sites to indicate the number of weeks post-RoR that a given survey or measure was administered to each participant. Sites were then tasked with implementing new mapping logic for as few as three, and as many as 25 new harmonized variables, depending on whether follow-up measures were administered according to the harmonized survey groups **(Figure S3.4)**. While not all change requests were this lengthy or involved, they cumulatively resulted in high demands on Data Wranglers and Project Managers throughout the harmonized measure mapping process.

To minimize burden placed on Data Wranglers and Project Managers due to change requests and to maximize transparency, the DCC maintained a "Change Log" page in the Data Hub, which listed the changes made between import template and data dictionary versions. During the last quarter of 2020, the DCC began distributing quarterly checklists that documented all new, removed, and modified variables for each quarterly data resubmission, and made these documents available for download on the Data Hub **(Lesson Learned 2b)**. Beginning in January

2021, the DCC also implemented a new "Change Request Schedule," which specified time intervals during which consortium members could make change requests and blocked off two-month intervals before each quarterly resubmission during which site data analysts could modify mapping pipelines without having to address incoming change requests. These strategies helped manage the computational and organizational burden of maintaining harmonized mapping pipelines, but nonetheless did not eliminate all tensions between site-level burden and consortium-level data sharing expectations **(Lesson Learned 2c)**.

3.5.2.b Sequence metadata harmonization

The AnVIL replaced dbGaP as the primary repository for NHGRI-funded genomic, phenotypic and survey datasets in mid-2019, during the CSER Phase II funding period [93]. While dbGaP provided data submitters with standardized templates and instructions for submitting sequence data and metadata to the platform, the AnVIL consortium was still developing standards when CSER commenced submissions. As a result, the CSER DCC was tasked with developing standardized metadata models that captured the necessary details without placing unreasonable burden on CSER sites. In mid-2020, the DCC convened a subgroup of CSER investigators (called the "Sequence Metadata Subgroup") with expertise in sequence data analysis to develop a harmonized set of sequence and sample metadata fields **(Lesson Learned 2d).** Prior to the first subgroup meeting, the DCC compiled a list of candidate variables using a combination of the dbGaP and TCGA standards. The DCC presented these variables to the Sequence Metadata Subgroup to assess the feasibility and descriptiveness of the proposed fields. Once the model was approved by the Sequence Metadata Subgroup, the Data Wranglers Working Group, and the

AnVIL team, the DCC developed the relevant import templates and data dictionaries and made these documents available for download on the Data Hub **(Table S3.2)**.

3.5.3 Informatics

The CSER DCC used the Data Hub platform to host data coordination resources in a centralized, secure, and easily accessible location. The Data Hub made it possible to link multiple data management platforms with one another **(Figure 3.4)** and to quickly distribute version-controlled resources to Data Wranglers and Project Managers **(Lesson Learned 3a)**. To develop and maintain the Data Hub, the DCC harnessed available information technology expertise and resources at the University of Washington ITHS **(Lesson Learned 3b)**. These resources took the form of one-on-one meetings and email exchanges with ITHS personnel, and computing resources for hosting the Data Hub website. However, they also relied heavily on informatics expertise within the DCC to develop the application itself and to provide troubleshooting support to CSER sites. Sample screenshots of the Data Hub user interface are shown in **Figures S3.5-S3.9**.

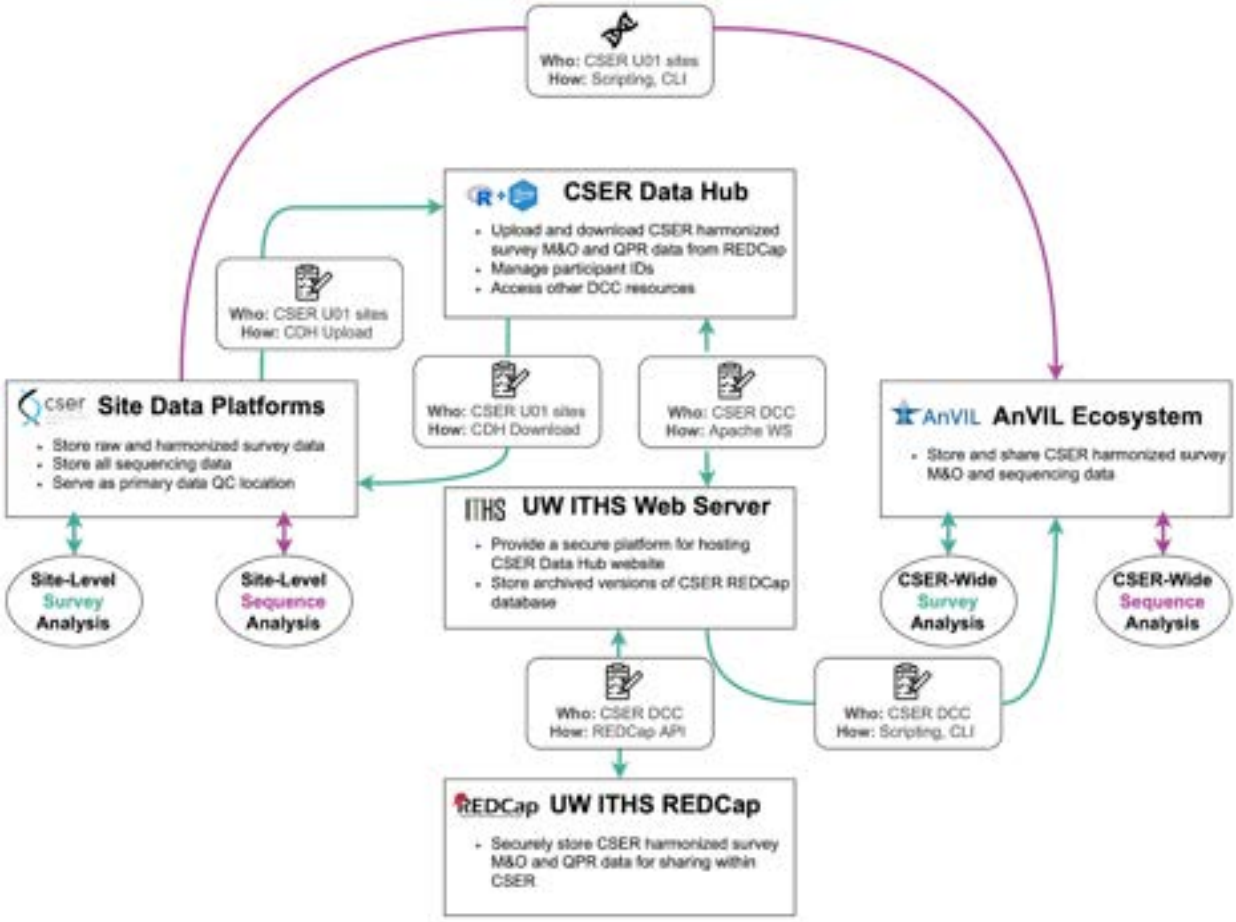


Figure 3.4. Movement of harmonized survey data (green) and sequence data (purple) between CSER data platforms. Abbreviations: CDH – CSER Data Hub; CLI – Command Line Interface; DCC – Data Coordinating Center; M&O – Measures and Outcomes; QPR – Quarterly Progress Report; WS – Web Services.

3.5.4 Data de-identification and security

Before submitting harmonized data to the Data Hub or sequence data to the AnVIL, all CSER sites were required to remove personally identifiable information (PII) from their datasets in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule [99]. To retain syntactic integrity of free text, sites were asked to redact all instances of PII and replace them with the category of identifier within brackets (e.g., "[date]," "[name]"). Measures

were also taken to protect local study identifiers for each participant. For each new record in the harmonized database, a unique CSER ID was randomly generated and linked with the participant's local study ID. Mappings between CSER IDs and local IDs were then stored within the DCC REDCap database, accessible only to members of the site from which each CSER ID originated.

Although the DCC took steps to prevent identifiable information from being uploaded to its platforms, multiple layers of security were built into the DCC informatics architecture to protect data in the unlikely event that sensitive, identifiable information were to be uploaded to a DCC platform **(Lesson Learned 3c)**. First, the Data Hub was deployed on a secure web server hosted by the University of Washington ITHS. All requested connections from client web browsers were established using the Apache HTTP Server software, and ITHS required that all hosted web applications establish encrypted connections between the server and the client browser. Second, all Data Hub users were required to log in to the Data Hub using University of Washington credentials, which were sponsored by the DCC team. Third, the Data Hub was designed in alignment with standards put forth by the HIPAA Security Rule, including the use of activity logs, password-protected access, automatic password timeout, and HIPAA-compliant data storage in REDCap. And fourth, the DCC developed standard protocols for removing records of participants that had withdrawn consent for sharing data, and continuously updated and distributed a list of CSER IDs that should be removed from previously downloaded datasets.

3.5.5 Consent group harmonization

CSER did not have a central study Institutional Review Board (IRB), and thus relied on IRBs at each CSER site (and in some cases additional IRBs at subsites) and the University of Washington—the Coordinating Center home institution—to make decisions about appropriate data sharing. All site and Coordinating Center PIs signed a Data Use Agreement in early 2019 detailing the data sharing terms between participating institutions in CSER, and the DCC used this document to broadly define the terms of data sharing across CSER sites and beyond the consortium.

While the use of local IRBs facilitated the implementation of varied clinical study designs across diverse patient populations at each site, the lack of a central CSER IRB also resulted in substantial heterogeneity in how data sharing consent groups were defined across CSER sites. Because the dbGaP Authorized Access System typically inherits consent group specifications from study Institutional Certifications [100], the DCC first surveyed all Institutional Certifications to determine if they sufficiently represented site-level consent groups. Following conversations with the CSER Project Managers, the DCC determined that while the Institutional Certifications provided high-level guidelines for how study data could be shared with non-CSER investigators, they did not fully represent subtleties of the permissions given by participants for sequence and/or survey data sharing during informed consent. For example, several CSER sites allowed participants to opt-out of broad data sharing (e.g., General Research Use or Health/Medical/Biomedical Research) and to restrict sharing to specified investigators, while other sites required study participants to consent to broad data sharing if they were to enroll in

the study. As a result, harmonized consent categories had yet to be developed when CSER sites were otherwise ready to share data.

To develop consortium-wide data sharing consent categories, the DCC convened a multidisciplinary "Data Access Subgroup" of data analysts, project managers and data ethicists to discuss key considerations and requirements for consent harmonization **(Lesson Learned 4a)**. The subgroup met twice over a period of two months in mid-2020 to develop a plan for mapping site-level consent categories to harmonized consent groups. Using a combination of standard NIH consent groups (e.g., General Research Use, Health/Medical/Biomedical research) and data use limitations (e.g., local IRB approval required, publication required) [101] indicated in the site Institutional Certifications, and more restrictive data use limitations gleaned from site-specific informed consents (e.g. CSER-only access), the Data Access Subgroup developed eight harmonized consent groups for survey and sequence data types **(Table S3.3; Lesson Learned 4b)** [102]. The Project Managers and Data Wranglers mapped participant-level consent groups to harmonized consent groups and submitted these consent assignments to the Data Hub in early 2021. These groups were used to determine how sequence and survey data could be stored and shared with non-CSER investigators in the AnVIL platform.

3.5.6 Cloud data sharing

The movement of data storage and computation to cloud platforms like Google Cloud, Amazon Web Services (AWS) or Microsoft Azure is widely regarded as a necessary next step in the field of genomics, given the large volume of genomic data generated daily, the increasing sophistication and scalability of cloud resources, and the need for extensive collaboration in

genomic research [103]. While the goal of this transition is to maximize the utility and impact of human-derived samples and phenotypic data, cloud technology is still relatively novel to most academic institutions—which have historically used privately managed, secure servers to store and process genomic data—and to many research participants contemplating broad data sharing. While the NIH has previously released guidance on best practices for cloud data sharing [104], the technical aspects of data security and administrative aspects of data privacy in the cloud are unfamiliar to many investigators. As a result, many institutions approach new cloud data sharing requirements with caution [105]. The CSER Consortium responded to cloud data sharing requirements by reviewing informed consent documents at each site and ensuring that research participants gave their consent to share data in NIH controlled-access repositories other than dbGaP. The DCC also collaborated with the AnVIL team to compile security documentation into a single resource that sites could use to personally assess the security of datasets submitted to the platform, particularly those restricted to use within the consortium. Consistent communication between the AnVIL team, NIH staff, the DCC, and CSER Working Groups was essential for building consortium-wide trust in this new technology, and for ensuring the ongoing privacy and security of de-identified genomic, phenotypic, and survey data in the new era of cloud storage and computing.

3.5.7 Analytics and documentation

3.5.7.a Harmonized survey data reliability

Given the heterogeneity in how common survey measures were modeled and administered at each CSER site, the DCC developed strategies to document differences in site-level measure

implementations. The DCC initially used separate Google Sheet data dictionaries for each site to document unique implementations of common measures. These site-level data dictionaries were then compiled into a single "Adaptation Dictionary," which documented the adaptations made to each harmonized variable across all CSER sites and was designed to highlight the degree to which each measure might be subject to data integration or reliability issues during analysis. To facilitate quick assessments of data reliability, the DCC implemented a cover sheet within the Adaptation Dictionary that indicated to what extent each measure was adapted **(Figure S3.10)**. Step-by-step instructions were also included on the first tab of the dictionary to help investigators consider how adaptations might affect their analyses. To increase adoption within CSER, the DCC provided a link to the Adaptation Dictionary on the Data Hub and advised CSER members to reference the dictionary before attempting any cross-site analyses. The Adaptation Dictionary was intended for use by investigators both within and beyond CSER and was designed to be shared on platforms like the AnVIL to enhance the usability of CSER data for future research. In addition to documenting adaptations to harmonized measures, the DCC developed a centralized help document for current and future users of CSER data. The document contained descriptions of all CSER projects, explanations for how key variables were harmonized, rationale for and descriptions of items that were added to the harmonized measures (e.g., vital status, survey completion dates), and FAQs related to database structure and use **(Lesson Learned 5a)**.

The DCC also implemented several automated, on-demand variable calculation features in the Data Hub to generate measures that could be programmatically derived from the harmonized measures. The CSER "Underserved Framework," developed by members of the CSER Ethical, Legal, and Social Implications and Diversity Working Group, employed different combinations

of demographic factors (including language, income, insurance status, residence, race, and ethnicity) to form nine distinct risk groups, indicating either direct barriers to medical care access or social factors that might indirectly impede access. Using the Data Hub download tool, consortium members could elect to download automatically calculated Underserved Framework variables along with documentation about how each variable was calculated.

3.5.7.b Using the NHGRI Analysis Visualization and Informatics Lab-space platform for analysis

The AnVIL platform seeks to enable users with scalable compute power, large-scale data access, and shared resources for analysis [77]. The AnVIL analysis environment was built using the Terra/Google Cloud platform, so users familiar with this system may experience shorter onboarding periods. Data exploration and analysis are supported through the use of Jupyter notebooks [106] and RStudio [107], which are commonly used tools in the field of data analytics and statistical analysis. AnVIL also supports genomics tools such as Galaxy [108] for users with less experience in programming who are interested in genomic analysis, and provides access to standard command line tools like GATK [109] to facilitate advanced data processing.

Although the potential benefits of using a platform like the AnVIL for sequence data storage, sharing, and analysis are numerous, the unfamiliarity of the platform may limit the ability of investigators to anticipate exactly how data might be shared and/or used and may therefore make early-stage decisions about data modeling and sharing difficult. For example, the automatic linkage of survey, phenotypic, and sequence data in a shared cloud workspace is a novel concept, and investigators will undoubtedly need to make challenging decisions regarding the best way(s)

to prepare, share and utilize such data. Large clinical genomics research consortia like the eMERGE Network and the Implementing Genomics in Practice (IGNITE) Consortium will likely face similar challenges to those experienced by CSER, and the AnVIL platform will be a valuable space for investigators from all disciplines to unite and support one another in this new generation of genomic data sharing and analytics **(Lesson Learned 5b)**.

3.6 Discussion

After dedicating much time and effort to developing and implementing strategies for harmonizing and coordinating consortium-wide datasets, the CSER Consortium is well-positioned to contribute an impactful and wide-reaching dataset to facilitate research in medical genomics. While the DCC developed tailored strategies to facilitate CSER data coordination, the principles behind these strategies are applicable to other research settings in which data are pooled from heterogeneous sources. **Table 3.3** lists 11 overarching needs and recommendations for conducting multi-site data coordination at the levels of Planning, Communication, Informatics, and Data Analytics. The following section explores these recommendations through the lens of four thematic domains that emerged from this work: 1. Transparency and translation; 2. Team morale, collaboration, and trust building; 3. Iterative design; and 4. Data governance. We also offer guidance on how these recommendations might generalize to projects of different sizes with diverse data coordination needs and capabilities.

Category	Needs	Recommendations
Planning	Clear expectations for internal and external data sharing	1. Build data sharing expectations into expected scope of work in funding announcements **(NIH)**
	Sufficient financial resources and time for data coordination	2. Budget for data coordination, management, and reporting at individual research sites **(NIH)**
	Integration between DCC and consortium	3. Establish DCC at start of funding period, if not before **(NIH, DCC)**
Communication	Consolidation of communication channels	4. Consolidate lines of communication from DCC to working groups, and assign action items appropriately **(DCC, Sites)**
	Technical specifications for data sharing	5. Maximize transparency of data coordination expectations and resources **(NIH, DCC)**
	Efficient use of diverse expertise available within the consortium	6. Facilitate translation of critical information between stakeholder groups **(DCC)**
Informatics	Consolidation of informatics platforms for data coordination	7. Deploy a secure, centralized web resource for data coordination **(DCC)**
	Flexibility in response to unforeseen events and changing analysis plans	8. Build flexibility into central databases and data management software **(DCC)**
	Correct implementation of site-level security and privacy agreements	9. Prioritize data privacy and security during platform design **(DCC)**
Analytics	High-quality and reliable data from heterogeneous sources	10. Provide clear and detailed documentation of shared data resources **(DCC, Sites)**
	Integration of research and clinical practice; Enhanced protection of data from vulnerable populations	11. Document approaches to data governance **(DCC, Sites)**

Table 3.3. Recommendations for consortium data coordination. Text in **bold** indicates which entities should be responsible for each recommendation.

3.6.1 Transparency and translation

Clear and consistent communication on the part of research leadership and data coordination teams should be a high priority, from project conception to completion. Ideally, Funding Opportunity Announcements (FOAs) issued by funding agencies should plan for and communicate data sharing expectations **(Planning, Recommendation 1)** to allow research sites to budget and plan for data coordination activities **(Planning, Recommendation 2)**. When possible, the DCC should be involved in the research planning phase and should continually facilitate conversations surrounding data collection, QA/QC, reporting, modeling, and sharing, so that research sites are sufficiently prepared to participate in data sharing at all project stages **(Planning, Recommendation 3)**. Given the availability of appropriate experience and expertise, the DCC may act as a stakeholder proxy across research sites and working groups and facilitate data coordination conversations and decision-making. As a liaison between project stakeholders, the CSER DCC was ideally positioned to assume the role of "translator" and facilitate adaptive communication between groups with unique roles and areas of expertise **(Communication, Recommendation 6)**. Translation should also take place between the consortium and the greater scientific community since data in controlled-access repositories is expected to have a lifespan beyond the consortium from which it originates. As such, clear documentation of shared data and resources should be developed to encourage appropriate data use, and alert users to any unusual or unique data elements prior to analysis **(Analytics, Recommendation 10)**.

The translator also has a responsibility to communicate data needs centrally and concisely. Separate lines of communication that request different (but related) data coordination action

items should be avoided, and requests should instead be aggregated and contextualized with one another **(Communication, Recommendation 4)**. The expected contributions of stakeholders to different data coordination activities should also be transparent, both to increase task accountability and to assess the equitable distribution of tasks across the consortium **(Communication, Recommendation 5)**. Stakeholder communication should be a two-way, responsive process in which DCC processes are adjusted in response to stakeholder feedback, and vice versa.

3.6.2 Team morale, collaboration, and trust building

An often-overlooked aspect of data coordination is the importance of interpersonal relationships and team morale within and between stakeholder groups. Making expectations transparent and achievable is critical to demonstrating respect and appreciation for team members' time and efforts **(Communication, Recommendation 5).** Similarly, giving team members the space and time to regularly voice ideas and concerns to the leadership and data coordination team is essential for maintaining a culture of mutual respect and understanding across stakeholder groups. Decisions that will impact research workflows and workloads of consortium members should be made mutually and transparently, both to demonstrate respect for one another's time and to avoid situations in which stakeholders must retrospectively address issues introduced earlier in the research process due to a lack of communication or collaborative planning. Strengthening these interpersonal relationships is essential for building a culture of trust within the research team and facilitating a positive data sharing experience.

3.6.3 Iterative design

Access points to important data coordination tools and resources should be consolidated to minimize burden placed on sites and improve resource transparency **(Informatics, Recommendation 7)**. Each resource should also be designed to withstand frequent modifications, both on the database and user interface ends, to accommodate inevitable changes in consortium needs **(Informatics, Recommendation 8)**. Building iterative design principles into the platform development process is far more effective at achieving a useful and usable system than deploying a static, pre-designed system [110]. Based on the Gould & Lewis principles of design [111], system development should involve: 1. Early focus on end-point users; 2. Early deployment and usability testing; and 3. Iterative system design. Employing these principles in practice will help end-users identify critical features and potential issues on a rolling basis and ensure that the resulting data coordination system is designed appropriately for the intended user base. However, platform security should remain the highest priority throughout the design process, and design decisions should never be made at the expense of security features **(Informatics, Recommendation 9)**.

3.6.4 Data governance

While there is an understanding among scientific communities worldwide that sharing research data is a necessary component of scientific progress, the mechanisms for protecting against potential harm while maximizing usefulness are not well-defined [112]. These two aims are often in tension and lend themselves to diverse data governance strategies across research projects within and between scientific disciplines. In genomics research studies, data governance frameworks that promote scientific progress should: 1. Enable data access; 2. Follow national

laws and international agreements; 3. Support appropriate data use; 4. Promote equity in the access and analysis of data; and 5. Use data for public benefit [71]. However, when operationalizing data governance frameworks within research consortia, major tensions exist in the areas of data access control, de-identification, and consent models. Combined with the technical challenges of cleaning, harmonizing, and annotating datasets, these tensions contribute to a disconnect between the intent to share data and real-life data sharing practices [113]. While it is tempting to trace this disconnect to a lack of clear guidance from national agencies and project funders, guidelines like those found in the NIH Genomic Data Sharing policy are left intentionally vague to account for vast contextual differences between research projects. To develop a reusable set of data governance guidelines that can accommodate different research settings and contexts, it may therefore be useful for research projects to document their own approaches to the five components of effective data governance frameworks listed above, and for funding agencies to then develop comprehensive guidelines that accommodate the unique data governance requirements of diverse research settings (**Analytics, Recommendation 11**).

One important tension that arises in clinical research is the need to accommodate varying data governance expectations across clinical and research settings, particularly for participant privacy and informed consent for data sharing. For example, the Federal Policy for the Protection of Human Subjects (also known as the "Common Rule") is a set of federal regulations that dictates requirements for the ethical management and distribution of data collected from human research subjects, while the HIPAA Privacy Rule is a federal law that enforces standards for the protection of patient medical data. While these regulations are intended to complement one another in clinical research settings, the details of how each set of rules should be applied to the

operational components of a data governance strategy are not well-defined, leading to potential gaps in data protections [114]. The US Department of Health & Human Services itself recognizes that "institutions, IRBs and investigators are frequently faced with applying both the Common Rule and the HIPAA Privacy Rule" when making decisions about clinical research protocols, since there are currently no formalized guidelines for merging these requirements [115]. The inclusion of genome and exome sequencing in clinical research further complicates questions of subject and biospecimen identifiability, for which guidance from the Common Rule and HIPAA is limited [116,117].

In the case of informed consent for data sharing, the details and implications of policies that govern data protections should be made transparent to clinical research participants who are asked to consent to broad data sharing, but researchers and policymakers themselves are still grappling with these details. For example, on the FAQ page of the NIH Genomic Data Sharing policy description, a common perception among genomic researchers is that the "NIH requires that investigators obtain consent for broad data sharing and that the participant is disqualified from participating in the study if consent is not obtained," although the NIH clarifies on the same page that this was not the intent of the policy [118]. In addition to questions of appropriate data sharing, the appropriate breadth and depth of information communicated during the informed consent process is challenging to pinpoint, given that it is extremely difficult—if not impossible—to predict exactly how genomic information will be used by researchers in the future. There is an even greater urgency for clarity in genomic data sharing consent procedures for patient populations that are historically marginalized and disadvantaged by biomedical research and medical practice [119]. For example, there is concern among US Indigenous

communities that participating in genomic research and sharing genomic data may lead to inappropriate use of that data in the future, leading to imbalanced societal benefits or even harm to those communities [120]. Data governance frameworks that support paradigms like data sovereignty for marginalized populations and dynamic consent procedures may help mitigate some of the risks posed by evolving consent details in medical genomics research [121]. Other suggestions for addressing misuse concerns include following documented Indigenous engagement practices, understanding worldviews unique to different Indigenous communities, and practicing complete transparency in all research partnerships with Indigenous communities [122]

3.6.5 Generalizability of recommendations

While these recommendations were designed to generalize to other multi-site research projects, we recognize that smaller or less well-funded projects may not be able–or even need–to implement all of the recommendations. For example, a smaller project with two homogenous research sites (e.g., similar participant populations, research aims, and institutional policies) may not need to establish a formal DCC **(Recommendation 3)** or deploy a multi-user web application **(Recommendations 7, 8, and 9)**. However, the same project would still benefit from having a dedicated group of investigators to oversee data coordination, encourage communication, and facilitate documentation **(Recommendations 4, 5, 6, 10, and 11)**. While the costs of these recommendations pale in comparison to funding an entire DCC or developing a web application, they are nontrivial. A "bare bones" implementation of a data coordination core would require part-time participation of at least one investigator at each site with data science expertise (similar to the CSER Data Wranglers), one investigator at each site with detailed knowledge about the

study (similar to the CSER Project Managers), and one central coordinator to facilitate communication and track progress. As funding agencies increasingly expect research projects to contribute high quality, harmonized data to public repositories, funders and researchers alike should recognize these dedicated groups as an essential component of any research program and provide appropriate budget support accordingly **(Recommendations 1 and 2)**.

Research projects should consider how the size, complexity, and privacy considerations of their anticipated datasets impact the relative importance of different data coordination needs (see the "Needs" column in **Table 3.2**) and implement recommendations accordingly. While dataset factors are partly influenced by the number of sites involved in a project, they are not defined by project size. For example, a project with 2 sites collecting 100 data types (variables, file types, etc.) might have a greater need for more robust data coordination tools than a project with 100 sites collecting 2 data types. Similarly, smaller consortia collecting data on a large number of participants at each site may have more complex needs than larger consortia collecting data on a small number of participants. However, as the CSER Consortium experienced, data coordination needs evolve as the project evolves. Projects should periodically re-evaluate how well their current approaches are addressing their needs and seek additional funding and/or personnel to help implement more rigorous coordination approaches as needed.

Finally, while these recommendations are most translatable to NIH-funded projects within the US, the basic principles still apply to non-NIH funded and multi-national projects. Other types of projects may have data sharing expectations and policies that differ considerably from those of NIH-funded projects but using well-reasoned communication and informatics practices is

ubiquitously beneficial for managing heterogeneous datasets. For example, a 2017 report by the Organisation for Economic Co-operation and Development identified common challenges across 32 international research data networks, including the need for clear roles and responsibilities, transparency, mutual respect, and clear data governance plans [123]. However, multi-national consortia like the Global Enteric Multicenter Study (GEMS) and the International Cancer Genome Consortium (ICGC) have cited additional challenges—like navigating differences in language, culture, and data transfer policies between countries—that the current recommendations do not address [30,80]. While privately funded projects may not be required to share data as a condition of funding, they will likely receive requests from peer-reviewed journals to share data before publishing. In this way, the evolving culture of data transparency within the scientific community itself necessitates data coordination.

3.6.6 Applications to the value-creating learning health system framework

The value-creating LHS framework, developed by Menear et al. (2019) [89], explicitly acknowledges the interconnectedness of social and technical factors in the LHS model This framework combines multiple LHS frameworks into a transtheoretical model that describes how different stakeholders can work together to achieve higher value care at lower costs. While this framework was originally developed to reflect the core values of the Canadian healthcare system (participatory leadership, equity, solidarity, inclusiveness, scientific rigor and personalization), these core values have long been a necessary component of healthcare reform internationally [124]. We therefore propose that this framework can be reasonably applied in the case of US healthcare reform. The framework builds upon the concept of rapid learning cycles, which consist of three core processes: 1. Converting data to knowledge; 2. Using knowledge to

influence care practices while documenting the impacts of new care practices on health outcomes; and 3. Generating new data from reformed healthcare practice [125]. The authors of the framework base their definition of "value" on the quadruple aim of healthcare (enhanced patient experience, improved population health, reduced costs, and improved working conditions for healthcare providers) [126], and argue that a variety of socio-technical factors should be considered throughout the iterative learning process in order to successfully generate value from an LHS.

Each of the core values and pillars of the value creating LHS framework emerged organically during the CSER data coordination process, highlighting the generalizability of the framework to different types of clinical research environments and to the LHS model. The natural alignment between the LHS pillars and our recommendations also underscores the importance of data coordination in both clinical research consortia and LHS-aligned clinical settings. **Table 3.4** shows how different recommendations from CSER data coordination can be applied to the Core Value and Pillar components of the value creating LHS framework.

Framework Component	Core Value or Pillar	Recommendation(s)
Core Values	Adaptability	8
	Cooperative and participatory leadership	1, 3
	Equity	11
	Inclusiveness	6
	Open innovation	6
	Person focused	4, 6, 11
	Privacy	9
	Scientific integrity	10
	Shared accountability	1, 3, 4, 6
	Solidarity	4, 6
	Transparency	1, 5
Pillars	Scientific	1, 2
	Social	3, 4, 6
	Technological	7, 8
	Political	1, 2, 3, 11
	Legal	9
	Ethical	11

Table 3.4. Recommendations applied to the Core Values and Pillars of the value-creating learning health system framework.

3.7 Limitations and future work

While the recommendations from this work are expected to be applicable to different settings in which data coordination is a key activity, the artifacts and experiences that informed those recommendations are still specific to the CSER Consortium. Integrating artifacts and experiences

across clinical research consortia could be useful for improving the generalizability of recommendations to different research environments with variable participant populations, study personnel, and financial resources. Additionally, while it is assumed that the recommendations can be applied to LHS settings based on the alignment between LHS goals and the goals of clinical research consortia, this work is not a complete substitute for similar analyses in actual LHS settings. Future studies of clinical research data coordination efforts should expand on and adapt the recommendations to LHS-aligned environments.

3.8 Conclusion

The artifact analysis methods used in this work uncovered the cultural aspects of data sharing that are essential for enabling the widely-sought "transition towards a culture of biomedical data sharing" (Piwowar et al. 2008, p. 1315) [82]. Data coordination is not simply a matter of algorithms and automation, but also of human communication, translation, mutual respect, and autonomy. These principles are particularly important to operationalize for projects that straddle the research-clinical interface, where the ethical and political aspects of data sharing are often in tension with one another. Identifying common challenges and new solutions to data coordination that are grounded in the experiences of clinical research projects is a key first step in defining community standards and expectations. The lessons learned and recommendations identified in this work reinforce previously identified challenges in clinical research projects and provide both context-specific and generalizable solutions that can guide the development of best practices moving forward. In the next chapter, we will transition from a researcher-focused view of data coordination to a clinician-focused view of knowledge generation. Medical geneticists operate at the cutting edge of clinically applied genomics research and can help identify the barriers and

enablers of moving coordinated clinical and research data into the realm of clinically motivated discovery.

CHAPTER 4: MEDICAL GENETICIST PERSPECTIVES ON CLINICALLY EMBEDDED GENOMIC DISCOVERY (AIM 2)

4.1 Introduction

In addition to coordinating clinical and genomic data for research purposes, generating new knowledge from clinical data is a central process in the LHS model [7]. Knowledge generation has also been identified by the NHGRI as a key part of developing "virtuous cycles in human genomics research and clinical care" (Green et al. 2020, p. 689) [9], in which new genomic discoveries are rapidly integrated into healthcare systems and outcomes data are used to assess the utility of genomic medicine and ultimately improve disease diagnosis and management. However, the barriers, drivers, and approaches to generating new knowledge in an LHS have been sparsely examined, especially in the context of important sociotechnical and ethical factors that affect research and clinical environments differently [14]. While the foundational characteristics of an LHS have been defined by the IOM, little has been done to assess the feasibility of implementing clinically-based discovery programs given the challenging realities of the US healthcare system [7]. Additional technical, ethical, and social complexities of genomic data collection and analysis are expected to make knowledge generation in GLHSs even more challenging to execute [8,31]. Understanding the perspectives of those who work closely with genetic information in clinical environments is an important step in assessing the feasibility of generating new genetic knowledge from clinical data, and for examining the implications of conducting genomic research in clinical environments.

In this aim, we explore the perspectives of board-certified medical geneticists on integrating genomic discovery research with clinical care. Using constructivist grounded theory methods, we identify perceived drivers and barriers for GLHS discovery, and offer an *a priori* conceptual model for understanding the technical, social, and ethical forces that influence the shifting boundaries between research and clinical care in genomics.

4.2 Related Work

4.2.1 The genomics-enabled learning health system

The concept of a "rapid-learning health system" was originally proposed by Lynn Etheredge in 2007 [127] as an approach for improving evidence-based medical care, advancing clinical research, and maximizing the value gained from healthcare spending in the US. Etheredge pinpointed the EHR as the driving technology for rapid learning in healthcare because it offered an inexpensive, queryable, clinically representative, and fast alternative to standard methods for gathering data in biomedical research. Similar arguments were made at a two-day workshop held by the IOM Roundtable on Evidence-Based Medicine in July 2007 called "The Learning Healthcare System," where participants acknowledged that "the nation needs a healthcare system that learns" in order to "[get] the right care to people when they need it and then [capture] the results for improvement" (IOM 2007, p. 3) [128]. During this workshop, participants identified several pressing needs of the LHS model (**Table 4.1**), and acknowledged that large, structural changes in the ways knowledge is developed and managed in clinical research are necessary to realize the full potential of an LHS. The proposals from this workshop were later formalized into a book published by the IOM Committee on the Learning Health Care System in America in

2013 called *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America* [7].

Need	Description
Adaptation to the pace of change	Adaptation to rapid developments in both technology and the information those technologies generate
Stronger synchrony of efforts	Coordination of responses to new knowledge to limit conflict and/or confusion
Culture of shared responsibility	Shared responsibility between patients, providers in the evolution of new knowledge
New clinical research paradigm	Better integration of clinical research and clinical practice
Clinical decision support systems	Information support for clinicians
Universal electronic health records	Comprehensive EHRs with all available capabilities
Tools for database linkage, mining, and use	Tools for searching and interpreting large, structured databases, and for linking multiple databases
Notion of clinical data as a public good	Resolutions surrounding the idea of data as a "proprietary good" and concerns about patient privacy
Incentives aligned for practice-based evidence	Aligning the incentives of research, clinical practice, and Information Technology (IT) to promote learning
Public engagement	Engagement of patients and healthcare professionals in generating and disseminating new evidence
Trusted scientific broker	A trusted entity that can guide movement and priorities in clinical research integration
Leadership	Guidance for developing and executing visions, strategies, and actions

Table 4.1. Needs for a learning health system identified during the 2007 Institutes of Medicine Roundtable on Evidence-Based Medicine: The Learning Healthcare System [128]. Needs are worded as seen in the original text, and descriptions are paraphrased from the text.

While both Etheredge and the 2007 IOM report acknowledge the potential for pharmacogenetics to improve patient-level treatment responses, neither discuss needs specific to implementing genetics more broadly in an LHS. The first draft of the human genome [129] had only been published several years before the LHS model had been proposed, and the model could not yet account for the rapid advancements in genomic technologies, including dramatic cost reduction, that would develop over the next two decades [130]. Healthcare providers, researchers, and policymakers alike soon recognized the importance of leveraging genomic data to improve disease prevention, diagnosis, and treatment [131]. However, most healthcare systems in the US were not equipped to routinely handle genomic data, and the original LHS model did not explicitly account for the additional complexities of these data. To address these opportunities and challenges, the IOM Roundtable on Translating Genomic-Based Research for Health hosted a workshop in December 2014 titled, "Genomics-Enabled Learning Health Care Systems: Gathering and Using Genomic Information to Improve Patient Care and Research" [8]. This single-day workshop laid the foundation for the concept of a GLHS as an extension of the original LHS model that accounted for additional complexities of genetic information in rapid learning, such as large file size and evolving analysis standards (IOM 2015, pp. 6-7). During the workshop, Lynn Etheridge himself acknowledged that genomic data is unique from other clinical data and must be incorporated into a "high-speed, high-performance research system" (IOM 2015, p. 5) that the current LHS model did not explicitly outline. Furthermore, Etheredge warned that failing to develop GLHSs quickly "may lead to massive amounts of genomic data being paid for by health systems but not being available for learning" (IOM 2015, p. 5). To further the goal of integrating genomics into the LHS model and achieving the full potential of genomic medicine, workshop participants compiled a set of needs and possible next steps for GLHSs

(**Table 4.2**). The proposed next steps largely focused on addressing the technical challenges of integrating genomic data into an LHS, but also acknowledged the important role of human factors in establishing complex information systems within healthcare settings.

Need	Next Steps
Interoperability of EHRs	• Ensure that genomic data is accessible and fit for clinical use • Support regulations to make EHRs interoperable with genomic information • Establish standards for genomic data • Demonstrate use cases for interoperability
Clinical Decision Support (CDS)	• Standardize allele nomenclature for CDS tools • Create and share CDS tool warehouses • Measure clinical outcomes of CDS interventions • Develop infrastructure to support CDS
Data Sharing	• Build information platforms with scalable and reusable components • Foster interoperable healthcare systems • Foster a "data donor" culture • Integrate data from around the world • Incorporate patient-provided data • Consider the use of "personally controlled health databanks" for secure data sharing • Support user interface research and development
Implementation	• Engage patients with a particular interest in genomics to demonstrate value • Measure and track healthcare outcomes and disparities • Conduct social sciences and behavioral research to understand human factors

Table 4.2. Needs and suggested next steps for developing a genomics-enabled learning health system, as defined during the 2014 Institutes of Medicine Roundtable on Translating Genomic-Based Research for Health workshop, "Genomics-Enabled Learning Health Care Systems: Gathering and Using Genomic Information to Improve Patient Care and Research" (IOM 2015, pp. 54-55) [8]. Needs are worded as seen in the original text, and next steps are paraphrased from the text.

Since the 2015 IOM roundtable report was published, little else has been published on the GLHS concept. The most active champion of the GLHS model thus far has been the Geisinger Health System in Pennsylvania, which publicly pledged to embrace the LHS model in 2014. They have since made great strides in the areas of patient-clinician engagement and informatics but have faced challenges when balancing research and clinical improvement incentives, and in developing a continuous learning culture [132]. Williams et al. (2018) [13] describes these successes and challenges in the context of Geisinger's experience with precision genomic medicine through the MyCode Community Health Initiative. Geisinger has achieved success in screening populations for well-known pathogenic and likely pathogenic variants using public engagement and alignment of incentives across the healthcare system. However, the Geisinger leadership are careful not to call themselves a fully realized LHS, and even go so far as to question whether such a goal is fully attainable because "the essence of learning and improvement is–and always will be–a moving target" (Williams et al. 2018, p. 9) [132]. Nonetheless, they recognize the utility of moving in closer alignment with LHS principles for the sake of advancing research and improving clinical care.

4.2.2 Clinical data to clinical knowledge

One of the more ill-defined aspects of the GLHS concept is the discovery process, through which clinical and genomic data are transformed into biomedical knowledge with potential care implications. The current body of LHS literature tends to focus on two extremes: the broad, structural components of the LHS model, and the individual components of the model (e.g., technical, cultural, or ethical) as they relate to discovery. What is missing from the literature is a Goldilocks understanding of the ways in which individual technical, social, cultural, ethical, and

political components of clinical discovery interact with one another to form a larger sociotechnical system [14]. The process of discovery is often thought of as largely technical in nature, but early implementations of LHS-aligned systems have demonstrated that technical innovations alone cannot support discovery as it is intended to be used in the LHS model: to improve patient care [133,134]. The contextual factors that surround the discovery process, such as system-wide alignment of goals, a learning culture, patient engagement, and a robust IT infrastructure, are known enablers of discovery, but little is known about how to align these factors with one another in practice. Additional complexities related to genomic data, such as privacy concerns, population representation, and questions of clinical validity and utility, make the operationalization of discovery in a GLHS all the more challenging [31]. Grounding this discussion in the experiences of clinicians (and clinician-researchers) who regularly work with genetic data in a healthcare setting is a reasonable approach for clarifying how the various dimensions of a GLHS might interact with one another.

Previous qualitative studies have identified strategies for sustainable LHS implementation in the Australian healthcare system [135] and challenges and drivers of implementing LHS models in social safety net health facilities [136]. However, none of the existing qualitative studies of LHSs have focused on the perspectives of healthcare providers, let alone genomic medicine providers. Despite the relative scarcity of qualitative studies in biomedical research, qualitative research methods such as semi-structured interviews and grounded theory analysis are ideal for characterizing cultural, social, and personal factors in healthcare that cannot easily be explored using quantitative methods [137]. Constructivist grounded theory, developed by Charmaz (2014) [25], diverges from the positivist lens of classical grounded theory in its use of a relativist lens,

which assumes that people participate in the construction of multiple realities, rather than take part in an orderly reality that can be objectively studied [138]. Constructivist grounded theory also recognizes the role of the researcher in constructing theory and employs techniques such as intensive interviewing and iterative data analysis to construct a plausible snapshot of different social realities, as opposed to offering a "window" into a single reality. Given the complex sociotechnical landscape of clinically embedded genomics research, constructivist grounded theory is a useful approach for holistically evaluating geneticist perspectives on the GLHS model.

4.3 Methods

4.3.1 Institutional review board approval and participant recruitment

The IRB application for this study was submitted to the UW Human Subjects Division on January 25th, 2022 and was approved with exempt status on January 27th, 2022. After obtaining IRB approval, a target sample of 20 study participants was recruited for interviews, based on the estimate that thematic saturation is typically reached between 20-30 interviews in grounded theory studies [139]. The inclusion criteria for study participants were as follows:

- MD-trained physician, preferably in a mid-to-senior level position
- An American Board of Medical Genetics and Genomics (ABMGG) certification in Clinical Genetics and Genomics
- Current or recent member of the eMERGE Network, CSER Consortium, and/or UW Medical Network

A preliminary list of interview candidates was compiled using a contact list provided by the doctoral committee Chair (for eMERGE participants), a private-facing web-based contact list (for CSER participants), and the public University of Washington (UW) Division of Medical Genetics faculty list (for UW participants). To determine whether each interview candidate was board certified in Clinical Genetics and Genomics by the ABMGG, the primary investigator (K.F.) conducted Google searches such as "[name] board certification" or "[name] clinical genetics and genomics" for each candidate and searched for information about board certifications on websites like DocSpot, Zocdoc, and home institution faculty pages. Board certifications were later confirmed by the participants themselves. The primary investigator then conducted a 30-minute Zoom call with two doctoral committee members who were familiar with the potential interviewees to confirm the contact information, home institution, and clinical specialty (or specialties) of each candidate. The final list consisted of 35 potential interviewees: 16 from the CSER Consortium, 6 from the eMERGE Network, and 13 from the UW Medical Network. The list was stored in a password-protected file, which was only shared with the two committee members involved in verifying the information of potential interviewees. Beginning in March 2022, potential interviewees were contacted via email with invitations to participate in the study. CSER members were contacted first, followed by eMERGE members, and then by members of the UW Medical Network. By July 2022, all 35 potential interviewees had been invited to participate, and 20 had accepted the invitation.

4.3.2 Interviews

The primary investigator developed a preliminary list of questions to ask during a one-hour, semi-structured interview, and reviewed these questions with the doctoral committee, the

Precision Medicine Informatics Group at UW, and a qualitative research expert. An initial interview guide (**Appendix A**) was developed based on these discussions. The primary investigator then conducted a pilot interview with the first study participant to assess the quality of the interview guide. After the fourth interview, a more comprehensive interview guide was developed to address emergent concepts of interest (**Appendix B**). This interview guide was used for interviews 5-11, after which a third interview guide was written to consolidate questions and concepts (**Appendix C**). The third interview guide was used for the remainder of the interviews.

Each interview was scheduled for one hour, which included time for introductions, study background, and informed consent. Interviews were conducted using the intensive interviewing method, as described in Charmaz 2014 [25]. This method encourages the interviewer to follow-up on interesting or important points made by the interviewee and is intended to generate rich and meaningful data on interview participants' perspectives. The key characteristics of intensive interviewing include (Charmaz 2014, p. 56) [25]:

- Selection of research participants who have first-hand experience that fits the research topic
- In-depth exploration of participants' experience and situations
- Reliance on open-ended questions
- Objective of obtaining detailed responses
- Emphasis on understanding the research participant's perspective, meanings, and experience

- Practice of following up on unanticipated areas of inquiry, hints, and implicit views and accounts of actions

The interview guides were therefore used as tools to guide conversations but were not intended to dictate the structure of each interview. Once all interviews had been completed, participants were asked to complete an anonymous REDCap demographics survey (**Appendix D**). Although the survey administration procedures were not included in the original IRB application, a modification request was submitted on September 15th, 2022, and was approved with exempt status on September 16th, 2022.

With the participant's verbal consent, each interview was recorded to the primary investigator's private Zoom Cloud. The audio (.MP4) file for each recording was exported to an encrypted device, then uploaded to a password protected Otter.ai account for transcription. Initial transcriptions were generated automatically using the Otter.ai program, then checked for accuracy by the primary investigator. Final transcripts were exported to an encrypted device, then uploaded to a local ATLAS.ti project.

4.3.3 Qualitative data analysis

The following sections describe the approach that was taken to analyze transcript data, including collaborative and iterative codebook development, inter-coder agreement (ICA) calculations, and thematic analysis. Although grounded theory studies do not typically involve multiple coders or ICA calculations, we chose to integrate the perspectives of multiple coders to increase the quality of codes and limit confirmation bias from the primary coder (K.F.) [140], given the potential policy implications of the resulting model. To maintain the iterative process of grounded theory

while involving multiple coders, a codebook development process similar to the one described in Tsai et al. (2020) [141] was used, where multiple coders developed the codebook and ICA was measured over successive iterations. Once ICA reached a satisfactory level, an ICA test was conducted using four additional transcripts that were randomly selected from the 16 transcripts not yet seen by the secondary coders. The codebook from iteration 4 was used to code all 20 of the transcripts, and thematic analysis was conducted using the final axial codes, memos, and semantic domains.

4.3.3.a Codebook development

4.3.3.a.i Initial and axial coding

An iterative process was used to develop the codebook and evaluate the consistency of coding on the 20 transcripts. For each transcript, the primary coder assigned initial codes in thematic units. As described in Burla et al. (2008) [142], thematic unit coding can be used instead of line-by-line coding when it is important to capture the context of each initial code. Each unit was defined as having a distinct meaning, message, or sentiment compared with surrounding data. Unit lengths ranged from several words to several sentences. For example, the contiguous excerpts from the interview with Participant 10 in **Table 4.3** show how three different initial codes with different meanings and unit lengths were assigned.

Initial Codes	Axial Codes	Excerpt
Even though there is more genetic actionability than we know about, the data to show that and act upon it is limited	Ensuring patient/research participant safety and wellbeing: Generating, collecting, and applying evidence for variant interpretation	"Well, yeah, I think, obviously, the data are flowing much more rapidly than our ability to digest it all. There's lots of instances where we get data back that we're not quite sure what it means."
There are a lot of genetic discoveries that could be very useful in real time	Ensuring patient/research participant safety and wellbeing: Turning new genetic associations and technologies into clinical interventions	"You know, that being said, I guess before going too far down that path, there's a lot of data that's exceedingly useful in real time."
Important to focus on genetic tests/results that can meaningfully change patient management (utility)	Ensuring patient/research participant safety and wellbeing: Determining variant actionability, utility, and returnability in the clinic and clinical labs	"And, you know, so I don't think we'd be doing this clinically if we didn't think there was a reasonable chance that we might come up with something that would actually answer the question."

Table 4.3. Initial and axial coding of three contiguous excerpts from an interview with Participant 10.

After each round of five (25%) interviews, the primary coder grouped initial codes into thematic categories, or axial codes. Categories were developed to reflect important or problematic aspects of clinically embedded genomic research that emerged from the transcript data. In the process of developing axial codes, initial codes were constantly compared with one another to reveal agreements and contradictions. Several examples of axial codes are shown in **Table 4.3,** alongside sample initial codes that contributed to axial code development. Axial codes were used to refine the interview guide and facilitate exploration of emergent categories that held theoretical promise. Theoretical codes were written in the form of memos throughout the study but were especially refined during the last two interview cycles as emergent categories began to reach saturation. Saturation was determined by the inductive thematic saturation approach [143],

which assumes that saturation is reached when there are few emergent codes or themes. Specifically, saturation was reached when new initial codes could be reasonably grouped into existing axial categories, and if new initial codes only marginally expanded on similar initial codes but did not diverge from existing themes.

4.3.3.a.ii Multiple coding

During each iteration of codebook development, one transcript was selected for multiple coding based on sufficient representation of axial codes. After initial and axial coding was completed by the primary coder, the selected transcript was coded by multiple reviewers in a five-step process:

1. The transcript was prepared for multiple coding using the following procedures:
    a. The entire ATLAS.ti project (**Version 0**) was duplicated (**Version 1**).
    b. All transcripts not selected for multiple coding were deleted from Version 1.
    c. All memos and identifying information were deleted and replaced with proxies.
    d. Initial codes were merged into axial codes, and all comment fields were cleared.
    e. Version 1 was saved and duplicated (**Version 2**).
    f. All code assignments from Version 2 were deleted, but the codebook and highlighted text segments remained the same.
    g. Version 2 was exported as an Atlas Version 22 (.ATLPROJ22) file and was shared with the two secondary coders via UW OneDrive.
2. Two secondary coders independently assigned one code from the codebook to each previously defined text segment in Version 2, using the methods described in O'Connor et al. (2020) [144]. For several rounds of codebook development, the primary coder

shared a document that described the scope and meaning(s) of each code (**Appendices E and F**).

3. Once both secondary coders uploaded their coded transcripts to separate UW OneDrive locations, the primary coder merged the two coded Version 2 projects into Version 1 to compare codings.
4. Krippendorf's Cu-alpha (α) [145]—the standard agreement measure offered in ATLAS.ti 22 Desktop—was used to assess agreement across semantic domains.
5. The secondary coders met with the primary coder to discuss disagreements, and the primary coder revised the codebook using feedback from the secondary coders.

This process continued until acceptable agreement ($0.667 \leq \alpha \leq 0.823$) [146] was reached across semantic domains. As described in Burla et al. (2008) [142], codes used for ICA analysis should "address substantive issues related to the research question" (Burla et a. 2008, p. 115). Because the semantic domains that axial codes were grouped into formed the basis of the thematic analysis, ICA was deemed most useful when assessing agreement across those domains.

4.3.3.a.iii Inter-coder agreement test

To evaluate coder consistency on the resulting codebook and on unseen data, four (20%) transcripts were randomly selected from the remaining transcripts that had not been coded by the secondary coders. One additional coder used the codebook to assign codes to these transcripts. Finally, simple percent agreement and Krippendorf's alpha were calculated across semantic domains between the two coders.

4.3.3.b Thematic analysis

All memo titles and axial codes were copied into text boxes in diagrams.net [147] to facilitate memo and code sorting, as described in Charmaz 2014 (pp. 216-224) [25]. First, relationships between memos were represented using unidirectional or bidirectional arrows between boxes, with connection descriptors used as needed. Multiple memo formations were created and assessed for data representativeness. Once the final memo formation was developed, axial codes were linked to the memo(s) that they best represented. This helped to ensure that the abstract theory was re-grounded using interview data and helped elucidate the ways in which thematic categories interacted with one another in the emergent model. The resulting flow model was further distilled into components that semantically described distinct groupings of clinical research operations.

4.4 Results

4.4.1 Participants

Twenty (20) individuals participated in phone or video interviews, which ranged from 27 minutes to 64 minutes, and lasted a median of 54 minutes. As described in **Table 4.4**, the majority (80%) of participants worked at academic medical centers and 50% of the participants were MD/PhD clinician scientists. In addition to a Clinical Genetics and Genomics board certification, many participants held an additional board certification in Pediatrics or other specialties such as Internal Medicine and Clinical Molecular Genetics.

Participant Characteristics	N (%)
Work Environment	
Academic Medical Center	16 (80%)
Integrated Care Organization	3 (15%)
Research-Only Hospital	1 (5%)
Credentials	
MD	10 (50%)
MD/PhD	10 (50%)
Board Certification(s)	
Clinical Genetics and Genomics	20 (100%)
Pediatrics	7 (35%)
Internal Medicine	4 (20%)
Clinical Molecular Genetics	4 (20%)
Clinical Cytogenetics and Genomics	3 (15%)
Medical Biochemical Genetics	2 (10%)
Obstetrics and Gynecology	1 (5%)
Psychiatry and Neurology	1 (5%)
Clinical Informatics	1 (5%)
Preventive Medicine	1 (5%)
Clinical Specialty	
Dysmorphology/Structural developmental abnormality	11 (55%)
Cancer	7 (35%)
Neurodevelopmental abnormalities (intellectual disability, autism)	6 (30%)
Cardiovascular disorders (cardiomyopathy, arrhythmia, vascular anomalies)	5 (25%)
CNS disorders (epilepsy, encephalopathy, structural brain malformations, neurodegenerative disease)	4 (20%)
Neuromuscular disorders (hypotonia, spasticity, neuropathy, myopathy)	3 (15%)
Immunodeficiency	2 (10%)
Metabolic disorders	2 (10%)
Skeletal dysplasias	1 (5%)
Population genomic screening	1 (5%)
Genodermatoses and Turner syndrome	1 (5%)
Race/Ethnicity	
White or European American	16 (80%)
Asian	2 (10%)
Middle Eastern of North African/Mediterranean	1 (5%)
American Indian, Native American, Alaska Native	0 (0%)
Black or African American	0 (0%)
Native Hawaiian/Pacific Islander	0 (0%)
Hispanic/Latino(a)	0 (0%)
Prefer not to answer	1 (5%)
Unknown/none of these fully describe me	2 (10%)
Gender	

Woman	10	(50%)
Man	9	(45%)
Prefer not to respond	1	(5%)
Non-binary/non-conforming	0	(0%)

Table 4.4. Characteristics of interview study participants (N = 20).

4.4.2 Identified themes and semantic domains

By the final iteration of codebook development, 1796 initial codes were linked to 2444 quotations across the 20 transcripts, and initial codes were assigned to 28 axial codes across 6 semantic domains: 1. Building a collaborative learning culture in medical systems (8 codes); 2. Building relationships with patients/research participants (4 codes); 3. Ensuring patient/research participant safety and wellbeing (5 codes); 4. Evaluating the role of genetics in medicine (6 codes); 5. Participant background (3 codes); and 6. Protecting patient/research participant rights to privacy and autonomy (2 codes). **Table 4.5** lists the axial codes and associated semantic domains, and **Tables S4.1-S4.6** contain detailed descriptions and sample quotes for each axial code in a given semantic domain.

Semantic Domain	Axial Code
Building a collaborative learning culture in medical systems	Benefits and drawbacks of using EHR data for research and equitably representing diverse populations
	Benefits, drawbacks, and realities of operating within integrated and universalized healthcare systems
	Challenges of operating within a stressed and fragmented US healthcare system
	Forming collaborations and support systems within and between healthcare systems

	Negotiating the roles of medical geneticists, genetic counselors, and non-genetics providers
	Paying for clinical sequencing and clinical research
	Sharing and recycling clinical and genomic data
	What are the differences (if any) between research, clinical care, and quality improvement?
Building relationships with patients/research participants	Building trust with patients, especially from minority communities
	Communicating with patients about research/clinical distinctions and navigating provider/researcher differences
	Engaging patients in the research process and being sensitive to their needs and motivations
	Providing incentives or clinical benefits to patients for participating in research
Ensuring patient/research participant safety and wellbeing	Determining variant actionability, utility, and returnability in the clinic and clinical labs
	Educating non-genetics providers about genetic medicine to prevent misuse and misinterpretation
	Ensuring appropriate clinical follow-up after genetic testing
	Generating, collecting, and applying evidence for variant interpretation
	Turning new genetic associations and technologies into clinical interventions
Evaluating the role of genetics in medicine	Considerations for using population-wide genetic screening in clinical care
	Deciding what types of genetics tests to order based on clinical indications
	Historical advancements in genomic research and technology
	Understanding genetic impacts on health and disease

	Using the EHR to represent genomic data and streamline clinical genomics
	Visualizing the best (and worst) uses for genomics in medicine going forward
Participant background	Types of patients they see or environments they do clinical work in
	Types of research they are or were involved in
	Where they trained, in what, and for how long
Protecting patient/research participant rights to privacy and autonomy	Challenges and strategies for ethical oversight and consent in clinical research
	Protecting the privacy and security of clinical data

Table 4.5. Semantic domains and associated axial codes after the fourth iteration of codebook development.

4.4.3 Inter-coder agreement

Overall ICA reached 75.1% for simple percent agreement between two coders on the four-transcript test sample, and Krippendorf's alpha reached 0.669 across semantic domains, which indicates acceptable agreement. **Table 4.6** shows the coding frequency and agreement between two coders for all semantic domains, both individually and overall. At the time this work was submitted, results from a third coder were pending. The ICA results in future publications will reflect agreement between all three coders.

Semantic Domain	Coding Frequency	Simple Percent Agreement	Krippendorf's Alpha
Building a collaborative learning culture in medical systems	114	66.8%	0.730
Ensuring patient/research participant safety and wellbeing	52	53.3%	0.647
Evaluating the role of genetics in medicine	46	57.2%	0.697
Building relationships with patients/research participants	68	49.5%	0.636
Participant background	42	76.1%	0.856
Protecting patient/research participant rights to privacy and autonomy	13	37.9%	0.543
Overall	335	75.1%	0.669

Table 4.6. Semantic domain coding frequency and agreement between two coders on the four-transcript sample.

4.5 Discussion

From a purely topical standpoint, this study reinforces many of the GLHS needs identified previously by the IOM, such as EHR interoperability, analysis and CDS tools, patient-participant engagement, an aligned learning culture, and structural support for combined clinical and research activities. However, the rich personal and experiential data collected during interviews offers a more nuanced look at the relationships between facilitators and inhibitors of clinical research in the context of iterative learning cycles. In the following sections, we describe an emergent theoretical model that captures these nuanced relationships and offers a high-level understanding of the essential processes involved in clinically embedded genomics research.

4.5.1 The Five R's of Clinical Genomics Research: Representation, Responsibility, Risks and Benefits, Relationships, and Resources

Previous LHS models have depicted ethical, technical, and social considerations as precursors to rapid learning cycles [45,89]. In the case of clinical genomic research, we argue that these elements are not static precursors to successful clinical learning, but rather are integral elements of dynamic relationships between learning processes. There are five core elements identified in this study that collectively represent different ethical, technical, and social considerations of clinical genomic research: Representation, Responsibility, Risks and Benefits, Relationships, and Resources. We further group these elements into three distinct groups that describe how they interact with one another and with the existing rapid learning processes: *negotiation processes* (Representation, Responsibility, Risks and Benefits), *binding factors* (Relationships), and *constraining factors* (Resources).

Figure 4.1 depicts an emergent theoretical model of rapid learning cycles in GLHS environments, where the data to knowledge, knowledge to practice, and practice to data processes are linked to one another through intermediate negotiations of representation, responsibility, and risks and benefits. The model also depicts community trust and learning cultures as the relationships that bind rapid learning processes to one another, while the structural and financial aspects of the healthcare system establish the bounds within which learning cycles must operate and adapt. The multi-directional, negotiative nature of the model suggests a constant reconstruction of what data, knowledge, and practice signify in the research and clinical contexts.

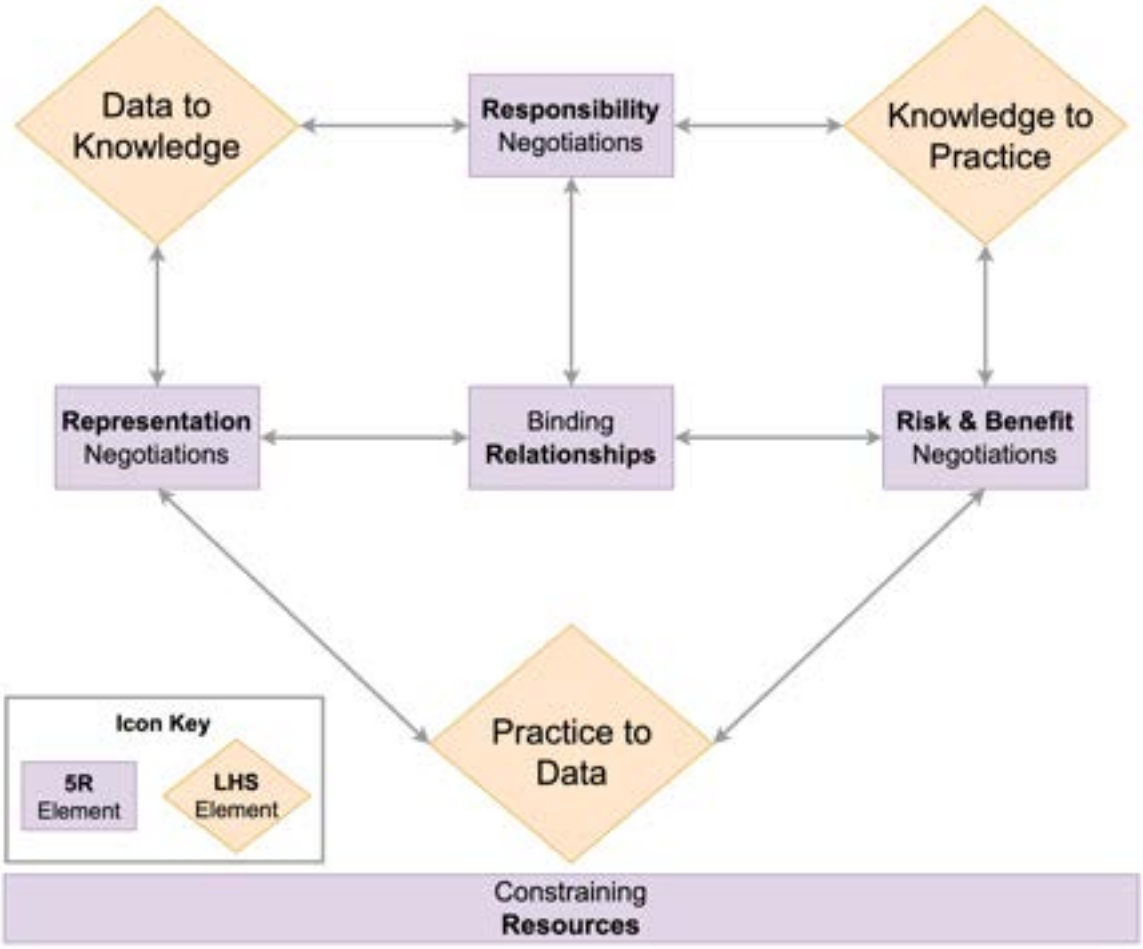


Figure 4.1. Schematic of the "5R" genomics-enabled learning health system conceptual model.

4.5.1.a Negotiation

This study originally sought to identify enablers and inhibitors of the data to knowledge process of rapid learning cycles in clinical genomics research, given that few studies had previously explored this topic in isolation. However, the interviews quickly revealed that it is not possible to separate knowledge generation from either data production or knowledge application when the work is being done in a clinical setting, with clinical data and clinical participants. When asked about the potential pros and cons of using clinical data for genomics research, discussions

naturally flowed both “backwards” towards the practice to data LHS process, and “forwards” towards the knowledge to practice LHS process, with neither direction seeming to take precedence over the other. In the former case, discussions revolved around the *representation* of clinical and genomic data in terms of quality, population characteristics, and sample size. In the latter case, discussions focused on the *responsibility* aspects of generating new knowledge within a clinical environment, where perceived and anticipated differences between research and clinical care imply different responsibilities of researchers, providers, and hospital systems in balancing the speed, evidence, quality, and safety of knowledge generation and testing. In both cases, discussions culminated in considerations of the *risks and benefits* that are constantly negotiated when conducting both research and clinical care, some shared and some distinct. These three types of negotiation collectively highlight the interconnectedness of rapid learning cycle processes and suggest that genomics-enabled learning cycles must be in constant flux as they evolve with the research and clinical enterprises.

4.5.1.a.i Representation

Interviewees universally agreed that EHR data is inherently messy and difficult to use for both clinical and research purposes, which has been well-documented in other studies [148–151]. Some commonly cited technical issues among interviewees were data missingness, difficulty of data entry, difficulty of data interpretation and extraction, redundancy, inconsistencies in medical terminologies or descriptions, and the sheer volume of both structured and unstructured data. However, they also tended to agree that EHR data was more representative of the “real world”

than many other types of data that could be used for clinical research, as Participant 16 explained:

> …if you're trying to **minimize ascertainment bias**, if you're trying to get **real world data**, if you're trying to understand like, where, where people are, how they're seeing what that **natural history** is, what that **patient journey and experience** is, like, I mean, that's the real world, you know that that's the way it is. And so that's incredibly [emphasized] valuable. (Participant 16)

Ideally, EHR systems would be redesigned to be more amenable to data representation for both research and clinical care, and the systems and devices that produce data would automatically integrate with EHRs. However, given the slow pace of technological development in medical information systems compared with that outside of healthcare, significant skill, time, and financial investment are needed to harness clinical data for research. Participant 1 captured several of the challenges of representing clinical data from the EHR:

> So doing any kind of any kind of research, when, when you're dealing with those types of data…you…are gonna have **missing data**, or you have to **pay someone who knows enough** to be able to abstract those data out of the chart, which is a pretty **specialized skill**. And, and which as a result can be **pretty expensive**. (Participant 1)

The trade-off between EHR data being a less-than-perfect information source while offering a broader picture of the lived healthcare experience is a common conflict in clinical research, and one that has received considerable attention in the healthcare informatics community. As Participant 16 noted, they are "confident that…smart people are figuring this out together." Arguably, the more pressing concern in genomics is the representation of diverse populations using clinically derived genomic data. While some research projects like CSER have actively recruited diverse populations for clinical genomic research [15], non-European populations are still largely unrepresented in genomic databases, which makes the potential benefits of precision

medicine far less accessible to individuals with diverse ancestral backgrounds [152]. Many interviewees voiced their concerns that certain populations in genomics would be "left behind" given the complex social and institutional conditions in the US that lead to underrepresentation in genomics research, including, but not limited to, systematic racism and the resulting mistrust in research and medicine, barriers in access to research and clinical care, and concerns about the privacy, security, and misuse of genomic data. Concerns about population representation are therefore inextricable from discussions of research engagement if precision medicine is to benefit the health of the entire population.

Interviewees also discussed the necessity of collating rich clinical and genomic datasets across healthcare institutions for the sake of conducting statistically viable and representative genomic analyses. Especially in the case of rare diseases–or common diseases with rare variation–large sample sizes must be accumulated to achieve sufficient power. For example, research in rare cancers has historically been driven by cooperative groups that share clinical data across institutions, as Participant 1 described:

> …because especially once you get to **rarer cancers**, the only way that we have learned about the treatment of those cancers is through **cooperative groups**. Because that's the only way to **accumulate enough information**. (Participant 1)

However, the barriers to sharing clinical and genomic data between, and even within, healthcare institutions are significant. There was little debate among interviewees that there are risks involved in sharing personal health information, but the degree to which those risks supersede the personal and/or scientific benefits of clinical genomic research was highly contested. Some interviewees argued that the current barriers to sharing data were unreasonable, given that many

patient-participants, in their experience, were willing to share their data between institutions, but that IRBs were not:

> And, and, and **data sharing**, too. That's, yeah, that's just **absolutely crucial** with this. And so yes, I mean, it's, it's the **only way forward** that I know [emphasized] of, and the barriers that we have erected between us and and that are really not smart. Much of it is due to an **unreasonable set of fears** that have been engendered by a lot of overreactions to genomics and genetics, and a lot of it is **ignorance and fear from IRBs**. (Participant 7)

Other interviewees took a more cautious stance, fearing that the risks of data breaches could be quite significant, especially for patient-participants who have been historically marginalized by the research and clinical enterprises. Participant 2 described the fears of patients in their institution's catchment area:

> So this all comes down to the interests of the patient because why would the patient give you their genome? And in [city name], there's **massive [emphasized] sensitivity** about this. I've actually had patient groups in the last week, saying, Yeah, you want to do genomics in [city name]. But what you're doing is you're **targeting people of color to get information that could ultimately be used to kill us**. So that is from a very sensitive group who have been **discriminated against traditionally and currently**. But everybody is going to have a shade of that, along the scale of, you know, relatively benign to incredibly concerned. (Participant 2)

Several interviewees discussed the possibility of using strategies like data anonymization to share data for research without risking the privacy of patient-participants, or continuing to share data in large, anonymized resources like gnomAD, ClinVar, the UK Biobank, and All of Us. While this is currently a standard practice to use these types of datasets for large scale genomic analyses, many interviewees recognized that there is a limit to the amount of clinical data that can be gleaned from anonymized datasets, thereby limiting the clinical scope of analyses. As Participant 4 described:

> So it's just to say that we're all trying to create structures so people can do the…initial research, but at some point...in many [emphasized] of these circumstances, that research still requires **a level of clinical information that's just not in the data**, or is only in the data, when, with enough identifying information that you have to have further consent. (Participant 4)

In clinical genomics, there is a constant negotiation between the scientific needs of the field for understanding the impacts of genomics on health and disease, and the need to protect the well-being and privacy of the patient-participants who are meant to be the beneficiaries of research. The concept of representation in data systems is central to this negotiation.

4.5.1.a.ii Responsibility

The prospect of generating new genomic knowledge within a clinical environment invariably caused interviewees to consider the different responsibilities of researchers, providers, and hospital systems in constructing and using that knowledge. One element at the heart of these differences is the tension between medical genomics as a clinical indication-based specialty, and the need for population-level data for research purposes. Aside from payment concerns associated with population-level sequencing (which are discussed in the section titled, "Constraining factors"), there is significant concern among geneticists that ordering broad tests like genomes or exomes on a wider clinical population would lead to more false positives that could unnecessarily concern participants, or even lead to inappropriate clinical actions. As Participant 17 described:

> I think, to me, it depends on how you phrase it, and how you would implement it, but the way it is right now, **we don't screen the population**. I mean, there's already a lot of issues…with regular screening in terms of **false positives**. You know, this is I know, there's a lot of papers and publications that look at this a lot. And I'm not up on the latest,

> but, you know, I still view **genetic testing as a test that requires an indication**… (Participant 17)

Even if results from larger genomic tests do pick up variations that are not false positives, there are ethical questions surrounding whether those results must be returned to the patient-participant in a clinical setting. This is a particularly challenging dilemma in the field of pediatrics, where large-scale screening could potentially save years of diagnostic odysseys but could also result in diagnoses that might not be entirely clinically meaningful, but nonetheless change the course of a child's life. Participant 17 described this challenge alongside their appreciation for the "huge range of human variation":

> And I am not the type of geneticist who just sequences the world or sequences everybody, I have to have a **good reason to decide to sequence a kid**. You know, I had, I had a pediatrics mentor, who…told me this line, and it stuck with me, which is, you know, **every kid has one chance to be normal**. And after you, you go in and do something, their parents view them as **not normal anymore**. And I have a great appreciation for the **huge range of human variation**. (Participant 17)

The prospect of delivering new genetic information to patient-participants is particularly controversial when the information is generated from a research study that is conducted using clinical data. Interviewees expressed a general sense of responsibility for divulging potentially clinically actionable information, but generally did not feel comfortable returning preliminary results that had not been functionally validated or backed by several publications. On the other hand, some interviewees recognized that more harm could potentially be caused by withholding information gleaned from clinical research than by providing information that could be potentially misleading. For example, Participant 4 described a situation in which a secondary research finding was identified using clinical data, and the finding strongly indicated that some patients in the dataset might have cancer. However, the patients were not consented to receive

research results, and the interviewee viewed the results as “potentially harmful secrets” that could needlessly endanger the wellbeing of patients. Other interviewees expressed more hesitancy, suggesting that the “right not to know” could be just as significant as the “right to know.” Participant 2 described this tension on the spectrum of pediatric and adult medicine:

> …it's very interesting when you start asking questions like the questions that we hear in genomics, like the **"right not to know,"** which is a big thing in pediatrics for genomics. It's also you know, somebody who's a young woman, and they don't need to be starting their breast mammography for another 10 years, **maybe just give them those 10 years**. But you then put in the search term, "Right not to know, cancer." And it's a very, very different discussion, because obviously, **it's a bit more acute. If you don't treat it, you're probably going to die.** But you know, there are situations where there are some people who are probably **sick enough anyway, that, you know, why would you tell them that they have cancer?** And for some reason, people are much more inclined to say, **You must tell them**. And I wonder whether that's where we will go in, in genetics and genomics, or whether it's, I think that it's a really interesting area…to look at. (Participant 2)

The possession of potentially clinically relevant information within a healthcare system–whether it was generated in the name of research or medicine–necessitates discussions about whether the information can or should be returned to the relevant patient-participants. A key factor in the returnability (and ultimately the clinical utility) of research-generated genetic variant information is the rigor of variant interpretation and validation. From a legal and procedural standpoint, interviewees noted that genomic results must be approved by a Clinical Laboratory Improvement Amendments Act of 1988 (CLIA)-certified diagnostic lab. However, they also noted that diagnostic labs do not necessarily interpret variants differently than research labs, and that standards for variant interpretation are remarkably difficult to implement. In fact, it has been well-documented that variant interpretation agreement between CLIA-accredited labs using the American College of Medical Genetics and Genomics and Association of Molecular Pathology (ACMG-AMP) published guidelines is around 35-50% [153,154]. There is a shared

responsibility between both researchers and clinicians to apply the best available evidence for variant interpretation, both for the sake of scientific credibility and patient-participant wellbeing:

> I think you have to have a standard of what the data [emphasized] is. There has to be a **minimum standard of the data to decide what should be returned and considered clinically relevant**. (Participant 13)

However, many interviewees identified a paradoxical relationship between the need for community standards and the relative impossibility of applying those standards exactly as they were intended to be applied. Participant 7 described this paradox in the context of validity and utility:

> And it in the end is a **judgment [emphasized] of a threshold of validity**, both the gene disease validity, analytic validity is kind of simple. That's just not that hard. The **gene disease validity is much more complex, and then the utility more complex still**. And again, it's **different for every disease and every gene**. And one has to apply, has to **genuinely endeavor to apply community accepted standards to that, even though you can't standardize it**. (Participant 7)

Ultimately, gene-disease associations and subsequent variant interpretations are a judgment call based on currently available evidence. This leads to a second paradox in the process of clinical genomic discovery: discovery validation and implementation requires accumulated evidence of validity and utility, but evidence of validity and utility are not fully informative without implementation. This lack of clinical evidence is a known issue in genomic medicine [155,156] and has been cited as one of the driving reasons for incorporating genomics into LHSs [8,157]. However, interviewees warned that research labs in LHS models would likely run into the same evidence paradox that geneticists and diagnostic labs run into every day in medical genetics practice. Providers, particularly clinician-researchers, who wish to push the envelope of genomic

medicine must make clinical judgment calls about new gene-disease associations that might be considered premature by some. As Participant 16 described:

> I've often described it as you know, you're **building the plane as you're flying it**, and you're trying to use the **best evidence and the best data that you have**, but you **don't always have complete information**. (Participant 16)

From a research perspective, there is an incentive to use new genomic information for clinical decision-making in an LHS if there is a reasonable expectation that the information could positively impact patient outcomes, thereby increasing the evidence for its clinical utility and creating a positive feedback loop of implementation. However, a clinician's duty is ultimately to the patient, and protecting the patient's wellbeing often requires the use of a higher bar for evidence of safety and utility. The responsibilities of researchers and clinicians to their roles are not mutually exclusive in this regard, since it is never the intention of the researcher to cause harm to the participant. Yet the risk benefit calculation appears to operate differently when the subject is a consented research participant, as opposed to a patient receiving routine clinical care.

4.5.1.a.iii Risks and Benefits

During interviews, discussions of data generation and knowledge implications naturally tended to shift to contemplations of the risk-reward tradeoff for conducting genomic research in a clinical setting. On the reward side, many interviewees included historical accounts of how genomic research has benefited both basic science and understandings of human health and disease. Genetic research in cancer was often cited as a shining example of rapid learning in a healthcare context, as described by Participant 10:

> And, you know it, reveals, you know, **much of what we know about cellular signaling pathways started from studying cancers** where those pathways were aberrantly activated, and that revealed the whole thing **cascade of genes** that then turn out to be important, **not just in cancer, but in normal cellular processes and in development**. (Participant 10)

However, there is a general understanding in research that neither the risks nor the benefits of research can be fully known, although there are protections in place to mitigate potential harms. The differences between ethical oversight of research and ethical oversight of clinical care were therefore a cause of concern for interviewees when considering oversight of clinically embedded research. Participant 1 noted that these differences might warrant the use of ethical expectations that are more aligned with clinical standards in an LHS:

> …it's harder to do research or there's **more oversight required for research [because] we don't know as much about potential harms. And we don't know as much about the potential benefits**. So [in an LHS] there does have to be, I think, an **establishment of we there's a very strong expectation, this is going to work** in certain populations. (Participant 1)

Participant 7 noted that research protections were largely established as "a reaction to malfeasance" in the aftermath of horrific, inhumane, and unethical medical research studies, such as the medical experiments in Nazi Germany and the Tuskegee syphilis study. Resulting ethical frameworks like the Nuremberg Code and the Belmont Report have set the expectation among researchers and participants that participants provide their voluntary consent to be involved in research, and that researchers and ethics committees do everything in their power to limit harms while maximizing scientific benefits. The expectation is that potential harms are communicated to participants, but not that potential harms are eliminated altogether. In the context of clinical medicine, however, ethical expectations are more straightforward yet less codified than in research. Interviewees frequently noted historical attempts to codify expectations of protections

and potential harms in clinical research, but no such attempts were mentioned for clinical medicine because the standard expectation is that the clinician has a duty to minimize harm and maximize benefit to the patient at all costs. However, medicine is inherently not risk free, and patients are at risk for many of the same harms they might experience as clinical research participants. It is instead the *expectation* of minimal to no harm among patients, and the sense of duty among clinicians, that separates care on the premise of medicine from care on the premise of research. There are constant risk-benefit calculations in genetic medicine, especially when clinicians must use partial information to make clinical decisions, as Participant 16 described:

> So I just, I guess every time I **balance risk and benefit and how important it is to the clinical care**, and if there's any other way of **validating what I'm doing**. (Participant 16)

There is an undeniable history of harm in clinical medicine, but those harms are largely classified as unintentional given the sacrosanct relationship between patient and provider, as Participant 1 noted:

> …we have **historically done some very foolish things**, things that **seemed reasonable at the time, and were very well intentioned**. But in retrospect, were you know, **didn't have the desired outcomes**. (Participant 1)

Risk-reward calculations in genomics-enabled learning healthcare are therefore highly dependent on the *expectations* of patient-participants, and the *relationships* between them and clinicians and researchers (and clinician-researchers) that shape those expectations. The following section describes how relationships are the focal point of all LHS processes and intermediate negotiations.

4.5.1.b Binding factors

The processes of negotiation that interviewees collectively described were all united by considerations of the relationships between those involved in learning healthcare. Although it is simpler to abstract the entities involved in LHS process into representative systems and institutions, systems and institutions are ultimately composed of people whose interactions with each other both form new meanings within systems and are products of the systems themselves.

In the case of representation, relationships between patient-participant communities and people who represent the clinical research and clinical medicine enterprises are the foundation of the local and shared clinical and genomic data systems that learning healthcare relies on. Without deeply rooted, trusting relationships between LHSs and the communities they are seeking to serve, genomic medicine will not advance in a way that benefits populations equitably, as it must. Interviewees suggested several ways to build these relationships, such as involving a more diverse and culturally sensitive workforce in patient-participant engagement, as Participant 6 explained:

> Yeah, I mean, like, if you're, if you're going to be recruiting from or doing this testing, you know, at the safety net hospital where most of the patients are Latina. And then you have a…**your research coordinator is a, you know, is Latina also, comes from that big community and can sort of explain why this is important. In Spanish. In a way that makes sense**. (Participant 6)

Some interviewees also emphasized that community-led data sharing efforts were likely to be more successful, at least at first. Data sharing for the purposes of research could be managed more directly by communities, as Participant 16 noted:

> And I've wondered to myself, is that a way to **dip your toe in and be able to get movement** and when you [share data], of course, to **have it from people in the community who are doing this**. So **they own the data**, they have the **grants that do this**, they have the **benefits**, they **govern it**, they **watch it**, they make sure that people are **using it responsibly**, you know what I mean, but it's coming **from the community for the community, in a way with limitations very much built into it**. (Participant 16)

Participant 16 also emphasized that building trust with patient-participant communities cannot be rushed, and that preliminary steps such as sharing anonymized might help gain traction with communities, even if the ultimate goal is to share identifiable EHR data to achieve the full potential of learning healthcare:

> …and it'll have to be **staged**. If you can start with [sharing anonymized data] and then **build the trust and show that you're a responsible partner**, you know, over maybe 20 years, but you know, over periods of time, can you build trust. (Participant 16)

The topic of informed consent for participation in clinical research was a recurring discussion during interviews, and the consensus was that conducting consent in stages throughout the learning process was preferable to obtaining broad, up-front consent from patient-participants, but that conducting consent this way was impractical for most health systems. Even if staged (i.e., dynamic) consent were more practical to implement, some interviewees noted that patient-participants might not want to be repeatedly asked for their consent to include their clinical and/or genomic data in new types of studies in an LHS, provided there is sufficient trust between the patient and the health system conducting the research. One interviewee who works at an integrated health institution that routinely conducts genomic research said of their patient-participants:

> …there was **sufficient trust** given that, you know, we, in many cases, because of the nature of our service area, have **long standing relationships with the individuals that we care for**. There's a very high degree of trust. And so they, when they heard about the overview of the program, they said, We think you guys **have the knowledge to be good**

> **stewards, and do right by us**. And so they were very comfortable with a **one time overarching broad consent**. (Participant 20)

The comfort level of patient-participants when sharing data for clinical research and consenting to their data being used for potentially unforeseen purposes, is highly dependent on their relationship with the local research and healthcare enterprises. If identifiable clinical and genomic data were to be shared across healthcare institutions, this symbolic relationship would need to extend to the research and clinical enterprises on a national scale.

In the process of producing new knowledge within a healthcare environment, several types of relationships underpin the notion of responsibility for both scientific and clinical excellence. First, the question of how geneticists, genetic counselors, and non-genetics healthcare providers should work together to order, interpret, and communicate genetic testing results (whether in the name of research or clinical care) was a topic of consistent disagreement. Although most interviewees acknowledged the inevitability that the demand for genomic medicine will likely expand beyond the capacities of genomic specialists, they often expressed discomfort with the idea of all healthcare providers ordering and interpreting genetic tests, especially if the results are preliminary or not backed by a wealth of experimental evidence. However, they also acknowledged that certain genetic conditions will become the "domain" of other specialties when the condition has specific indications within that medical specialty, as Participant 19 described:

> And so for a lot of patients, if they have already seen a neurologist who said that this is a myopathy, or a neuropathy, like, **I'm not going to come in and change that diagnosis after meeting them for 15 minutes, compared to a neurologist who has followed them over time has sent their own studies has their own area of expertise in that**

> **area.** And so if that person needs a neuropathy panel, because a neurologist says they have a neuropathy, I feel like that, **that's kind of their lane.** (Participant 19)

Along with genetics becoming a routine practice in other medical specialties, interviewees imagined that genetic specialists would be more highly involved in the cases that eluded current knowledge of genetics. In this way, genetics specialists would transition into more of a "research" role when working with patients, for the sake of solving difficult cases and pushing the boundaries of what is known in genomic medicine, as Participant 16 noted:

> I still think there are going to need to be at the core...people will, there needs to be people that are going to be at the **cutting edge** and think about the first way of doing things. And those are likely to be people who are still, you know, I think of them as the faithful…who **know enough about the broad field** to be able to know what you can pull in and how to think about this broadly and how to **think about the implementation challenges and not do things recklessly. So those first movers, I think, largely, are going to be, you know, medical geneticists.** (Participant 16)

During interviews, discussions of the relationship between genetics specialists and non-genetics care providers then naturally shifted to the question of the researcher-clinician relationship. In the field of genomic medicine, it has long been the case that geneticists play a dual research-clinical role in the sense that many of their patients present with indications that suggest a genetic etiology of disease, but for which the current body of evidence does not provide a clear diagnosis or treatment pathway. The genetic provider must then make judgment calls to understand the case and make the best decisions possible for the patient, as Participant 11 described:

> You know, one of our previous chairs once said, that **we should not see a single patient in our genetics clinic that was not a research subject.** And what he meant by that was that everybody [emphasized] that we see, **we should be thinking about in a way that we learn from them, and we move forward from them.** And that if you know that, that we certainly have still questions to be answered about the diseases with which we are dealing or the situations with which we're dealing or the treatment that we, that **every**

> **[emphasized] single patient should be, should provide something that most moves us forward.** And I think that that's actually true. (Participant 11)

In this case, the definition of research is "unofficial" in the sense that research is defined as a way of thinking or approaching a clinical problem. The patient-participant-provider relationship becomes more complicated when the research is "official," i.e., the provider is involved in an IRB-monitored research study and using clinical data for research purposes. Interviewees who were involved in both research and clinical care in the official sense mentioned that they must be very transparent about their dual roles when interacting with patients, so as not to compromise either the patient-provider relationship or the participant-researcher relationship. As Participants 1 and 5 described:

> I, when I've done this with patients that I have seen in both a clinical and research perspective, I try to be almost almost **ridiculously clear** about, I will, in part, probably partially my pediatric training, but I will say, Okay, I'm **taking off my clinician hat now** and, I will like mime I'm taking off the hat and I'm **putting on my researcher hat.** And in this point, I'm **talking about things where I have less certainty about what it means**, I'm talking about things that are **completely optional for you to participate in**, it's not going to affect your, you know, it's **not going to affect how I would take care of you as a as a clinical patient**. And I do that, because I really want to emphasize those points. (Participant 1)

> And so I think this is like, the critical thing you always have to imagine is that when you're doing research, **you have [emphasized] to truly try your hardest to disentangle your hat as their provider from your hat as their researcher, because those are two very, very different sort of roles that you're in, and they sometimes have conflicting incentives.** (Participant 5)

As described in the previous section ("Responsibility"), there are different expectations of the patient-provider relationship than of the participant-researcher relationship, in part due to different historical contexts. Even if the potential risks and benefits of participating in genomics research are not entirely different from those of receiving genomics-based clinical care, the

symbolic meanings of research and clinical care–and the implications for the relationship between the patient-participant and research and healthcare professionals–must be considered when engaging patient-participants in clinical research.

However, in a learning health model, those conducting research and those providing clinical care are not necessarily the same people. Although it is useful for dually trained clinician-scientists to be involved in the rapid learning process, it is not feasible for all people who are conducting research to also be trained in medicine, and vice versa. Instead, the boundaries between clinicians and researchers must be broken down to achieve seamless clinical research integration. While there are certainly structural facilitators for increasing collaboration between research and clinical personnel–such as dedicated time and funding for such collaborations–the relationships themselves are ultimately not dictated by structure or policy. Instead, they are driven by a shared curiosity and excitement for learning, mutual understanding of peoples' respective roles and expertise, and passion for the wellbeing of patient-participants and populations. These are essential components of what has previously been referred to as a "continuous learning culture" in an LHS (Davis et al. 2020, p. 3) [132]. Participant 18 described the researcher-clinician relationship as such:

> I think at academic places, there's probably a lot of **collaboration between the clinical people and the researchers, because they both need each other probably, right?** Like, the researchers need the clinical people for samples and for, you know, clinical information and things like that. And I think the clinicians, it's nice for them to have the researchers involved to **help them best care for their patients and learn new things about them.** (Participant 18)

Fostering a continuous learning culture is also highly dependent on the relationship between institutional leadership and those who work at a clinical research institution. Because it is not

typical for researchers and clinicians to work in a truly integrated fashion with one another, the nature of those relationships should be facilitated by a leadership team with guiding values, as Participant 14 described:

> And I find…that **people are resistant to change overall**, you know, **if it's not broken, don't fix it.** Right. It's more, more that and in **leadership**, I see this happen a lot, where you have to **get the team motivated for the why [emphasized] before you even start doing anything. And that takes a lot of TLC [laughs]. It takes a lot of people owning it before you start kind of shaking their trees, you know.** (Participant 14)

A key part of facilitating these collaborations is integrating ethical oversight of research and clinical care. When considering the potential risks and benefits of clinical genomics research, IRBs determine the scope of research, whereas a clinician's commitment to the patient determines the scope of clinical practice. Many interviewees expressed frustration with the process of obtaining IRB approval for conducting clinical research, acknowledging that although IRBs were in place for very good reasons, they sometimes constricted clinical research to the point of complete obstruction. As Participant 11 described, the discordance between traditional ethical oversight of research and the goals of clinical research raises questions of whether traditional methods of research oversight are truly protecting patient-participants, if potentially life-saving research is obstructed completely:

> Well, I think…the **IRB whose mission is, is laudable**…on the ground is, as one of my colleagues said, **inimical to research.** It is so [emphasized] **time consuming**, and, and **frustrating**. And I think that that, I mean, sometimes you think I'd like to do this project, but **I'm just not going to do the IRB. So I'm not going to do the project.** It's just, it's just too much work…[provides example]. So that's the kind of thing where you go, You know, seriously, guys? **Who are we protecting here?** What are we doing here? (Participant 11)

However, some interviewees cited instances in which ethical oversight of clinically integrated research worked particularly well at their institution. Participant 12 described the role of one

person who facilitates the relationship between the institution's IRB and the clinicians and researchers leading clinical research studies:

> …she kind of coordinates human subjects, including for studies that employ genomic data. And that's all she does. And she has kind of core responsibilities, and then people can pay her for the hours that she devotes to particular projects. And she stays abreast of all this and has a **strong working relationship with the IRB**. So the interface with the IRB has been pretty good at [institution name]. And I would say that that's not been a barrier. In general. It's handled pretty well. (Participant 12)

If there is a strong relationship between IRBs and those leading clinical research studies–either through a facilitator who has working knowledge of and relationships with the ethical and clinical research teams, or through the people who compose both teams–the perceived discordance between the responsibilities of ethics committees and the responsibilities of those conducting clinical research and working with patients can begin to be resolved. However, the two entities must be willing to evolve with one another as the pace and clinical implications of genomics research continue to change.

4.5.1.c Constraining factors

Systems are composed of people. However, people interact with their environments in ways that are not extricable from the characteristics of the larger systems in which they operate. In general, the US healthcare and research enterprises are financially separate from one another, which dictates the ways in which researchers and healthcare providers can bill research and clinical activities. The financial separation necessitates distinguishing research from clinical care, even if there is significant overlap between the two approaches in their processes and intentions. There was near consensus among participants that achieving true research-clinical integration was

nearly impossible without merging payers and providers. Participants 16 and 17 discussed this issue in the context of billing fraud, and the ways in which billing dictates the boundaries of clinical practice:

> Well, I think that's actually **the most sensitive issue**. And I think it has, and I give people the benefit of the doubt, but it has to do with **billing fraud**, right? And there have historically been cases, right, where there were issues of double dipping. So a single patient where they were double billed, **their insurance was billed for something and then a grant was billed for something and the institution was double dipping**, and double billing, and, you know, concerns on both sides, that there was fraud being committed. **And so if there's any place where people are really [emphasized] concerned, it's actually around that issue.** And, anyway, it's just, that's the way I've seen it evolve in terms of people being very, very careful about that issue specifically. (Participant 16)

> …so when you want to fix a lot of this stuff, all you...it's a very simple solution, all you need to do is **radically change the healthcare system of the United States, probably into a single payer system**. And that'll fix everything. So I'm just kidding. But I mean, I do think that's kind of a part of it, like, you want to know, what the barriers are, is that like, **a lot of our practice of medicine ends up being dictated by the billing.** (Participant 17)

Because the US healthcare system is driven by a mixed financial model of privatized health insurance coverage and publicly funded insurance coverage, there is little incentive for payers to invest in clinical genomics research. As several interviewees described, different payers are consistently aiming to pass costs off onto one another because it is very likely that patients will switch between insurance providers throughout their lifetime. Single payer models have the luxury of deciding how funds get distributed between research and clinical activities, whereas separate payer-provider models require that funding for research come from an entity other than the healthcare institution. This results in a disincentive for healthcare payers to invest in research that could potentially improve the efficiency of care and decrease costs in the long run, since

those cost benefits will likely benefit a different insurance provider, as Participants 7 and 20 described:

> And so, in our healthcare system, of course, **everyone is trying to pass off costs onto someone else. That's how the healthcare game is played, right?** Etna is hoping that Cigna will pay for the genome so they don't have to. Now in a country like England, **where you have a national health service, they're not playing that game.** And that's why **they're so far ahead of us**. And they're thinking, look, we own these people's health care for their entire lifetime, every problem, every healthcare problem they have until they die, is **our [emphasized] problem.** How are we going to address that problem? Well, one of the solutions to that is to **know as much about their health care liabilities as you can**, because they're all your problem, you might as well, you're better off knowing. Whereas Cigna says, I don't want to know, because odds are in five years, **they're going to be on somebody else's insurance plan. And it's not my problem.** (Participant 7)

> But if you're parsed out, if you have the insurer over here, and the hospital over here, you can't play those games, **everybody's out to try and maximize the margin that they're [emphasized] making**. And so the insurance says, Okay, we're going to impose these, you know, y'all have to do these guidelines. **The hospitals lose money, the insurance company is great.** And in other situations, you know, the hospitals do stuff that, you know, costs insurance companies money, and I mean **it's just this insanity, but it is our system.** (Participant 20)

The mixed insurance model in the US also propagates the need to maintain indication-based testing for genomic medicine, because insurance companies generally do not reimburse large-scale genomic tests without having a clear clinical indication for doing so. Therefore, while the tension between medical genetics as an indication-based specialty and the need for population-wide research genomes and exomes is partially driven by the Bayesian logic of limiting false positives, it is also exacerbated by the fact that genomes and exomes are not routinely collected due to insurance limitations. If one of the goals of learning healthcare is to use clinically generated data for research purposes, this payment model limits the number of patient-participants whose data can be used for research, and may also limit the ways in which the

existing data can be used for research. For example, Participants 11 and 19 noted that insurance companies must be "convinced" both of the clinical appropriateness of ordering a genomic test, and of the utility of conducting research using that data for the sake of maximizing the insurance company's profit:

> And if we're, you know, **if we get it under clinical [emphasized] dollars, it's because it's clinically appropriate to do but you might do something with that information that would be research based.** So for example, we did exomes in 27 people sequentially, because they were indicated. And the outcomes were, we got four results. And you might say, Okay, let's write a paper about screening in the clinic. And what's going to be, you know, where are we likely to get results? And where don't we get results? Or what are the hitches in getting that testing done? Or what did I, is it research if you go, I'm writing a paper about how insurance companies dealt with saying yes or no to getting exomes? So you know, it's, you know, **if the insurance company covers it, it's because we've managed to convince them that it's clinically appropriate to do so.** (Participant 11)

> I love [emphasized] research genomes and just the availability of, kind of being able to really look at the data in new and different ways. But for things like my clinical utility question, where we're trying to show the like, the, both the utility but also the financial aspects and the economic utility of genome sequencing, I think we have to think about the clinical arm of that. And so i**f you're trying to say that insurance should pay for this, then you need to model your study around the product that insurance is getting.** (Participant 19)

An additional layer of complexity in the discussion of healthcare and research reimbursement is the question of how funding should be distributed within and between research and clinical efforts, regardless of the source of funding. Interviewees agreed that money was a relatively finite resource, and that funding limitations informed many of the cost-benefit tradeoffs that are routinely made in medicine. There is a spectrum of costs in clinical testing and management, and within that spectrum there are patients on their own spectrum of lowest clinical need to highest clinical need, as Participants 17 and 6 described:

> I heard someone give a talk and they said, they were talking about a health care system and they said, **Well, no one's healthy, they don't care, and it's certainly not a system.** And I think that's, I think with something like genetics when you get into tests that are really expensive, **it's unfortunate when they get overused because every health care dollar comes from someplace else**. (Participant 6)

> And, you know, when you ask, **I don't think that research and, and clinical management are two different things.** But **when it comes to genetic testing, we do make a big distinction between them**, because, you know, we have, we don't have enough money for everyone right now to get everything that they need. **So we have to ration it out to the people who need it the most.** (Participant 17)

While the costs of exome and genome sequencing have dropped significantly in the past decade, and will continue to drop, interviewees expressed concerns that obtaining population-wide genomic data would overwhelm the healthcare system in terms of the follow-up implications of that data. Providers have a responsibility to provide the best care possible to their patients given the best possible information, but the implications of having the best possible data for all people in a healthcare system are morally sticky when there are limited resources to ensure appropriate clinical follow-up. As Participant 8 described:

> …if you have a population of 700,000, and one in 100, or one in 200 has a BRCA 1 or 2 pathogenic variant, what does that mean when you identify all those patients, in terms of **the bolus that's going to come to your surgical teams, to your screening teams, and all of that**…so I've seen extensive spreadsheets, about, you know, people being so **concerned about what that bolus is.** (Participant 8)

Finally, the ways in which clinical and genomic data can be shared for research are partially dictated by the national funding and data governance model. In a monolithic system where data is shared across many participating institutions, it is easier to conduct population-wide research using those data. Interviewees cited programs like Deciphering Developmental Disorders (DDD) in the UK as examples of national research projects that have harnessed vast repositories of

clinical and genomic data shared across a nationalized health system, and subsequently contributed potentially clinically actionable discoveries at a rapid pace:

> And so the **huge GWAS studies…**which might have over a million participants, are **only possible in large governmental health care models that we don't have in the states.** (Participant 12)

However, as previously discussed, the legal, technical, and financial aspects of sharing clinical and genomic data are far from the only challenges of sharing data. Arguably, the more pressing challenge in the US is gaining the trust and willing participation of the people whose data would be shared both within and across institutions. In this way, there are constant interactions between the structural, ethical, and social aspects of conducting clinical research that necessitate a continuous process of evaluating the local and national ideas of what healthcare and research stand for as institutions.

4.5.2 Dynamic meanings of data, knowledge, and practice

Tensions at the research-clinical interface are at the root of the challenges in a GLHS model. From the symbolic interactionist perspective, "language and symbols play a crucial role in forming and sharing our meanings and actions" (Charmaz 2014, p. 262) [25,158]. Interviewees offered several alternate definitions of research and clinical care to explain the motives and goals behind them. Phrases such as "[studying] a question," "[increasing] the knowledge for the field as a whole," "[hypothesizing] something," and "[doing] something novel" were used to describe research. The only definition of clinical care that was offered was "trying to save [a] life." The collection of definitions that describe research is not necessarily separable from the Hippocratic

oath of medicine when the goal of research is to improve human health. External structures may constrain clinical research integration, but people with a common goal can work together to adapt to structural constraints, as the symbolic interactionist perspective suggests:

> **Structures exist and persist but some individuals may resist, circumvent, or ignore these constraints or use them for their own purposes**. Institutionalized values and practices precede and **constrain individuals** and **set the conditions for possible actions**, although **how they respond** to these conditions can vary (Charmaz 2014, p. 269) [25].

Genomic medicine may not be the singular saving grace of humanity, but if it is guided in a direction where the research is well-designed, clinically relevant, and representative of all populations, it can vastly improve medicine. Participant 16 described their vision for the future of genomic medicine as follows:

> So I do have this fantasy, that I'm trying to make reality, but this fantasy that **the next generation will grow up differently.** And so it'll start with the diagnosis, **early diagnosis** at a time when you can actually **prevent and treat conditions.** So it won't just be reactive, but it'll be more **proactive**. (Participant 16)

4.6 Limitations and future work

Although the data gathered during this project were in-depth for each participant, a relatively small number of participants were interviewed. However, code groups did appear to reach theoretical saturation after about 15 interviews, which indicates that including a larger number of participants may not have altered the dataset significantly. A lack of racial and ethnic diversity among interviewees was also a limitation of this work, given that 80% of the participants self-identified as White. While this is reflective of an overall lack of diversity in the human genetics and genomics workforce [159], efforts should be made to include a diverse group of geneticists, patient-participants, researchers, and other stakeholders in future conversations. It is also

important to recognize the role of the primary investigator (K.F.) in this project. Although the final model was grounded in the available data, the interpretation of those data was highly dependent on the ways in which the primary investigator interpreted and synthesized the data. The model presents one possible interpretation of the data, but there are many more possible interpretations.

4.7 Conclusion

The conceptual model developed during this study offers a novel approach to understanding the research-clinical interface in genomics. The tensions at the interface of clinical care and research in genomics are the basis for the GLHS model. They manifest in questions of data and human representation, in questions of ethical and occupational responsibilities, and in questions of risk-reward tradeoffs. They are embodied in the real and symbolic relationships that people form as they occupy roles that are created by and for them. In the end, the research-clinical interface is defined by those who participate in constructing its meaning and is bounded by the structures and cultural expectations that emerge from history to right historical wrongs. In the next chapter, we will demonstrate the power of using merged clinical and genomic data for gene-disease association discoveries, using *C. diff.* infection as a clinical use case. These types of studies could be conducted on a routine basis in LHS environments and could rapidly offer insights into potential biological mechanisms and treatment targets for common and complex diseases.

CHAPTER 5: DISCOVERY OF GENETIC RISK FACTORS FOR CLOSTRIDIOIDIES DIFFICILE INFECTION USING MERGED CLINICAL AND GENOMIC DATA (AIM 3)

5.1 Introduction

One major goal of integrating research and clinical care in genomics is to more rapidly and accurately detect new gene-disease associations, which are critical for advancing genomic medicine. In contrast with rare monogenic disorders, susceptibility to common health issues and diseases like diabetes and hypertension is primarily driven by multiple genetic and environmental factors [160]. Characterizing gene-disease associations is an important first step in identifying causal variants associated with complex diseases, and ultimately in developing targeted therapies and treatments for those diseases [161]. Genome-wide association studies (GWAS), which screen for gene-disease or disease-trait associations across the entire genome without a prior hypothesis, are commonly used to detect new associations [162]. Larger sample sizes, enhanced genome annotations, and improved sequencing and analysis technologies are expected to drive the prevalence and impact of GWAS. Additionally, analyzing genetic data with EHR data can allow for richer and more cost-effective GWAS [36]. Research programs like the Electronic Medical Records and Genomics (eMERGE) Network have demonstrated the utility of using EHR data to detect genetic loci associated with conditions like hypothyroidism and type 2 diabetes [163–165] and with clinical traits like erythrocyte sedimentation rate and white blood cell count [166,167]. The network has developed and validated 68 clinical phenotypes across multiple EHR systems since its inception in 2007 [16].

CDI is one such clinical phenotype developed by the eMERGE Network. In this aim, we use merged genetic and clinical data from the eMERGE Network to conduct a logistic regression based GWAS of CDI cases and controls to identify common genetic variants associated with higher risk of developing CDI. We also demonstrate the utility of using clinical data for gene-disease association studies and provide a practical example of clinical genetic discovery in action.

5.2 Related work

5.2.1 History and future directions of gene-disease associations

Genetic association studies are designed to identify genetic variants that are associated with a particular disease or phenotype, typically by comparing genotypes in affected and unaffected individuals using a case-control design [168]. Since the advent of large-scale genotyping and genomic sequencing, it has become clear that genetic contributions to human health and disease are extremely variable, especially in the case of complex and common disease [169]. Evolutionary forces have generally caused variants with large phenotypic consequences to be removed from the population and have allowed variants with small individual phenotypic effects but large cumulative effects to reach higher population frequencies [161]. Although there are exceptions to this trend, it has nonetheless impacted the ways that variants associated with rare and common diseases are typically detected. Because the variants contributing to rare monogenic (Mendelian) diseases with severe phenotypic effects are most commonly found in coding regions of the genome, exome sequencing is commonly used to determine the causal variants for Mendelian diseases [170,171]. Detecting the causative gene typically involves duo, trio, or

extended family sequencing [172]. Common disease risk, on the other hand, is driven by many variants that may or may not be in coding regions of the genome, which necessitates the use of genome or DNA microarray data for risk variant identification [173,174]. GWAS are currently the standard method for detecting variants associated with non-Mendelian disease because they offer a relatively unbiased approach to identifying common marker variants in disease [175]. Although other methods must be used to infer variant causality and identify rare variants associated with common disease, GWAS results lay the groundwork for targeted analyses of potential genetic drivers in many common diseases [176].

5.2.2 Gene-disease associations using electronic health record data

Given the small effect sizes of variants associated with common diseases, large sample sizes are required to run sufficiently powered GWAS [177]. Traditionally, GWAS are conducted using "purpose-built cohorts" where high-quality genetic and phenotypic data are collected using "self-report questionnaires and/or clinical staff" (Wei & Denny 2015, p. 1) [178]. Although the costs of large-scale genotyping and genome sequencing have decreased over the past decade, this prospective approach can be time-consuming, expensive, and yield insufficient sample sizes. Large, pre-existing biorepositories, like the UK Biobank [74], can be used to conduct genetic association studies much more cost-effectively, and using sufficiently large patient cohorts [174]. Although public repositories like the UK Biobank typically offer granular phenotype information like International Classification of Disease (ICD) codes, complex phenotypes may not be accurately captured using only a de-identified subset of clinical data [179]. Linking genetic data with EHR data has therefore been proposed as a cost-effective and clinically relevant approach to prospective gene-disease association research [36,178].

Although EHRs are not immediately amenable to research due to data quality and accessibility limitations, they contain a wealth of information that can be extracted using an interdisciplinary set of tools. Because most EHRs were designed to support billing and routine clinical care, rather than research, they support a patchwork of structured and unstructured data in the form of billing codes, laboratory test results, ICD diagnosis codes, procedure codes, prescription information, and narrative reports [180]. Each of these data types can be leveraged to construct rich clinical phenotypes that can be used in case-control GWAS. The eMERGE Network has led this field of research by demonstrating the value of iterative phenotyping algorithm development and validation, during which informaticists, clinical content experts, epidemiologists, and geneticists collaborate to refine the algorithm and enhance its accuracy [181]. EHR-driven genetic analyses are expected to become more routine in the coming decades, but additional studies demonstrating their utility and laying out best practices in EHR phenotyping are required to advance research in this area [36].

5.2.3 Pathophysiology and genetic susceptibility to *C. diff.* infection

CDI is the leading infectious cause of nosocomial diarrhea in North America and is associated with a high global burden of disease [37]. Once acquired, this reemerging, Gram-positive, spore-producing bacteria secretes a toxin that causes watery diarrhea, and can progress to severe pseudomembranous colitis, toxic megacolon, and sepsis [182]. In the early 2000s, the emergence of *C. diff.* strain NAP1/BI/027 led to increased incidence, prevalence, morbidity, and mortality associated with CDI [183,184]. This epidemic strain produces more toxin, has a higher resistance to common treatments, and causes more recurrent infections than other common *C. diff.* strains. Despite aggressive antibiotic treatment (e.g. vancomycin, metronidazole) and fecal transplant

[185,186], outcomes of NAP1/BI/027 CDI include significant morbidity across all age groups, 5% mortality in individuals older than 65 years of age, and an estimated $1.1 billion dollars per year in health care costs [182].

Asymptomatic colonization with *C. diff.* is common among patients in acute care and long-term care settings, with an estimated prevalence of 3%-26% in younger adults and 5%-7% in older adults [187]. Progression from *C. diff.* colonization to acute CDI is generally associated with one or more risk factors [188], including new exposure to *C. diff.*, older age, hospitalization or nursing home residency, chemotherapy, severe comorbid illness, proton pump inhibitor, transplant medication or corticosteroid use, or prior use of high-risk antibiotics such as fluoroquinolones or cephalosporins [189–191]. Antibiotic use and proton pump inhibitor use are also risk factors for recurrent CDI [192]. Despite having one or more risk factors, some *C. diff.* carriers either do not develop CDI or successfully clear an initial infection, while some individuals are burdened by severe and/or recurrent CDI. This differential susceptibility may have a genetic component, given that genetic variation underlying susceptibility to infectious disease is well documented for other infections, including enteric infections such as *Helicobactor pylori* [193]. Identification of genetic susceptibility loci could yield methods for prevention and/or treatment of this important pathogen [194,195].

Previous studies have identified candidate risk loci for primary and recurrent CDI in small patient populations using a combination of genetic and clinical data. Apewokin et al. (2018) [196] performed a genome-wide logistic regression analysis of CDI in 646 patients (57 cases; 589 controls) undergoing stem cell transplantation for multiple myeloma, and found several

single nucleotide variants (SNVs) in the *RLBP1L1*, *ASPH,* and *P7B* genes that were associated with higher risk of CDI. Shen et al. (2020) [197] identified two alleles in in the extended major histocompatibility complex (MHC; *HLA-DRB1*07:01* and *HLA-DQA1*02:01*) that were associated with a reduction in CDI recurrence among 704 patients who achieved initial clinical cure with bezlotoxumab treatment in the MODIFY clinical trials. Several studies have also suggested that common SNVs in the promoter region of the interleukin-8 (IL-8) gene may confer increased risk for CDI by altering neutrophil recruitment during disease pathogenesis [198,199]. While these results are collectively suggestive of genetic involvement in CDI risk, the aforementioned studies had small sample sizes and did not always control for major risk factors such as previous antibiotic use or corticosteroid use in their association models. GWAS that properly control for known risk factors and include a large number of participants are needed to identify risk loci with sufficient power and reliability. One such study identified 16,464 patients (1,160 cases; 15,304 controls) from the Geisinger MyCode cohort [38] using a *C. diff.* phenotyping algorithm developed by the Electronic Medical Records and Genomics (eMERGE) Network. The authors identified several MHC variants with predicted functional impacts on nearby genes among European-ancestry patients treated with antibiotics, but these variants did not reach genome-wide significance. Additional validation studies in other large patient cohorts are needed to evaluate the role of factors in CDI risk.

5.3 Methods

5.3.1 Participants

Cases and controls were selected from among the ~99,000 participants of the eMERGE Network. Participating sites included the following: 1. The Children's Hospital of Philadelphia, Philadelphia, PA; 2. Cincinnati Children's Medical Hospital, Cincinnati, OH; 3. Columbia University, New York, NY; 4. Geisinger, Danville, PA; 5. Mass General Brigham, Boston, MA; 6. Kaiser Permanente Washington (formerly Group Health Cooperative) and University of Washington partnership, Seattle, WA; 7. Marshfield Clinic, Marshfield, WI; 8. Mayo Clinic, Rochester, MN; 9. Meharry Medical College, Nashville, TN; 10. Mount Sinai, New York, NY; 11. Northwestern University, Evanston, IL; and 12. Vanderbilt University, Nashville, TN. Informed consent was obtained from participants by each eMERGE site. The eMERGE study was approved by each participating site's institutional review board.

5.3.2 Case-control selection using *C. diff.* phenotyping algorithm

C. diff. cases and controls were selected using a variety of information contained in the EHR, including International Classification of Disease (ICD) Clinical Modification (CM) codes 9th and 10th editions, lab and medication data, and clinician progress notes. The *C. diff.* phenotyping algorithm used in this study was designed collaboratively by University of Washington, Group Health and Vanderbilt as part of the eMERGE Network and was published in the Phenotyping KnowledgeBase (PheKB) in 2012 [200,201]. Case/control selection and exclusion criteria are depicted as a flowchart in **Figure 5.1**.

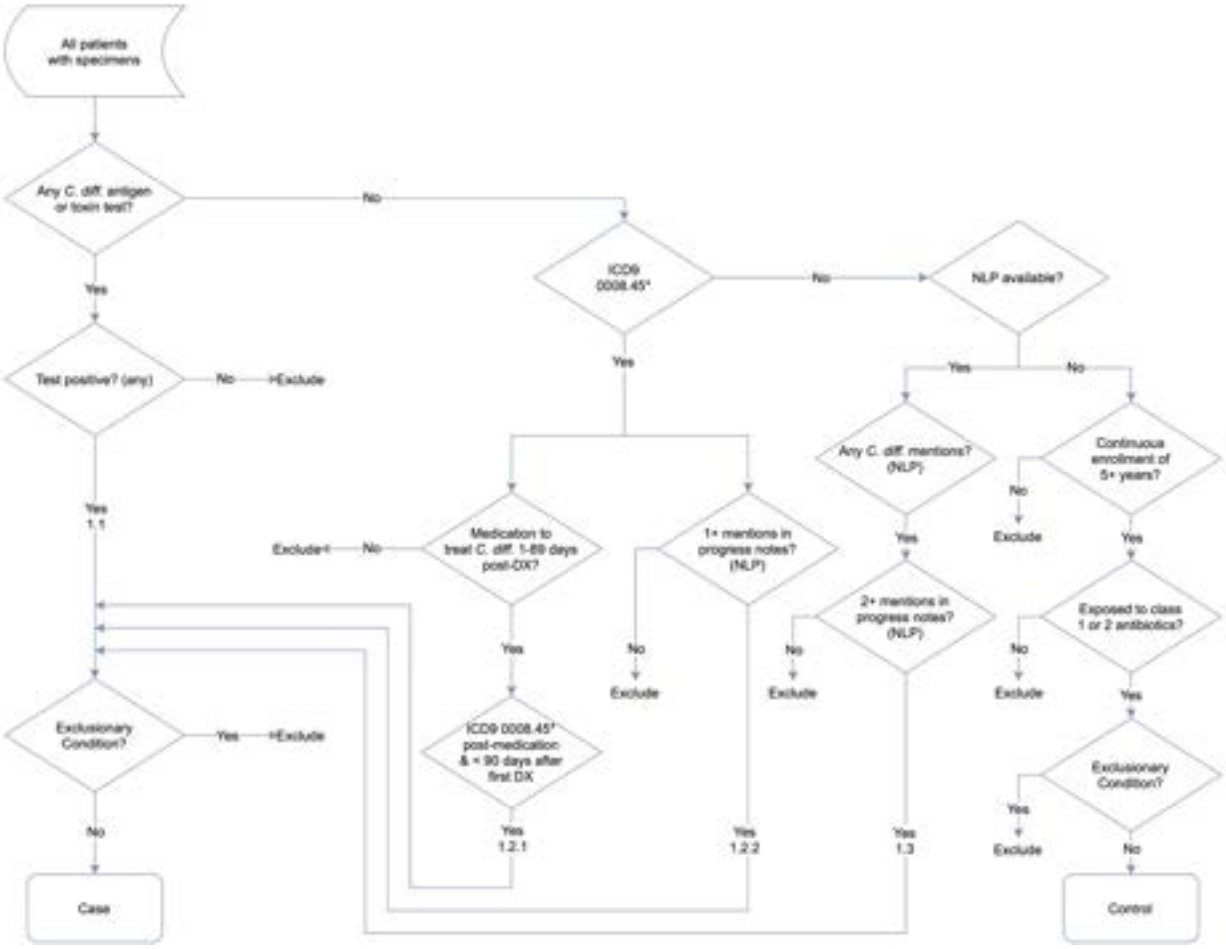


Figure 5.1. eMERGE *C. diff.* phenotyping algorithm flowchart.

For participants aged two years or older, there were four combinations of EHR data considered for case selection. First, individuals with a positive *C. diff.* antigen or toxin test were selected. Second, those with one or more inpatient or outpatient diagnoses of *C. diff.* (ICD-9-CM code 008.45; ICD-10-CM code A047), followed by one or more days of medication for treatment (metronidazole, oral vancomycin, fidaxomicin, or linezolid), followed by another inpatient or outpatient *C. diff.* diagnosis code, were selected. Third, individuals with at least one *C. diff.* ICD-

CM code combined with at least one affirmative mention (unqualified by negation, uncertainty, or historical reference) of *C. diff.* infection in a clinical progress note as identified through natural language processing (NLP), were selected. The *C. diff.* mentions used by the NLP algorithm are listed in **Table S5.1**. Finally, individuals with two or more affirmative mentions of *C. diff.* infection on separate calendar days in clinical progress notes, identified by NLP, were selected. To exclude severely immune-compromised participants from the test population, participants meeting one of the four above criteria were excluded from being cases if they had a diagnosis of bone marrow cancer in the two-year period prior to their *C. diff.* case index date (i.e. the first positive lab test, diagnosis code or progress note mention), or within seven days following their index date. Participants were also excluded from being cases if they had received chemotherapy in the 180-day period prior to their *C. diff.* index date, or within seven days following their index date. Using these criteria, 1,598 cases were selected.

Controls were selected from eMERGE participants two years of age or older who had no known test for and no diagnosis codes for *C. diff.* in their records. Since *C. diff.* toxin tests have sensitivities ranging from 60 to 70% [202], a single test does not rule out disease, and multiple tests could signal a concern that disease exists. Additionally, controls must have had at least one hospital admission with a prior exposure to a high-or moderate-risk antibiotic (**Table S5.2**) in the 7 to 62-day period before admission. Alternatively, they must have had exposure to a high or moderate-risk antibiotic and had five or more years of documented clinical visits following exposure with no mention of *C. diff.* infection in their progress notes. Participants meeting the control criteria were excluded if they had chemotherapy or bone marrow cancer in the 180-day period prior to the *C diff.* control index date (i.e., the earliest hospital admission with antibiotic

exposure or earliest antibiotic exposure with five years of follow-up), or within seven days following the index date. These criteria resulted in the selection of 23,061 eMERGE participants as controls. We excluded 202 cases and 2,723 controls that were missing genotype data. An additional 31 cases and 889 controls were excluded because the genotype imputation quality failed to meet our quality control (QC) threshold (mean $R^2 > 0.3$) [33].

Cryptic relatedness was assessed in all participants by calculating the probabilities of sharing alleles identical by descent (IBD), where Z0 is the probability of sharing zero alleles IBD and Z1 is the probability of sharing one allele IBD. Families were constructed when sample pairs had $Z0 < 0.83$ and $Z1 > 0.1$ [33]. When study participants were found to be in the same family, we prioritized the inclusion of cases. In situations where two or more cases or two or more controls were found to be in the same family, one participant was selected at random, and the others were excluded. For participants selected via the *C. diff.* phenotyping algorithm, 9 cases and 937 controls were excluded due to cryptic relatedness.

5.3.3 Covariates identified for phenotyping algorithm sample

The following covariates were identified for all cases and controls using structured EHR data: 1. Age at index date (index age); 2. Body mass index (BMI); 3. Sex; 4. Genetically determined ancestry; 5. Nursing home status (y/n); 6. Chemotherapy (y/n); 7. Diabetes mellitus (y/n); 8. Human immunodeficiency virus (HIV) positive status (y/n); 9. Any transplant medications (y/n); 10. Any corticosteroid medications (y/n); and 11. Any medium or high-risk antibiotic exposure (y/n). We used the median BMI record for the age year that matched most closely to the participant's index age. Nursing home status was determined either by structured data on skilled

nursing facility residence, or by mentions of nursing home status in social work and case management notes, as identified by NLP (**Table S5.3**). We flagged chemotherapy using Current Procedural Terminology (CPT) codes 96400, 96408, 96409, 96411-96425, 96520, and 96530. We flagged participants as having diabetes mellitus if they had at least two of the following three indications: 1. An ICD-CM code from ICD-9-CM 250.* or ICD-10-CM E08-E13.*; 2. Prescriptions for diabetes medications including insulin (**Table S5.4**); or 3. A hemoglobin A1C (HbA1C) reading > 6.5% or a glucose reading of > 200 mg/dL. Participants were flagged as having HIV infection if they had one instance of ICD-9-CM 042.*, ICD-10-CM B20-B24.* or Z21.*. Patients were flagged as having been exposed to transplant or corticosteroid medications if any medication listed in **Table S5.4** was administered outside of the exclusionary time range.

5.3.4 Genotyping and imputation

Genotypes for all participant samples from eMERGE-I, eMERGE-II and eMERGE-III were imputed using the Michigan Imputation Server [203]. The server uses the Minimac3 algorithm to impute missing genotypes and uses the Haplotype Reference Consortium reference panels [204] (HRC1.1) as the reference set. The majority of samples from the 13 eMERGE sites were genotyped on the Human 660 Quad (eMERGE-I). Other genotyping platforms included the CytoSNP-850K BeadChip, the OmniExpress chip, the Affymetrix 6.0 array, and the Illumina MEGA among others. In this analysis, variants with an allelic R^2 >= 0.3 and minor allele frequency (MAF) >= 0.05 were included. Additional QC filters were applied as described in case-control selection.

5.3.5 Genetically determined ancestry

The set of ~99,000 unique imputed samples was analyzed by Principal Component Analysis (PCA) using the PLINK 2.0 software [205]. Variants with >= 0.05 MAF, missingness of <= 0.1 and LD-pruned R^2 threshold of 0.7 were included in the multisample analysis. K-means clustering of Principal Component (PC) 1 and PC2 identified three groups (corresponding to African ancestry, Asian ancestry, and European ancestry) was used to find genetically determined ancestry (GDA) of each sample. GDA and self-described ancestry were checked for concordance, and samples were ultimately grouped into African ancestry, Asian ancestry and European ancestry clusters, respectively. IBD was calculated for all pairwise sample comparisons using the plink --genome function, and cryptic relatedness between samples was assessed as described in case/control selection.

5.3.6 Genome-wide association study

To identify genetic variants associated with CDI, we performed logistic regression-based association analyses for the case/control curated phenotype using PLINK 1.90 [206]. All covariates and genotypes were used in the joint analysis of all participants, whereas the PC1 and PC2 covariates for the African and European ancestry-stratified analyses were derived from ancestry specific PCA analyses. An additive genotypic model of SNV genotypes coded as 0, 1 or 2 copies of the minor allele was used. The regional linkage disequilibrium (LD) plots of the index SNV were created using the LocusZoom web-based tool [207]. Following the initial stratified analyses, an additional logistic regression-based association analysis was performed in

the European sample using the index SNV as a covariate to determine whether this SNV was truly driving the risk association.

5.3.7 Human leukocyte antigen association analyses

Classical HLA alleles were imputed against four ancestry-specific reference panels (African, Asian, European, and Hispanic) using the HIBAG software [208]. *HLA-DRB3*, *4* and *5* gene dosages were inferred based on the *HLA-DRB1* alleles present in each individual, as described in Habets et al. (2018) [209]. Calls were quality-filtered for a HIBAG posterior probability of > 0.5.

To test for haplotype-specific effects of the most significantly associated SNVs, four overlapping participant subgroups were selected from the European ancestry sample based on the presence of at least one of the following: (1) *DRB3* gene; (2) *DRB4* gene; (3) *DRB5* gene; or (4) any of the above genes in each participant. Haplotype subgroups were further divided into DR15 and DR16 haplotype carriers (stemming from the *DRB5* gene carriers, or DR51 haplotype family), and DRB1*15:01 carriers (stemming from the DR15 haplotype). Logistic regression-based association analysis was performed separately in each haplotype subgroup, using the same covariates described in “Methods: GWAS” for the European ancestry sample.

To test for HLA alleles driving the association, case-control logistic regression-based association analysis was performed in the European ancestry population sample for 276 classical HLA alleles, using the same covariates described in “Methods: GWAS” for the European ancestry sample. The CEU Chromosome 6 LD dataset from the HapMap 3 project was used to assess LD of the most significantly associated SNVs among classical HLA alleles.

5.4 Results

5.4.1 Demographics

After all exclusions, there were 1,349 cases and 18,512 controls identified via the eMERGE *C. diff.* phenotyping algorithm (**Table 5.1**). Approximately 74% of cases and controls self-identified as White, and 19% self-identified as Black or African American. Although older age is a known risk factor for *C. diff.* infection [191], controls tended to be older than cases (z=14.37, P=2.20 x 10^{-16}), which reflected the patient populations of the participating eMERGE study sites. Controls also tended to have higher BMIs than cases (z=14.58, P=2.20 x 10^{-16}). Cases had slightly higher exposure to Class 1 (high-risk) antibiotics than controls (28% vs. 21%), yet they had much less exposure to Class 2 (moderate risk) antibiotics than controls (11% vs. 26%). More cases received chemotherapy outside of the exclusionary time period than did controls.

N	**Case n=1,349**	**Control n=18,512**	**Overall n=19,861**
Site			
Children's Hospital of Philadelphia	11% (149)	1.4% (265)	2.1% (414)
Cincinnati Children's Medical Hospital	1.0% (14)	0.0% (0)	0.1% (14)
Columbia	5.6% (76)	0.5% (88)	0.8% (164)
Geisinger	4.2% (57)	4.9% (899)	4.8% (956)
Kaiser Permanente/UW	4.2% (57)	11% (2128)	11% (2185)
Mass General Brigham	3.5% (47)	8.8% (1623)	8.4% (1670)
Mayo Clinic	7.2% (97)	17% (3127)	16% (3224)
Marshfield	2.4% (32)	4.7% (861)	4.5% (893)
Mt. Sinai	7.9% (106)	15% (2776)	15% (2882)
Northwestern	5.6% (76)	2.0% (362)	2.2% (438)
Vanderbilt	47% (638)	34% (6383)	35% (7021)
Sex (female)	51% (690)	55% (10232)	55% (10922)
Median BMI (kg/m^2)*	20.8,25.2,29.8	24.4,28.1,32.9	24.2,28.0,32.8
Median age*	39.7,57.3,70.0	51.1,64.9,76.1	50.4,64.4,76.0
Self-identified race			
American Indian or Alaska Native	0.2% (3)	0.2% (40)	0.2% (43)

Black or African American	15% (196)	19% (3562)	19% (3758)
Asian	0.8% (11)	0.8% (142)	0.8% (153)
Native Hawaiian or other Pacific Islander	0.07% (1)	0.02% (2)	0.02% (3)
White	75% (1008)	74% (13716)	74% (14724)
Unknown	9.2% (124)	5.0% (933)	5.3% (1057)
Not reported	0.4% (6)	0.6% (117)	0.6% (123)
Self-reported ethnicity			
Hispanic or Latino	6.0% (81)	4.8% (895)	4.9% (976)
Not Hispanic or Latino	88% (1193)	92% (17120)	92% (18313)
Unknown	5.6% (75)	2.7% (497)	2.9% (572)
Genetically determined ancestry			
African	17% (235)	21% (3849)	21% (4084)
Asian	2.4% (32)	1.6% (287)	1.6% (319)
European	80% (1082)	78% (14376)	78% (15458)
>=1 HLA-DRB3, 4 OR 5 gene	71% (955)	72% (13336)	72% (14291)
>=1 HLA-DRB3 gene (DR52)	41% (559)	50% (8328)	45% (8887)
>=1 HLA-DRB4 gene (DR53)	36% (507)	40% (7356)	40% (7863)
>=1 HLA-DRB5 gene (DR51)	22% (299)	21% (3831)	21% (4130)
Antibiotic exposure (Within 7-62 days prior to index date)	28% (376)	21% (3832)	21% (4208)
High risk	11% (147)	26% (4838)	25% (4985)
Moderate risk	1.9% (25)	1.5% (284)	1.6% (309)
Low risk	59% (801)	52% (9558)	52% (10359)
No exposure			
Cancer (First record to index date + 7 days)	20% (272)	14% (2520)	14% (2792)
Chemotherapy (Before 180 days prior to index date, after 7 days following index date)	20% (270)	12% (2263)	13% (2533)
Diabetes Mellitus (Ever)	24% (326)	25% (4700)	25% (5026)
HIV (Ever)	3.0% (44)	2.0% (302)	2.0% (346)
Nursing Home Status (Within 90 days prior to index date)	11% (147)	2.0% (393)	3.0% (540)
Corticosteroid medications (Within 21 days prior to index date)	17% (227)	10% (1848)	10% (2075)
Transplant medications (First record to index date + 7 days)	19% (250)	6.0% (1059)	7.0% (1309)

Table 5.1. Summary statistics of demographic data and phenotypes for *C. diff* cases and controls selected using the *C. diff* phenotyping algorithm. Significant differences between case and control distributions (as determined by chi-squared test for binary variables and two-sided Z-tests for continuous variables) are shown in **bold**. *The three numbers for body mass index (BMI) and age represent the 25th, 50th and 75th quartiles of the distribution.

5.4.2 Genome-wide association study

Table 5.2 summarizes the logistic regression association results that reached genome-wide significance in the combined and European ancestry-only samples, with corresponding summary statistics for those findings in the African ancestry-only sample. A strong association in the human leukocyte antigen (HLA) region was found in the European and joint ancestry samples (**Figure 5.2; Figure S3.2**) but was not found in the African ancestry sample. The lack of association in the African ancestry sample could be due to either insufficient detection power as a result of small sample size or different haplotype or LD structures compared to individuals of European ancestry. Manhattan plots and corresponding QQ plots for the European, joint, and African ancestry GWAS analyses are provided (**Figures S5.1-S5.5**). The five most significantly associated SNVs driving the association in the European sample (rs68148149, P=8.06 x 10^{-14}; rs3828840, P=9.96 x 10^{-14}; rs35882239, P=8.18 x 10^{-12}; rs35882239, P=5.12 x 10^{-11}; rs35222480, P=9.88 x 10^{-11}) mapped to the intergenic region between the *HLA-DRB5* and *HLA-DRB1* genes in the beta block of the MHC Class II region. Three of the five most significant SNVs (rs3828840, rs35882239, and rs35222480), with minor allele frequencies (MAFs) of 0.17, 0.17 and 0.20, respectively, also mapped to the 3' end of the HLA-DRB6 pseudogene.

Chr	SNV	Ref	Alt	CA	BP	Joint CAF (n=19861)	Logistic Joint P-value	EUR CAF (n=15458)	Logistic EUR P-value	Logistic EUR SNV-controlled P-value	AFR CAF (n=4084)	Logistic AFR P-value
							OR (95% CI)		OR (95% CI)	OR (95% CI)		OR (95% CI)
							Beta		Beta	Beta		Beta
6	rs68148149	C	A	C	32511725	0.17	**6.83 x 10-9**	0.17	**8.06 x 10-14**	0	0.18	7.2 x 10-1
							1.36 (1.06-1.74)		**1.56 (1.13-2.15)**	0		0.93 (0.80-1.13)
							0.13		**0.2**	0		-0.02
6	rs3828640	T	C	T	32520907	0.17	**8.42 x 10-9**	0.17	**9.98 x 10-14**	0	0.18	7.1 x 10-1
							1.36 (1.06-1.74)		**1.56 (1.13-2.15)**	0		0.95 (0.79-1.13)
							0.13		**0.2**	0		-0.02
6	rs35882239	A	G	A	32522576	0.2	**1.32 x 10-8**	0.21	**8.18 x 10-12**	9.80 x 10-1	0.2	6.7 x 10-1
							1.34 (1.05-1.70)		**1.49 (1.10-2.00)**	1.00 (1.00-1.00)		0.94 (0.78-1.13)
							0.13		**0.17**	0		-0.03
6	rs71534541	C	T	C	32513076	0.08	7.98 x 10-7	0.07	**5.13 x 10-11**	2.30 x 10-1	0.1	8.2 x 10-1
							1.38 (1.04-1.80)		**1.62 (1.12-2.33)**	1.15 (0.90-1.46)		0.96 (0.81-1.14)
							0.14		**0.21**	0.06		-0.02
6	rs35222480	A	T	A	32522813	0.08	8.41 x 10-7	0.08	**9.88 x 10-11**	2.20 x 10-1	0.1	5.0 x 10-1
							1.37 (1.04-1.80)		**1.59 (1.11-2.26)**	1.14 (0.90-1.44)		0.89 (0.66-1.19)
							0.14		**0.2**	0.06		-0.05
6	rs138603449	C	T	T	32595194	0.21	**6.59 x 10-9**	0.21	**5.42 x 10-10**	**4.54 x 10-9**	0.22	8.73 x 10-2
							1.31 (1.05-1.62)		**1.39 (1.07-1.80)**	**1.37 (1.06-1.77)**		1.24 (0.90-1.70)
							0.12		**0.14**	**0.14**		0.09
6	rs9270896	A	G	G	32571876	0.41	1.27 x 10-5	0.42	1.21 x 10-5	**6.08 x 10-9**	0.33	3.96 x 10-2
							1.19 (1.01-1.40)		1.22 (1.01-1.47)	**1.32 (1.05-1.65)**		1.26 (0.92-1.74)
							0.08		0.09	**0.12**		0.1
6	rs9270894	A	G	G	32571872	0.26	1.17 x 10-5	0.24	1.06 x 10-6	**1.13 x 10-8**	0.32	1.16 x 10-1
							1.22 (1.01-1.47)		1.29 (1.03-1.63)	**1.37 (1.06-1.77)**		1.20 (1.90-1.58)
							0.09		0.11	**0.14**		0.08
6	rs9270895	C	T	T	32571873	0.45	5.95 x 10-5	0.44	5.39 x 10-5	**2.32 x 10-8**	0.42	3.54 x 10-2
							1.17 (1.00-1.37)		1.21 (1.00-1.45)	**1.31 (1.05-1.64)**		1.26 (0.92-1.73)
							0.07		0.08	**0.12**		0.1
6	rs616095	G	A	A	32574736	0.26	5.05 x 10-7	0.25	2.69 x 10-6	**3.71 x 10-8**	0.36	1.19 x 10-2
							1.26 (1.03-1.53)		1.29 (1.02-1.62)	**1.35 (1.05-1.73)**		1.32 (0.94-1.87)
							0.1		0.11	**0.13**		0.12

Table 5.2. Index SNV results from logistic regression-based genome wide analysis for joint ancestry (n=19,861), European ancestry (n=15,458), and African ancestry (n=4,084) samples. An additive model was used to assess the disease susceptibility impact of the minor (coded) allele at each position, while controlling for age, BMI, sex, ancestry, nursing home status, chemotherapy, diabetes, HIV, transplant medications, corticosteroids, and medium or high-risk antibiotic exposure as covariates. Results meeting the genome-wide significance threshold ($P < 5 \times 10^{-8}$) are displayed in bold. Abbreviations: Chr = Chromosome; SNV = Single Nucleotide Variant; Ref = Reference Allele; Alt = Alternate Allele; CA = Coded Allele; BP = Base Pair; CAF = Coded Allele Frequency; OR = Odds Ratio.

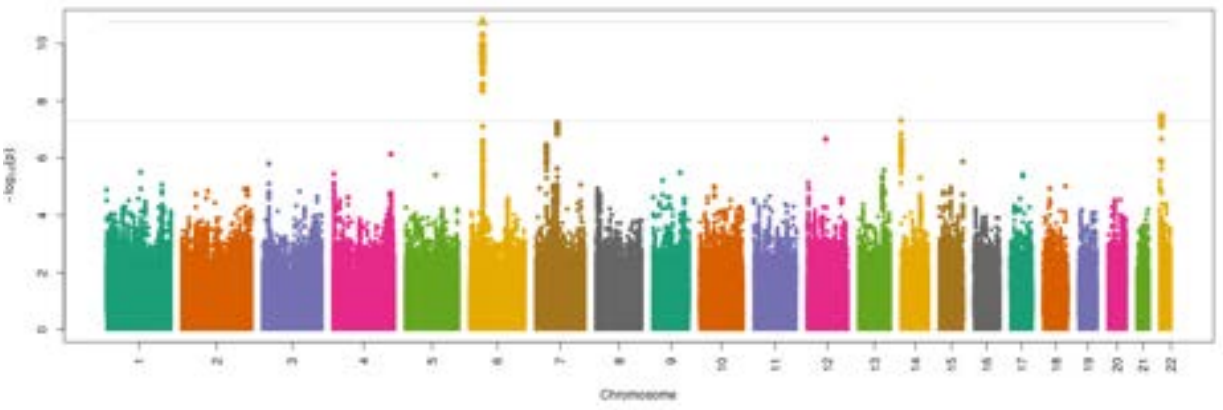


Figure 5.2. Manhattan plot of *P*-values generated using logistic regression analysis in the European ancestry sample (n=15,458). An additive model was used to assess the disease susceptibility impact of the minor (coded) allele at each position, while controlling for age, BMI, sex, ancestry, nursing home status, chemotherapy, diabetes, HIV, transplant medications, corticosteroids, and medium or high-risk antibiotic exposure as covariates. Genomic coordinates are displayed along the X-axis, and the negative logarithm of logistic regression *P*-values are displayed on the Y-axis. Each dot represents a SNV in the regression model, with associated *P*-values plotted accordingly, while the triangle represents the most significantly associated SNV. The dotted line represents the negative logarithm of the genome-wide significance threshold ($P < 5 \times 10^{-8}$). Colors are used to distinguish between SNVs in adjacent chromosomes.

Given the well-known presence of high LD within the HLA region [210], a regional LD plot with reference to the index SNV (rs68148149) was generated using *P*-values from the European logistic regression analysis and using the 2014 1000 Genomes European superpopulation as a reference group (**Figure 5.3**). This step was taken to assess the possibility that variants other than the index SNV might better explain disease association in terms of functional impact. While the second two most significant SNVs were in high LD with the index SNV ($R^2 > 0.8$), the index SNV had the highest regulatory potential among the most significantly associated SNVs, as annotated by RegulomeDB [211]. To assess the possibility that the lack of disease association in

the African ancestry sample is a result of different regional LD structures, a regional LD plot with reference to the index SNV was generated using the 1000 Genomes African superpopulation as a reference (**Figure S5.6**). The second two most significant SNVs in the European-ancestry sample were also in high LD with the index SNV in the African-ancestry superpopulation, but higher LD was observed with more SNVs in the HLA-DRB1/5 intergenic region in the African superpopulation ($R^2 > 0.4$) than in the European superpopulation ($R^2 > 0.2$). On the other hand, lower LD was observed with SNVs in the region spanning HLA-DRB1 and HLA-DQA1 in the African superpopulation ($R^2 > 0.6$) than in the European superpopulation ($R^2 > 0.8$). Differences in regional LD patterns between the European-ancestry and African-ancestry samples could therefore have contributed to the observed differences in gene-disease association patterns, in addition to insufficient detection power.

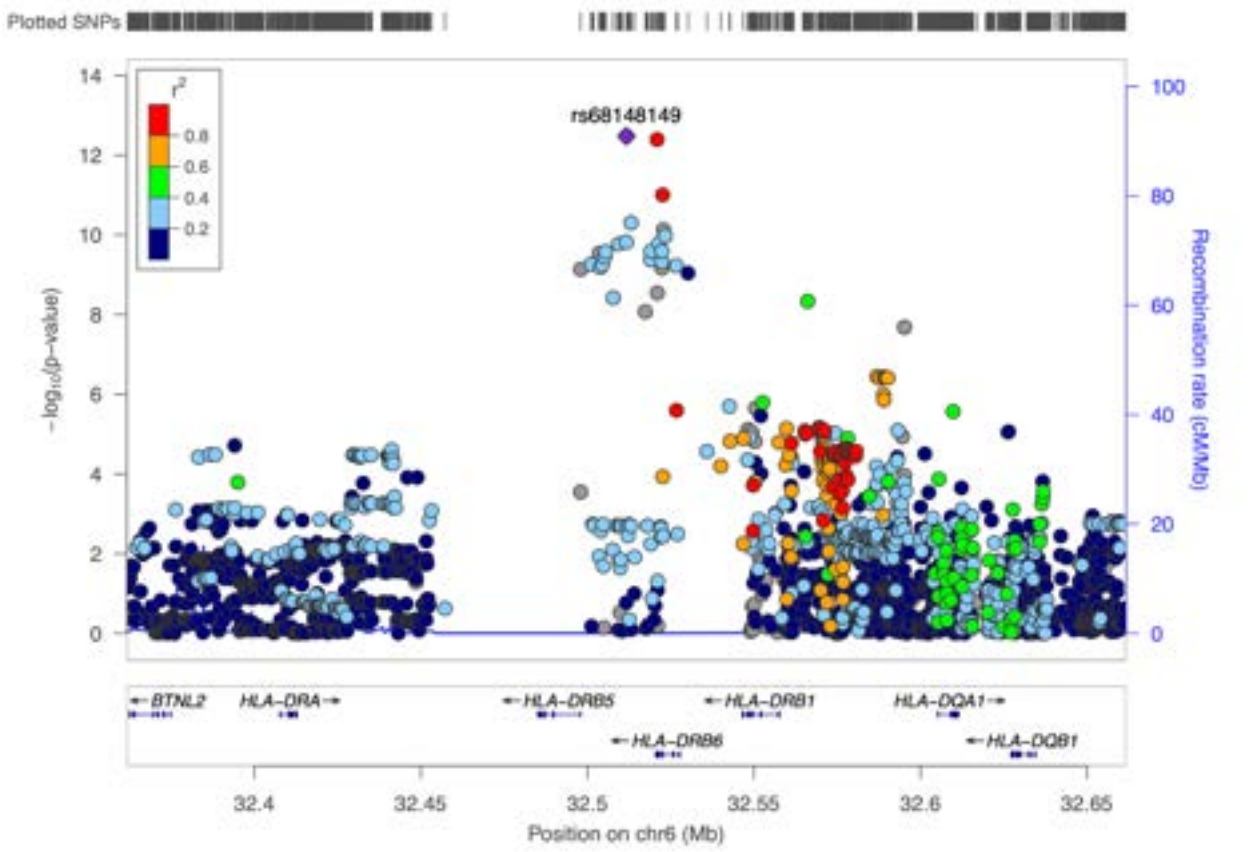


Figure 5.3. Regional LD plot of SNVs evaluated in the European-ancestry logistic regression analysis, using the European 1000 Genomes superpopulation as a reference group. Genomic coordinates spanning the HLA-DRB region and surrounding genes are shown on the X-axis in both subplots. Negative logarithms of *P*-values from the European-ancestry logistic regression analysis are shown on the Y-axis in the upper sublot, and annotated gene transcripts are distributed along the Y-axis in the lower subplot. Each dot represents a SNV in the regression model, with associated *P*-values plotted accordingly. SNVs in high LD with reference to the index SNV (rs68148149) are colored in red. The LD plot was generated with the LocusZoom [207] tool using default parameters and the 1000 Genomes Project 2014 EUR reference panel.

A follow-up GWAS using the index SNV as a covariate revealed several new SNVs associated at genome-wide significance (rs116603449, P=4.54 x 10^{-9}; rs9270896, P=6.09 x 10^{-9};

rs9270894, P=1.12 x 10^{-8}; rs9270895, P=2.32 x 10^{-8}; rs618095, P=3.71 x 10^{-8}) (**Table 5.2; Figures S5.7-S5.8**). While suggestive peaks were observed in chromosomes 14 and 22 using the unadjusted model, the elimination of these peaks in models that included the genome-wide significant index SNVs suggests that they were spuriously associated with the tagged region in chromosome 6. However, no SNVs of interest on chromosomes 14 or 22 were in high LD with any the index SNVs on chromosome 6, therefore the nature of the associated remains unknown.

5.4.3 Human leukocyte antigen association analyses

All 14,620 European ancestry participants had high quality imputed HLA genotypes available for association analyses. **Table 5.1** summarizes the number of participants in the European ancestry group possessing at least one *HLA-DRB3*, *4* and/or *5* gene (corresponding to haplotype families (HLA-)DR52, 53 and 51, respectively) [212] (**Figure S5.9**). The most significant SNVs from the GWAS reached genome-wide significance among individuals with at least one *DRB3*, *4* or *5* genes collectively (rs68148149, P=1.26 x 10^{-13}; rs3828840, P=1.49 x 10^{-13}; rs35882239, P=2.37 x 10^{-11}; rs71534541, P=1.67 x 10^{-11}; rs35222480, P=3.17 x 10^{-11}), and among individuals with at least one *DRB5* gene only, or DR51 haplotype carriers (rs68148149, P=1.55 x 10^{-11}; rs3828840, P=1.72 x 10^{-11}; rs35882239, P=2.62 x 10^{-10}; rs71534541, P=1.56 x 10^{-11}; rs35222480, P=4.68 x 10^{-11}) (**Table 5.3, Figure S5.10**). Among DR51 haplotype carriers, the most significantly associated SNVs only reach genome-wide significance among carriers of the DR15 haplotype (rs68148149, P=2.08 x 10^{-11}; rs3828840, P=2.27 x 10^{-11}; rs35882239, P=4.14 x 10^{-10}; rs71534541, P=1.75 x 10^{-12}; rs35222480, P=5.81 x 10^{-12}), and more specifically, carriers

of the HLA-DRB1*15:01 allele (rs68148149, P=7.45 x 10^{-11}; rs3828840, P=8.11 x 10^{-11}; rs35882239, P=1.42 x 10^{-9}; rs71534541, P=7.37 x 10^{-12}; rs35222480, P=1.43 x 10^{-11}). No SNVs reached genome-wide significance among participants with at least one *DRB3* or *DRB4* gene only, suggesting that the HLA-DR51 haplotype in combination with variants in the HLA-DRB1/5 intergenic region may singularly drive genetic risk for CDI in the European ancestry population. However, examining the risk allele frequencies of the index SNV (rs68148149) in cases and controls across DR51, DR52, and DR53 haplotype-enriched groups showed that the risk allele frequency was higher in European-ancestry cases than controls in all haplotype groups, suggesting that the SNV may indeed drive risk in all HLA-DR haplotype groups but that the low frequency in the DR52 and DR53 haplotype groups limits the power to detect the association in these groups (**Figure S5.11**). The same pattern was not observed in African-ancestry cases and controls, indicating that haplotype differences between ancestry groups may indeed play a role in differentially conferring risk.

Chr	SNV	Ref	Alt	CA	BP	DR01(+), DR02(+), or DR10(+)		DR01(+)		DR02(+)		DR05(+)		DR10(+)		DR01*15-01(+)	
						CAF (n=14291)	Logistic P-value OR (95% CI) Beta	CAF (n=4130)	Logistic P-value OR (95% CI) Beta	CAF (n=8887)	Logistic P-value OR (95% CI) Beta	CAF (n=7962)	Logistic P-value OR (95% CI) Beta	CAF (n=3791)	Logistic P-value OR (95% CI) Beta	CAF (n=3606)	Logistic P-value OR (95% CI) Beta
6	rs6914849	C	A	C	32111725	0.13	**1.86 x 10-10** **1.59 (1.34-2.23)** **0.2**	0.32	**1.35 x 10-11** **1.97 (1.14-3.28)** **0.99**	0.08	1.30 x 10-4 1.33 (1.00-2.34) 0.18	0.08	1.11 x 10-7 1.79 (1.07-2.93) 0.25	0.32	**3.08 x 10-11** **2.01 (1.18-3.42)** **0.8**	0.32	**7.54 x 10-11** **2.00 (1.17-3.41)** **0.8**
6	rs3829640	T	C	T	32130607	0.13	**1.49 x 10-12** **1.59 (1.34-2.22)** **0.2**	0.32	**1.72 x 10-11** **1.97 (1.14-3.28)** **0.99**	0.08	1.70 x 10-4 1.32 (1.00-2.33) 0.18	0.08	1.21 x 10-7 1.76 (1.07-2.93) 0.25	0.32	**3.27 x 10-11** **2.01 (1.18-3.41)** **0.8**	0.32	**8.11 x 10-11** **2.00 (1.17-3.40)** **0.8**
6	rs3829642	A	G	A	32122576	0.13	**2.37 x 10-11** **1.59 (1.33-2.04)** **0.18**	0.27	**2.62 x 10-10** **1.95 (1.10-3.29)** **0.99**	0.1	1.20 x 10-3 1.42 (0.97-2.07) 0.35	0.1	4.02 x 10-5 1.86 (1.01-2.60) 0.59	0.26	4.14 x 10-10 **1.99 (1.16-3.40)** **0.8**	0.27	**1.42 x 10-9** **1.97 (1.15-3.38)** **0.8**
6	rs7154541	C	T	C	32113676	0.06	**1.67 x 10-11** **1.84 (1.15-2.88)** **0.21**	0.26	**1.55 x 10-11** **2.060 (1.19-3.56)** **0.82**	0.04	7.80 x 10-3 1.45 (0.95-2.27) 0.16	0.05	9.84 x 10-6 1.82 (1.04-3.16) 0.26	0.24	**1.75 x 10-12** **2.95 (1.88-5.01)** **0.85**	0.24	**7.37 x 10-12** **2.93 (1.83-4.65)** **0.85**
6	rs3222480	A	T	A	32122613	0.06	**5.17 x 10-11** **1.81 (1.12-2.31)** **0.21**	0.27	**4.58 x 10-11** **2.01 (1.18-3.44)** **0.8**	0.05	1.20 x 10-3 1.41 (0.92-2.16) 0.15	0.05	8.33 x 10-6 1.72 (1.04-2.87) 0.24	0.23	**5.81 x 10-12** **2.18 (1.22-3.91)** **0.84**	0.23	**1.43 x 10-11** **2.18 (1.21-3.93)** **0.84**

Table 5.3. Index SNV results from logistic regression-based analysis of the HLA region in European samples enriched for each HLA-DRB haplotype or haplotype family: DR51, DR52, DR53, DR15, DRB1*15:01, and any of the above. An additive model was used to assess the disease susceptibility impact of the minor (coded) allele at each position in the genomic region that yielded highly associated SNVs in the genome-wide analysis (chr6:32400001-32600000). Age, BMI, sex, ancestry, nursing home status, chemotherapy, diabetes, HIV, transplant medications, corticosteroids, and medium or high-risk antibiotic exposure were included as covariates in the model. Results meeting the genome-wide significance threshold ($P < 5 \times 10^{-8}$) are displayed in bold. Abbreviations: Chr = Chromosome; SNV = Single Nucleotide Variant; Ref = Reference Allele; Alt = Alternate Allele; CA = Coded Allele; BP = Base Pair; CAF = Coded Allele Frequency; OR = Odds Ratio.

To assess the possibility that one or more HLA alleles themselves were driving the risk association in the European ancestry sample, rather than the most significantly associated SNVs identified in the GWAS, we performed a separate logistic regression analysis using the HIBAG-imputed HLA genotypes in the European ancestry sample. None of the imputed HLA alleles reached genome-wide significance. Using the classical HLA tags identified by de Bakker et al. (2006) [213] and the NCI LDMatrix tool [214], it was also confirmed that none of the GWAS-identified SNVs were in high LD ($R^2 > 0.5$) with any classical HLA alleles in either the European ancestry or African ancestry 1000 Genomes superpopulations. The index SNV was in moderate LD with the tag SNV for the *DRB1*15:01-DRB5*01:01* haplotype in the European ancestry superpopulation (rs3135388; R^2=0.186) and low LD with the tag SNV in the African ancestry superpopulation (rs443623; R^2=0.002).

5.5 Discussion

Using a robust EHR-based phenotyping algorithm, we identified a large, multi-institutional corpus of patients with a history of at least one episode of CDI and controls without CDI. Our results suggest that genetic variation in the *(HLA-)DRB* locus of the HLA region may increase risk of infection in European ancestry populations. In this study, European participants who possessed the minor allele among the most significantly associated SNVs had 56% greater odds of having at least one episode of CDI. As the key beta-subunits of MHC Class II surface receptors on antigen presenting cells (APCs), the proteins encoded by *DRB* genes play a critical role in stimulating the host adaptive immune response against foreign peptides and are therefore excellent candidates for future studies of host immunity to *C. diff.* [215].

The MHC (HLA) Class I and II loci are among the most polymorphic coding regions in the human genome, and *DRB* genes are particularly variable in copy and combination. Although there is only one monomorphic *DRA* gene per (HLA-)DR haplotype, there are five common DR haplotype families composed of different combinations of protein coding *DRB* genes (*DRB1, DRB3, DRB4* and *DRB5*) and pseudogenes (*DRB2, DRB6, DRB8* and *DRB9*) [212]. *DRB1* is present in all haplotypes, but any given individual may have as few as two protein coding *DRB* genes (2 copies of *DRB1*), or as many as four genes (2 copies of *DRB1* + 1 or 2 copies of *DRB3*, *4* or *5*) between homologs. The unique combination of *DRB* genes on each haplotype is remarkably conserved and has been maintained in ancestral DNA since before the divergence of human and gorilla lineages over 5 million years ago [216]. Although having a diverse set of MHC II molecules may confer a selective advantage against infection [217], each additional

DRB gene is nonetheless susceptible to intragenic and/or regulatory mutations in the highly polymorphic HLA region and may paradoxically increase susceptibility to other diseases. In the case of gastrointestinal infections, protective effects of the *HLA-DRB1*04:05* allele against enteric infection caused by *Salmonella typhi* or *Salmonella paratyphi* have been observed in Vietnamese and Nepalese patients [218]. Conversely, the *DRB1* gene has also been implicated in increasing host susceptibility to a number of inflammatory diseases, including Crohn's disease, type I diabetes mellitus, rheumatoid arthritis, multiple sclerosis (MS), ulcerative colitis and Alzheimer's disease, primarily in European populations [219–224].

Haplotype effects appear to play a critical role in conferring risk for CDI. In this study, CDI susceptibility was highly correlated with the most significantly associated SNVs only among individuals possessing at least one haplotype in the rarer DR51 haplotype family, in which a *DRB1*15* or **16* allele is paired with a coding *DRB5* gene (**Table 5.3, Figure S5.9**). The risk association was exclusively observed in individual carrying at least one copy of the *DRB1*15:01-DRB5*01:01* haplotype [225], and individuals in this group had 200% higher odds of developing CDI on average. These results indicate that the *DRB1*15:01-DRB5*01:01* haplotype is involved in conferring CDI risk among individuals with common genetic variants in the tagged *DRB1-DRB5* intergenic region (**Figure S5.12**). However, it is also possible that the comparatively low risk allele frequency in the DR52 and DR53 haplotype groups limited the power to detect a true risk association in other HLA-DR haplotype groups (**Figure S5.11**). One possible explanation for increased CDI risk among these individuals is that differential MHC II gene expression impacts the baseline composition of their gut microbiota, thereby influencing colonization resistance to opportunistic enteric pathogens like *C. diff.* Secretory

Immunoglobulin A (IgA) antibodies play an essential role in shaping an individual's gut microbial community and maintaining a homeostatic balance of microbes within the mucosal immune system [226], and the interactions between APCs and CD4+ T-follicular helper (Tfh) cells are key to driving the production of IgA by plasma cells [227]. Studies in mouse models have previously demonstrated that MHC II polymorphisms directly affect antibody-mediated microbiota composition, and that the unique microbial communities formed under the influence of different MHC genotypes can impact an organism's susceptibility to opportunistic pathogens like *Salmonella enterica typhimurium* when treated with antibiotics [228,229]. The *DRB1*15:01-DRB5*01:01* haplotype has also been identified as the major genetic risk factor for MS--a disease that has been increasingly associated with taxa imbalances within the gut microbiome [225,230–232]. This association lends support to the hypothesis that the gut microbiota mediates susceptibility to CDI in a genetically determined manner, assuming that the composition of the microbiota is indeed a key driver of resistance to disease in both CDI and MS. However, it is currently unknown exactly which symbiotic microbe lineages or consortia might contribute to colonization resistance against *C. diff.* after a major disruption to the gut microbiota [233]. Understanding the unique interactions between commensal microbe antigens presented by APCs, the MHC II molecules encoded by the *DRB1*15:01-DRB5*01:01* haplotype, and Tfh cells may provide valuable insights into how host genetics impact the composition of gut microbial communities in individuals susceptible to enteric infection, compared with those who are resistant to infection.

Alternatively, increased CDI risk among these individuals may be driven by differential T-cell mediated responses to the TcdA and TcdB toxins produced by *C. diff.* bacteria. In addition to

sculpting the host microbiota, high affinity IgA helps to neutralize bacterial toxins [234]. Unique interactions between T-cells and *C. diff.* toxins specifically bound by *DRB1*15:01-DRB5*01:01* MHC II molecules may impact the host anti-toxin IgA response differently than other T-cell-MHC II interactions, thus influencing the host's ability to clear circulating toxins. Recent Phase III, placebo-controlled clinical trials of the monoclonal antibody treatments actoxumab (anti-TcdA) and bezlotoxumab (anti-TcdB) showed that TcdB toxin neutralization alone could decrease CDI recurrence by 38% among patients receiving standard antibiotic therapy for initial or recurrent CDI [235]. Naturally occurring anti-TcdB antibodies in the placebo group also conferred protection against recurrent CDI, recapitulating the importance of neutralizing TcdB in controlling infection [236]. However, other studies have failed to replicate these results when comparing healthy controls with CDI patients, suggesting that anti-toxin antibody concentrations may not fully explain susceptibility to initial and/or recurrent infection [237].

Although the MHC II region is strongly associated with CDI in this study, the SNVs that confer risk are neither located in coding regions, nor in high LD with SNVs in coding regions, suggesting that the mechanism for altered gene expression may be regulatory. One possible mechanism for altered expression of the *DRB1*15:01-DRB5*01:01* haplotype is allele-specific DNA methylation (ASM) of the *DRB1* and/or *DRB5* regulatory regions, given that the two most significantly associated SNVs (rs68148149 and rs3828840) overlap with CpG dinucleotides and may therefore be involved in altering DNA methylation patterns in those regions. It is well known that cytosine residues at CpG sites are disproportionately targeted for DNA methylation, which directly impacts gene expression at the level of transcription [238,239]. SNVs that overlap with CpG sites account for 38%-88% of ASM regions [240], and disruptions to normal DNA

methylation patterns have been known to modulate susceptibility to a number of human diseases [241,242]. For example, in the case of *DRB1*15:01-DRB5*01:01*-associated MS, DNA hypermethylation in exon 2 of *HLA-DRB1* confers protection against the major risk allele and is driven by several SNVs in high LD with one another that overlap with CpG sites [243]. It is possible that disrupted methylation patterns at or near the regulatory regions of *DRB1*15:01* and/or *DRB5*01:01* also contribute to differential expression of these MHC II proteins, thus impacting the landscape of the host adaptive immune response via microbiome-mediated and/or toxin-mediated mechanisms. To test this hypothesis, local bisulfite sequencing or methylation quantitative trait loci (mQTL) analysis of the HLA region could be performed in *DRB1*15:01-DRB5*01:01* heterozygotes to assess differential methylation patterns in the *DRB1-DRB5* intergenic region. These experimental data could then be superimposed on GWAS data to determine whether the GWAS peaks identified in this study are suggestive of true regulatory SNVs, and to subsequently prioritize these SNVs for downstream validation experiments in animal models [244]. It is also worth noting that the additional SNVs identified using the top SNV-corrected model were all located in the *DRB1-DQA1* intergenic region near several histone H3K27ac marks, which are often located near active regulatory elements [245]. This observation lends additional support to the hypothesis that MHC molecules involved in CDI pathogenesis are transcriptionally regulated.

Our findings suggest that genetic variation in the MHC II locus of the HLA region drives susceptibility to CDI and highlights the importance of the adaptive immune response in combating opportunistic pathogens. To better understand how host genetics might confer microbiome-mediated risk for opportunistic enteric infections, future studies should explore the

mechanisms of interaction between commensal microbe antigens presented by APCs and the MHC II molecules encoded by the *DRB1*15:01-DRB5*01:01* haplotype. Interactions between *DRB1*15:01-DRB5*01:01* MHC II, *C. diff.* exotoxins and T-cells may alternately play a critical role in CDI pathogenesis, and additional work is needed to understand whether and how the host IgA response is differentially impacted by the combined effects of haplotype and transcriptional modifications. Finally, future work should address the possibility that ASM is a driver of epigenetic transcriptional regulation of the *DRB1* and/or *DRB5* genes. If this mechanism is experimentally validated, therapeutics that modulate MHC II molecule transcription levels could potentially be developed to decrease the incidence of CDI among individuals that carry the risk genotype.

5.6 Limitations and future work

This study has several important limitations. First, sample size and statistical power were severely limited among non-European ancestry samples, which may have contributed to the lack of significant associations in the African ancestry analyses. Second, replicate studies are needed to confirm the identified association. However, the large, multi-site biobank of linked EHR and genotype data used in this study supports the replicability and reliability of these results, and future association studies would benefit immensely from these types of biobanks. Third, *C. diff.* cases were not stratified by primary and recurrent CDI and is it possible that the genetic variants driving pathogenesis are different between these two forms of infection. Fourth, the length and severity of infection were not considered in the current study, but future analyses would benefit from continuous trait regression association analyses to identify genetic variants associated with increased CDI severity, rather than susceptibility. Additionally, *C. diff.* cases in this study

partially included individuals with a positive antigen test as their only criterion for infection. The *C. diff.* antigen test cannot accurately distinguish between toxigenic and non-toxigenic strains and may falsely identify asymptomatic carriers as *C. diff.* cases. Finally, the specific toxigenic ribotype that each case was exposed to was not included in the analysis, and it is possible that different *C. diff.* ribotypes are associated with different genetically determined host responses.

5.7 Conclusion

In this study, we identified a potential genetic driver for CDI in the HLA-DRB locus, offered several directions for future functional studies, and demonstrated the utility of merging genetic and EHR data for gene-disease association studies of infectious disease. Routinely conducting genetics association studies using EHR data is a promising avenue for advancing our understanding of how common genetic variation impacts human health and disease. In the next chapter, we will explore how the existing LHS literature characterizes the barriers and enablers of conducting routine genomic discovery studies in clinical settings and integrate these results with the 5R GLHS model developed in the previous chapter.

CHAPTER 6: DEVELOPMENT OF AN INTEGRATIVE SOCIOTECHNICAL MODEL FOR GENOMICS-ENABLED LEARNING HEALTH SYSTEM DISCOVERY (AIM 4)

6.1 Introduction

The types of genomic research studies that can be conducted using merged clinical and genomic data hold great promise for the future of genomic medicine and human health, but routine implementation of genomics research in clinical environments remains elusive. While the current literature on genomic discovery in LHS-aligned models appears limited, its sheer complexity speaks to the practical challenges of achieving the learning healthcare vision. In the closing remarks of their 2020 progress update, investigators at Geisinger–one of the world's most fully-formed LHS-aligned healthcare institutions to date–recognized the eternal challenge of implementing a successful LHS:

> We will close by reflecting on our position and our prospects as we seek to move along "the developmental path toward a fully realized LHS." Although we do, indeed, hope and intend to move further along that path, **we have come to question whether the goal of a fully realized LHS is ever fully attainable.** For we suspect that the reality is that in light of the ongoing dynamic evolution of technologies, the growth of evidence, and other forces of change, **the goal of a fully realized LHS, much like the paradox of Achilles and the tortoise, can never fully be achieved because the essence of learning and improvement is—and always will be—a moving target** (Davis et al. 2020, p. 9) [132].

Although there may be no discernable endpoint in the LHS model, the iterative process of learning and improvement can only benefit a healthcare system that is in great need of change. To facilitate a broader understanding of the GLHS concept and move towards actual implementations of learning cycles, it is useful to ground available evidence in a conceptual model. In this case, the vast majority of insights on clinically embedded genomic discovery exist

in qualitative studies. Qualitative evidence synthesis (QES) is a collection of methods that can be used to integrate findings from qualitative studies to "establish a greater understanding of issues, often of a subtle or sensitive nature, that primary qualitative research frequently addresses" (Flemming et al. 2018, p. 1) [39]. The multiplicity of technical, social, ethical, political, and structural elements that support a GLHS may individually be moving targets in the context of constant shifts in the US healthcare and research enterprises, but this movement should not preclude researchers, clinicians, and policymakers from seeking a more cohesive understanding of how these elements interact with one another. QES methods can help achieve this cohesive vision and ground the complexity of the GLHS concept in a conceptual model to inspire tangible changes in approaches to healthcare research and delivery.

The objectives of this aim are twofold. First, we conduct a systematic literature review of studies that have identified enabling factors of genomic discovery and validation research in the LHS model and describe this literature landscape using a theory of change model. Second, we use best-fit framework synthesis (BFFS) to synthesize the *a priori* 5R model from Aim 2 with themes identified in Aim 1 and the systematic literature review to create an integrative sociotechnical model for GLHS discovery.

6.2 Related Work

6.2.1 Systematic reviews of enabling factors for clinical genomic discovery research in learning health systems

Previous systematic literature reviews of LHS models have focused on the outcomes of such models and their impacts on different aspects of care. Enticott et al. (2021) conducted a systematic review of studies across 23 LHS environments in six different countries [11]. They investigated the reported health impacts achieved through LHS-aligned healthcare models and found that such systems yielded benefits such as improved longitudinal patient tracking, enhanced access to personal health records, and improved adherence to clinical guidelines. Other in-progress reviews are investigating the impacts of LHS models on pediatric health outcomes [40], and investigating strategies used to implement LHS models in existing healthcare systems [42]. Lim et al. (2022) [41] conducted a systematic review of data analytics approaches in LHS-aligned models and found that challenges were widely faced when implementing EHR data analytics in an LHS. In the most recent review of LHS literature, Ellis et al. (2022) [43] used the PubMed and Scopus databases to survey the available LHS research through an implementation science lens. They found that, unsurprisingly, there is little empirical research on LHS implementations and outcomes since very few LHS-aligned systems exist worldwide. Systematically investigating the enablers and barriers of clinically embedded discovery is an important precursor to implementation and outcomes measurement. However, no systematic reviews have investigated barriers and drivers for accelerating genomic discovery in LHS models.

6.2.2 Qualitative evidence synthesis and framework development in learning health systems research

Several studies have previously used QES (or similar) methods to seek clarity from the complex body of LHS literature. Enticott et al. (2021) [45] developed an LHS framework for the Australian health system by synthesizing evidence from expert panels, stakeholder workshops, and a systematic literature review of studies showing explicit health impacts from LHS-aligned implementations. While the authors report utility in integrating multiple perspectives for developing a sustainable and scalable framework, they do not describe how the evidence synthesis was conducted. Easterling et al. (2022) [44] recognizes that the "LHS concept has been defined in broad terms, which makes it challenging for health system leaders to determine exactly what is required to transform their organization into an LHS" (Easterling et al. 2022, p. 1). To address this gap in proposed requirements, they developed a 94-part taxonomy of LHS elements, then calculated the frequency of each element in 79 publications that discussed organizational characteristics or actual implementations of LHSs. This process aligns with the description of framework synthesis as described in Flemming & Noyes (2021) [246]. Their approach successfully integrated salient results from a large, complex body of research, and clearly revealed "specific types of work that need to be launched and supported in order to operate according to the principles of an LHS" (Easterling et al. 2022, p. 12). Although few studies have addressed the LHS concept using QES methods, the two aforementioned studies demonstrate that framework synthesis is a useful approach for clarifying and integrating complex concepts for the sake of inspiring action within healthcare organizations.

6.3 Methods

6.3.1 Systematic literature review

The systematic literature review plan was developed using the Cochrane Reviews of Interventions guidelines [247]. However, because the goal of the review was not to assess the outcomes of comparative-effectiveness research, but rather to integrate published perspectives and experiences, some suggested procedures such as statistical meta-analysis and systematic bias assessment were not conducted. Given the known paucity of literature on actual implementations of LHSs [11], and even more limited evidence on LHS factors that enable genomic discovery efforts specifically, the scope of this review was left intentionally broad and qualitative to incorporate as much evidence as possible in the analysis.

6.3.1.a Scope

The purpose of the review was to survey the proposed and observed enabling factors for accelerating translational genomic discovery research in LHSs. Specifically, the review centered around the following question: **What technical, social, political, and/or cultural factors enable and improve genomic innovation, discovery, or validation research in an LHS-aligned model?** The Population, Intervention, Comparison, and Outcome (PICO) [248] strategy of the review was defined as follows:

1. **Population:** Any population receiving healthcare in a country where LHSs have been proposed as a model for improving care, advancing research, and decreasing healthcare costs.

2. **Intervention:** Technical, social, political, and/or cultural changes to healthcare systems or research operations within healthcare systems that are intended to enhance clinically meaningful genomic discovery.
3. **Comparison:** Enabling factors for genomic discovery research that occurs outside the context of a healthcare providing organization.
4. **Outcome:** Expected (or observed) improvements and/or accelerations of clinically meaningful genomic discovery research.

6.3.1.b Ethical considerations

Because this field of research is in its infancy and there is little empirical evidence to support the perspective pieces that comprise the majority of this body of research, it is important to frame the results of this review as a survey of *potential* next steps for integrating genomics research into the LHS model, rather than as fully supported evidence of effective interventions. These potential next steps should be systematically applied and tested in combination with one another in different healthcare and research contexts to empirically identify enablers of clinical genomic discovery in LHSs. There is also a risk of disseminating suggestions that are not feasible to implement in other countries or in communities in the US that lack adequate financial or political support, potentially worsening health disparities. It is therefore important to consider questions of health equity as a core category of the analysis, and to prioritize the inclusion of papers with a focus on health equity in LHS discovery research.

6.3.1.c Search strategy and data collection

Databases used in the literature search included PubMed, Embase (via Elsevier), the Public Affairs Information Service (PAIS, via ProQuest), the Health Technology Assessment Database, the Cumulative Index to Nursing and Allied Health Literature (CINAHL), Web of Science, PsycINFO, and Medline. Initial searches were limited to peer-reviewed, English language articles that had been published since 2008–the year following the publication of Etheredge's rapid-learning health system concept [127]. Given the scope of the review, articles were also required to include mentions of the LHS concept and of genetics or genomics, because it is known that requirements for conducting research with genomic data in a healthcare setting are similar to but distinct from requirements for conducting research with other types of health-related data [8,31]. **Table 6.1** displays the search queries used to identify relevant literature in each of the eight surveyed databases.

Database	Search Query
PubMed	(("learning health system") OR ("learning healthcare system") OR ("learning health care system")) AND ((genomic) OR (genome) OR (genetic) OR (gene) OR (genes)) AND ("2008"[Date - Publication] : "3000"[Date - Publication]) AND (English[Language])
Embase	('learning health system' OR 'learning healthcare system' OR 'learning health care system') AND (genomic OR genetic OR gene OR genes OR genome) AND [english]/lim AND [2008-2022]/py
PAIS	(("learning health system") OR ("learning healthcare system") OR ("learning health care system")) AND ((genomic) OR (genetic) OR (gene) OR (genes) OR (genome))
Health Technology Assessment Database	("learning health system") OR ("learning healthcare system") OR ("learning health care system")
CINAHL	(("learning health system") OR ("learning healthcare system") OR ("learning health care system")) AND ((genomic) OR (genetic) OR (gene) OR (genes) OR (genome))
Web of Science	(("learning health system") OR ("learning healthcare system") OR ("learning health care system")) AND ((genomic) OR (genetic) OR (gene) OR (genes) OR (genome))
PsycINFO	(("learning health system") OR ("learning healthcare system") OR ("learning health care system")) AND ((genomic) OR (genetic) OR (gene) OR (genes) OR (genome))
Medline	(("learning health system") OR ("learning healthcare system") OR ("learning health care system")) AND ((genomic) OR (genetic) OR (gene) OR (genes) OR (genome))

Table 6.1. Search queries used to identify eligible articles in each database. For searches where publication date and/or language could not be included in the search wuery (all databases other than PubMed and Embase), results were manually filtered by English language and publication date after the initial search.

Following each search, reference lists were exported as Research Information Systems (.RIS) files, which were then imported into the EPPI-Reviewer Web 4.0 literature review management

system [249]. Duplicate articles were removed using the EPPI-Reviewer duplicate checking tool. All remaining screening and data extraction activities were conducted by one reviewer (K.F.) using EPPI-Reviewer.

6.3.1.d Inclusion and exclusion criteria

Publications were first screened by title and abstract content to exclude publications that clearly met the exclusion criteria. Titles and abstracts were **excluded** if they met one or more of the following conditions:

1. No mention of the LHS model
2. No mention of genetic or genomic data
3. Conference abstract
4. Table of contents
5. Dissertation/thesis
6. Protocol article
7. Review article
8. No peer review
9. Not in English
10. Published before 2008

Publications that passed the title and abstract screening phase were then screened on the full text. Articles were **excluded** if they met one or more of the following conditions:

1. Minimal to no discussion of discovery research in the context of the LHS model
2. No suggested needs, actions, or opportunities for discovery research identified

3. Minimal to no discussion of genetic or genomic data

Publications that passed both the title and abstract screening and the full text screening were included for data extraction.

6.3.1.e Data extraction

Data extraction was conducted in two phases: 1. Background extraction; and 2. Content extraction. During the background extraction phase, the following information was gathered:

1. Article type (e.g., Special Report, Commentary, Methodology)
2. Study design
3. Country
4. Home institution name
5. Institution type
6. Medical domain
7. Source(s) of funding
8. Conflicts of interest

During the content extraction phase, enabling factors explicitly identified by authors throughout the text (with phrases such as, “this would require…” or “crucial to this approach is…”) or summarized in lists or tables were identified. Codes representing these factors were iteratively created and sorted into topical categories, such as “Funding and incentives” or “Policy and governance.” While codes were re-used between publications when possible, new codes with similar sentiments to existing codes were written when there were nuances in the publication that

the existing code did not capture. This process continued until all publications had been evaluated for content. Separate reports were generated for Study Characteristics/Background and Outcomes using the EPPI-Reviewer configurable reports tool.

6.3.1.f Data synthesis

Because the studies included in the review involved a variety of methods and did not report comparable quantitative findings, narrative synthesis was used to combine the results into a textual narrative. The analysis was conducted using the process for narrative synthesis proposed by Popay et al. (2006) [250]: 1. Preliminary synthesis; 2. Theory of change development; 3. Relationship exploration; and 4. Assessment of robustness.

1. **Preliminary synthesis:** Initial data synthesis was conducted during the data extraction process, during which emerging codes were grouped into descriptive themes. Once data extraction was complete, codes were re-grouped based on their similarities and differences with respect to their relationships with the outcome of interest (clinical research integration), and descriptive themes were rephrased to better describe codes assigned to each theme.
2. **Theory of change development:** Flow diagramming was used to develop an initial theory of change using the descriptive themes identified during preliminary synthesis, as described in Weiss 1998 [251] and White 2017 [252]. Weiss describes the theory of change as “the chain of causal assumption that [links] programme resources, activities, intermediate outcomes and ultimate goals” that is used to better understand “how the

intervention works, why, and for whom" (Weiss 1998, as cited in Popay et al. 2006, p. 12). The initial theory was modified as needed during the relationship exploration.

3. **Relationship exploration:** Concept mapping was used to relate descriptive themes and the properties of those themes (e.g., codes) with one another. Relationships identified through concept mapping were incorporated into an updated theory of change model, which was then used to write the final narrative synthesis.
4. **Assessment of robustness:** Methodological limitations of each study were identified and described in aggregate. The most common sources of bias and assumption were also identified, both for the publications included in the review and for the investigator conducting the review. Discrepancies and uncertainties between study results were also considered, in addition to contextual factors of each study that may have influenced outcomes.

6.3.2 Qualitative evidence synthesis

BFFS is a QES method that has previously been used to address "applied policy or clinical questions in a specific setting or context" (Flemming & Noyes 2021, p. 6) [246], and QES methods have been well established as effective aids in health policy and healthcare decision-making [253,254]. BFFS was first described by Carroll, Booth, & Cooper in 2011 [255], and an updated method was described by Carroll et al. (2013) [256]. The following measures were taken based on proposed steps of BFSS: 1. Framework identification; 2. Systematic literature review; 3. Evidence comparison; and 4. Evidence synthesis.

1. **Framework identification:** An *a priori* model for GLHSs was developed using the grounded theory approach, as described in Aim 2.

2. **Systematic literature review:** Primary research studies were chosen for inclusion in the evidence synthesis using the methods described in section 6.3.1 ("Systematic Literature Review").
3. **Evidence comparison:** Individual codes generated from the systematic literature review were imported into ATLAS.ti, along with thematic codes generated from the Recommendations section in Aim 1. All codes were compared with the primary themes of the *a priori* model using the constant comparison method [257]. Codes that were sufficiently related to the *a priori* model themes were grouped accordingly, and codes that did not sufficiently relate to existing themes were grouped into new thematic categories.
4. **Evidence synthesis:** Concept mapping was used to identify relationships between existing themes in the *a priori* model and new themes identified during evidence comparison. New properties of the *a priori* themes were also identified using the code groupings from the evidence comparison. A final diagram of the synthesized model was created, along with a narrative description of the model.

6.4 Results

6.4.1 Systematic review

Of a total of 291 records identified through database searches, 152 duplicates were removed. Of the 139 remaining records, 75 were excluded on Title & Abstract screening. Of the 64 recordings remaining after screening, 64 records were included for eligibility assessment. After 21 records

were excluded due to either insufficient discussion of genomics or LHS models, or insufficient identification of needs or opportunities, 43 records were included in the review (**Figure 6.1**).

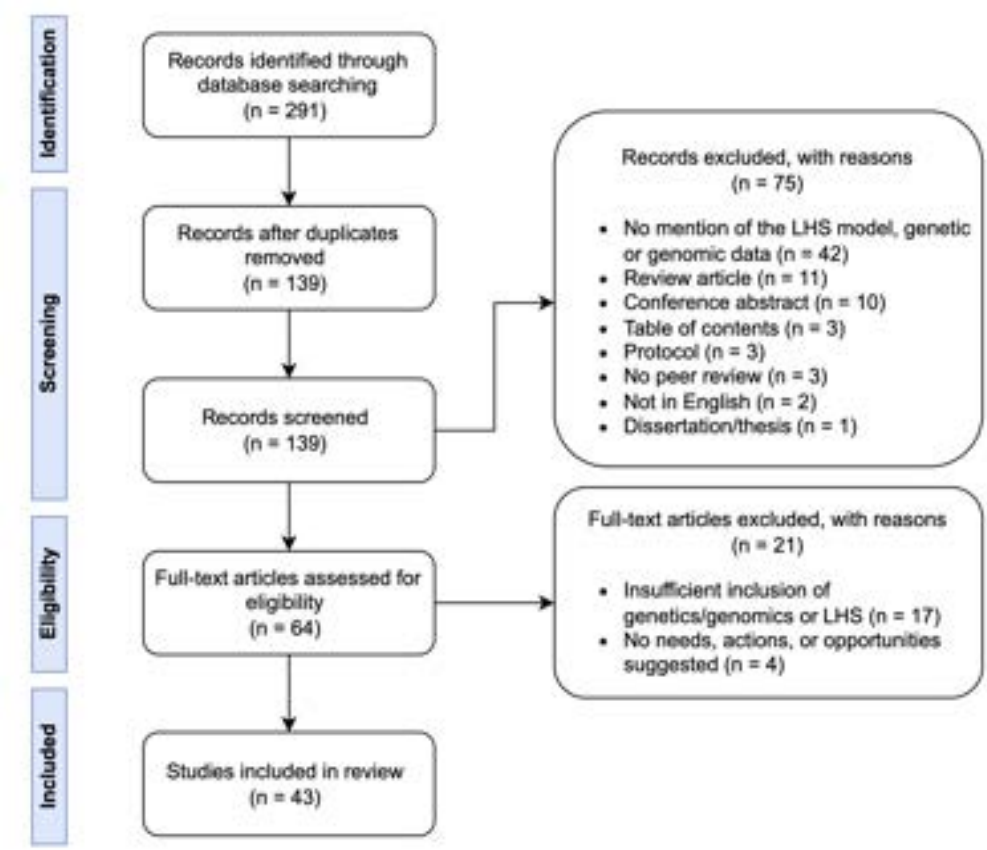


Figure 6.1. PRISMA [258] flow diagram of study Identification, Screening, Eligibility, and Inclusion for the systematic literature review.

Detailed study characteristics and results can be found in **Table S6.1**. Thirty-six (36) of the studies were written by investigators solely in the United States [13,31,132,259–291]. Two (2) were written by an international group of authors [292,293], 2 were written by Canadian authors [294,295], and 3 publications were written about health systems in Australia, Denmark, and the Netherlands, respectively [10,296,297]. Nineteen (19) publications were qualitative expert analyses [10,270–282,293,295,297], 8 were experience self-assessments from actual LHS

implementations [13,132,265–269,292], 6 were conference or workshop summaries [260–264,294], 4 were qualitative analyses of original interview or focus group data [283–286], 4 were system development and evaluation studies [287–290], 1 was a pilot study [291], and 1 was a case study [259]. Twenty-nine (29) publications were condition-agnostic, 11 focused on oncology, and individual studies focused on aneurysms, asthma, and inflammatory bowel disease.

During outcomes data extraction, 14 descriptive themes were iteratively identified among 291 codes. These themes were grouped into 4 higher-level analytical themes: 1. Data and standards; 2. Culture and acceptance; 3. Engaging with and protecting patients; and 4. Political and institutional support (**Table 6.2**).

Analytical Theme	Descriptive Theme	Number of Codes
Data and standards (**Table S6.2**)	Collecting, integrating, and sharing genomic and phenotype data	26
	Analyzing data for discovery	27
	Developing standards	17
Culture and acceptance (**Table S6.3**)	Building a collaborative learning culture	27
	Demonstrating value and feasibility	14
	Aligning learning with existing healthcare improvement models	10
Engaging with and protecting patients (**Table S6.4**)	Advancing health equity	16
	Prioritizing patient-centeredness	15
	Obtaining consent for clinical research	13
	Safety measures and outcomes monitoring	19
	Privacy and security protections	9
Political and institutional support (**Table S6.5**)	Funding and incentives	17
	Policy and governance	23
	Building institutional capacity for genomic medicine and learning	58

Table 6.2. Analytical and descriptive themes generated during systematic literature review content extraction.

A high-level exploration of the 14 descriptive themes identified during the literature review revealed how the current literature characterizes the inputs, activities, and intermediate outputs of clinical genomic discovery (**Figure 6.2**). Funding, policy, and governance were identified as requisite inputs to the activities that comprise clinical genomic discovery, which span the areas of support-building, cultural acceptance, patient-participant engagement, and data analysis and standardization. Demonstrating the value and feasibility of clinically embedded genomics

research to stakeholders was identified as an intermediate output of these activities that could potentially fuel the input of additional resources into the change process. The desired outcome was defined as the enhancement of clinically meaningful genomic discovery, as specified by the study PICO. Desired impacts were based on the expected outcomes of an effectively implemented LHS: better patient outcomes, better scientific understandings of health and disease, and decreased healthcare costs [7].

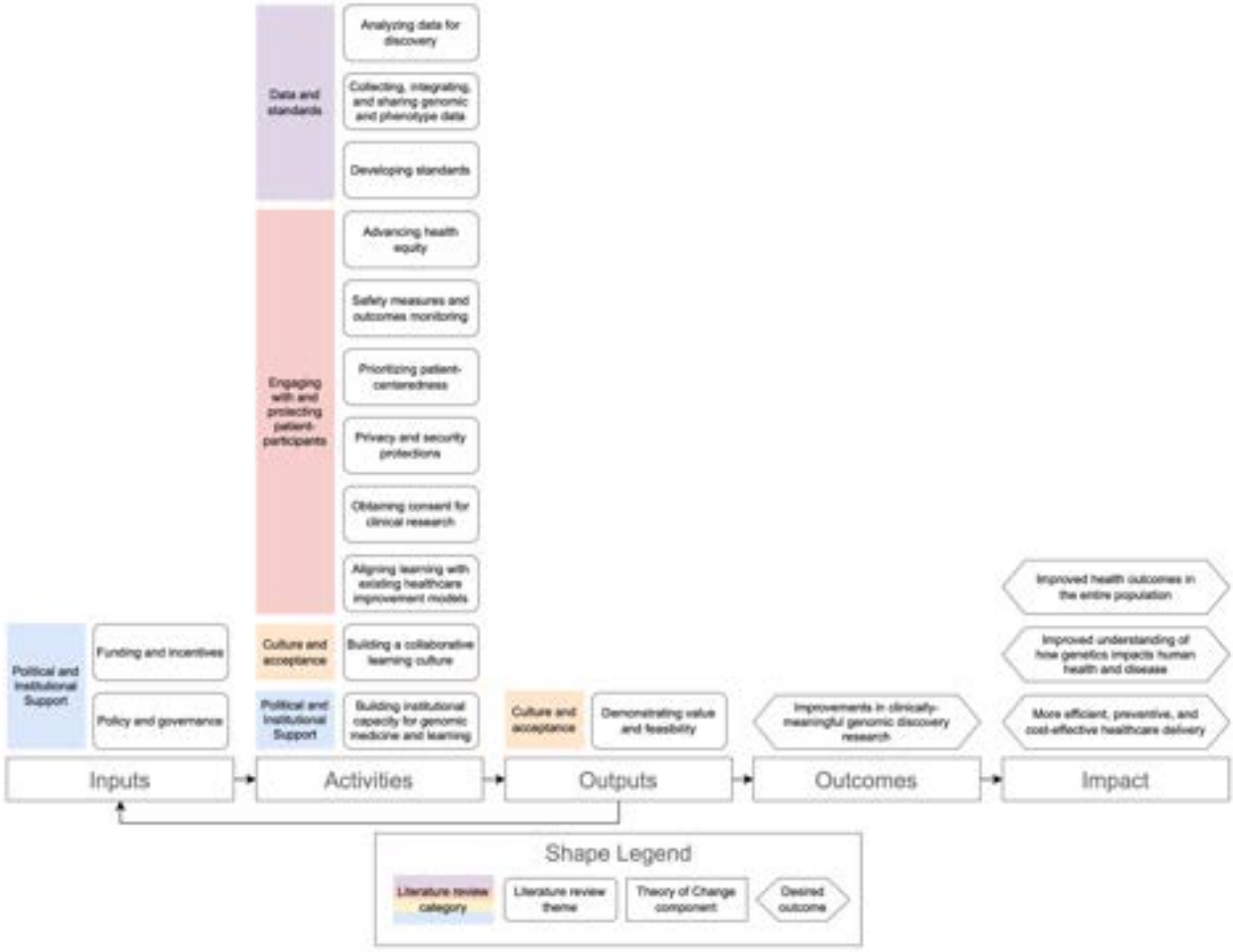


Figure 6.2. Initial theory of change diagram, based on 14 descriptive themes identified during the systematic literature review and desired outcomes and impacts defined by the study PICO.

After exploring the properties of each theme and identifying relationships between them through concept mapping, a new theory of change model was developed to better represent the spectrum of identified enabling factors for clinical genomic discovery in the literature (**Figure 6.3**). The literature tended to characterize themes as requisite inputs for a functioning LHS, rather than components of learning activities themselves. Enablers of learning activities depended on diverse combinations of different inputs. The intermediate output of LHS activities (demonstration of value) did not change between **Figure 6.2** and **Figure 6.3**.

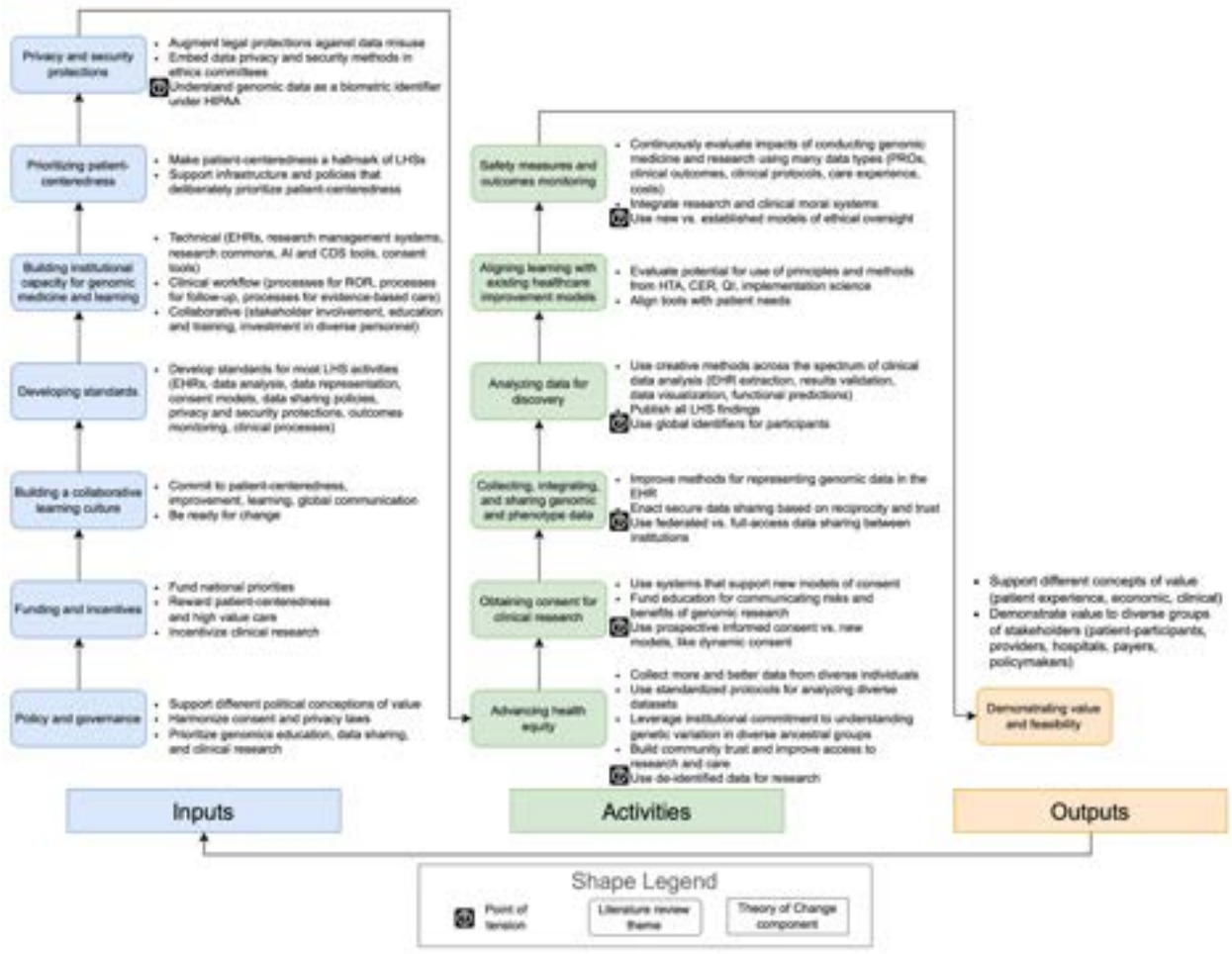


Figure 6.3. Updated theory of change diagram, based on property and relationship exploration of 14 descriptive themes identified during the literature review.

6.4.2 Best-fit framework synthesis

A total of 291 qualitative codes were created using the codes identified during the systematic literature review, and 12 codes were created from the "Needs" column in **Aim 1, Table 3.3**. These codes were grouped into the 5 major themes from the *a priori* 5R model developed in Aim 2 (Representation, Responsibility, Risks and Benefits, Relationships, and Resources), and new themes were iteratively created to accommodate codes that did not sufficiently relate to the *a priori* model themes. Four (4) new thematic codes were created after constant comparison of framework themes and the 303 total codes from Aim 1 and the literature review. All other codes were grouped into the 5 existing *a priori* themes. Descriptions of the new themes are included in **Table 6.3**.

Theme	Codes	Description
Analysis approaches	18	Suggestions for specific analysis tools and approaches that can be used to represent clinical and genomic data, such as rigorous statistical models, scalable data extraction methods, platform-agnostic tools, and functional effect prediction
Automation approaches	15	Suggestions for automated approaches to implementing new knowledge in clinical practice, such as mass customization, FHIR tools, predictive models, and usability labs
Outcomes monitoring approaches	16	Suggestions for approaches to monitoring the outcomes of clinical research and implementation, such as the use of patient-reported outcomes, mechanisms for routinely following up with patient-participants, and definition of consensus outcomes measures
Standardization approaches	21	Different approaches to standardizing aspects of clinically embedded research, such as unified data architectures, standards for capturing diversity, semantic interoperability of biomarker data, and regulated diagnostic approaches using sequencing technologies
Value assessment approaches	7	Tools for assessing and demonstrating the value of genomic medicine to diverse stakeholders, including patients, organizations, and payers

Table 6.3. New themes identified using best-fit framework synthesis.

6.5 Discussion

6.5.1 Systematic literature review

Overall, the volume of peer-reviewed literature on the enabling factors of clinically embedded genomic discovery was limited, but the suggestions made across articles were broad and spanned many disciplines (**Table S6.1**). The vast majority of articles were perspective pieces or commentaries based on limited implementation experiences or content expert suggestions.

Several articles were written by investigators from countries with nationalized health systems such as Canada, Australia, Denmark, and the Netherlands, but the majority were written in the context of the US healthcare system. Few articles focused solely on the data to knowledge aspect of LHSs, but rather discussed discovery in the context of other learning processes. The lack of comparative-effectiveness research on enablers of clinical genomic discovery suggests that this area of research is still in its infancy and speaks to the immense challenge of implementing such systems in practice.

As defined in the scope of this review, the ultimate vision of genomics-enabled learning healthcare is to improve population health, increase the public understanding of how genomics impacts human health and disease, and increase the efficiency of the healthcare system by enhancing disease diagnosis and prevention [8]. Unsurprisingly, the literature collectively suggested a lofty and complex set of inputs and activities that would likely be necessary to achieve this ambitious vision. While the chain of literature review themes depicted in **Figure 6.3** is not strictly linear, it approximates the ways in which inputs and activities build upon one another in the process of conceptualizing a model of clinically embedded research, preparing to implement the model, implementing the model, and demonstrating the value of the implementation.

The literature unanimously asserted the importance of national policy, funding, and incentive systems in providing a foundation on which learning systems could be built (**Table S6.5**). Several publications suggested that harmonizing government policies relating to health data use and data sharing [261,294] could relieve the burden placed on both health institutions and

patient-participants to decode separate but overlapping policies. Policy prioritization was also identified as having significant potential for enabling desired downstream effects, such as implementation of patient-centered systems, genomics education initiatives, data sharing networks, and global collaboration in clinical research. One publication also suggested that health research policies recognize the various dimensions and conceptions of value that an LHS could provide [263]. The concept of value was a recurring theme in many publications and was identified as a key component in creating virtuous learning cycles. Policymakers have the authority to both define measures of value and judge whether value is being produced in LHSs, which makes it all the more important to encourage nuanced understandings of value in the clinical, research, and policy communities. Initiatives deemed as having value, either potential or observed, receive funding, which is the lifeblood of the research enterprise and publicly financed healthcare institutions. Several publications suggested that incentive systems should be used to fuel desired innovations in healthcare research and practice [263,272,291], such as tools and processes for maximizing patient-centeredness. The literature collectively suggests that policy and funding provide both the means and the motivation for institutions to begin preparing for clinical research integration.

Because the preparation and activities involved in genomics-enabled learning healthcare are labor intensive and sometimes require controversial decisions, a cultural commitment to learning, communication, and improvement within and between healthcare institutions is widely recognized as necessary for LHS development (**Table S6.3**). Only once institutions are invested in change can they be earnestly involved in developing and adopting standards, building institutional capacity, and adopting and implementing privacy and security measures. Many

studies deemed standards essential for the full spectrum of LHS preparations and activities, given the current heterogeneity in research and clinical approaches across healthcare institutions in the US and the need for large scale collaboration (**Table S6.2**). The technical, procedural, and collaborative systems that institutions develop to enable learning healthcare are ideally based on these standards to maximize the safety and efficiency of such systems. Many studies also suggested making a patient-participant-centered approach the backbone of all learning health infrastructures and activities, which necessitates a focus on building technical and procedural capacities for privacy and security protections of patient-participant data. However, the current legal protections of genomic data under HIPAA are still unclear and should be clarified [261] (**Table S6.4**).

All subsequent LHS activities identified in the literature build upon the basic tenets of funding, prioritization, culture, standardization, patient-centeredness, institutional infrastructure, and data security. However, the suggested approaches to different activities are diverse and sometimes in tension with one another. Several publications focused on the role of clinical learning environments in advancing (or impeding) health equity [31,272,274,276,277], suggesting that significant efforts should be made to form trusting and sustainable relationships between health systems and diverse patient-participant communities. With increased engagement of diverse individuals will come more and potentially higher quality clinical and genomic data from diverse communities, which can be used to advance clinical genomics research that truly gives back to those communities. Using appropriate analysis methods for diverse genomic datasets should become a standard practice in LHSs, but it is possible that individuals from backgrounds that have historically been marginalized and mistreated by the medical and research enterprises will

only feel comfortable sharing data for research if it is de-identified. One publication suggested that new models be developed to enhance the research utility of de-identified data [31], but there is debate among the literature over whether the full potential of LHS research can be reached using de-identified clinical data. A related tension is the debate over whether federated data sharing of clinical and genomic data can be used in place of all-access data sharing for genomic discovery research [291]. Generating high-quality, representative clinical and genomic datasets is widely recognized as an essential part of the learning process, but the policies and moral systems that dictate the level of detail available to different clinical researchers remain highly debated.

Different perspectives exist regarding the appropriate methods of consenting patient-participants for clinically embedded genomics research. Some publications advise the use of prospective, broad consent to reap the full benefits of clinical research [262,268], while other advise the use of new and creative models of consent, such as dynamic consent, to best respect the wishes of patient-participants [10,13,294,297]. However, most publications agreed that widespread education on the potential risks and benefits of genomic data sharing should be publicly funded, and that communication between patient-participants and those consenting them to clinical research be as clear as possible, regardless of the consent model used.

The literature was in broad agreement that more advanced, representative, and accurate tools for clinical and genomic data analysis are needed. Methods should be developed by an interdisciplinary network of researchers and clinicians, given the interdisciplinary origins of the data and broad implications for use in a medical context. Some publications also stressed the importance of communicating research findings between LHSs and the broader research

community through peer-reviewed publications [263,276,283]. However, for LHS research that is more quality-improvement oriented, many publications suggested using existing principles and tools from comparative-effectiveness research (CER), quality improvement (QI), health technology assessment (HTA), and implementation science, rather than developing new tools altogether [132,261,263,267,271,279,282,294,296]. The only stipulation of these publications is that existing tools be evaluated for effectiveness in a learning health environment, and that they cater to the patient-centered approach of GLHSs.

The literature was in wide agreement that any changes made to clinical practice based on research findings should be heavily documented and monitored. For patient safety reasons, centralized ethical oversight should be used to weigh the risks and benefits of clinical research. However, this would require a complex merging of moral systems used to differentially guide research and clinical decision-making [277]. While the current models of ethical oversight for clinical research are deeply embedded in research culture, some publications propose that new models of ethical oversight for routine, clinically-embedded research be developed [10,132,261]. For the purposes of measuring and demonstrating value, a diverse set of clinical, economic, and patient-reported outcomes should be continuously measured in an LHS. If definitions of value are to be dynamic and context-specific, the outcome measures used to inform value assessments should be dynamic as well. Ultimately, demonstrating value to a diverse group of stakeholders could help fuel the necessary inputs of LHS development, leading to virtuous cycles of clinically embedded genomics research.

As previously mentioned, most articles included in the literature review were perspectives and experience reports (**Table S6.1**). This limits the empirical strength of the assertions made in the literature review, and necessitates that all results be viewed as suggestive, rather than prescriptive. Many studies also disclosed sources of funding that were not strictly academic, or disclosed conflicts of interest such as involvement with pharmaceutical companies or for-profit healthcare organizations. While this partially speaks to the understanding in the literature that partnerships with for-profit organizations make learning healthcare more feasible [13], it also warrants a critical examination of whether such partnerships are truly necessary, or whether company involvement in academic discussions biases the literature towards that understanding. It is also important to recognize that very few publications referenced the "genomics-enabled learning health system" as a singular concept in their analyses. Instead, the primary investigator assumed that publications that both considered the LHS model and recognized genomic data as unique from other clinical data could be referred to as studies that discussed the core concepts of a GLHS. Additional studies that discuss the GLHS model as a discrete entity are needed to fully capture the challenges and opportunities of the model.

6.5.2 Best-fit framework synthesis

Evidence from the literature review and Aim 1 could largely be grouped into the existing themes of the *a priori* model from Aim 2, suggesting that the 5R model is a reasonable representation of the factors involved in learning healthcare. However, two new insights were gained from the BFFS process: 1. The literature offers potential tools that can be used to facilitate negotiations; and 2. The inputs of the theory of change model developed from the literature review do not necessarily need to be complete prior to LHS implementation (**Figure 6.4**).

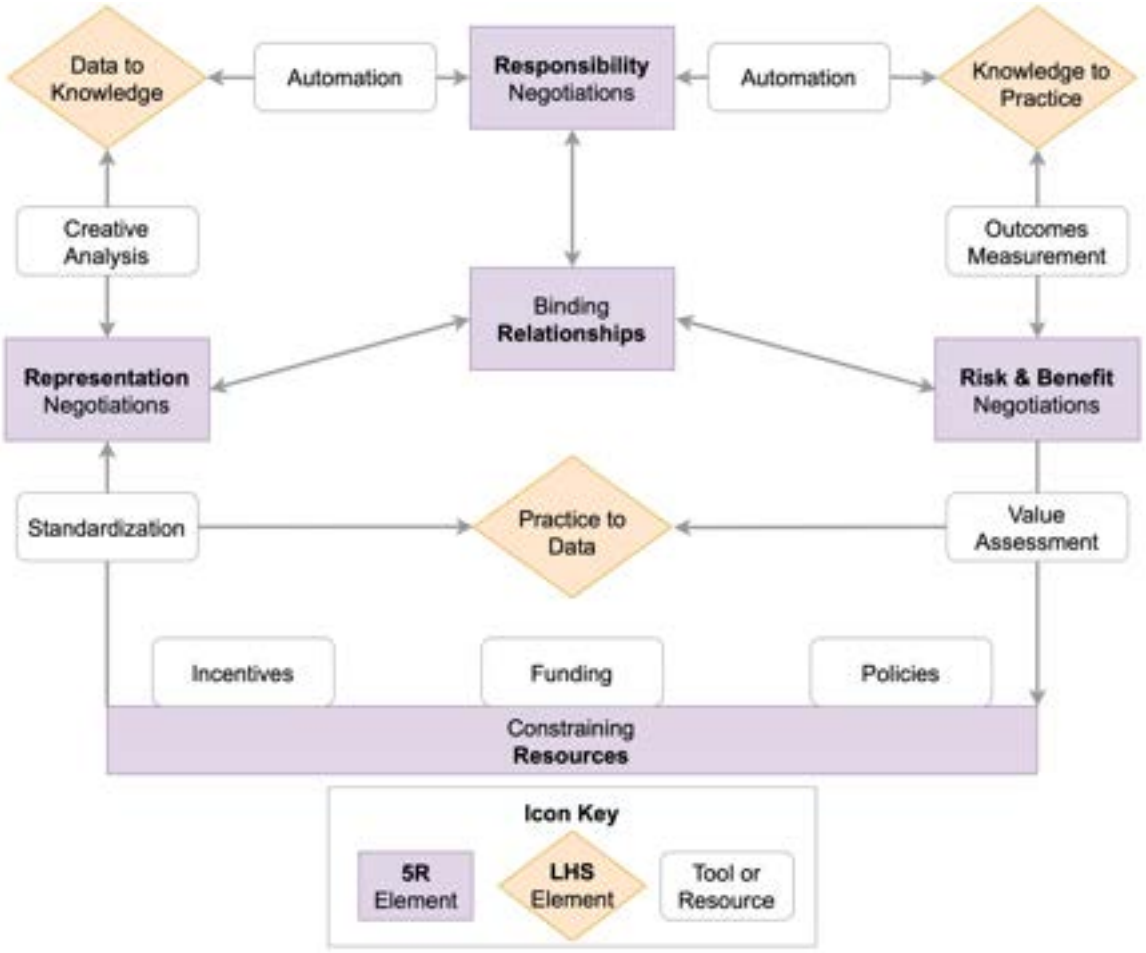


Figure 6.4. 5R sociotechnical model of discovery in a genomics-enabled learning health system, based on best-fit framework synthesis of the *a priori* model from Aim 2, a systematic literature review, and Aim 1 results.

While the in-depth interview study approach is useful for characterizing relationships between LHS processes and deconstructing the roles, expectations, and tensions involved, surveying the available literature is useful for brainstorming approaches to resolving tensions and reconstructing the operational picture of an LHS. There was widespread consensus in the literature that standards can facilitate many aspects of the learning process, including infrastructure development, analysis, clinical and research protocols, implementation, and outcomes measurement. The recommendations from CSER data coordination in Aim 1

corroborated this finding, given that standardizing and streamlining both the informatics and communication aspects of data coordination were key to successful implementation. If the space between data generation and knowledge generation is considered an "entry point" into LHS learning cycles, standardization efforts could be focused at that entry point and then propagated forward through the learning process. Novel approaches to clinical and genomic data analysis were also suggested across the literature, including ontology-driven approaches to omics data extraction and analysis, scalable and platform-agnostic tool development, methods for using de-identified clinical and genomic data, and improved data visualization techniques. In this way, both standardization and creative data analysis could facilitate the scientific and workflow-dependent aspects of representation. However, building relationships between patient-participant communities and health research institutions is still at the core of improving representation in clinical genomics research.

The literature identified automation as the primary tool for facilitating the path from knowledge to practice in an LHS. The use of genomic CDS tools and predictive models to sustain innovation and encourage evidence-based practice was widely suggested, which aligned with a recurring suggestion from interviewees in Aim 2 that CDS tools could help curb inappropriate use of genomic knowledge in clinical decision-making. Automated processes for variant interpretation that systematically use the best available evidence, even if limited or preliminary, could assist researchers and clinicians in adjudicating variant validity, actionability, and utility. However, automation will not eliminate differences in the meaning of "research" given different research and clinical histories and expectations in genomics. Ethical limitations in the production, validation, and use of new knowledge are ultimately at the whims of social evolution

in the research and clinical traditions, and automation will likely have little to do with the mutual understandings that must form between researchers and clinicians when negotiating responsibility for patient-participant wellbeing.

In the case of negotiating risks and benefits, systematically measuring a variety of outcomes was identified as the primary method for determining both the risks and benefits of conducting genomic research in a healthcare setting. Because value is defined differently by different stakeholders (e.g., patient-participants, researchers, providers, payers, policymakers), multiple types of information that could be used to measure value should be captured (e.g., patient-reported outcomes, scientific outcomes, clinical outcomes, economic outcomes). These measurements can be used to inform risk and reward negotiations on the part of researchers and clinicians in future learning cycles and can also be used to justify policies that sustain funding and incentives for learning healthcare. The question of how ethical oversight of clinical research could be improved, however, is still dependent on the evolving relationships between embedded ethics committees and those conducting clinical research.

While the literature tends to characterize processes like standards development and institutional capacity building as prerequisites for conducting learning activities in an LHS, the updated GLHS model proposed in this study suggests that these processes do not necessarily need to be fully fleshed out before beginning the learning process, in much the same way that "clinical research need not be complete prior to implementation" (Williams et al. 2018, p. 763) [13]. Instead, preliminary implementations of genomics-enabled LHS cycles that incorporate early versions of the technical, social, ethical, and structural components identified broadly in the

literature can be used as a way to dip the world's proverbial toe into clinically embedded research. This approach should be taken with great caution at first, because there should be a widely held understanding that learning cycles will improve with time as components of the learning process generate better understandings of themselves. Heeding the processes of negotiation that will likely shape the evolution of learning systems can help instill a culture of respect and vigilance in those systems and can help ensure that research and clinical traditions evolve together into something that is greater than the sum of their parts. They may never reach a state of true equilibrium, but they can continually learn from one another for the sake of improving human health.

6.6 Limitations and future work

6.7 Conclusion

In this aim, we demonstrated that the number of studies addressing the topic of clinically embedded genomic discovery is small, but that the complexity of suggestions made by those studies is disproportionately large. Proposed enablers of clinical genomics research can be roughly organized into a set of interdependent inputs and activities that enable continuous learning that is participant-centric, value-oriented, and equitable. The integrative 5R sociotechnical GLHS model developed using BFFS methods offers a conceptual basis for the ways in which proposed enablers of clinical genomics research integration can facilitate negotiations between the (sometimes conflicting) priorities of research and clinical care. Progressively developing and testing the suggested components of genomics-enabled learning cycles can elicit additional resources and fuel virtuous cycles of learning.

CHAPTER 7: CONCLUSIONS AND SUMMARY OF CONTRIBUTIONS

In this work, we derived an integrative conceptual model for GLHS discovery (the "5R" model) that represents clinical research integration in genomics as an iterative and multidirectional process involving constant negotiations between research and clinical stakeholders, exploratory informatics tool development and adoption, and relationship building. The 5R GLHS model offers a genomics-specific enhancement of the original LHS model and provides a conceptual foundation upon which future GLHS implementation frameworks can be built. Through careful consideration of the sociotechnical factors involved in building virtuous cycles of learning and improvement in health systems, genomics research and genomic medicine can co-evolve to improve population health equitably, safely, and effectively.

This work offers several contributions to the fields of biomedical informatics, immunogenetics, and learning health systems research:

1. Recommendations for best practices in multi-institutional clinical and genomic data coordination work that can be generalized to projects with diverse patient-participant populations, sizes, and funding capacities (**Aim 1**).
2. A novel conceptual model for understanding the research-clinical interface in the context of GLHSs, from the perspective of genomic medicine experts (**Aim 2**).
3. A novel gene-disease association in the HLA-DRB locus that may predispose individuals to *C. diff.* infection, and mechanistic hypotheses for the association that span multiple immunological perspectives (microbiome-mediated, T-cell mediated, methylation-mediated) (**Aim 3**).

4. A novel, integrative genomics-enabled learning health system conceptual model that incorporates enabling factors for clinical genomic discovery identified by a Cochrane-style systematic literature review (**Aim 4**).

APPENDIX

Appendix A. First interview guide for Aim 2 interview study.

Intro

1. Background
    a. Genomic medicine is becoming more widely recognized as useful and important, but widespread clinical adoption is lacking
    b. Genomic discovery activities are typically conducted in research settings using research data, rather than clinical settings using clinical data
    c. Embedding genomic research within clinical environments could both increase pace of clinically meaningful discoveries, and increase evidence for utility needed for wider adoption
        i. In a genomic learning healthcare system model:
            1. Implementation of new genomic medicine practices → collection and analysis of outcomes data (and genomic data?) → new genomic knowledge → quality improvement strategies → cycle starts over
        ii. Idea is to rapidly move genomic discoveries into clinical care, then bring clinical observations back to research setting, then use those observations to inform discovery efforts, and so on
    d. The only question is: how?
2. Informed consent for participation
3. Informed consent for interview recording

Discussion Points

- Personal background (clinical training and focus, work setting, etc.)
- What kinds of genetic/genomic tests do you currently use, if any? In what types of clinical situations?
- Are there any clinical areas that might particularly benefit from rapid genomic discovery efforts (and subsequent applications)?
- How might clinically-based genomic discovery research (e.g. GWAS, PHEWAS) that is conducted as a by-product of clinical care *differ* from discovery work that is conducted in purely research settings?
- What are some benefits and/or drawbacks of embedding genomic discovery research programs within clinical environments?

- What are some of the supporting elements that would need to be in place to facilitate clinically-based genomic discovery research (technical, legal, social, ethical, structural, etc.)?
- What information would you need to determine if a new clinical genomic discovery (e.g. a gene-disease association) is ready to be put to clinical use? What safeguards would need to be in place?

Appendix B. Second interview guide for Aim 2 interview study.

Intro

1. Background
    a. We know there is a lack of evidence for the outcomes of implementing genomic medicine (exactly the kind of thing that CSER is trying to address…large, multi-site research consortia can help address this evidence gap)
    b. I'm interested in exploring another way that we can start to close this evidence gap and bring useful discoveries into the clinic more quickly - through a learning healthcare system model, where genomic research is just embedded within clinical environments, and new discoveries are used to iteratively inform care, and we can collect outcomes data through the healthcare system
    c. Interested in exploring
        i. Is this a good idea? Are there any major red flags we should be looking for?
        ii. And if some healthcare systems do start to implement this model, how? There are likely many ethical, technical, social components that need to be considered, but it's not exactly clear what those components are and how they should be addressed
2. Informed consent for participation
    a. This will be completely anonymous, your name won't be used in relation to any of your responses, either within the study team or in writeups
    b. You can remove yourself from the study at any time (including during this interview), and request that your responses be removed from the study as long as it's possible to extract them from aggregate analyses
3. Informed consent for interview recording
    a. Will transcribe and qualitatively code the interview data, which will be anonymized
    b. You can choose to stop the recording any time

Discussion Points

- Background
    - Can you tell me a bit about your **personal background as a clinician-researcher** (clinical training and focus, work setting, day to day, etc.)?
        - What kinds of **genetic/genomic tests do you currently use**? In what types of clinical situations?
    - Can you tell me more about your **research** at [institution name]?
- Challenges related to the data itself
    - What are some **challenges** you might run into when using **clinical data for research** (both genomic and otherwise) vs. using data collected as part of a dedicated research process? An example of research could just be something like a GWAS or PHEWAS, looking at genotype-phenotype correlations.
    - Conversely, what are some **potential benefits** of using this type of data for research?
    - What might make i**t difficult to collect clinical outcomes** of clinical decision-making that has potentially changed due to a new genomic discovery, for example a new gene-disease or variant-disease association?
    - Where does that evidence for actionability come from? How would this work in a LHS?
- Potential for clinical benefit?
    - Thinking about the LHS model that I mentioned earlier, are there any clinical areas that might **particularly benefit from rapid genomic discovery efforts** (and subsequent applications)?
    - Conversely, are there clinical areas where it might be **more dangerous and/or challenging** than others to introduce things like new gene-disease associations into clinical practice more quickly?
    - What are some **benefits and/or drawbacks** of embedding genomic discovery research programs **more broadly within clinical environments**?
        - We already see a lot of this type of discovery work done in oncology. Can (and should) we use clinical genetics research in oncology to inform research models in other clinical areas?
    - **Is the LHS model a reasonable one for increasing the evidence base** for genomic medicine, where genomic discoveries might be implemented more swiftly than they otherwise would?
    - Do you think genomic **surveillance or population screening should be implemented more widely in a healthcare system?** An example of such an existing system is prenatal screening, which seems pretty widely applied.
- Differences in discovery practice between research and clinic

    - How might clinically-based genomic discovery research (e.g. GWAS, PHEWAS, gene-drug interactions, etc.) that is **conducted as a by-product of clinical care *differ* from discovery work that is conducted in purely research settings?**
    - Do you think there is a reason that genomics research and medical genomics practice **should be carried out in two separate venues,** potentially using **different funding sources?** Why or why not?
        - Is there anything specific about genomics that would make it more or less amenable to integrating research into a clinical setting?
- Requirements for implementing clinically-based genomics discovery programs
    - What are some of the **supporting elements** that would need to be in place to facilitate clinically-based genomic discovery research (technical, legal, social, ethical, structural, etc.)?
        - What do we need to make data better?
    - What information would you need to determine if a new clinical genomic discovery (e.g. a gene-disease association) is **ready to be put to clinical use/actionable?** What safeguards would need to be in place?
        - How would/should this **"clinical use" be identified, verified, tested, and ultimately validated?**
    - How should the **roles of patient and research participant be balanced** in something like a learning healthcare system, especially for genomics research?
        - How might consent models need to change to accommodate this dual role?
        - What do you think about dynamic consent? Should it be just forward based? Or always backwards based?
- Ethics in genomics
    - How do you think medical genomics fits in with the notion of **distributive justice** in healthcare system (the assumption that if you're spending money one place, you're not spending it somewhere else, where it might be needed more urgently)
- "Futuristic" thinking
    - What is your **personal vision for the future of genomic medicine?** How can it best be used to **equitably improve healthcare?**
    - How do you think third party vendor data should fit into the healthcare system?

Appendix C. Third interview guide for Aim 2 interview study

Intro

1. Background
    a. We know there is a lack of evidence for the outcomes of implementing genomic medicine, and like most areas of research, it takes a long time for new discoveries to come to fruition in a clinical setting
    b. I'm interested in **exploring one model we might be able to use** to start to close this evidence gap and bring useful discoveries into the clinic more quickly:
        i. Through a learning healthcare system model (originally proposed by IOM/NAM in the early 2000s), where genomic research is just embedded within clinical environments, and new discoveries are used to iteratively inform care, and we can collect outcomes data through the healthcare system
    c. Interested in exploring
        i. Is this a good idea? Are there any major red flags we should be looking for?
        ii. And if some healthcare systems do start to implement this model, how? There are likely many ethical, technical, social components that need to be considered, but it's not exactly clear what those components are and how they should be addressed
2. Informed consent for participation
    a. This will be completely anonymous, your name won't be used in relation to any of your responses, either within the study team or in writeups
    b. You can remove yourself from the study at any time (including during this interview), and request that your responses be removed from the study as long as it's possible to extract them from aggregate analyses
3. Informed consent for interview recording
    a. Will transcribe and qualitatively code the interview data, which will be anonymized
    b. You can choose to stop the recording any time
4. Any questions for me before we get started?

Discussion Points

Background

- Could you tell me about your professional background? Clinical training, research training (if any), clinical specialties, etc.

- Are you currently seeing patients in a medical genetics clinic? If so, what types of genetic tests do you typically order, and for what indications?
- Do you currently do any clinical research? If so, what does that research look like?

Using clinical data for research

- What are some of the challenges you might run into when using clinical data for research, as opposed to using data collected as part of a dedicated research study or data in public or controlled access databases?
- Conversely, what would make clinical data better for genetics research than research data or public/controlled-access data?
- If "healthy" individuals were to be broadly sequenced in a healthcare system, what are some downstream effects that should be considered (if any)?
- What are the pros and cons of using clinical data to monitor longitudinal patient outcomes?

Consent for clinical research

- Are there other models of consent (other than broad, up-front consent) that should be considered for clinically embedded genetics research? If so, what would those models look like? If not, why not?
- Should patients receive incentives (monetary or otherwise) to participate in clinical research, or to consent to data sharing in a clinical research institution? If so, why, and what types of incentives? If not, why not?

Use and return of research results

- If you found a potential disease-associated variant during a clinical research study, what would need to happen to deem it clinically actionable? Where would the evidence for that decision come from? Who would/should be making that decision?
- Should new genetic discoveries be used to impact care in a healthcare system? If so, how? If not, why not?
- Do research patients typically expect results to come back to them? If so, should there be efforts on the part of researchers and/or clinicians to manage patient expectations?
- If there is a secondary finding from a purely clinical test (e.g., NIPT), should those results be returned to patients? If so, under what protocols (if any)? If not, why?

Roles

- If genetic research activities were conducted more routinely in clinical settings, what would need to be done (if anything) to reconcile the dual role of patients as both patients and research participants? Do you foresee any challenges or benefits of this dual role?
- Who should be conducting genetic research in clinical settings? What types of

collaborations would need to be in place to make clinical research successful (if any)?

Payment and distributive justice

- Where do you think medical genetics stands with regards to distributive justice (the idea that if you are spending money one place, you are not spending it somewhere else that may or may not need it more) in the US healthcare system?
- Do you think there should continue to be separate funding sources for research and clinical care if the activities become synchronous? If so, why? If not, what might new funding models look like?

Defining "clinical research"

- How would you distinguish between research and routine quality improvement if genetics research were embedded in clinical environments (if at all)?
- Do you feel like the research and clinical enterprises are separate entities? If yes, in what ways? If not, why?

The future of genetics in medicine

- Do you have any concerns about genetics being used more broadly in other medical disciplines? If so, what are your concerns? If not, why?
- Should certain types of genetics research be prioritized to improve population health? If so, what might that prioritization look like? If not, why not?
- What is your personal vision for the future of genetics in medicine? How do you think genetics can be used to have the greatest impact on individual and population health?

Appendix D. REDCap demographics survey for Aim 2 interview study.

Clinical Genomics Discovery Project - Interview Participant Information

Thank you for participating in my interview study! I am collecting some basic demographic information to include in a summary table for the project write-up. Please see below for a few quick questions, which will be collected anonymously.

What type of medical and/or research environment do you currently work in?

- ◯ Academic Medical Center
- ◯ Integrated Care Organization
- ◯ Research-Only Hospital
- ◯ Other (please specify)

Other (please specify the type of medical and/or research setting you currently work in):

_____

What are your credentials?

- ◯ MD
- ◯ MD, PhD
- ◯ Other (please specify)

Other (please specify your credentials):

_____

What board certification(s) do you have in addition to Medical/Clinical Genetics and Genomics (if any)? Please check all that apply:

- ☐ Clinical Cytogenetics and Genomics
- ☐ Clinical Molecular Genetics
- ☐ Internal Medicine
- ☐ Medical Biochemical Genetics
- ☐ Obstetrics and Gynecology
- ☐ Pediatrics
- ☐ Preventive Medicine
- ☐ Psychiatry and Neurology
- ☐ Other (please specify)

Other (please specify any board certification(s) you have in addition to Medical/Clinical Genetics and Genomics)

_____

What type(s) of genetic disorders do you specialize in? Please check all that apply:

- ☐ Cardiovascular disorders (cardiomyopathy, arrhythmia)
- ☐ CNS disorders (epilepsy, encephalopathy, structural brain malformations, neurodegenerative disease)
- ☐ Dysmorphology/Structural developmental abnormality
- ☐ Immunodeficiency
- ☐ Neurodevelopmental abnormalities (intellectual disability, autism)
- ☐ Neuromuscular disorders (hypotonia, spasticity, neuropathy, myopathy)
- ☐ Metabolic disorders
- ☐ Skeletal dysplasias
- ☐ Cancer
- ☐ Other (please specify)

Other (please specify the type(s) of genetic disorder that you specialize in):

_____

What category or categories best describe you? Check all that apply:

- ☐ American Indian, Native American, Alaska Native
- ☐ Asian
- ☐ Black or African American
- ☐ Native Hawaiian/Pacific Islander
- ☐ White or European American
- ☐ Middle Eastern of North African/Mediterranean
- ☐ Hispanic/Latino(a)
- ☐ Prefer not to answer
- ☐ Unknown/none of these fully describe me

How do you describe yourself?

- ◯ Woman
- ◯ Man
- ◯ Non-binary/non-conforming
- ◯ Prefer not to respond
- ◯ Other (please specify)

Other (please specify how you describe your gender identity):

_____

Appendix E. Axial code descriptions for iteration 3 of codebook development.

Code Group	Code	Definition
Consent	Avoiding coercion and respecting patient wishes during consent	Strategies for protecting patients during the research consent process
	Considering the pros and cons of broad, up-front consent	Descriptions of what **broad** consent might look like in clinical research settings, and the potential implications for people and research processes
	Considering the pros and cons of dynamic consent	Descriptions of what **dynamic** consent might look like in clinical research settings, and the potential implications for people and research processes
	Maintaining transparency and setting expectations in consents	Considerations for explaining research objectives to patients and documenting the consent process
	Merging ethical oversight of research and clinical care	Thoughts about whether it would be feasible or useful to combine the forces of IRB monitoring and clinical oversight for clinical research studies
	Using technology to aid the consent processes	Suggestions for using technology to our advantage when consenting patients for clinical research
	Working with IRBs to conduct clinical research	Observations from working with IRBs on clinical research projects in the past: the good, the bad, and the ugly
Current Practices	Deciding between broader and narrower tests for different indications	How things like Bayesian logic and uncertain results factor into clinicians' decisions about ordering certain types of genetic tests
	Doing clinical research in integrated and universal health systems	Personal experiences and observations of what it's like to do clinical research in settings like Kaiser, Mayo, and Geisinger, or nationalized healthcare systems
	Doing clinical research in non-integrated healthcare systems and research hospitals	Personal experiences and observations of what it's like to do clinical research and clinical care in the fragmented US healthcare system
	Ordering different types of genetic tests	The types of genetic tests that geneticists typically order for different indications
Data	Getting data into the EHR	Problems with how genetic and clinical data gets into commercial EHRs
	Getting data out of the EHR	Problems with finding and extracting useful data out of the EHR

	Protecting the privacy and security of clinical data	Techniques for anonymizing patient data, and concerns about who can and should access patient data
	Sharing and recycling clinical and genomic data	Experiences with accessing and transferring patient data across institutions
	Using clinical diagnostic lab data for secondary research	Experiences and perceived pros/cons of using data from clinical diagnostic labs for research
	Using commercial genomic data in the clinic	Considerations for using direct-to-consumer testing reported by patients in a clinical setting
	Using EHR and clinical genomic data for research	Things that make EHR data both better and worse for research than other types of clinical data
	Using large databases for genomic research	Considerations for using large, de-identified genetic databases (like gnomAD, UK Biobank and All of Us) for clinical research
	Using traditional clinical research data for research	Things that make it easier to use data collected as part of a dedicated research project for clinical research
Discovery	Contextualizing current knowledge with past discoveries	Musings on where we are now in genetic research vs. the very recent past
	Studying rare vs. common genetic variation/disease	Arguments about the merits and drawbacks of studying common diseases vs. rare diseases with a genetic etiology
	Understanding genetic impacts on health and disease	Allusions to the vastness of potential genetic impacts on human health and how much we still don't know
	Using clinical tests for secondary research	How purely clinical tests (like prenatal genetic screening) can or cannot be used for secondary research
Engagement	Educating communities about genetics	Issues with teaching basic genetic concepts to the general population, and strategies for doing so
	Educating non-genetics specialists about genetic medicine	Needs and suggestions for getting more non-genetic medical specialists interested in comfortable with using genetics as a tool
	Engaging underrepresented communities in genetics research	Observations and ethical considerations for engaging underserved and underrepresented communities in genetics research
	Forming collaborations within and between hospital systems	Current challenges with forming collaborations within healthcare research institutions, and strategies for building relationships

	Generating excitement and interest in clinical research	Reasons why researchers and clinicians do and don't get involved in clinical research
	Incentivizing and compensating research participants	Pros and cons of altruism vs. using incentives to engage and/or compensate research participants
Implementation	Comparing and contrasting research and clinical care	Opinions about whether or not the research-clinical divide is real or imagined
	Developing clinical guidelines from genetic discoveries	Observations about the current state of translating new genetic discoveries into clinical guidelines that impact care practices
	Distinguishing between research and quality improvement	Observations of differences and/or similarities between research and clinical quality improvement
	Embedding genetic research in routine clinical care	Potential pros and cons of embedding genetic discovery research within clinical environments
	Following up with patients after genetic testing	Downstream implications of genetic tests or screens, and what patients are owed in terms of clinical follow-up
	Misusing and misinterpreting genetic tests for clinical care	Concerns about genetic tests being inappropriately ordered or interpreted by either non-genetics or genetics specialists
	Testing new clinical interventions	Current standards for testing out new clinical interventions, and concerns about implementing new knowledge or tools too quickly
	Using population-wide genetic screening in clinical care	Pros and cons of using population-de genetic screening to guide clinical care and improve public health
	Using the EHR to streamline clinical genomics	Current challenges with integrating genomic CDS into EHRs, and hopes for the future of CDS integration
	Utilizing remote medicine and eConsults in genomics	Pros and cons of using remote medicine and/or eConsults to practice genetic medicine
Participant Background	Clinical Practice	Current and past experiences in clinical practice
	Leadership, Teaching, and Entrepreneurship	Current and past experiences with leadership, teaching, or industry
	Research	Current and past experiences with research
	Training	Past areas of study in research and/or medicine

Payment and Reimbursement	Comparing and contrasting single-payer and multi-payer healthcare funding models	Observations or personal experiences in accountable care organizations or nationalized health systems, and how those opportunities and services compare with most health systems in the US
	Evaluating the role of genetics in the investment and distribution of healthcare funds	Discussions about the ethics and utility of spending money on genetic services and research within the US healthcare system
	Funding research and clinical testing through outside organizations and companies	Observations of non-government organizations (pharma companies, philanthropy, etc.) funding genetics research
	Negotiating costs from integrated clinical research	Discussions about who is or should be responsible for different clinical or research-related genetic costs in a healthcare system
Returning Results	Clinical regulations for returning research results	Current clinical standards for returning genetic results to patients in a clinical context
	Deciding what types of results to return to patients, and when	Clinician and researcher considerations for returning genetic results to patients/research participants
	Managing patient expectations and understanding before and after testing	Observations about how patients typically react to receiving genetic testing results, and those reactions can be managed with pre and post-return of results counseling
Roles	Comparing and contrasting the duties and motivations of researchers and clinicians	Understandings about how researchers and clinicians have different stakes in the research and clinical processes, and how those motivations should influence types of involvement
	Navigating the medical system as both a patient and a research participant	Hopes and concerns about the overlap between patient and research participant roles in learning health systems
	Negotiating the roles of medical geneticists, genetic counselors, and non-genetics providers	Debates about how involved each type of clinician should be in the genetic medicine process (test ordering, interpretation, return of results)
Utility	Adopting genetics into other medical domains	Arguments for integrating genomics more broadly into medicine, rather than having it remain its own specialty
	Understanding personal utility of genetics for patients	Reasons why patients themselves might want to get genetic testing done, clinical or otherwise
	Using genetics vs. other medical tests or interventions	Decisions that go into ordering genetic tests or using genetic interventions as opposed to other

		"standard" tests or interventions in medicine
	Visualizing the best uses for genomics in medicine	Hopes for the future of genomics in medicine, and predictions of how it can be used to maximally benefit individual and population health
Variant Actionability and Validity	Determining variant actionability and utility in the clinic and clinical labs	Current processes for interpreting genetic variants given other clinical information
	Generating, collecting, and applying evidence for variant interpretation	Processes and needs for collecting information that can be used for variant interpretation and actionability assessments
	Standardizing and curating variant interpretations	How external bodies and resources like ACMG, ClinGen, and ClinVar contribute to variant interpretation and actionability assessments
	Weighing analytic validity, gene-disease validity, and utility	Complexities associated with determining clinical validity, analytic validity, actionability, utility, etc. of genetic variants

Appendix F. Axial code descriptions for iteration 4 of codebook development.

Code Group	Code	Definition
Building a collaborative learning culture in medical systems	Benefits and drawbacks of using EHR data for research and equitably representing diverse populations	Benefits, challenges, and implications of collecting and using routine clinical data and genomic data for research, and how that data may or may not be representative of the populations that should benefit from that research
	Benefits, drawbacks, and realities of operating within integrated and universalized healthcare systems	Observations or personal experiences with doing clinical research and/or clinical care in integrated US healthcare system (like Kaiser, Geisinger, Mayo, the VA), or in countries with universalized healthcare systems
	Challenges of operating within a stressed and fragmented US healthcare system	Personal experiences with doing clinical research and/or clinical care under the typical US healthcare model, which is generally disconnected and under resourced
	Forming collaborations and support systems within and between healthcare systems	Examples and observations of healthcare providers, researchers, and leadership working together (or not) to conduct genomic medicine and/or research
	Negotiating the roles of medical geneticists, genetic counselors, and non-genetics providers	Discussions of who should or could be ordering/interpreting genetic tests among genetic specialists (GCs and geneticists) and non-genetic specialists (neurologists, oncologists, cardiologists, etc.)
	Paying for clinical sequencing and clinical research	Discussions of who (healthcare systems, insurers, federal/state governments, commercial entities, patients) should be paying for different types of clinical genetic research or care, and observations of what types of funding models currently exist in genetics
	Sharing and recycling clinical and genomic data	Benefits and challenges of sharing participant-level genomic and clinical data within and between institutions
	What are the differences (if any) between research, clinical care, and quality improvement?	Discussions of how research and clinical care overlap and/or diverge, and how routine quality improvement might be distinct from both
Building relationships	Building trust with patients, especially from minority communities	How researchers and healthcare providers can respectfully engage with

with patients/research participants		patients/research participants, especially from backgrounds that have been historically disadvantaged in medicine and/or genetics research
	Communicating with patients about research/clinical distinctions and navigating provider/researcher differences	How researchers and clinicians do, can, or should help research participants/patients navigate the research-clinical boundary, including clarifying the roles of researchers vs. providers
	Engaging patients in the research process and being sensitive to their needs and motivations	Discussions of how involved patients should be during the research process, particularly for receiving preliminary research results or bringing their own third-party data (e.g. from 23andme) to the table. This code also addresses why people might be interested in genetic testing in the first place, and how they can or can't access genetic medicine resources
	Providing incentives or clinical benefits to patients for participating in research	Discussions of whether people should receive monetary or healthcare incentives or compensation for participating in clinical research, or if they should be participating in research altruistically (or a mix of both, depending on the situation)
Ensuring patient/research participant safety and wellbeing	Determining variant actionability, utility, and returnability in the clinic and clinical labs	Current clinical processes for deeming genetic variants clinically actionable (e.g., through a CLIA lab), and what criteria are or should be used to determine if a variant is clinically actionable (e.g. it could impact their care in a meaningful way) and/or should be returned to a patient
	Educating non-genetics providers about genetic medicine to prevent misuse and misinterpretation	Observations of how genetic medicine is currently misused by healthcare providers, and strategies of training and aiding providers to prevent misuse from happening
	Ensuring appropriate clinical follow-up after genetic testing	Considerations for what clinical follow up is needed after genetic testing
	Generating, collecting, and applying evidence for variant interpretation	Discussion of current authoritative bodies that develop variant interpretation standards (e.g., ClinGen, ACMG), and how accumulated evidence of variant pathogenicity can and should be used to aid variant interpretation
	Turning new genetic associations and technologies into clinical interventions	Benefits, challenges, and safety considerations for "fast tracking" potentially

		actionable genetic variants and tools into clinical use, either using standard clinical trial methods or other implementation models
Evaluating the role of genetics in medicine	Considerations for using population-wide genetic screening in clinical care	Pros and cons of doing routine, population-wide genomic screening
	Deciding what types of genetics tests to order based on clinical indications	Current practices in ordering genetic tests for specific indications (e.g., developmental delay, family history), and considerations of whether broader (e.g. exome) or narrow (e.g. targeted panel) tests should be ordered in different clinical situations
	Historical advancements in genomic research and technology	Ways that genomic research and genomic medicine have progressed over the past ~50 years, and how those advancements have impacted other scientific discoveries and developments
	Understanding genetic impacts on health and disease	Discussions of how much we do or don't know about how genetics impacts human health and disease, and why that knowledge is important for science and for healthcare in general
	Using the EHR to represent genomic data and streamline clinical genomics	Examples of genomics CDS in EHRs (e.g., through the Epic genomics module), and current challenges with getting genetics data into and out of the EHR
	Visualizing the best (and worst) uses for genomics in medicine going forward	Considerations of trade-offs between genetic testing and other medical tests, and predictions of the best uses for genomics in advancing science and population health
Participant Background	Types of patients they see or environments they do clinical work in	The participant's typical patient populations (e.g., adults, pediatrics, oncology, OBGYN), and where/how they used to or currently work (e.g. institution name, institution type, position)
	Types of research they are or were involved in	Past and current areas of research (e.g., data science, family communication, implementation science), and how they split time between research and clinical care
	Where they trained, in what, and for how long	Institution names, types of degrees, lengths of degrees, people they trained with, reasons for choosing certain career paths, etc.

Protecting patient/research participant rights to privacy and autonomy	Challenges and strategies for ethical oversight and consent in clinical research	Benefits and challenges of different consent models (e.g., broad consent, dynamic consent) for merging research and clinical care, and experiences working with IRBs to do clinical research
	Protecting the privacy and security of clinical data	Considerations for protecting the privacy and security of clinical and genetic data that is used for research in clinical settings

SUPPLEMENTAL FIGURES

Figure S3.1. CSER projects, site populations and sequencing modalities.

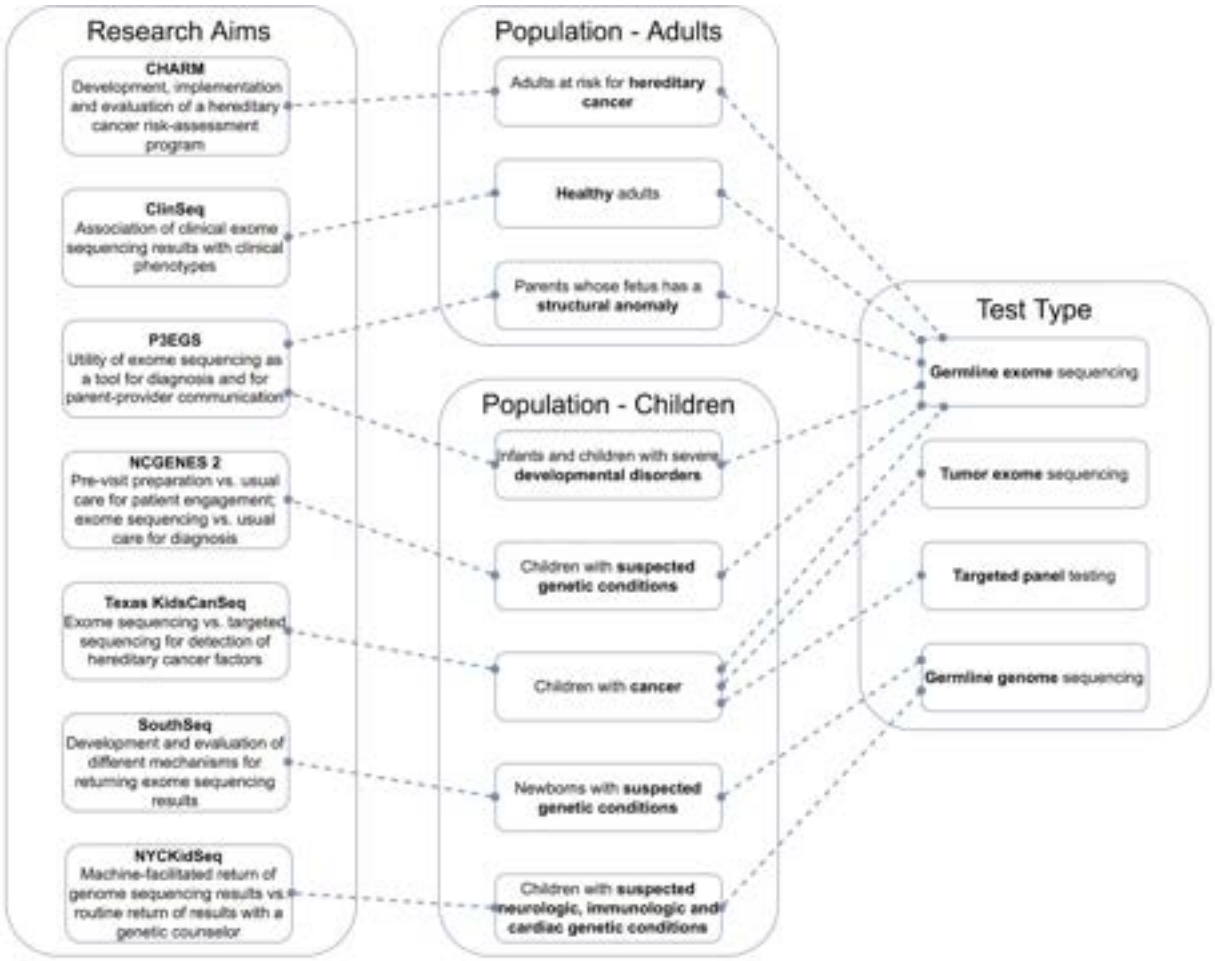


Figure S3.2. Survey administration timepoints for CSER harmonized survey measures.

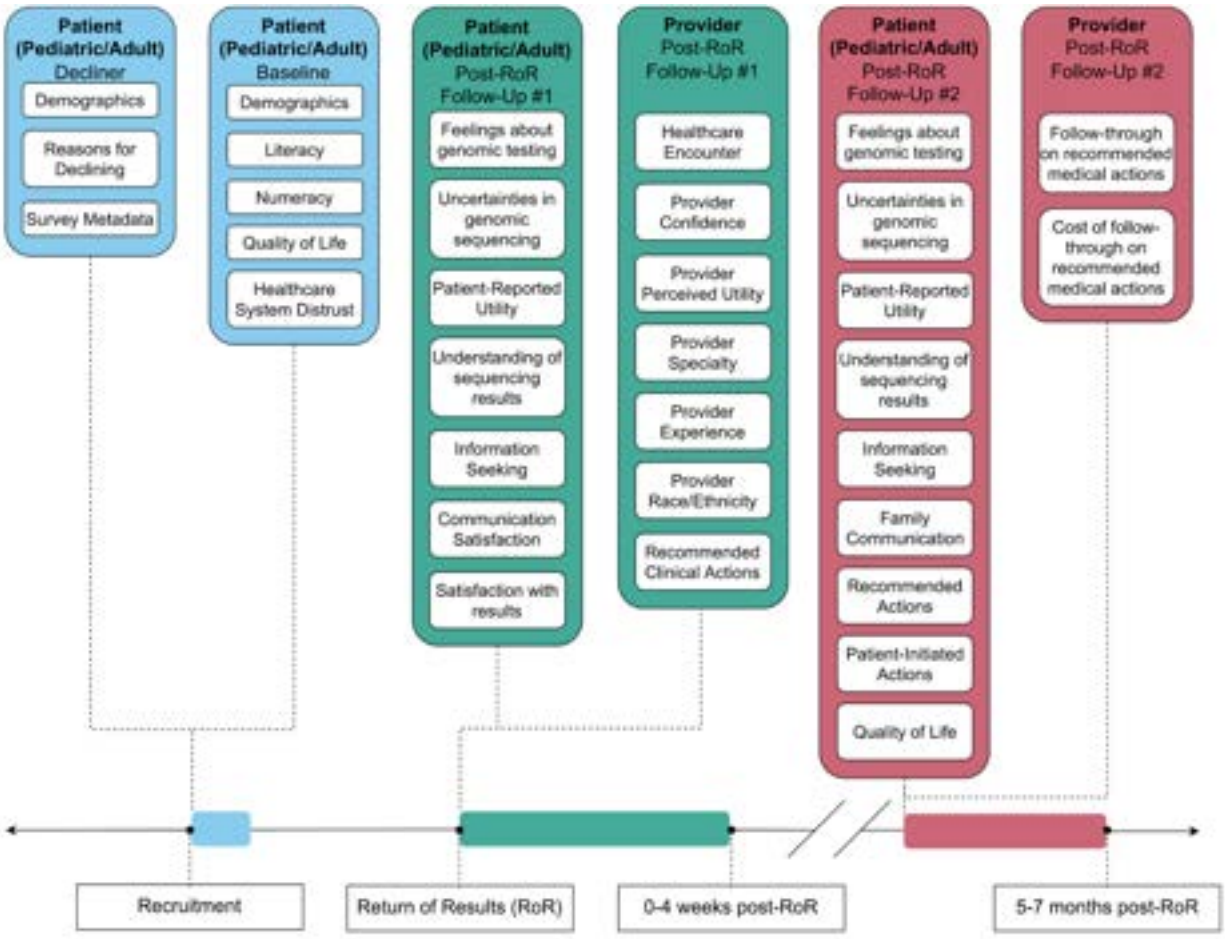


Figure S3.3. Reporting timepoints for genomic sequencing results, both at the participant level and at the case level.

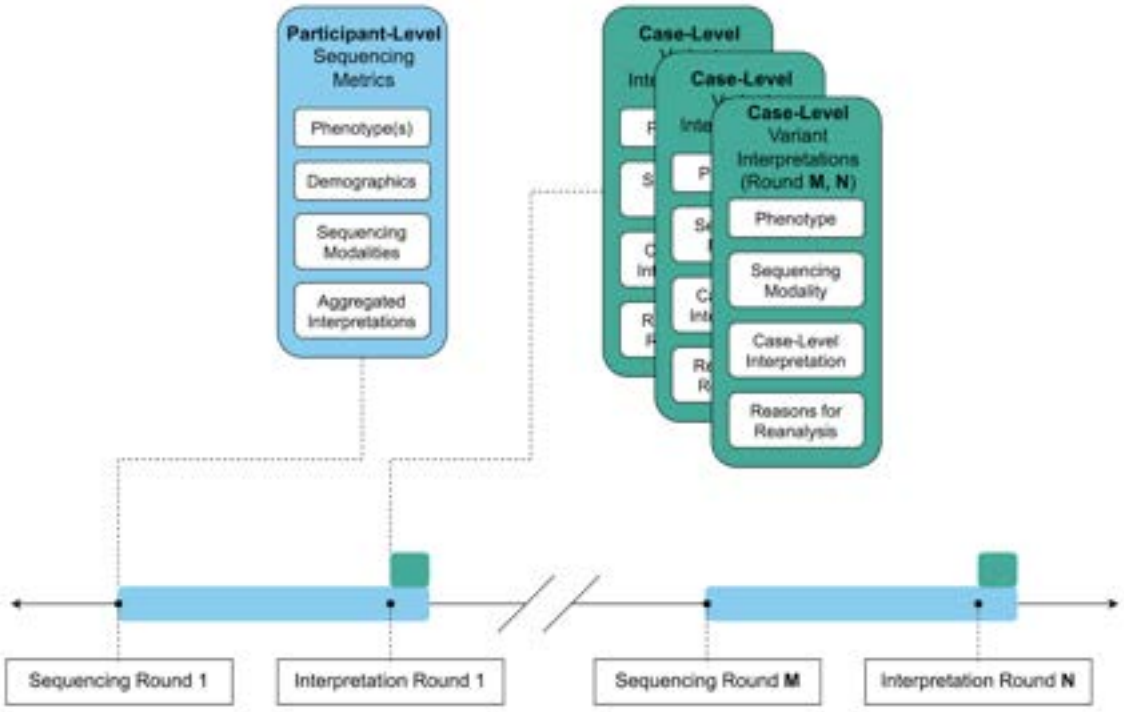


Figure S3.4. Timeline of the harmonized measure change proposal process and implementation of the post-Return of Results (RoR) to follow-up survey elapsed time variables.

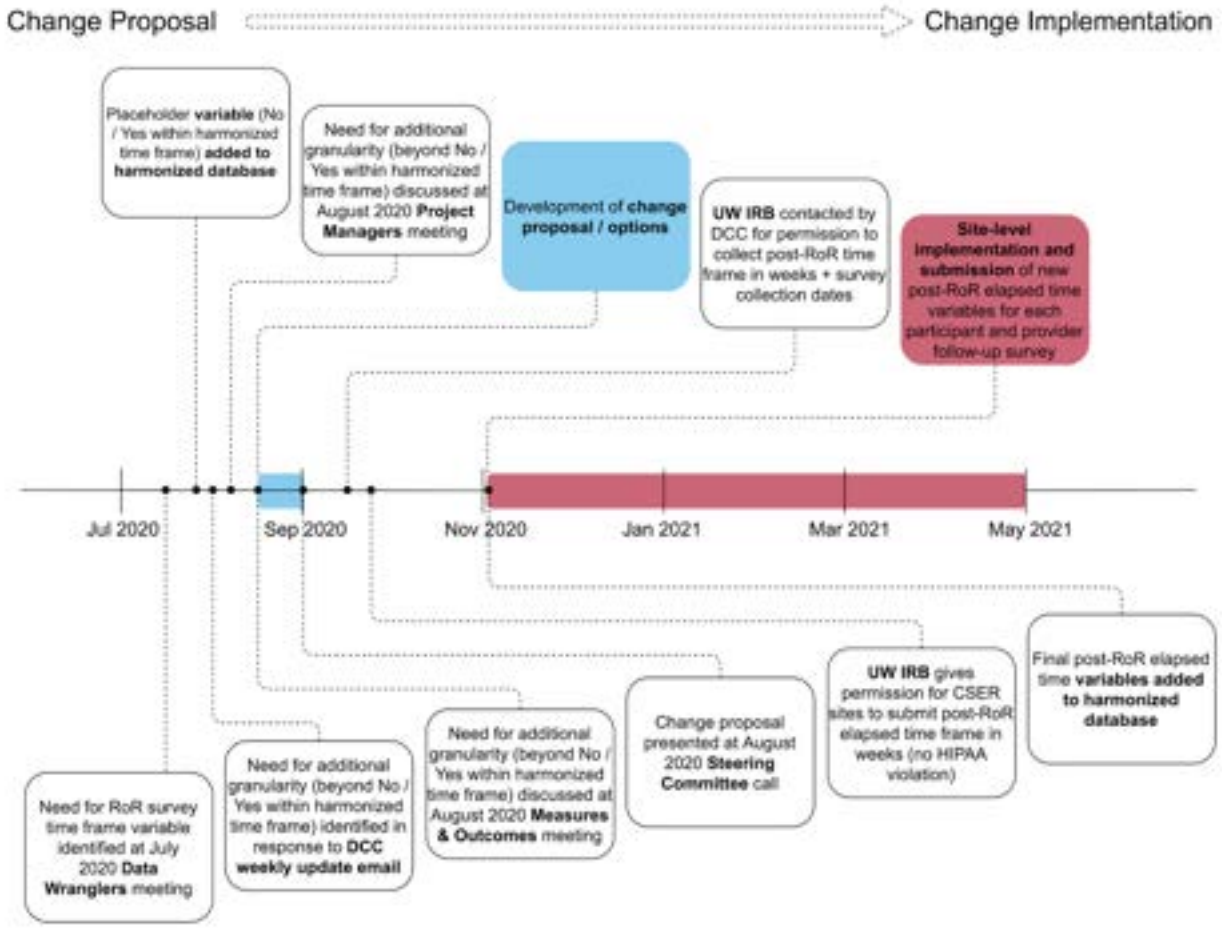


Figure S3.5. Data upload interface on the CSER Data Hub website.

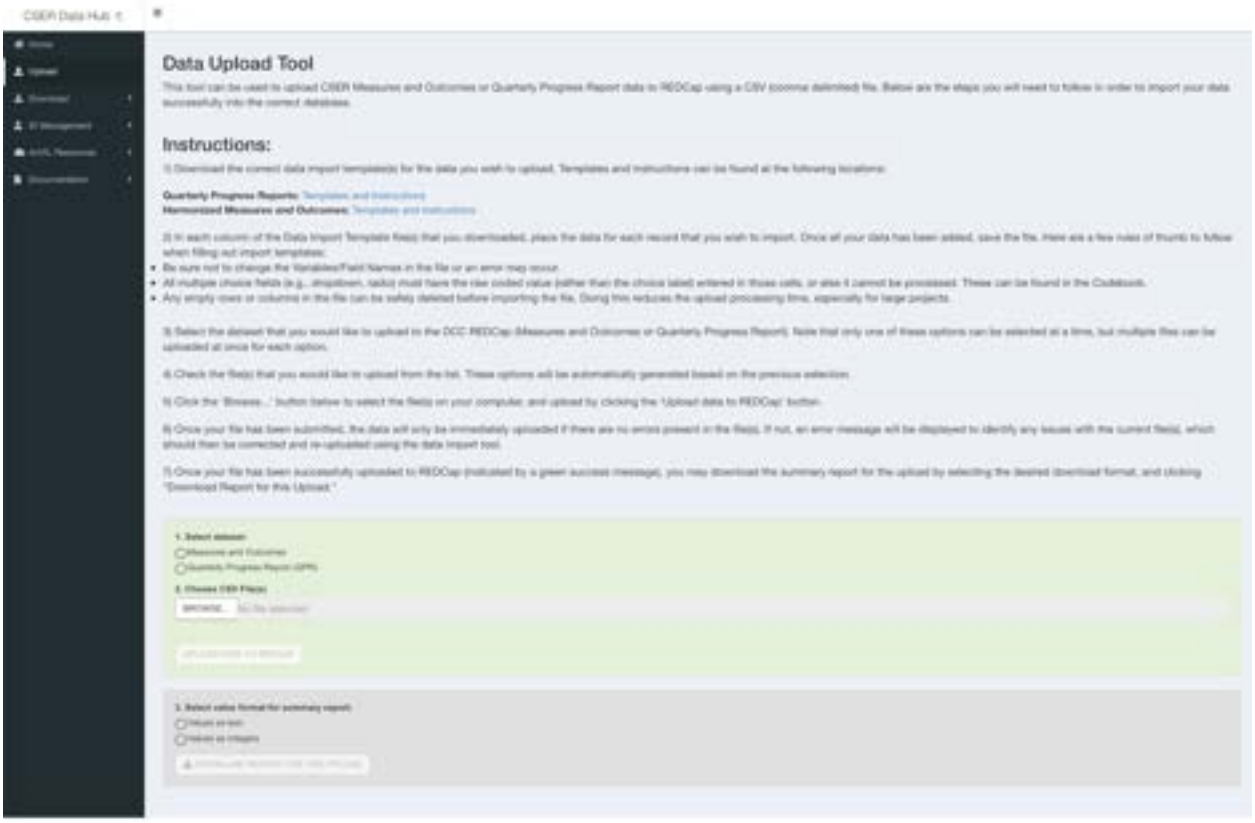


Figure S3.6. Multi-site harmonized data download interface on the CSER Data Hub website.

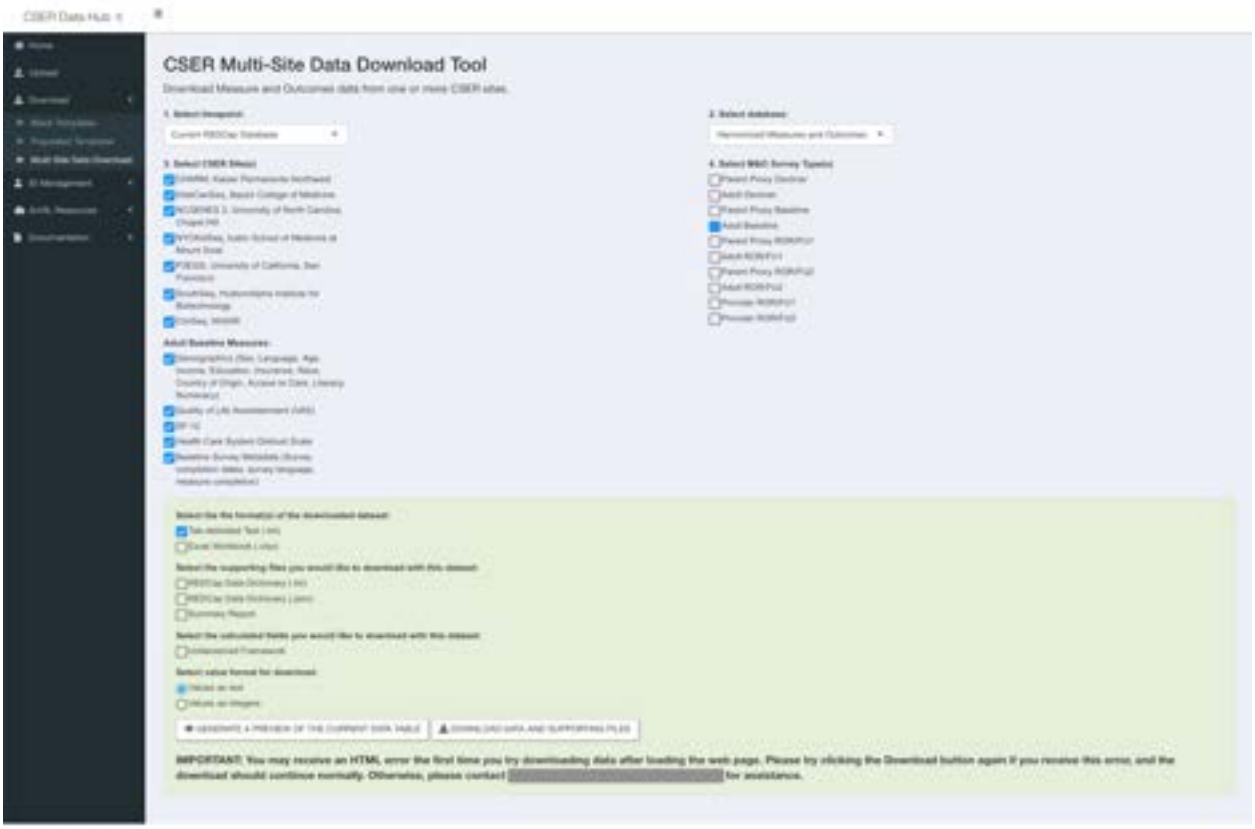


Figure S3.7. CSER ID management interface on the CSER Data Hub website.

Figure S3.8. Sequence data upload instructions on the CSER Data Hub website.

Figure S3.9. Change log documentation on the CSER Data Hub website.

Figure S3.10. Reference sheet for Baseline Measures in the CSER cross-site Adaptation Dictionary.

BASELINE						
	CHARM	Baylor	UCSF	HUDSONALPHA	Sinai/NYCKidSeq	NCGENES/UNC
Gender	A	A	A	D	SA	A
DOB						
Age year					SA	
Language	A	A	A	?	SA	A
Income	B	B			SA	
Education	B	B	B		SA	
Insurance	SA	SA	SA	SA	SA	
Country of origin	B	D?		D	A	D
Access	A			D	SA	
Literacy		A			SA	
Numeracy			D		SA	
Race/Ethnicity			D		SA	
R/E parent 1	NA		A			
R/E/parent 2	NA	D	A			D
Zip						
VAS			A	A		SA
SF-12		NA	NA	NA	NA	NA
PEDSQL	NA	D	D	D	SA	

Legend		Notes
Identical to harmonized item and response scale =		No changes to harmonized items or response scale
Identical to harmonized item and response scale. Brief version used =	B	Collapsed/Brief harmonized version used for this item
Slight Adaptation=	SA	Slight change in question, question format or response scale. Ex: changed formatting, slight changes to wording of question or response scale
Adaptation=	A	Significant change in question, question format or response scale. Ex: dropped or changed items, changed responses
Dropped=	D	Whole survey dropped
Removed from data base=		Items were removed due to changes in privacy policy
Scale not applicable to study population=		Ex: Adult scale in pediatric population

Figure S5.1. Quantile-Quantile (Q-Q) plot for logistic regression analysis in the European ancestry sample (n=15,458). Expected *P*-values from a theoretical $\chi 2$-distribution are plotted on the X-axis and observed *P*-values for each SNV in the logistic regression model are plotted on the Y-axis. The red line represents the null hypothesis that the theoretical and observed *P*-values correspond with one another.

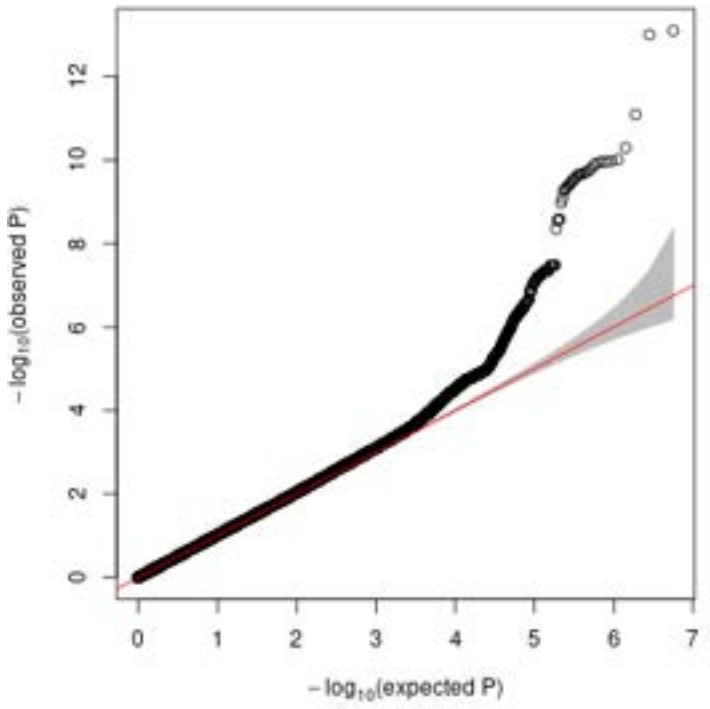


Figure S5.2. Manhattan plot of P-values generated using logistic regression analysis in the joint ancestry sample (n=19,861). An additive model was used to assess the disease susceptibility impact of the minor (coded) allele at each position, while controlling for age, BMI, sex, ancestry, nursing home status, chemotherapy, diabetes, HIV, transplant medications, corticosteroids, and medium or high-risk antibiotic exposure as covariates. Genomic coordinates are displayed along the X-axis, and the negative logarithm of logistic regression P-values are displayed on the Y-axis. Each dot represents a SNV in the regression model, with associated P-values plotted accordingly, while the diamond represents the most significantly associated SNV. The dotted line represents the negative logarithm of the genome-wide significance threshold ($P < 5 \times 10^{-8}$). Colors are used to distinguish between SNVs in adjacent chromosomes.

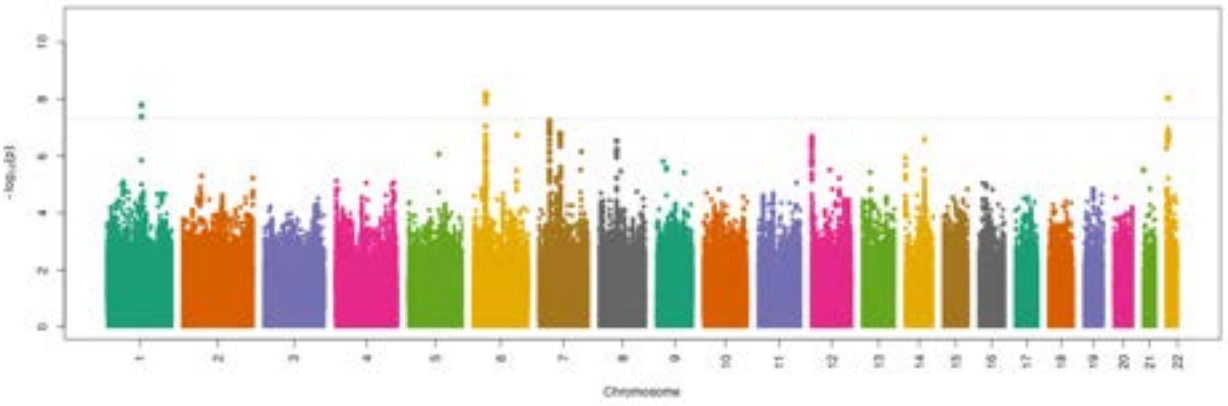


Figure S5.3. Q-Q plot for logistic regression analysis in the joint ancestry sample (n=19,861). Expected *P*-values from a theoretical χ2-distribution are plotted on the X-axis and observed *P*-values for each SNV in the logistic regression model are plotted on the Y-axis. The red line represents the null hypothesis that the theoretical and observed *P*-values correspond with one another.

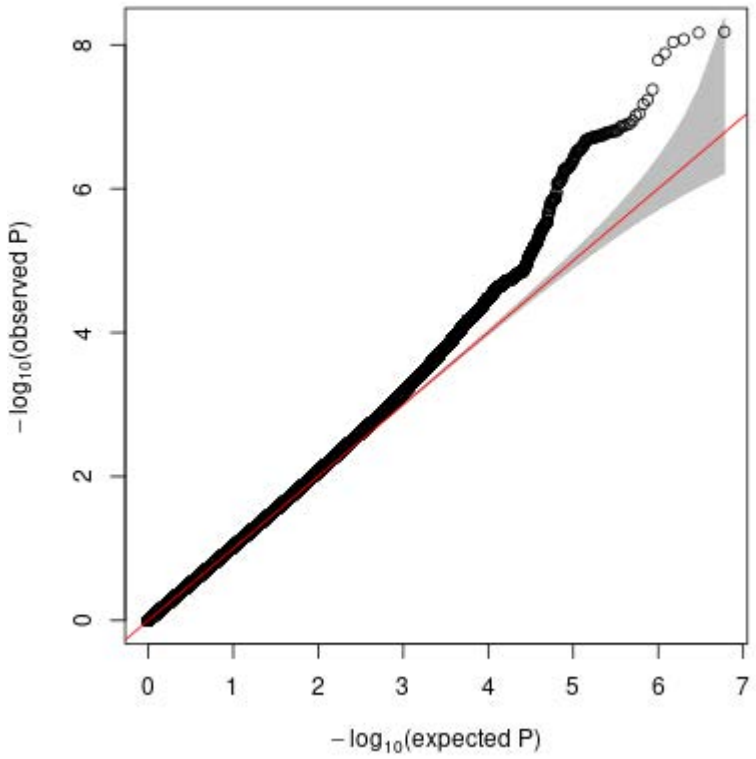


Figure S5.4. Manhattan plot of *P*-values generated using logistic regression analysis in the African ancestry sample (n=4,084). An additive model was used to assess the disease susceptibility impact of the minor (coded) allele at each position, while controlling for age, BMI, sex, ancestry, nursing home status, chemotherapy, diabetes, HIV, transplant medications, corticosteroids, and medium or high-risk antibiotic exposure as covariates. Genomic coordinates are displayed along the X-axis, and the negative logarithm of logistic regression *P*-values are displayed on the Y-axis. Each dot represents a SNV in the regression model, with associated *P*-values plotted accordingly. The dotted line represents the negative logarithm of the genome-wide significance threshold ($P < 5 \times 10^{-8}$). Colors are used to distinguish between SNVs in adjacent chromosomes.

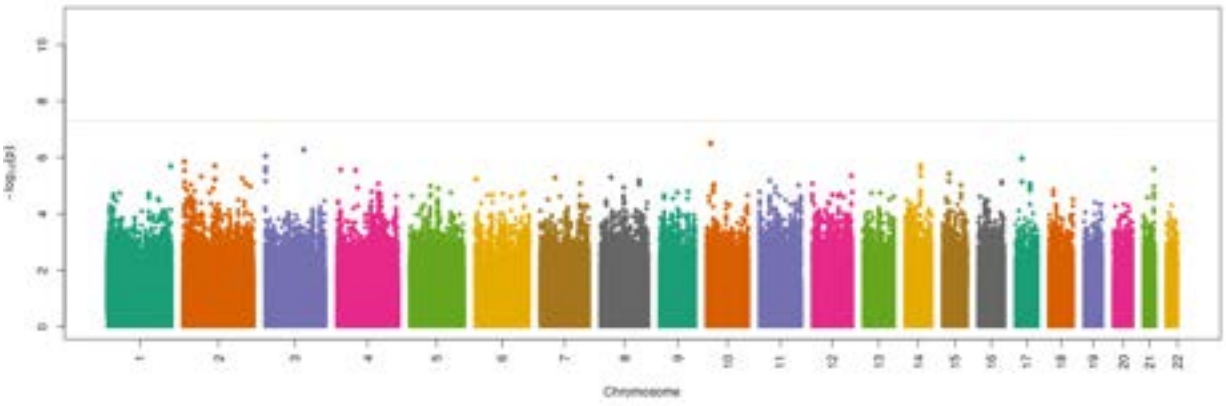


Figure S5.5. Q-Q plot for logistic regression analysis in the African ancestry sample (n=4,084). Expected *P*-values from a theoretical $\chi 2$-distribution are plotted on the X-axis and observed *P*-values for each SNV in the logistic regression model are plotted on the Y-axis. The red line represents the null hypothesis that the theoretical and observed *P*-values correspond with one another.

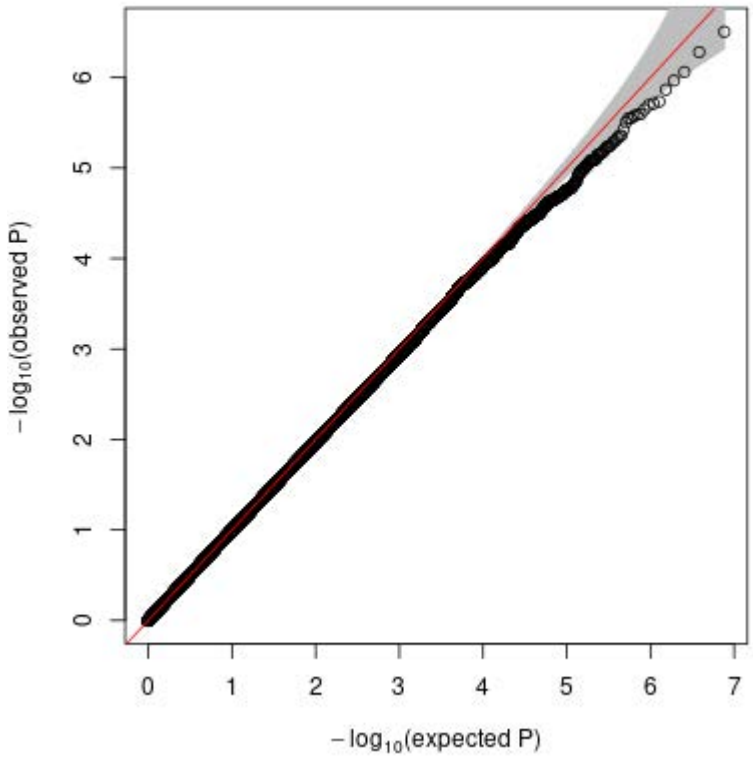


Figure S5.6. Regional LD plot of SNVs evaluated in the African-ancestry logistic regression analysis, using the African 1000 Genomes superpopulation as a reference group. Genomic coordinates spanning the HLA-DRB region and surrounding genes are shown on the X-axis in both subplots. Negative logarithms of *P*-values from the African-ancestry logistic regression analysis are shown on the Y-axis in the upper sublot, and annotated gene transcripts are distributed along the Y-axis in the lower subplot. Each dot represents a SNV in the regression model, with associated *P*-values plotted accordingly. SNVs in high LD with reference to the index SNV (rs68148149) are colored in red. The LD plot was generated with the LocusZoom [207] tool using default parameters and the 1000 Genomes Project 2014 AFR reference panel.

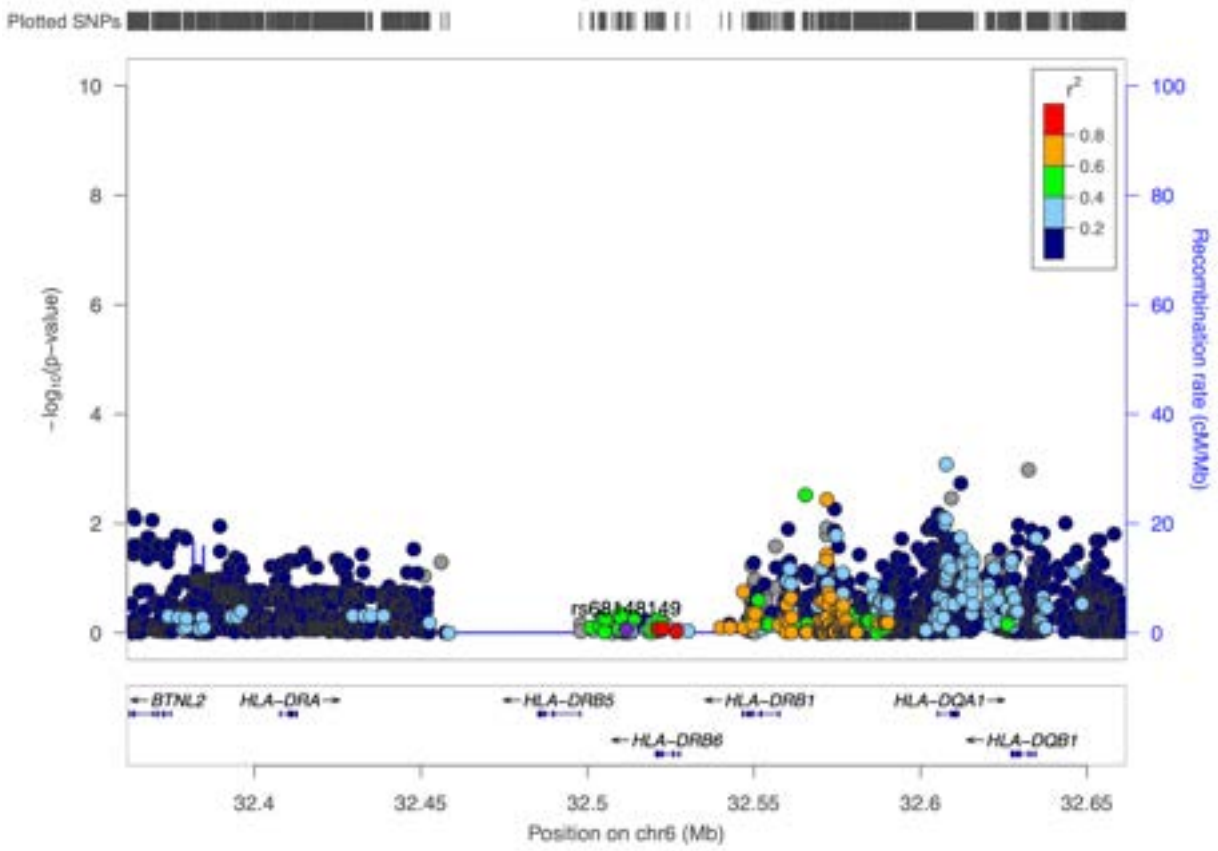


Figure S5.7. Manhattan plot of *P*-values generated using logistic regression analysis in the European ancestry sample (n=15,458), controlling for the index SNV identified in the joint and European-ancestry genome-wide logistic regression analyses (rs68148149). An additive model was used to assess the disease susceptibility impact of the minor (coded) allele at each position, while controlling for the index SNV, age, BMI, sex, ancestry, nursing home status, chemotherapy, diabetes, HIV, transplant medications, corticosteroids, and medium or high-risk antibiotic exposure as covariates. Genomic coordinates are displayed along the X-axis, and the negative logarithm of logistic regression *P*-values are displayed on the Y-axis. Each dot represents a SNV in the regression model, with associated *P*-values plotted accordingly. The dotted line represents the negative logarithm of the genome-wide significance threshold ($P < 5 \times 10^{-8}$). Colors are used to distinguish between SNVs in adjacent chromosomes.

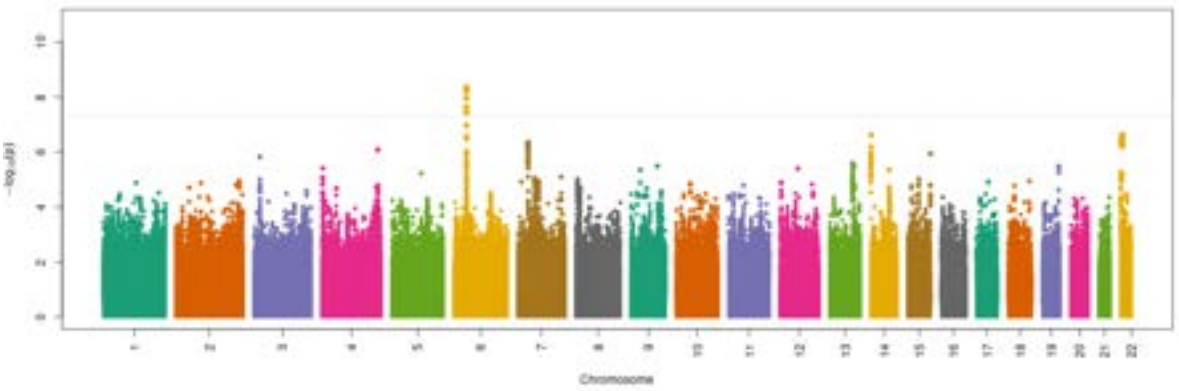


Figure S5.8. Q-Q plot for logistic regression analysis in the European ancestry sample (n=15,458), controlling for the index SNV identified in the joint and European-ancestry genome-wide logistic regression analyses (rs68148149). Expected *P*-values from a theoretical $\chi 2$-distribution are plotted on the X-axis and observed *P*-values for each SNV in the logistic regression model are plotted on the Y-axis. The red line represents the null hypothesis that the theoretical and observed *P*-values correspond with one another.

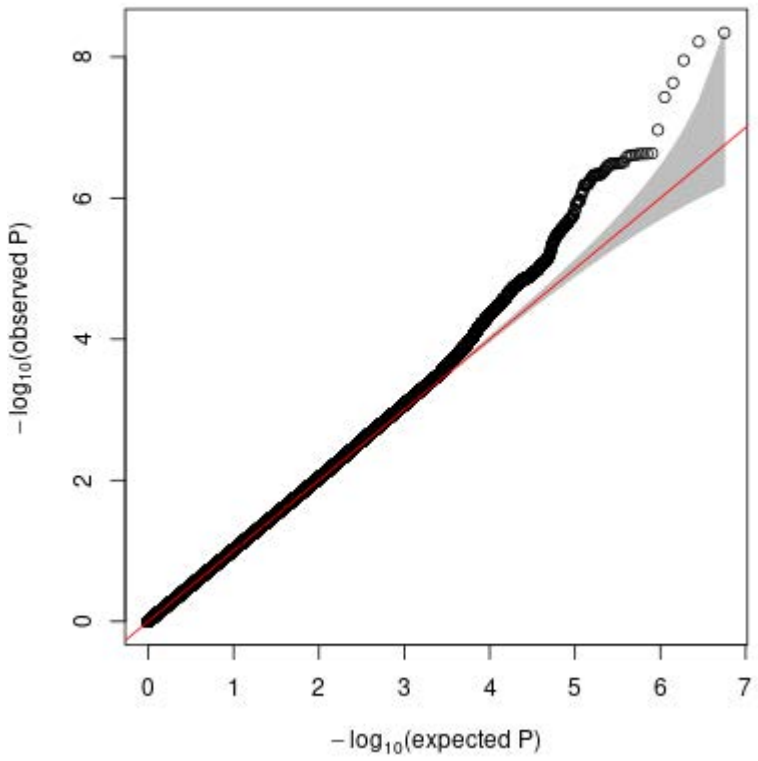


Figure S5.9. Regional Manhattan plot of *P*-values generated using logistic regression analysis of SNVs in the chr6:32400001-32600000 region for 4 participant groups: participants with >=1 copies of the DR51, 52 or 53 haplotype (top left, n=14,291), participants with >=1 copies of the DR51 haplotype (top right, n=4,130), participants with >=1 copies of the DR52 haplotype (bottom left, n=8,887), and participants with >=1 copies of DR53 haplotype (bottom right, n=7,863). An additive model was used to assess the disease susceptibility impact of the minor (coded) allele at each position within each participant group, while controlling for age, BMI, sex, ancestry, nursing home status, chemotherapy, diabetes, HIV, transplant medications, corticosteroids, and medium or high-risk antibiotic exposure as covariates. Genomic coordinates are displayed along the X-axis, and the negative logarithm of logistic regression *P*-values are displayed on the Y-axis of each plot. Each dot represents a SNV in the regression model, with associated *P*-values plotted accordingly. The red line in each plot represents the negative logarithm of the genome-wide significance threshold ($P < 5 \times 10^{-8}$), and the blue line represents a suggestive genome-wide significance threshold ($P < 5 \times 10^{-6}$). Significantly associated SNVs from **Table 5.2** are colored in green.

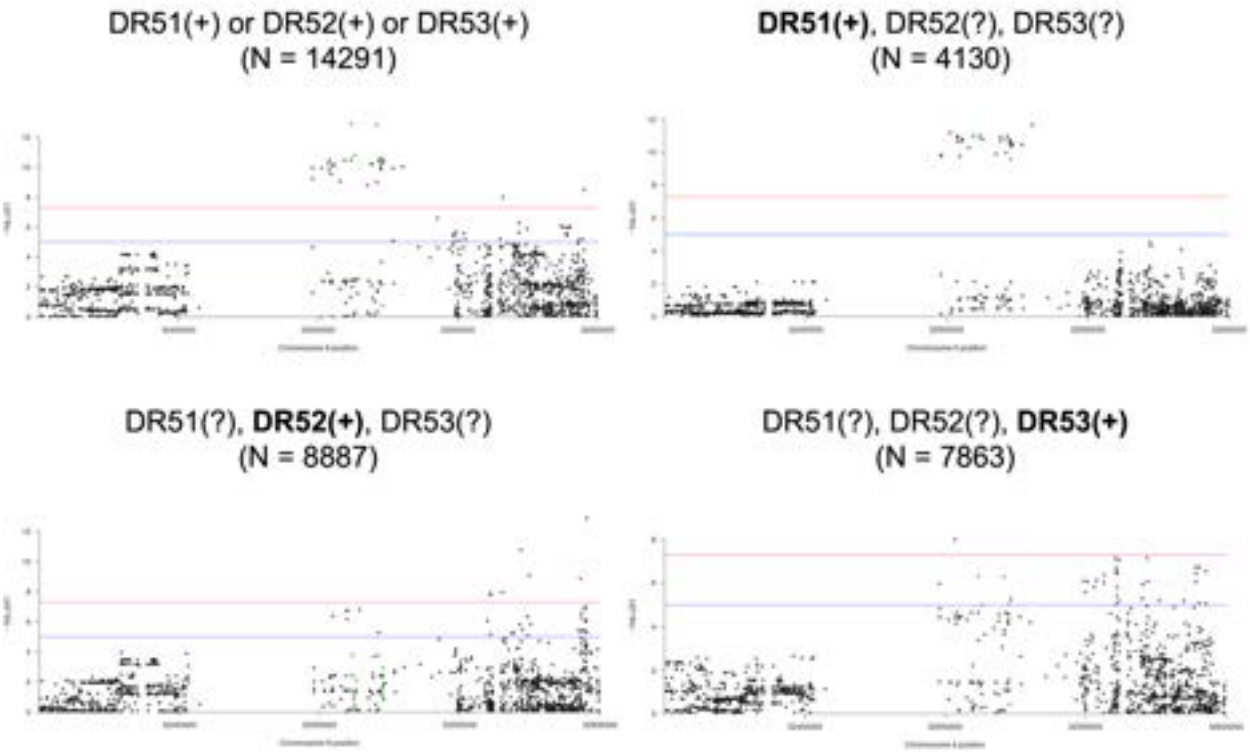


Figure S5.10. Flowchart of regional Manhattan plots of *P*-values generated using logistic regression analysis of SNVs in the chr6:32400001-32600000 region, categorized by the following haplotype subsamples: DR51(+) (n=4,130), DR15(+) (n=3,791), DR16(+) (n=381), and DRB1*15:01(+) (n=3,608). An additive model was used to assess the disease susceptibility impact of the minor (coded) allele at each position within each participant group, while controlling for age, BMI, sex, ancestry, nursing home status, chemotherapy, diabetes, HIV, transplant medications, corticosteroids, and medium or high-risk antibiotic exposure as covariates. Genomic coordinates are displayed along the X-axis, and the negative logarithm of logistic regression *P*-values are displayed on the Y-axis of each plot. Each dot represents a SNV in the regression model, with associated *P*-values plotted accordingly. The red line in each plot represents the negative logarithm of the genome-wide significance threshold ($P < 5 \times 10^{-8}$), and the blue line represents a suggestive genome-wide significance threshold ($P < 5 \times 10^{-6}$). Significantly associated SNVs from **Table 5.2** are colored in green.

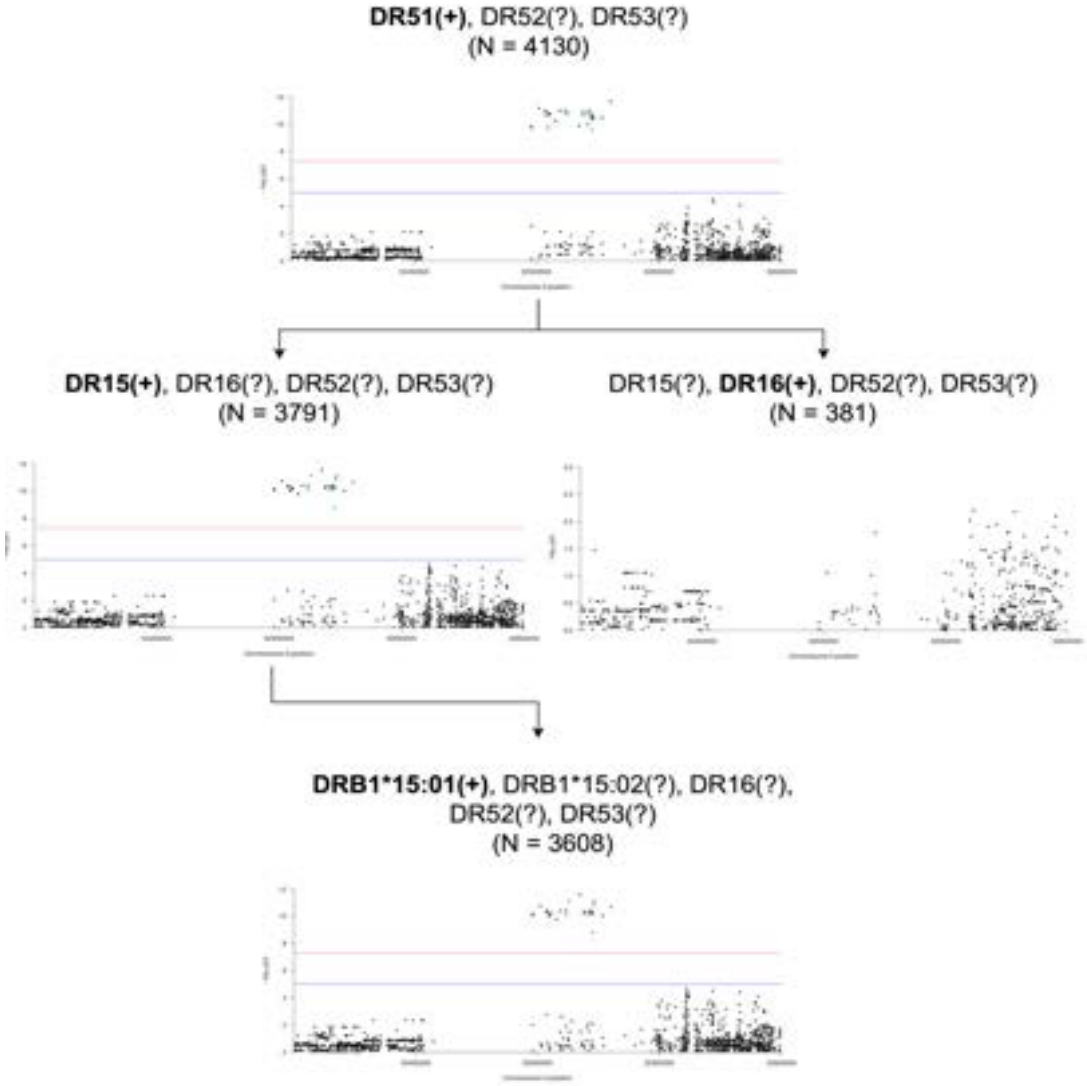


Figure S5.11. Flowchart of coding allele frequencies (CAFs) of the index SNV identified in the joint and European-ancestry genome-wide logistic regression analyses (rs68148149) in different HLA-DR haplotype-enriched groups (DR51, DR52, and DR53).

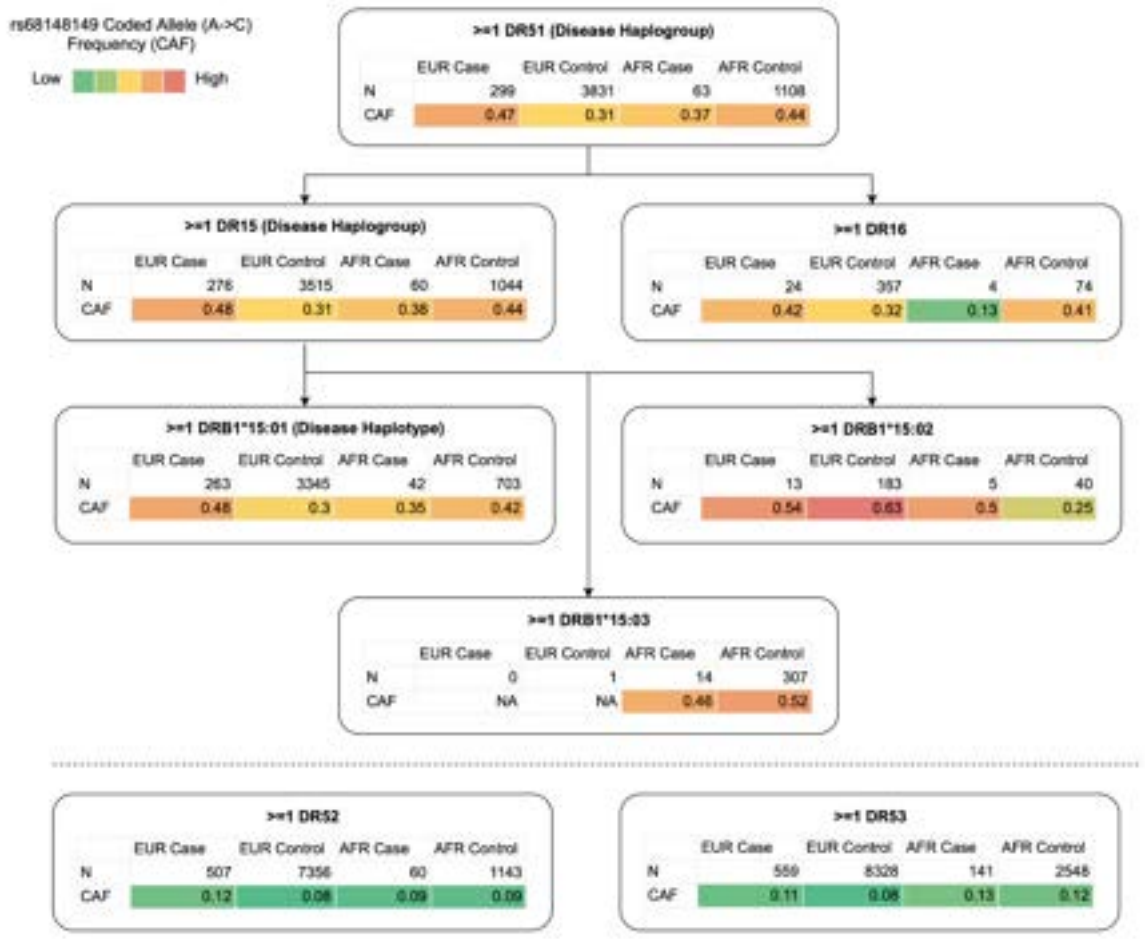
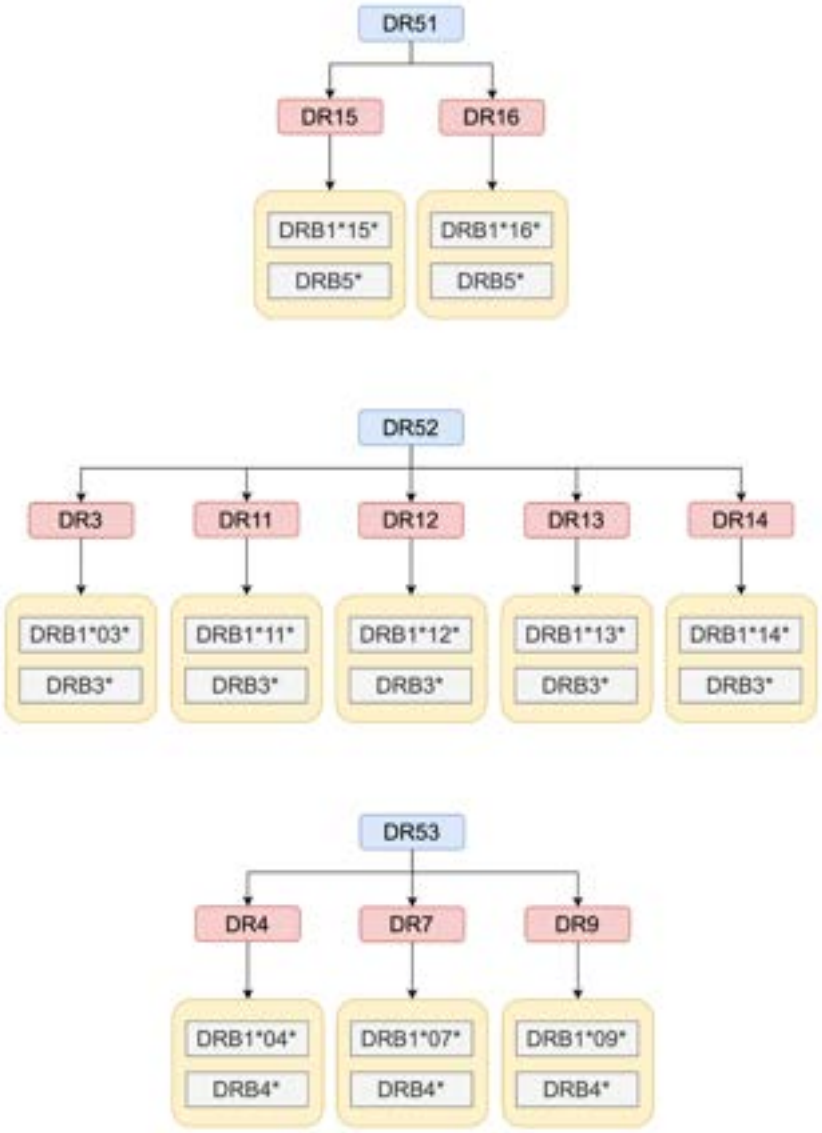


Figure S5.12. Relational flowchart of the *HLA-DRB* haplotypes identified in the eMERGE *C. diff.* cohort.



SUPPLEMENTAL TABLES

Table S3.1. Examples of modifications, additions, and transformations to the harmonized CSER survey measures and outcomes database.

Change Type	Field Type	Details	Rationale
Modification	Multiple fields	Fix typo(s) in variable name	Enhance interpretability and consistency of database
		Fix typo(s) in response scale	Ensure consistency of common data model
		Change radio buttons to checkboxes in response list	Allow sites to report multi-selection responses, even if others collected single responses
	Pediatric age	Replace original harmonized age variable in pediatric survey (truncated year) with two variables: 1) numeric value, traditional rounding, 2) units for numeric value (months or years)	Accurately capture pediatric ages with appropriate granularity, given range of participant ages across CSER sites (newborns to young adults)
	Prenatal status	Add field that distinguishes between pediatric and prenatal probands	Prenatal testing should be distinguished from pediatric testing due to unknown effects on construct validity of harmonized measures
Addition	Pregnancy status	Add field that indicates whether a participant's pregnancy is ongoing or terminated at follow-up	Participant's undergoing prenatal testing and administered different surveys, depending on whether the pregnancy is ongoing or terminated
	Survey/measure completion date	Add field that indicates the completion date for each survey type, or for each measure if surveys were not grouped according to harmonized groupings	Allow for future research on how the COVID-19 pandemic may or may not have impacted follow-up responses
	Elapsed time since return of results (RoR) at time of survey/measure	Add field that indicates the number of weeks post-RoR that a survey or measure was administered	Allow for future use of elapsed time variable as a covariate or exclusionary criterion
	Language (Spanish)	Add field that indicates whether a survey was originally administered in Spanish	Capture impacts of translation/survey language on survey results, if any
	Vital status	Add field that indicates whether the proband is alive, dead, stillborn (pregnancy) or terminated (pregnancy), and at what age the proband died	Allow data analysts to distinguish between loss to follow-up due to death, and loss to follow-up due to other reasons
	Consent group	Add field that indicates which harmonized survey/sequence consent group a participant is in (see "Consent Group Harmonization" for details)	Centrally track how survey and sequence data should be access-controlled in the AnVIL platform
	Patient or parent	Add field that distinguishes between parent and patient records	Centrally track study IDs for parents with sequence samples
	Provider ID	Add field that uniquely identifies a clinical provider within a CSER site	Allow for future analyses of provider characteristics and effects, while avoiding duplicates

Continued on next page

Continued from previous page

Change Type	Field Type	Details	Rationale
	Data collection method	Add field that indicates whether a provider follow-up survey was administered to providers or collected via chart review	Allow for analysis adjustments based on data collection method
Transformation	Zip Code	Replace zip code with rural/urban designation using the Health Resources & Services Administration Federal Office of Rural Health Policy lookup table	Avoid collection of identifiable information

Table S3.2. CSER harmonized sequence and sample metadata model.

Table(s)	Field	Description	Field Type	Priority	Enumerations	Delimiter	Example
Key: Subject and Sample	subject_id	Subject CSER ID	identifier	1	N/A	N/A	"123456789CSER"
Sample	subject_id_local	Subject Local ID (Optional; ONLY provide if sufficiently de-identified)	identifier	3	N/A	N/A	"subject1"
Sample and Sequence	sample_id	Identifier for sample	identifier	1	N/A	N/A	"114443" "Sample_20"
Sample	submitter_id	CSER site name	enumeration	1	P3EGS TexasKidsCanSeq NYCKidSeq SouthSeq NCGENES2 CHARM ClinSeq	N/A	
Sample	dbgap_sample_id	Sample identifier used in dbGaP (if previously submitted)	identifier	1	N/A	N/A	
Sample	sample_source	Tissue type of sample (i.e. melanocytes, keratinocytes, buccal cells, embryonic stem cells)	string	2	N/A	N/A	"basal cell carcinoma" "skin"
Sample	body_site	Collection site of the sample (i.e. skin, breast, peripheral blood, inner oral cavity)	string	1	N/A	N/A	"lymph node"
Sample	tissue_affected_status	If applicable to disease, is the tissue from an affected source or an unaffected source?	boolean	2	Yes No	N/A	
Sample	is_tumor	Is this sample from a tumor?	boolean	1	Yes No	N/A	
Sample	primary_metastatic_tumor	Primary tumor, metastasis, or transformed cell line (if applicable)	enumeration	1	Primary Metastasis Transformed	N/A	
Sample	primary_tumor_location	Primary tumor location (if applicable)	string	1	N/A	N/A	"right-side colon"
Sample	tumor_stage	Tumor stage of sample (if applicable)	string	1	N/A	N/A	"stage IIIA"
Sample	tumor_grade	Tumor grade of sample (if applicable)	integer	1	N/A	N/A	2

Continued on next page

Continued from previous page

Table(s)	Field	Description	Field Type	Priority	Enumerations	Delimiter	Example
Sequence	target_depth	Targeted sequencing depth	integer	1	N/A	N/A	20
Sequence	sequencing_strategy	Paired vs single end (or mate pair)	string	1	N/A	N/A	"paired end"
Sequence	read_length	Sequenced read length (bp)	integer	1	N/A	N/A	100
Sequence	number_of_independent_libraries	Number of independent libraries sequenced	integer	1	N/A	N/A	1

Table S3.3. CSER harmonized consent groups. DUC = dbGaP Data Use Category; DUR = Data Use Restriction; GRU = General Research Use; HMB = Health/medical/biomedical research; IRB = Ethics Approval Required.

Consent Group	Survey DUC	Survey DUR	Sequence DUC	Sequence DUR
1	GRU	None	GRU	None
2	GRU	IRB	GRU	IRB
3	GRU	IRB	N/A	CSER-ONLY
4	GRU	IRB	N/A	N/A
5	N/A	CSER-ONLY	GRU	IRB
6	N/A	CSER-ONLY	N/A	CSER-ONLY
7	N/A	CSER-ONLY	N/A	N/A
8	HMB	None	HMB	None

Table S4.1. Descriptions and quotation examples of axial codes in the "Building a collaborative learning culture in medical systems" semantic domain.

Axial Code	Description	Quotations
Benefits and drawbacks of using EHR data for research and equitably representing diverse populations	Benefits, challenges, and implications of collecting and using routine clinical data and genomic data for research, and how that data may or may not be representative of the populations that should benefit from that research	"Oh, the data is terrible. EHRs are god-awful. So the amount of, how data is structured in an electronic medical record is terrible." (Participant 5) "So you know, unfortunately, I do worry [emphasized] to be quite honest, that certain populations are going to be left behind." (Participant 16)
Benefits, drawbacks, and realities of operating within integrated and universalized healthcare systems	Observations or personal experiences with doing clinical research and/or clinical care in integrated US healthcare system (like Kaiser, Gesinger, Mayo, the VA), or in countries with universalized healthcare systems	"And Kaiser. I mean, I think, you know, Kaiser is a model where they can study some of these things over time, because patients tend to stay in the Kaiser system." (Participant 6) "And Estonia actually has a small genome project. And so they have several, several publications out about what they're doing. And it's, I mean, it's, it's pretty darn cool." (Participant 20)
Challenges of operating within a stressed and fragmented US healthcare system	Personal experiences with doing clinical research and/or clinical care under the typical US healthcare model, which is generally disconnected and under resourced	"But it's not it's not an area that I have enough energy, psychic energy to really be, to think about it really critically. Beyond the system sucks, we got to do something better." "And so the, you know, the nihilists among us are saying, well it's going to require a system collapse and rebuild for this to actually get fixed. But we'll we'll see." (Participant 20)
Forming collaborations and support systems within and between healthcare systems	Examples and observations of healthcare providers, researchers, and leadership working together (or not) to conduct genomic medicine and/or research	"It's part of the, you know, I feel like, you know, part of what happens in healthcare is getting these silos. Yeah, you know, silo the researchers away from the clinicians." (Participant 8) "I think it's, I think it's going to require several people I do think, I mean, certainly the, the clinical person who's a part of it would probably need to have their hand in research, or, or some other research folks kind of associated with it." (Participant 18)
Negotiating the roles of medical geneticists, genetic counselors, and non-genetics providers	Discussions of who should or could be ordering/interpreting genetic tests among genetic specialists (GCs and geneticists) and non-genetic specialists (neurologists, oncologists, cardiologists, etc.)	"Not at all. You can't have genomic specialists doing this, there's not enough, plus there's not enough training prospects for getting enough of us." (Participant 5) "I mean, they're always going to need geneticists to help them interpret it, and the nuances to the result." (Participant 13)
Paying for clinical sequencing and clinical research	Discussions of who (healthcare systems, insurers, federal/state governments, commercial entities, patients) should be paying for different types of clinical genetic research or care, and observations of what types of funding models currently exist in genetics	"So if I want to order a test, it doesn't get approved by an insurance company with pre authorization because it gets approved by our department pathology, which means that I can order anything." (Participant 2) "Well, because, you know, the whole premise of health insurance is that you enroll a bunch of people because you don't know who's likely to get sick and you spread the costs across the people who aren't likely to get sick versus the people who are likely to get sick." (Participant 7)

Continued on next page

Continued from previous page

Axial Code	Description	Quotations
Sharing and recycling clinical and genomic data	Benefits and challenges of sharing participant-level genomic and clinical data within and between institutions	"So, they've now figured out how to make that data available to you, without violating the consents of the original thing. But again, it's a very large effort with a whole bunch of software engineers and geneticists and stuff, that's setting up the infrastructure." (Participant 4) "We are glad to share it now, we just don't have the resources to do so on the scale that that would require." (Participant 10)
What are the differences (if any) between research, clinical care, and quality improvement?	Discussions of how research and clinical care overlap and/or diverge, and how routine quality improvement might be distinct from both	"And the answer is, Well, really, it's both and they're not, they're not separable." (Participant 11) "Yeah, I feel like we're a bit of an odd...I feel like clinical and research is so well integrated where we are for the patients that are doing clinical studies." (Participant 15)

Table S4.2. Descriptions and quotation examples of axial codes in the "Building relationships with patients/research participants" semantic domain.

Axial Code	Description	Quotations
Building trust with patients, especially from minority communities	How researchers and healthcare providers can respectfully engage with patients/research participants, especially from backgrounds that have been historically disadvantaged in medicine and/or genetics research	"I mean, I think you don't know and I think it's very, it would have to be fairly trusting on the family's behalf. You know what I mean to say, Hey, you can have my data and it doesn't really matter what you do with it." (Participant 3) "Right. Right. You do have to you have to get the engagement, for historical, ethical reasons." (Participant 13)
Communicating with patients about research/clinical distinctions and navigating provider/researcher differences	How researchers and clinicians do, can, or should help research participants/patients navigate the research-clinical boundary, including clarifying the roles of researchers vs. providers	"Right, and the jargon is awful [emphasized]. Papers are written in jargon. It's all of those things. I think if we can, if we had a chance of cleaning that up, it would be great." (Participant 9) "And so sometimes, that's a little bit difficult with with just like trying to help families understand the process to publication, and how there are cases where it's taken years to, to move things forward. And and so we try to set those expectations early." (Participant 19)
Engaging patients in the research process and being sensitive to their needs and motivations	Discussions of how involved patients should be during the research process, particularly for receiving preliminary research results or bringing their own third party data (e.g. from 23andme) to the table. This code also addresses why people might be interested in genetic testing in the first place, and how they can or can't access genetic medicine resources	"So the patient, the patient has two roles to fulfill here. One is just to self advocate, as a patient, and to survive the inefficiencies and so on. And then they're also having to deal with the potential issues and benefits that are involved with having given their DNA to the health system." (Participant 2) "Um, I've seen it both ways before. I think that really speaks to the importance of pretest counseling and like, expectation setting." (Participant 19)
Providing incentives or clinical benefits to patients for participating in research	Discussions of whether people should receive monetary or healthcare incentives or compensation for participating in clinical research, or if they should be participating in research altruistically (or a mix of both, depending on the situation)	"Well, It's implicit that everyone is there, and contributing to generalizable knowledge for everyone else's benefit. I mean, that's why they're all there." (Participant 7) "And you're gonna know [emphasized], and knowledge is power. Even if you do nothing with it, it's still your gene, you own it, you know what it is and makes a difference to people." (Participant 11)

Table S4.3. Descriptions and quotation examples of axial codes in the "Ensuring patient/research participant safety and wellbeing" semantic domain.

Axial Code	Description	Quotations
Determining variant actionability, utility, and returnability in the clinic and clinical labs	Current clinical processes for deeming genetic variants clinically actionable (e.g. through a CLIA lab), and what criteria are or should be used to determine if a variant is clinically actionable (e.g. it could impact their care in a meaningful way) and/or should be returned to a patient	"So I could live with the fact that it takes a while to make these, you know, to get to the point where these variants are viewed as as reliable and clinically useful." (Participant 10) "Nevertheless, there's a tremendous role for the art [emphasized] of medicine as well, and using clinician discretion and intuition and educated guesses." (Participant 12)
Educating non-genetics providers about genetic medicine to prevent misuse and misinterpretation	Observations of how genetic medicine is currently misused by healthcare providers, and strategies of training and aiding providers to prevent misuse from happening	"And that's hard, you know, that process is not easy for everybody, even the geneticists, they don't grow up thinking about proteins and that sort of stuff. So, but I think that those are important." (Participant 9) "Yeah, I think we'll never have be able to train enough of us to be clinical geneticists are genetic counselor. So we do have to have some sort of training for other providers." (Participant 13)
Ensuring appropriate clinical follow-up after genetic testing	Considerations for what clinical follow up is needed after genetic testing	"One of the other problems will be determining, you know, what are the downstream implications if it's now relevant to health? Do you have to do the monitoring, you have to do surveillance? Do you have treatment? Do you have a clinical trial? Do you have a swap group?" (Participant 3) "Yeah, if you have a BRCA one, the recommendations are completely different. And your risk is quite, you know, it's, it's out of proportion, right to what you would if you were to be in the general population. So if you get a mammogram age 40 and up, it's not, it's not going to do the trick." (Participant 14)
Generating, collecting, and applying evidence for variant interpretation	Discussion of current authoritative bodies that develop variant interpretation standards (e.g. ClinGen, ACMG), and how accumulated evidence of variant pathogenicity can and should be used to aid variant interpretation	"But in genetics, what you know really, I think that these tests should be sort of investigated and studied and explored initially, in probably, you know, sort of centers of excellence or academic places where there is [emphasized] the requisite expertise to sort of understand and where data is collected." (Participant 6) "And sometimes it's, you know, sometimes it's one off and you have some biochemical data, or you have some, you know, have some functional data that it makes sense." (Participant 9)
Turning new genetic associations and technologies into clinical interventions	Benefits, challenges, and safety considerations for "fast tracking" potentially actionable genetic variants and tools into clinical use, either using standard clinical trial methods or other implementation models	"And maybe that's the maybe that's, you know, I think that the fact that you're not doing quote unquote research, doesn't mean that you still don't need to set ahead of time, what are your expected outcomes?" (Participant 1) "And I guess one of the things you do about it is you try to quickly policy make some kind of, as quickly as possible make some kind of guideline about it." (Participant 4)

Table S4.4. Descriptions and quotation examples of axial codes in the "Evaluating the role of genetics in medicine" semantic domain.

Axial Code	Description	Quotations
Considerations for using population-wide genetic screening in clinical care	Pros and cons of doing routine, population-wide genomic screening	"And what happens in medicine is that you could have a perfect test, but it needs to be thoroughly evaluated by everyone in terms of actually showing that this is a test that should be used in a screening methodology in this population, and that you actually have some sort of tangible outcomes." (Participant 5) "And yeah, that would that would serve people's health. Because then you, you spare them the diagnostic odyssey, the proverbial Odyssey." (Participant 15)
Deciding what types of genetics tests to order based on clinical indications	Current practices in ordering genetic tests for specific indications (e.g. developmental delay, family history), and considerations of whether broader (e.g. exome) or narrow (e.g. targeted panel) tests should be ordered in different clinical situations	"But I, I tend towards the side of, let's be careful about what we test for. Let's be careful about what we test for whether it's because of the clinical indication or what, or if it's because of a population, population based risk." (Participant 1) "So we, you know, we're doing less and less targeted testing, more and more, more panels and more general exomes and genomes. So that's at the clinical interface." (Participant 11)
Historical advancements in genomic research and technology	Ways that genomic research and genomic medicine have progressed over the past 50 years, and how those advancements have impacted other scientific discoveries and developments	"So the ability, of course, the ability to study cancer at the genomic level has also exploded with sequencing. Because, you know, we were really still in the one gene at a time mode, but many of those genes stood out." (Participant 10) "Well people didn't want to put that much money behind this, because you had like five patients a year. Which, which I think is changing quite a bit now, especially given that the the different models for genetic, you know, implications, genetic...how genetics is involved in disease, common disease and rare disease right?" (Participant 14)
Understanding genetic impacts on health and disease	Discussions of how much we do or don't know about how genetics impacts human health and disease, and why that knowledge is important for science and for healthcare in general	"You know, I feel like it's like the beginning of, you know, mapping the world." (Participant 8) "I don't think it's going to happen in my [emphasized] lifetime, but I think should our species survive all the other challenges that await it, that the definition of what it means to be human is going to be redefined in part by lots of technologies including bio technologies." (Participant 12)

Continued on next page

Continued from previous page

Axial Code	Description	Quotations
Using the EHR to represent genomic data and streamline clinical genomics	Examples of genomics CDS in EHRs (e.g. through the Epic genomics module), and current challenges with getting genetics data into and out of the EHR	"So, you know, what you really need to make this work well, is some kind of, you know, sort of turnkey system that takes a lot of the friction out, you know, and where a result comes back automatically to the EHR. And it what is important to allow these kinds of things to happen, is extracted automatically and placed into a place where it's available to everybody." (Participant 10) "The second piece was we encouraged the system to invest in the genomic indicators module that Epic has that allows us to represent variant and gene data as structured data so that we have the ability then to search." (Participant 20)
Visualizing the best (and worst) uses for genomics in medicine going forward	Considerations of trade-offs between genetic testing and other medical tests, and predictions of the best uses for genomics in advancing science and population health	"But I think that I think that medicine will have much more of a genetic component to it. Or genetics will have much more of a medicine component to it." (Participant 9) "A lot of people now are talking about polygenic risk scores, and maybe we could use those more to assess risk and that obviously could help with lifestyle changes and things that you need." (Participant 13)

Table S4.5. Descriptions and quotation examples of axial codes in the "Participant background" semantic domain.

Axial Code	Description	Quotations
Types of patients they see or environments they do clinical work in	The participant's typical patient populations (e.g. adults, pediatrics, oncology, OBGYN), and where/how they used to or currently work (e.g. institution name, institution type, position)	"Yes, so I do both inpatient and outpatient genetic services, and we do all gamut of genetic testing from karyotype, chromosomal microarrays, small panels, single gene testing, large panels, exomes." (Participant 13) "Okay, so I'm a medical geneticist in pediatric dermatology and I trained in pediatrics first, then medical genetics and then dermatology and the majority of my practicing career has been bifurcated into standard pediatric dermatology and the other half has been medical genetics." (Participant 11)
Types of research they are or were involved in	Past and current areas of research (e.g. data science, family communication, implementation science), and how they split time between research and clinical care	"And I've been involved in a number of studies that sort of try to look at this interface of clinical medicine versus research, using clinical data to try to make new gene discoveries as well as giving results back to physicians." (Participant 4) "But my lab now is focused on post, the impact of post-zygotic mutations, not just on cancer, but on congenital malformations or birth defects." (Participant 17)
Where they trained, in what, and for how long	Institution names, types of degrees, lengths of degrees, people they trained with, reasons for choosing certain career paths, etc.	"Then after that, I went out to [city name] and did a combined internal medicine and medical genetics clinical training out there, through [hospital name] in the [institution name] combined genetics training program." (Participant 5) "And so I've had both formal training and informal exposure to change management and quality improvement and all that sort of stuff, which of course, has then morphed more into the research focus of implementation science." (Participant 20)

Table S4.6. Descriptions and quotation examples of axial codes in the "Protecting patient/research participant rights to privacy and autonomy" semantic domain.

Axial Code	Description	Quotations
Challenges and strategies for ethical oversight and consent in clinical research	Benefits and challenges of different consent models (e.g. broad consent, dynamic consent) for merging research and clinical care, and experiences working with IRBs to do clinical research	"But if you're really talking about trying to do a big study, there's certainly evidence that people want to please their doctors, and you need to have some ability for despite you saying, Oh, it's fine if you don't participate, right? If they sense that it's your study and your name's on the consent form and everything, then, you know, they're going to feel a certain amount of coercion." (Participant 4) "It is so [emphasized] time consuming, and, and frustrating. And I think that that, I mean, sometimes you think I'd like to do this project, but I'm just not going to do the IRB. So I'm not going to do the project. It's just, it's just too much work." (Participant 11)
Protecting the privacy and security of clinical data	Considerations for protecting the privacy and security of clinical and genetic data that is used for research in clinical settings	"Um, and so that's what I mean by truly [emphasized] anonymizing. And truly anonymizing obviously has downstream effects, like we couldn't go back and then offer to enroll, even tell them that they had it, or offer to, you know, study, you know, learn something new about what was then a relatively newly described variant." (Participant 4) "And they seem to be able to, at one point, another hack a lot of things, yeah. You can say, Oh, I promise, it's all secure. And I'm like, is it as secure as you can make it? Because that's not 100%." (Participant 17)

Table S5.1. *C. diff.* progress note mentions used by the natural language processing algorithm. The commonly used abbreviation for clostridioides/clostridium is the single letter "c." This is difficult to implement in a word search or dictionary look up and was therefore omitted from the NLP algorithm.

Mentions
difficile colitis
diff colitis
dif colitis
difficile diarrhea
diff diarrhea
dif diarrhea
difficile infection
diff infection
dif infection
difficile enteritis
diff enteritis
dif enteritis

Table S5.2. Class 1 (high risk) and Class 2 (moderate risk) antibiotics, as defined by the eMERGE *C. diff.* phenotyping algorithm [201].

Drug Name	Risk Category	Risk Code
amox	Moderate Risk	2
amoxicillin	Moderate Risk	2
amoxicillin-clavulanate	Moderate Risk	2
amoxil	Moderate Risk	2
ampicillin	Moderate Risk	2
AMPICILLIN / MEROPENEM	Moderate Risk	2
AMPICILLIN SODIUM	Moderate Risk	2
ampicillin-sulbactam	Moderate Risk	2
ancef	Moderate Risk	2
augmentin	Moderate Risk	2
AVALOX	High Risk	1
avelox	High Risk	1
azactam	Moderate Risk	2
azithromycin	Moderate Risk	2
azithromycin : zithromax	Moderate Risk	2
aztreonam	Moderate Risk	2
biaxin	Moderate Risk	2
BIAXIN / PENICILLIN	Moderate Risk	2
BIAXIN XL	Moderate Risk	2
BICILLIN	Moderate Risk	2
ceclor	Moderate Risk	2
cedax	High Risk	1
cefaclor	Moderate Risk	2
CEFADROXIL	Moderate Risk	2
cefazolin	Moderate Risk	2
CEFAZOLIN / CLINDAMYCIN	High Risk	1
cefdinir	High Risk	1
CEFDINIR : OMNICEF	High Risk	1
cefepime	High Risk	1
cefixime	High Risk	1
CEFOTAN	Moderate Risk	2
cefotaxime	High Risk	1
cefotetan	Moderate Risk	2
cefoxitin	High Risk	1
cefpodoxime	High Risk	1
CEFPROZIL	Moderate Risk	2
ceftazidime	High Risk	1
ceftin	Moderate Risk	2
ceftriaxone	High Risk	1
ceftriaxone w/lidocaine	High Risk	1
cefuroxime	Moderate Risk	2
cefuroxime : ceftin	Moderate Risk	2
cefuroxime axetil	Moderate Risk	2
cefuroxime axetil (ceftin)	Moderate Risk	2
cefzil	Moderate Risk	2
cephalexin	Moderate Risk	2
CEPHALEXIN (KEFLEX)	Moderate Risk	2
CEPHALEXIN HCL	Moderate Risk	2
CEPHALOTHIN	Moderate Risk	2
cipro	High Risk	1
CIPRO / LEVOFLOXACIN	High Risk	1
CIPRO XR	High Risk	1
CIPROFLAXACIN	High Risk	1
ciprofloxacin	High Risk	1
ciprofloxacin : cipro	High Risk	1

Continued on next page

Continued from previous page

Drug Name	Risk Category	Risk Code
CIPROFLOXACIN (CIPRO)	High Risk	1
CIPROFLOXIN	High Risk	1
claforan	Moderate Risk	2
clarithromycin	Moderate Risk	2
CLARITHROMYCIN (GENERIC)	Moderate Risk	2
CLARITHROMYCIN / AMIKACIN	Moderate Risk	2
CLAVULANATE (AUGMENTIN)	High Risk	1
cleocin	High Risk	1
cleocin t	High Risk	1
clindamycin	High Risk	1
clindamycin : cleocin	High Risk	1
clindamycin hcl	High Risk	1
CLINDAMYCIN HCL (CLEOCIN)	High Risk	1
CLINDAMYCIN PHOSPHATE	High Risk	1
dicloxacillin	Moderate Risk	2
e-mycin	Moderate Risk	2
ees	Moderate Risk	2
ertapenem	Moderate Risk	2
ERYTHROCIN	Moderate Risk	2
erythromycin	Moderate Risk	2
erythromycin base	Moderate Risk	2
ERYTHROMYCIN ETHYLSUCCINATE	Moderate Risk	2
ERYTHROMYCIN LACTOBIONATE	Moderate Risk	2
erythromycin stearate	Moderate Risk	2
floxacillin	Moderate Risk	2
floxin	High Risk	1
fortaz	High Risk	1
imipenem	Moderate Risk	2
imipenem / cilastatin	Moderate Risk	2
imipenem-cilastatin	Moderate Risk	2
imipenem-cilastatin injection	Moderate Risk	2
invanz	Moderate Risk	2
keflex	Moderate Risk	2
kefzol	Moderate Risk	2
KETEK	Moderate Risk	2
levaquin	High Risk	1
levaquin / ibuprofen	High Risk	1
levaquin leva-pak	High Risk	1
levofloxacin	High Risk	1
levofloxacin : levaquin	High Risk	1
LORABID	Moderate Risk	2
maxipime	High Risk	1
MEFOXIN	High Risk	1
MERONEM	Moderate Risk	2
meropenem	Moderate Risk	2
MEROPENEM : MERREM	Moderate Risk	2
merrem	Moderate Risk	2
methicillin	Moderate Risk	2
moxifloxacin	High Risk	1
nafcillin	Moderate Risk	2
omnicef	High Risk	1
oxacillin	Moderate Risk	2
pen vk	Moderate Risk	2
PEN-VEE K	Moderate Risk	2
PEN-VK	Moderate Risk	2

Continued on next page

Continued from previous page

Drug Name	Risk Category	Risk Code
penicillin	Moderate Risk	2
PENICILLIN G	Moderate Risk	2
PENICILLIN G BENZATHINE	Moderate Risk	2
PENICILLIN G POTASSIUM	Moderate Risk	2
penicillin v potassium	Moderate Risk	2
penicillins	Moderate Risk	2
piperacillin	Moderate Risk	2
piperacillin / tazobactam	Moderate Risk	2
piperacillin-tazobactam	Moderate Risk	2
piperacillin-tazobactam inj	Moderate Risk	2
primaxin	Moderate Risk	2
rocephin	High Risk	1
suprax	Moderate Risk	2
tequin	High Risk	1
ticar	Moderate Risk	2
ticarcillin	Moderate Risk	2
ticarcillin / clavulanate	Moderate Risk	2
timentin	Moderate Risk	2
trimox / amox	Moderate Risk	2
TROVAFLOXACIN	High Risk	1
TROVAN	High Risk	1
ULTRACEF	Moderate Risk	2
unasyn	Moderate Risk	2
vanc / cefepime	High Risk	1
VANC / DORIPENEM	Moderate Risk	2
vanc / rocephin	High Risk	1
vanc / zosyn	Moderate Risk	2
VANCOMYCIN / CEFOTAXIME	High Risk	1
vancomycin / doripenem	Moderate Risk	2
vancomycin / ertapenem	Moderate Risk	2
vantin	High Risk	1
ZARTAN	Moderate Risk	2
zinacef	Moderate Risk	2
zithromax	Moderate Risk	2
ZITHROMAX (ZPAK)	Moderate Risk	2
zithromax / rocephin	High Risk	1
zithromax z-pak	Moderate Risk	2
zosyn	Moderate Risk	2
zosyn / cipro	High Risk	1
zosyn / nsaids	Moderate Risk	2
ZPACK	Moderate Risk	2
zpak	Moderate Risk	2

Table S5.3. Nursing home mentions used by the natural language processing algorithm.

Name Type	Examples
Generic	NH NSH nursing home SNF skilled nursing facility Hospice NHC
Proper (area specific)	Cumberland Manor Ida Culver House etc.

Table S5.4. Medications used for case-control exclusion and covariate analysis.

Medication Class	Examples
Transplant Medications	Cellcept
	munoloc
	mycophenylate mofetil
	Tacrolimus
	fk-506
	fk5
	k506
	tacarolimus
	tacrolimus hydrate
	fujimycin
	lcp-tacro
	prograf
	protopic
	Cyclosporine
	ciclosporin
	cyclosporin
	cyclosporin a
	gengraf
	neoral
	restasis
	sandimmune
	sangcya
	azothioprine
	azathioprin
	azathioprine sodium
	azatioprin
	azamun
	azanin
	azasan
	ccucol
	imuran
Corticosteroids	Cortisone
	Cortisone Acetate
	Hydrocortisone
	Hydrocortisone Sodium Phosphate
	Hydrocortisone Sodium Succinate
	Hydrocortisone Acetate
	Hydrocortisone Cypionate
	Prednisone
	Prednisolone
	Prednisolone Sodium Phosphate
	Methylprednisolone
	Methylprednisolone Sodium Succinate
	Methylprednisolone Acetate
	Triamcinolone
	Triamcinolone Acetonide
	Triamcinolone Diacetate
	Triamcinolone Hexacetonide
	Dexamethasone
	Dexamethasone Acetate Dexamethasone Sodium
	Phosphate
	Betamethasone
	Betamethasone Sodium Phosphate
	Betamethasone Acetate

Continued on next page

Continued from previous page

Medication Class	Examples
Diabetes Mellitus	Insulin
	glucagon
	glucagon-like peptide-1 (GLP-1)
	receptor agonists
	biguanides
	sulfonylurea
	thiazolidinediones
	meglitinides
	biguanides
	α-glucose inhibitor
	DPP-4 inhibitors
	SGLT2 inhibitors
	Cycloset

Table S6.1. Study characteristics of references included in the systematic literature review.

Short Title	Article Type	Study Design	Country	Medical Domain(s)	Conflicts of Interest
Abernethy (2014)	Special Report	Conference or workshop summary	US	Oncology	Board member for healthcare or pharmaceutical company
Blizinsky (2018)	Experience Report	Expert determination	US	Multiple	None disclosed
Braithwaite (2020)	Opinion	Expert determination	Australia	Multiple	None disclosed
Bubela (2019)	Experience Report	Conference or workshop summary	Canada	Multiple	None disclosed
Chambers (2016)	Viewpoint	Expert determination	US	Multiple	None disclosed
Chuong (2018)	Expert Analysis	Expert determination	Canada	Inflammatory Bowel Disease	None disclosed
David (2015)	Reflection	Conference or workshop summary	US	Multiple	Scientific advisor for healthcare company
Davis (2021)	Experience Report	Experience self-assessment	US	Multiple	None disclosed
Etheredge (2009)	Expert Analysis	Expert determination	US	Oncology	None disclosed
Etheredge (2014)	Expert Analysis	Expert determination	US	Multiple	None disclosed
Finlayson (2016)	System Evaluation	System development, implementation, and evaluation	US	Oncology	None disclosed
Ginsburg (2018)	Overview	Expert determination	US	Multiple	None disclosed
Glasgow (2018)	Special Report	Expert determination	US	Multiple	None disclosed
Hindorff (2018)	Perspective	Expert determination	US	Multiple	Scientific advisor for pharmaceutical company
Hirsch (2012)	Perspective	Expert determination	US	Multiple	Scientific advisor for healthcare company Scientific advisor for pharmaceutical company Equity holder of healthcare company
Holm (2017)	Symposium	Expert determination	Denmark	Multiple	None disclosed

Continued on next page

Continued from previous page

Short Title	Article Type	Study Design	Country	Medical Domain(s)	Conflicts of Interest
Hsu (2015)	Experience Report	System development, implementation, and evaluation	US	Aneurysms	None disclosed
IOM (2011)	Workshop Series Summary	Conference or workshop summary	US	Multiple	None disclosed
Jones (2020)	Interview Study	Qualitative analysis of interview or focus group data	US	Oncology	Scientific advisor for healthcare company Scientific advisor for pharmaceutical company Equity holder of healthcare company Employee of healthcare or pharmaceutical company
Jones (2022)	Interview Study	Qualitative analysis of interview or focus group data	US	Oncology	Scientific advisor for healthcare company Scientific advisor for pharmaceutical company Equity holder of healthcare company Employee of healthcare or pharmaceutical company
Kehl (2019)	Original Investigation	System development, implementation, and evaluation	US	Oncology	Scientific advisor for healthcare company Scientific advisor for pharmaceutical company Equity holder of healthcare company Research support from healthcare or pharmaceutical company
Key (2018)	Commentary	Experience self-assessment	US	Multiple	None disclosed
Khalifa (2021)	Interview Study	Qualitative analysis of interview or focus group data	US	Multiple	Employee of healthcare or pharmaceutical company Research support from healthcare or pharmaceutical company

Continued on next page

Continued from previous page

Short Title	Article Type	Study Design	Country	Medical Domain(s)	Conflicts of Interest
Khalifa (2021)	Interview Study	Qualitative analysis of interview or focus group data	US	Multiple	Employee of healthcare or pharmaceutical company Research support from healthcare or pharmaceutical company
Krumholz (2014)	Perspective	Expert determination	US	Multiple	Research support from healthcare or pharmaceutical company
Mandl (2020)	Pilot Study		US	Multiple	Board member for healthcare or pharmaceutical company Scientific advisor for healthcare company Scientific advisor for pharmaceutical company Research support from healthcare or pharmaceutical company
McGinnis (2021)	Perspective	Expert determination	US	Multiple	None disclosed
McInnes (2021)	Perspective	Expert determination	US	Multiple	None disclosed
Nwaru (2017)	Correspondence	Expert determination	International	Asthma	Board member for healthcare or pharmaceutical company
Potter (2020)	Software	Experience self-assessment	US	Oncology	Equity holder of healthcare company Employee of healthcare or pharmaceutical company Research support from healthcare or pharmaceutical company
Preston (2022)	Software	System development, implementation, and evaluation	US	Multiple	Scientific advisor for healthcare company Employee of healthcare or pharmaceutical company
Schwartz (2018)	Original Investigation	Experience self-assessment	US	Multiple	Board member for healthcare or pharmaceutical company Research support from healthcare or pharmaceutical company
Scollen (2017)	Perspective	Expert determination	US	Multiple	None disclosed

Continued on next page

Continued from previous page

Short Title	Article Type	Study Design	Country	Medical Domain(s)	Conflicts of Interest
Shaikh (2014)	Panel Discussion Summary	Conference or workshop summary	US	Oncology	Employee of healthcare or pharmaceutical company
Simon (2020)	Case Study	Case study	US	Oncology	Scientific advisor for healthcare company Scientific advisor for pharmaceutical company Research support from healthcare or pharmaceutical company
Trifiletti (2015)	Perspective	Expert determination	US	Oncology	None disclosed
Wallace (2014)	Experience Report	Experience self-assessment	US	Multiple	Employee of healthcare or pharmaceutical company
Wiley (2016)	Perspective	Conference or workshop summary	US	Multiple	None disclosed
Williams (2018)	Experience Report	Experience self-assessment	US	Multiple	None disclosed
Williams (2019)	Experience Report	Experience self-assessment	US	Multiple	Research support from healthcare or pharmaceutical company
Wouters (2021)	Perspective	Expert determination	Netherlands	Multiple	None disclosed
Yang (2019)	Experience Report	Experience self-assessment	International	Multiple	None disclosed
Yu (2015)	Perspective	Expert determination	US	Oncology	None disclosed

Table S6.2. Data and standards study outcomes of references included in the systematic literature review.

Short Title	Collecting, integrating, and sharing genomic and phenotype data	Analyzing data for discovery	Developing standards
Abernethy (2014)		Methodologically rigorous analysis	Standards for communicating research findings between the public and private sectors
Blizinsky (2018)	Inclusion of additional sociodemographic, psychologic, behavioral, and environmental data in EHRs Flow of information between the biomedical research community and LHSs	Development of CDS that considers social, environments, ancestral, genetic factors	International standards for classifying populations
Bubela (2019)	Networked approach to data sharing Clear processes and pathways for data access, integration, and use within and between systems Creative and collaborative strategies for data integration Data sharing between government agencies and research entities		
Chuong (2018)	Considerations of how big data can be integrated and stored in the EHR		
David (2015)	Combination of multiple biorepositories across institutions Easy access to genomic information in the EHR	Use of assays that can survey multiple important genotypes effectively and inexpensively	Informed consent standards for genome sequencing Standardized protocols for data return Standardized order sets and ordering protocols Standardized protocols for evaluation genomic medicine applications
Etheredge (2009)	Biobanks that link clinical, genetic, and environmental data		
Etheredge (2014)	Large, de-identified, shareable datasets		Development of a strategy to standardize, store and protect genomic data Standardized federally subsidized EHRs

Continued on next page

Continued from previous page

Short Title	Collecting, integrating, and sharing genomic and phenotype data	Analyzing data for discovery	Developing standards
Finlayson (2016)	Combination of multiple biorepositories across institutions	Inclusion of cohort selection, outcomes analysis, and examination of raw data as core functionalities of analysis Appropriately designed statistical models Expert engagement in interpreting results Comparison of results with current literature	
Ginsburg (2018)	Combination of multiple biorepositories across institutions Data democratization		
Hindorff (2018)			Standards for using case reports and observational studies to improve clinical decision making Standardized criteria for including multiple lines of evidence from underrepresented populations Standards for measuring clinical utility Standards for capturing diversity in health system data Standards for collecting social and environmental data Development and use of standards for data privacy
Hirsch (2012)	Combination of multiple biorepositories across institutions Input of genomic data reliably into the EHR Integration of biological data with patient-reported and wearable device data Alternative data collection platforms that are flexible and affordable	Improvement of data visualization techniques	

Continued on next page

Continued from previous page

Short Title	Collecting, integrating, and sharing genomic and phenotype data	Analyzing data for discovery	Developing standards
Hsu (2015)	Mechanisms to routinely follow up with patients to collect additional relevant data	Ontology-driven approaches to data extraction, standardization, and analysis Consideration of context when integrating longitudinal data	
IOM (2011)			Standards for distributed queries across systems Consensus on standards for care, quality, public health, and research Development and use of standards for data privacy
Jones (2022)		Publication of all LHS findings	Development and use of standards for data privacy
Kehl (2019)		Scalable methods for extracting data from EHRs	
Khalifa (2021)	Genetic reports that are both clinically useful (e.g. contain patient case characteristics) and computationally friendly		Prioritization of data to be standardized
Khalifa (2021)			Prioritization of data to be standardized
Krumholz (2014)		Development of criteria that guide interpretation of enormous data sets Results validation	Development and use of standards for data privacy
Mandl (2020)	Data sharing agreements that are based on reciprocity and interoperability Use of federated data sharing, where individual sites have access to their own data but other sites have access to de-identified data		Prioritization of data to be standardized
McGinnis (2021)			Standards for communicating research findings between the public and private sectors
McInnes (2021)	Transparent data sharing expectations across all levels of participation	Alternative methods for determining and predicting functional effects of genetic variants Use of whole-population datasets	Development and use of standards for data privacy

Continued on next page

Continued from previous page

Short Title	Collecting, integrating, and sharing genomic and phenotype data	Analyzing data for discovery	Developing standards
Nwaru (2017)	Frameworks for data harmonization, standardization, transformation, and linkage	Methodologically rigorous analysis Improvement of data visualization techniques	Development and use of standards for data privacy
Potter (2020)	Input of genomic data reliably into the EHR Frameworks for data harmonization, standardization, transformation, and linkage	Improvement of data visualization techniques	
Schwartz (2018)		Preliminary bioinformatics assessments in house with multiple partner labs	Development and use of standards for data privacy
Scollen (2017)	Combination of multiple biorepositories across institutions Sharing of genomic data internationally using FAIR (findable, accessible, interoperable, reusable) principles		
Wallace (2014)		Methodologically rigorous analysis Use of analytics techniques used by other disciplines, like businesses Plans for dissemination of research findings as a part of the research project's design Validate research findings through peer review	
Wiley (2016)	Requirement that omics data be returned in computer-readable formats as part of the Clinical Laboratory Improvement Amendment certification	Use of a standing expert committee to identify necessary metadata elements for omic data reanalysis and reinterpretation as new technologies emerge Research on the impact of documentation errors on the reuse of medical record data by computational methodologies Definition of who bears ethical and legal responsibilities for reanalysis of raw data Research adequacy of existing ontologies and identify additional needs to capture omics-related metadata and interpretations	

Continued on next page

Continued from previous page

Short Title	Collecting, integrating, and sharing genomic and phenotype data	Analyzing data for discovery	Developing standards
Williams (2018)		Processes to re-analyze previously analyzed sequences	Development and use of standards for data privacy
Williams (2019)	Data sharing to combine genomic knowledge		
Yang (2019)		Use of a global unique identifier for each participant	
Yu (2015)	Integration of biomarker data into the EHR data in a way that allows for CDS and population health studies	Appropriately designed statistical models Semantic interoperability of biomarker data	

Table S6.3. Culture and acceptance study outcomes of references included in the systematic literature review.

Short Title	Building a collaborative learning culture	Demonstrating value and feasibility	Aligning learning with existing healthcare improvement models
Abernethy (2014)	Alignment of value assessments with patient needs and rapid scientific advancements Communication and involvement across stakeholder groups A collaborative multidisciplinary ecosystem Partnerships between stakeholders	New approaches for defining and demonstrating "value" in healthcare	Alignment of CER and HTA with patient needs, values, and characteristics
Braithwaite (2020)	Commitment to improvement Readiness and preparedness for change Recognition of the capacities and barriers to progress		Network and complexity science-driven understanding of health systems Understanding of available implementation strategies
Bubela (2019)	Culture of trust and mitigation of risk aversion by data users Processes that sustain trust at the individual and institutional levels	Proof of concept models/case studies using existing health records	Implementation checklists
Chambers (2016)	Commitment to improvement Communication and involvement across stakeholder groups		Considerations of context and theoretical models from implementations science
David (2015)	Institutional advisory committees with senior leadership Partnerships with medical subspecialists with content expertise	Demonstration that the cost of testing is not necessarily prohibitive Use of institutional quality improvement analysis to assess value Gradual demonstration to patients and healthcare communities of the value of genomic medicine	
Davis (2021)	Strong commitment to making culture change a central part of the LHS transition Close, sustained alignment between multiple levels of leadership	Use of patient-reported outcomes to measure progress and success Communication about initiatives that successfully implement cyclic improvement	Alignment of research, quality improvement, innovation, and the clinical enterprise
Etheredge (2009)		Use of patient-reported outcomes to measure progress and success	
Etheredge (2014)	Commitment to improvement Public-private and international collaboration	Proof of concept models/case studies using existing health records	Definition of CER priorities

Continued on next page

Continued from previous page

Short Title	Building a collaborative learning culture	Demonstrating value and feasibility	Aligning learning with existing healthcare improvement models
Ginsburg (2018)		Emphasis on the ability of precision medicine to benefit the entire population ("precision public health") Evidence of value Demonstration of economic value to both patients and organizations Collaboration between payers and industry to develop evidence base for economic value of precision medicine	
Hindorff (2018)	Communication and involvement across stakeholder groups Participation of different types of healthcare systems in the LHS		
Hirsch (2012)		Proof of concept models/case studies using existing health records	
Hsu (2015)	Trust among stakeholders		
IOM (2011)	Communication and involvement across stakeholder groups	Proof of concept models/case studies using existing health records New approaches for defining and demonstrating "value" in healthcare	
Jones (2020)	Trust among stakeholders		
Jones (2022)	Trust among stakeholders		
Key (2018)	Trust among stakeholders		
Khalifa (2021)		Increased motivation for data standards adoption through increased clinical demand	
Krumholz (2014)	A clinical research community that realizes the promise of big data Clinician comfort with evidence generated from big data Communication and involvement across stakeholder groups	Demonstration of economic value to both patients and organizations	

Continued on next page

Continued from previous page

Short Title	Building a collaborative learning culture	Demonstrating value and feasibility	Aligning learning with existing healthcare improvement models
Mandl (2020)	Communication and involvement across stakeholder groups Close, sustained alignment between multiple levels of leadership		
McGinnis (2021)	Cultural commitment to learning Communication and involvement across stakeholder groups Local and global communication Accountability for quality care		
Nwaru (2017)	Communication and involvement across stakeholder groups		
Schwartz (2018)	Organizational dedication to creating an LHS Communication and involvement across stakeholder groups		
Shaikh (2014)	Communication and involvement across stakeholder groups Recruitment of new, interdisciplinary communities of investigators into biomedical research		
Trifiletti (2015)			Combination of big data with CER
Wallace (2014)	Development of clinical champions Engagement between a diverse group of stakeholders across the healthcare ecosystem		
Wiley (2016)			Mechanisms to transition QI projects to research designations
Williams (2018)	Multidisciplinary working group that represents key organizational functions Coupling of genomics/multidisciplinary expertise with a non-traditional communication strategy that crosses institutional boundaries	Demonstration that research does not need to be complete prior to implementation	

Continued on next page

Continued from previous page

Short Title	Building a collaborative learning culture	Demonstrating value and feasibility	Aligning learning with existing healthcare improvement models
Williams (2019)		Engagement of diverse stakeholders to understand the value proposition for genomic medicine	Use of implementation science frameworks to understand the barriers and facilitators of using genomic medicine in the clinic
Yu (2015)	Communication and involvement across stakeholder groups		

Table S6.4. Engaging with and protecting patients study outcomes of references included in the systematic literature review.

Short Title	Advancing health equity	Prioritizing patient-centeredness	Obtaining consent for clinical research	Safety measures and monitoring	Privacy and security protections
Abernethy (2014)		Deliberate shift to patient-centered care Development of a clear patient-centered research agenda Shared patient-provider decision making using CDS tools		Longitudinal outcomes measurement	Sharing of data in a privacy-protected manner
Blizinsky (2018)	Reflection of diversity and clinical complexity in the EHR Use of more complex measures in place of crude proxies for racial and ethnic categories Commitment to understanding genetic variation among ancestral groups Development of new models to enhance the use of de-identified clinical data from diverse populations Development of sustainable, respectful relationships with diverse communities to encourage research participation, and develop appropriate recruitment strategies Improved capture of population diversity measures in EHRs Monitoring of available genotype and phenotype data from diverse populations				

Continued on next page

Continued from previous page

Short Title	Advancing health equity	Prioritizing patient-centeredness	Obtaining consent for clinical research	Safety measures and monitoring	Privacy and security protections
Bubela (2019)		Participant-centric approach	Creative and ethical approaches to consent		
Chuong (2018)		Patient and family engagement			Careful consideration of patient privacy and confidentiality
David (2015)		Improved strategies for communication with at-risk families	Prospective obtainment of informed consent Research on reasons for refusal of confirmatory testing		
Davis (2021)			Communication with patients about the importance of patient-related and provided data for learning and improvement	Augmentation of the IRB with an advisory body that facilitates and tracks the spectrum of learning activities	
Etheredge (2009)				Information capture about clinical protocols	
Ginsburg (2018)	Understanding of how precision medicine works to increase or decrease historical health disparities				
Glasgow (2018)		Political commitment to providing patient-centered, personalized care using the best available evidence Infrastructure that supports the value of patient-centered, personalized care		Investigation of best uses for patient-reported measures and outcomes	

Continued on next page

Continued from previous page

Short Title	Advancing health equity	Prioritizing patient-centeredness	Obtaining consent for clinical research	Safety measures and monitoring	Privacy and security protections
Hindorff (2018)	Development of LHSs in underserved health systems Improvement of access to care, including for uninsured populations Inclusion of data from diverse individuals in routine analysis and observational studies Development of practices within clinical labs to include ancestry in analyzing results Understanding of how precision medicine works to increase or decrease historical health disparities Inclusion of more and better data from underrepresented populations Correct documentation of race, ethnicity, gender identity, social determinants of health More and better data from diverse individuals				Sharing of data in a privacy-protected manner Consideration of heterogeneity of priorities and challenges across the spectrum of US healthcare institutions when exchanging data and protecting patient privacy
Hirsch (2012)		Patient and family engagement		Collection of patient-reported outcomes	Careful consideration of patient privacy and confidentiality

Continued on next page

Continued from previous page

Short Title	Advancing health equity	Prioritizing patient-centeredness	Obtaining consent for clinical research	Safety measures and monitoring	Privacy and security protections
Holm (2017)			Infrastructure that supports new models of consent (e.g. dynamic consent, meta-consent) Consideration of a meta-consent model that allows participants to choose how they should be consented		
IOM (2011)				Definition of consensus outcomes measures	Careful consideration of patient privacy and confidentiality Use a consortium approach to make patients securely identifiable
Jones (2020)			Clear explanation of data security measures to patients during consent Communication of societal benefits to patients during consent		
Jones (2022)		Distribution of LHS informational materials to patients Focus on transparency and communication with patients to improve trust			
Key (2018)		Community engagement on a continuum during the research process Focus on transparency and communication with patients to improve trust			

Continued on next page

Continued from previous page

Short Title	Advancing health equity	Prioritizing patient-centeredness	Obtaining consent for clinical research	Safety measures and monitoring	Privacy and security protections
McGinnis (2021)	Improvement of access to care, including for uninsured populations	Infrastructure that supports the value of patient-centered, personalized care			
McInnes (2021)	Comparison of screened individuals with wider unequal social system Inclusion of more and better data from underrepresented populations		Balance between individual control and public good	Better integration of the moral systems embedded in research and clinical care	
Nwaru (2017)					Careful consideration of patient privacy and confidentiality
Schwartz (2018)			Broad consent that allows results to be returned over time	Careful selection of medically actionable genes for return Stringent variant interpretation to minimize false positives	
Simon (2020)				Systematic integration of care experience from patient navigators into 4R (right formation, treatment, patient, time) care sequence templates	
Wallace (2014)		Community engagement on a continuum during the research process			Scrutinization of existing legal foundations for privacy protection for their applicability to learning environments

Continued on next page

Continued from previous page

Short Title	Advancing health equity	Prioritizing patient-centeredness	Obtaining consent for clinical research	Safety measures and monitoring	Privacy and security protections
Wiley (2016)		Clarification of the patient's right under HIPAA to access raw biomolecular data collected by care providers when those data are not stored in the medical record National discussion on the rights of patients to go beyond reading their medical records as assured by HIPAA to having the ability to add data to the record to identify and correct errors without going through a physician intermediary	Use of public education funds from the Department of Health and Human Services to develop public awareness campaigns to accurately communicate benefits and risks of data sharing	Movement towards centralized IRB solutions	Classification of non-interventional research as appropriate use of PHI under HIPAA regulations Clarification of whether omics data are considered biometric identifiers under HIPAA Augmentation of legal protections to safeguard deidentified data from misuse and attempted reidentification of subjects
Williams (2018)		Patient engagement Maintenance of trust with the community by involving the community	Process for re-consenting participants, if need be	Evaluation of the impact of reporting variants to patient-participants and to the system Collection of outcomes data to incorporate into economic models and evaluate cost-effectiveness	

Table S6.5. Political and institutional support study outcomes of references included in the systematic literature review.

Short Title	Funding and incentives	Policy and governance	Building institutional capacity for genomic medicine and learning
Abernethy (2014)	Funding for basic and applied research in the public and private sectors Development of evidence-based tools and incentives for patient-centered care Discussions about the cost of care	Policies that recognize and support different dimensions of clinical value (cost/clinical efficacy, QOL, productivity, patient preference) Policies that encourage innovation Policies that are aligned with the dynamic and fast-pace nature of scientific discovery Policies that stimulate private-public partnerships Policies that address concerns about integrating clinical research and clinical care Policies that facilitate data liquidity	Tools to help effectively disseminate clinical information to patients and physicians Support for clinical data infrastructure for research
Braithwaite (2020)	Allocation of resources to fast-paced learning		
Bubela (2019)		Balance between over and under-centralization of data sharing policies Harmonized government policy for health data use and development of innovative technologies	Single health system Platforms that enable system learning Infrastructure that supports real-time and real-world data analytics
Chambers (2016)	Allocation of resources to fast-paced learning		Strategies to implement evidence-based practices Infrastructure that supports real-time and real-world data analytics
Chuong (2018)		Adaptable governance approaches	

Continued on next page

Continued from previous page

Short Title	Funding and incentives	Policy and governance	Building institutional capacity for genomic medicine and learning
David (2015)	Use of internal pilot study funding to conduct pilot studies and increase acceptance Funding for interim testing between discovery and adoption Internal funding to prevent patients from getting charged	New genomic education objectives set by the NHGRI	Investment in research personnel and resources to ensure research quality is equal to clinical quality Allocation of sufficient personnel to manage consent Support for biobanking to aid clinical confirmatory sequencing Anticipation of rises in interpretive (rather than testing) costs Establishment of expected involvement of healthcare institution with family members after testing proband Establishment of GCs and geneticists in non genetics clinical services Use of genomic medicine teams, rather than primary care clinicians, to follow up with patients after testing Education for non-genetics personnel on how to order, interpret, and act on genetic tests Clinician oversight for trainees who are ordering genetic tests Expanded institutional CLIA-compliant genotyping Development of a usability lab to assess usability of genomic medicine applications in the EHR Development of CDS tools in interdisciplinary groups Focus groups with patients, clinicians, and other key stakeholders to identify educational needs

Continued on next page

Continued from previous page

Short Title	Funding and incentives	Policy and governance	Building institutional capacity for genomic medicine and learning
Davis (2021)			Automated processes Unified data architecture Definition of enterprise-wide strategies for AI/ML tools Involvement of patients and clinicians in the development process to build empathy into systems
Etheredge (2009)	Payment models that reward high-quality care	Requirements for publicly-funded studies to report de-identified data to a national research database Legislative authority and financing of research and payment reforms Research priorities and measures set by HHS Prioritization of research assessments of new technologies	
Etheredge (2014)	Funding for research programs that address national priority questions	Attention to Medicare and Medicaid needs Development of governing principles, priorities, system specifications, and cooperative strategies	Modernization of clinical trial and registry systems Development of predictive models to help clinicians at the point of care
Ginsburg (2018)	Incentives for data sharing	Development of an approach for regulating diagnostics that incorporate sequencing technologies by FDA and CMS Global leadership and perseverance	Development of secure and interoperable genomics-enabled IT systems in health care and community settings
Glasgow (2018)	New funding priorities		Understanding of how to train patients and healthcare workers to make the best use of new tools and data
Hindorff (2018)			Maximization of the use of existing health infrastructure while building new capacities Familiarization of providers with authorization processes for genetic testing using different insurance providers

Continued on next page

Continued from previous page

Short Title	Funding and incentives	Policy and governance	Building institutional capacity for genomic medicine and learning
Potter (2020)			Implementation of FHIR in EHRs
Preston (2022)			Infrastructure for guiding researchers through complex variant curation guidelines and standards
Schwartz (2018)			Clinical processes that support clinicians and patients post return of genetic results Implementation of mass customization to speed up results return but maintain subject-specific factors
Scollen (2017)			Research commons where investigators can share and refine tools and data
Shaikh (2014)	Unconventional models for incentivizing innovation, like prizes and challenges		Research commons where investigators can share and refine tools and data
Wallace (2014)			Increased awareness of the nuances and challenges of the care delivery process among researchers Balance between overall goal of generalizable knowledge and institutional needs for rapidly implementable knowledge Reframing of innovations to fit specifics of different healthcare environments Alignment of interventions with the priorities and capabilities of the healthcare system
Wiley (2016)		Harmonization of state and federal laws on consent requirements to reduce the burden placed on patients who are willing to share their data	
Williams (2018)			Coordination with the payer to make sure that recommended treatments after testing are covered
Williams (2019)			Development of systems that support the use of genomic medicine in EHRs
Yang (2019)			Use of a protocol tracking and management system to track and manage the life cycle of clinical research Use of an electronic data capture and study management system to collect, integrate, and standardize research data Use of a specimen tracking system

Continued on next page

Continued from previous page

Short Title	Funding and incentives	Policy and governance	Building institutional capacity for genomic medicine and learning
Hirsch (2012)			Continued adoption of EHRs Familiarization of providers with authorization processes for genetic testing using different insurance providers
Holm (2017)			Training for researchers in research ethics
Hsu (2015)			Single health system
IOM (2011)		Development of governing principles, priorities, system specifications, and cooperative strategies	Consideration of the ultra-large-system (ULS) approach Interdisciplinary collaboration to develop IT infrastructure
Jones (2022)	Funding for additional transparency and communication needs		Inclusion of relationship-building and shared decision making education in medical school curricula
Krumholz (2014)	Funding for "unconventional" research and new types of clinical expertise	Attention to implementation on the part of leaders	Research commons where investigators can share and refine tools and data
Mandl (2020)	Incentives for development of computational phenotypes Provision of local benefit to participating sites	Decentralized governance model	Development of both technical and academic frameworks for sustaining collaborations Central tracking of proposed research projects so that there are "no surprises"
McGinnis (2021)			Unified data architecture Development of secure and interoperable genomics-enabled IT systems in health care and community settings Research commons where investigators can share and refine tools and data
Nwaru (2017)			Development of "knowledge to practice" infrastructures Shared digital infrastructure that supports both individual needs and improves the overall system Enhancement of current platforms to enable them to capture and integrate multiple forms of information Definition of enterprise-wide strategies for AI/ML tools Platform-agnostic analysis tools

Continued on next page

Continued from previous page

Short Title	Funding and incentives	Policy and governance	Building institutional capacity for genomic medicine and learning
Yu (2015)		Attention to legal frameworks from other countries to understand the ethical complexity of biomarker testing amid shifting societal attitudes	Use of a specimen tracking system

REFERENCES

1. McCarthy JJ, McLeod HL, Ginsburg GS. Genomic medicine: a decade of successes, challenges, and opportunities. Sci Transl Med. 2013;5: 189sr4.

2. Manolio TA, Chisholm RL, Ozenberger B, Roden DM, Williams MS, Wilson R, et al. Implementing genomic medicine in the clinic: the future is here. Genet Med. 2013;15: 258–267.

3. Khoury MJ. Dealing with the evidence dilemma in genomics and personalized medicine. Clin Pharmacol Ther. 2010;87: 635–638.

4. Angrist M, Jamal L. Living laboratory: whole-genome sequencing as a learning healthcare enterprise. Clin Genet. 2015;87: 311–318.

5. Manolio TA, Rowley R, Williams MS, Roden D, Ginsburg GS, Bult C, et al. Opportunities, resources, and techniques for implementing genomics in clinical care. Lancet. 2019;394: 511–520.

6. Berkman BE, Hull SC, Eckstein L. The unintended implications of blurring the line between research and clinical care in a genomic age. Per Med. 2014;11: 285–295.

7. Committee on the Learning Health Care System in America, Institute of Medicine. Best Care at Lower Cost: The Path to Continuously Learning Health Care in America. Smith M, Saunders R, Stuckhardt L, McGinnis JM, editors. Washington (DC): National Academies Press (US); 2014.

8. Roundtable on Translating Genomic-Based Research for Health, Board on Health Sciences Policy, Institute of Medicine. Genomics-Enabled Learning Health Care Systems: Gathering and Using Genomic Information to Improve Patient Care and Research: Workshop Summary. Washington (DC): National Academies Press (US); 2015.

9. Green ED, Gunter C, Biesecker LG, Di Francesco V, Easter CL, Feingold EA, et al. Strategic vision for improving human health at The Forefront of Genomics. Nature. 2020;586: 683–692.

10. Wouters RHP, van der Graaf R, Rigter T, Bunnik EM, Ploem MC, de Wert GMWR, et al. Towards a Responsible Transition to Learning Healthcare Systems in Precision Medicine: Ethical Points to Consider. J Pers Med. 2021;11. doi:10.3390/jpm11060539

11. Enticott J, Johnson A, Teede H. Learning health systems using data to drive healthcare improvement and impact: a systematic review. BMC Health Serv Res. 2021;21: 200.

12. Wouters RHP, van der Graaf R, Voest EE, Bredenoord AL. Learning health care systems:

Highly needed but challenging. Learn Health Syst. 2020;4: e10211.

13. Williams MS, Buchanan AH, Daniel Davis F, Andrew Faucett W, Hallquist MLG, Leader JB, et al. Patient-Centered Precision Health In A Learning Health Care System: Geisinger's Genomic Medicine Experience. Health Affairs. 2018. pp. 757–764. doi:10.1377/hlthaff.2017.1557

14. Platt JE, Raj M, Wienroth M. An Analysis of the Learning Health System in Its First Decade in Practice: Scoping Review. J Med Internet Res. 2020;22: e17026.

15. Amendola LM, Berg JS, Horowitz CR, Angelo F, Bensen JT, Biesecker BB, et al. The Clinical Sequencing Evidence-Generating Research Consortium: Integrating Genomic Sequencing in Diverse and Medically Underserved Populations. Am J Hum Genet. 2018;103: 319–327.

16. eMERGE Consortium. Lessons learned from the eMERGE Network: balancing genomics in discovery and practice. Human Genetics and Genomics Advances. 2021;2: 100018.

17. Byrd JB, Greene AC, Prasad DV, Jiang X, Greene CS. Responsible, practical genomic data sharing that accelerates research. Nat Rev Genet. 2020;21: 615–629.

18. Wolf SM, Amendola LM, Berg JS, Chung WK, Clayton EW, Green RC, et al. Navigating the research–clinical interface in genomic medicine: analysis from the CSER Consortium. Genet Med. 2017;20: 545–553.

19. Zhang D, Prabhu VS, Marcella SW. Attributable Healthcare Resource Utilization and Costs for Patients With Primary and Recurrent Clostridium difficile Infection in the United States. Clinical Infectious Diseases. 2018. pp. 1326–1332. doi:10.1093/cid/cix1021

20. Johnson M, O'Hara R, Hirst E, Weyman A, Turner J, Mason S, et al. Multiple triangulation and collaborative research using qualitative methods to explore decision making in pre-hospital emergency care. BMC Med Res Methodol. 2017;17: 11.

21. Denzin NK. The research act: A theoretical introduction to sociological methods, Aldine Pub. Co. Chicago. 1970.

22. Braithwaite J, Runciman WB, Merry AF. Towards safer, better healthcare: harnessing the natural properties of complex sociotechnical systems. Quality and Safety in Health Care. 2009. pp. 37–41. doi:10.1136/qshc.2007.023317

23. Gray K, Sockolow P. Conceptual Models in Health Informatics Research: A Literature Review and Suggestions for Development. JMIR Med Inform. 2016;4: e7.

24. Charmaz K. Constructing Grounded Theory: A Practical Guide through Qualitative Analysis. SAGE; 2006.

25. Charmaz K. Constructing Grounded Theory. SAGE; 2014.

26. Blasimme A, Fadda M, Schneider M, Vayena E. Data Sharing For Precision Medicine: Policy Lessons And Future Directions. Health Aff . 2018;37: 702–709.

27. Borgman CL. The conundrum of sharing research data. J Am Soc Inf Sci Technol. 2012;63: 1059–1078.

28. Raza S, Hall A. Genomic medicine and data sharing. Br Med Bull. 2017;123: 35–45.

29. Mai PL, Sand SR, Saha N, Oberti M, Dolafi T, DiGianni L, et al. Li-Fraumeni Exploration Consortium Data Coordinating Center: Building an Interactive Web-Based Resource for Collaborative International Cancer Epidemiology Research for a Rare Condition. Cancer Epidemiology Biomarkers & Prevention. 2020. pp. 927–935. doi:10.1158/1055-9965.epi-19-1113

30. Biswas K, Carty C, Horney R, Nasrin D, Farag TH, Kotloff KL, et al. Data management and other logistical challenges for the GEMS: the data coordinating center perspective. Clin Infect Dis. 2012;55 Suppl 4: S254–61.

31. Blizinsky KD, Bonham VL. Leveraging the Learning Health Care Model to Improve Equity in the Age of Genomic Medicine. Learn Health Syst. 2018;2. doi:10.1002/lrh2.10046

32. Crawford DC, Sedor JR. Biobanks Linked to Electronic Health Records Accelerate Genomic Discovery. Journal of the American Society of Nephrology: JASN. 2021. pp. 1828–1829.

33. Stanaway IB, Hall TO, Rosenthal EA, Palmer M, Naranbhai V, Knevel R, et al. The eMERGE Genotype Set of 83,717 Subjects Imputed to ~40 Million Variants Genome Wide and Association with the Herpes Zoster Medical Record Phenotype. Genet Epidemiol. accepted 28-Aug-2018.

34. Kullo IJ, Fan J, Pathak J, Savova GK, Ali Z, Chute CG. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. J Am Med Inform Assoc. 2010;17: 568–574.

35. Bellenguez C, Küçükali F, Jansen IE, Kleineidam L, Moreno-Grau S, Amin N, et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. Nat Genet. 2022;54: 412–436.

36. Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet. 2011;12: 417–428.

37. Balsells E, Shi T, Leese C, Lyell I, Burrows J, Wiuff C, et al. Global burden of Clostridium difficile infections: a systematic review and meta-analysis. J Glob Health. 2019;9: 010407.

38. Carey DJ, Fetterolf SN, Davis FD, Faucett WA, Kirchner HL, Mirshahi U, et al. The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. Genet Med. 2016;18: 906–913.

39. Flemming K, Booth A, Garside R, Tunçalp Ö, Noyes J. Qualitative evidence synthesis for complex interventions and guideline development: clarification of the purpose, designs and relevant methods. BMJ Glob Health. 2019;4: e000882.

40. Jane Shrapnel, Nan Hu, Nora Samir, Michael Hodgins, Ingrid Wolfe, James Newham, Melanie Keep, Raghu Lingam. What implementation strategies and outcome measures are used in transforming health care organizations into learning health systems? A mixed methods systematic review. Available: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42019153775

41. Lim HC, Austin JA, van der Vegt AH, Rahimi AK, Canfell OJ, Mifsud J, et al. Toward a Learning Health Care System: A Systematic Review and Evidence-Based Conceptual Framework for Implementation of Clinical Analytics in a Digital Hospital. Appl Clin Inform. 2022;13: 339–354.

42. Somerville M, Cassidy C, Curran J, Rothfus M, Sinclair D, Rose AE. What implementation strategies and outcome measures are used to transform health care organizations into learning health systems? A mixed methods review protocol. 2022. doi:10.21203/rs.3.rs-1269601/v1

43. Ellis LA, Sarkies M, Churruca K, Dammery G, Meulenbroeks I, Smith CL, et al. The Science of Learning Health Systems: Scoping Review of Empirical Research. JMIR Med Inform. 2022;10: e34907.

44. Easterling D, Perry AC, Woodside R, Patel T, Gesell SB. Clarifying the concept of a learning health system for healthcare delivery organizations: Implications from a qualitative analysis of the scientific literature. Learn Health Syst. 2022;6: e10287.

45. Enticott JC, Melder A, Johnson A, Jones A, Shaw T, Keech W, et al. A Learning Health System Framework to Operationalize Health Data to Improve Quality Care: An Australian Perspective. Front Med. 2021;8: 730021.

46. Surkis A, Read K. Research data management. J Med Libr Assoc. 2015;103: 154–156.

47. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3: 160018.

48. Gray J. Data Management: Past, Present, and Future. arXiv [cs.DB]. 2007. Available: http://arxiv.org/abs/cs/0701156

49. Weiner MG, Embi PJ. Toward reuse of clinical data for research and quality improvement:

the end of the beginning? Annals of internal medicine. 2009. pp. 359–360.

50. Muenzen KD, Amendola LM, Kauffman TL, Mittendorf KF, Bensen JT, Chen F, et al. Lessons learned and recommendations for data coordination in collaborative research: The CSER consortium experience. HGG Adv. 2022;3: 100120.

51. McCarthy MI, MacArthur DG. Human disease genomics: from variants to biology. Genome Biol. 2017;18: 20.

52. Auffray C, Balling R, Barroso I, Bencze L, Benson M, Bergeron J, et al. Making sense of big data in health research: Towards an EU action plan. Genome Med. 2016;8: 71.

53. ACMG Board of Directors. 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. 2017;19: 721–722.

54. Eisenstein M. Big data: The power of petabytes. Nature. 2015;527: S2–4.

55. Kruse CS, Goswamy R, Raval Y, Marawi S. Challenges and Opportunities of Big Data in Health Care: A Systematic Review. JMIR Med Inform. 2016;4: e38.

56. Reisman M. EHRs: The challenge of making electronic data usable and interoperable. P T. 2017;42: 572–575.

57. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. Summit Transl Bioinform. 2010;2010: 1–5.

58. Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. J Med Internet Res. 2018;20: e185.

59. Geneviève LD, Martani A, Mallet MC, Wangmo T, Elger BS. Factors influencing harmonized health data collection, sharing and linkage in Denmark and Switzerland: A systematic review. PLoS One. 2019;14: e0226015.

60. Institute of Medicine, Board on Health Care Services, Board on Health Sciences Policy, Roundtable on Translating Genomic-Based Research for Health, National Cancer Policy Forum, Forum on Neuroscience and Nervous System Disorders, et al. Sharing Clinical Research Data: Workshop Summary. National Academies Press; 2013.

61. Daniels H, Jones KH, Heys S, Ford DV. Exploring the Use of Genomic and Routinely Collected Data: Narrative Literature Review and Interview Study. J Med Internet Res. 2021;23: e15739.

62. Shabani M, Marelli L. Re-identifiability of genomic data and the GDPR: Assessing the re-identifiability of genomic data in light of the EU General Data Protection Regulation. EMBO Rep. 2019;20: e48316.

63. Kaye J. The Tension Between Data Sharing and the Protection of Privacy in Genomics Research. Annual Review of Genomics and Human Genetics. 2012. pp. 415–431. doi:10.1146/annurev-genom-082410-101454

64. Oliver JM, Slashinski MJ, Wang T, Kelly PA, Hilsenbeck SG, McGuire AL. Balancing the Risks and Benefits of Genomic Data Sharing: Genome Research Participants' Perspectives. Public Health Genomics. 2012. pp. 106–114. doi:10.1159/000334718

65. Knoppers BM, Harris JR, Budin-Ljøsne I, Dove ES. A human rights approach to an international code of conduct for genomic and clinical data sharing. Hum Genet. 2014;133: 895–903.

66. Bentley AR, Callier S, Rotimi CN. Diversity and inclusion in genomic research: why the uneven progress? J Community Genet. 2017;8: 255–266.

67. Kaye J, Heeney C, Hawkins N, de Vries J, Boddington P. Data sharing in genomics — re-shaping scientific practice. Nature Reviews Genetics. 2009. pp. 331–335. doi:10.1038/nrg2573

68. Boland MR, Karczewski KJ, Tatonetti NP. Ten Simple Rules to Enable Multi-site Collaborations through Data Sharing. PLoS Comput Biol. 2017;13: e1005278.

69. Anagnostou P, Capocasa M, Milia N, Bisol GD. Research data sharing: Lessons from forensic genetics. Forensic Sci Int Genet. 2013;7: e117–e119.

70. Liao Z, Quintana Y. Challenges to Global Standardization of Outcome Measures. AMIA Jt Summits Transl Sci Proc. 2021;2021: 404–409.

71. O'Doherty KC, Shabani M, Dove ES, Bentzen HB, Borry P, Burgess MM, et al. Toward better governance of human genomic data. Nat Genet. 2021;53: 2–8.

72. Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. UCLA Law Rev. Available: https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/uclalr57&section=48&casa_token=u4a5lK0w32oAAAAA:IkthJYK5nC-6sbjg0PdMRhlotvxGj6ZXE_jhptNiHCIlB4viNow27z88Qgjkqgzk3C2iKQ

73. Genomes Project Consortium 1000. A global reference for human genetic variation. Nature. 2015. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/pmc4750478/

74. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562: 203–209.

75. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, et al. NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic Acids Res. 2014;42: D975–9.

76. Investigators TA of URP, The All of Us Research Program Investigators. The “All of Us” Research Program. New England Journal of Medicine. 2019. pp. 668–676. doi:10.1056/nejmsr1809937

77. Schatz MC, Philippakis AA, Afgan E, Banks E, Carey VJ, Carroll RJ, et al. Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space. Cell Genomics. 2022;2: 100085.

78. Quinn SC, Garza MA, Butler J, Fryer CS, Casper ET, Thomas SB, et al. Improving informed consent with minority participants: results from researcher and community surveys. J Empir Res Hum Res Ethics. 2012;7: 44–55.

79. Spencer K, Sanders C, Whitley EA, Lund D, Kaye J, Dixon WG. Patient Perspectives on Sharing Anonymized Personal Health Data Using a Digital System for Dynamic Consent and Research Feedback: A Qualitative Study. J Med Internet Res. 2016;18: e66.

80. Phillips M, Molnár-Gábor F, Korbel JO, Thorogood A, Joly Y, Chalmers D, et al. Genomics: data sharing needs an international code of conduct. Nature. 2020;578: 31–33.

81. Knoppers BM, Harris JR, Tassé AM, Budin-Ljøsne I, Kaye J, Deschênes M, et al. Towards a data sharing Code of Conduct for international genomic research. Genome Med. 2011;3: 46.

82. Piwowar HA, Becich MJ, Bilofsky H, Crowley RS, caBIG Data Sharing and Intellectual Capital Workspace. Towards a data sharing culture: recommendations for leadership from academic health centers. PLoS Med. 2008;5: e183.

83. Rolland B. Designing and Developing Coordinating Centers as Infrastructure to Support Team Science. In: Hall KL, Vogel AL, Croyle RT, editors. Strategies for Team Science Success: Handbook of Evidence-Based Principles for Cross-Disciplinary Science and Practical Lessons Learned from Health Researchers. Cham: Springer International Publishing; 2019. pp. 413–417.

84. Goddard KAB, Angelo FAN, Ackerman SL, Berg JS, Biesecker BB, Danila MI, et al. Lessons learned about harmonizing survey measures for the CSER consortium. Journal of Clinical and Translational Science. 2020;4: 537–546.

85. Fleming EM. Artifact Study: A Proposed Model. Winterthur Portf. 1974;9: 153–173.

86. Lyman P, Kahle B. Archiving digital cultural artifacts. D-lib Magazine. 1998;4. Available: http://mirror.dlib.org/dlib/july98/07lyman.html

87. Fang Y, Neufeld D, Zhang X. Knowledge coordination via digital artefacts in highly dispersed teams. Inf Syst J. 2022;32: 520–543.

88. Beyer H, Holtzblatt K. Contextual design. Interactions. 1999;6: 32–42.

89. Menear M, Blanchette M-A, Demers-Payette O, Roy D. A framework for value-creating learning health systems. Health Res Policy Syst. 2019;17: 79.

90. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009;42: 377–381.

91. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: Building an international community of software platform partners. J Biomed Inform. 2019;95: 103208.

92. Chang W, Cheng J, Allaire JJ, Sievert C, Schloerke B, Xie Y, et al. Web Application Framework for R [R package shiny version 1.6.0]. 2021 [cited 24 Jun 2021]. Available: https://CRAN.R-project.org/package=shiny

93. National Human Genome Research Institute (NHGRI). Notice of New NIH-Designated Data Repository: NHGRI's Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL). NIH Grants, https://grants.nih.gov/grants/guide/notice-files/NOT-HG-19-024.html. 2019. Available: https://grants.nih.gov/grants/guide/notice-files/NOT-HG-19-024.html

94. Global Alliance for Genomics and Health. GENOMICS. A federated ecosystem for sharing genomic, clinical data. Science. 2016;352: 1278–1280.

95. NIH Office of Science Policy. NIH GDS Policy Oversight. In: NIH Office of Science Policy [Internet]. 2021. Available: https://osp.od.nih.gov/scientific-sharing/policy-oversight/

96. Terra Team. Terra Home Page. Terra, https://app.terra.bio. 2021 [cited 12 Apr 2021]. Available: https://app.terra.bio

97. Clinical Sequence Evidence-Generating Research Consortium. CSER Research Materials. CSER Website, https://cser-consortium.org/cser-research-materials. 2018. Available: https://cser-consortium.org/cser-research-materials

98. Nutter B, Lane S. redcapAPI: accessing data from REDCap projects using the API. CRAN, https://cran.r-project.org/web/packages/redcapAPI/index.html. 2018.

99. Office for Civil Rights (OCR). Summary of the HIPAA privacy rule. HHS.gov, https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html. 2008 [cited 20 Aug 2021]. Available: https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html

100. dbGaP Team. dbGaP study submission guide. NCBI, https://www.ncbi.nlm.nih.gov/gap/docs/submissionguide/. 2022 [cited 1 Mar 2022]. Available: https://www.ncbi.nlm.nih.gov/gap/docs/submissionguide/

101. NIH Office of Science Policy. Points to Consider in Developing Effective Data Use Limitation Statements. In: NIH Office of Science Policy [Internet]. 2015. Available: https://sharing.nih.gov/sites/default/files/flmngr/NIH_PTC_in_Developing_DUL_Statements.pdf

102. Dyke SOM, Philippakis AA, Rambla De Argila J, Paltoo DN, Luetkemeier ES, Knoppers BM, et al. Consent Codes: Upholding Standard Data Use Conditions. PLoS Genet. 2016;12: e1005772.

103. Langmead B, Nellore A. Cloud computing for genomic data analysis and collaboration. Nat Rev Genet. 2018;19: 325.

104. National Human Genome Research Institute (NHGRI). Notice for Use of Cloud Computing Services for Storage and Analysis of Controlled-Access Data Subject to the NIH Genomic Data Sharing (GDS) Policy. NIH Grants, https://grants.nih.gov/grants/guide/notice-files/not-od-15-086.html. 2015. Available: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-086.html

105. Carter AB. Considerations for Genomic Data Privacy and Security when Working in the Cloud. J Mol Diagn. 2019;21: 542–552.

106. Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, et al. Jupyter Notebooks - a publishing format for reproducible computational workflows. In: Loizides F, Scmidt B, editors. Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press; 2016. pp. 87–90.

107. RStudio Team. RStudio: Integrated Development Environment for R. Boston, MA: RStudio, PBC; 2021. Available: http://www.rstudio.com/

108. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 2018;46: W537–W544.

109. Van der Auwera GA, O'Connor BD. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. O'Reilly Media; 2020.

110. Nielsen J. Iterative user-interface design. Computer . 1993;26: 32–41.

111. Gould JD, Lewis C. Designing for usability: key principles and what designers think. Commun ACM. 1985;28: 300–311.

112. Tenopir C, Rice NM, Allard S, Baird L, Borycz J, Christian L, et al. Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. PLoS One. 2020;15: e0229003.

113. Danchev V, Min Y, Borghi J, Baiocchi M, Ioannidis JPA. Evaluation of Data Sharing

After Implementation of the International Committee of Medical Journal Editors Data Sharing Statement Requirement. JAMA Netw Open. 2021;4: e2033972.

114. Wolf LE, Hammack CM, Brown EF, Brelsford KM, Beskow LM. Protecting Participants in Genomic Research: Understanding the "Web of Protections" Afforded by Federal and State Law. J Law Med Ethics. 2020;48: 126–141.

115. US Department of Health & Human Services Office for Human Research Protections, The Secretary's Advisory Committee on Human Research Protections (SACHRP). Attachment D: FAQ's Terms and Recommendations on Informed Consent and Research Use of Biospecimens. HHS.gov, https://www.hhs.gov/ohrp/sachrp-committee/recommendations/2011-october-13-letter-attachment-d/index.html. 2011. Available: https://www.hhs.gov/ohrp/sachrp-committee/recommendations/2011-october-13-letter-attachment-d/index.html

116. Evans BJ, Jarvik GP. Impact of HIPAA's minimum necessary standard on genomic data sharing. Genetics in Medicine. 2018. pp. 531–535. doi:10.1038/gim.2017.141

117. Fisher CB, Layman DM. Genomics, Big Data, and Broad Consent: a New Ethics Frontier for Prevention Science. Prev Sci. 2018;19: 871–879.

118. NIH Office of Science Policy. Genomic Data Sharing Policy FAQs: Consent for Broad Sharing. In: NIH Office of Science Policy [Internet]. 2015. Available: https://sharing.nih.gov/faqs#/genomic-data-sharing-policy.htm

119. Norstad M, Outram S, Brown JEH, Zamora AN, Koenig BA, Risch N, et al. The difficulties of broad data sharing in genomic medicine: Empirical evidence from diverse participants in prenatal and pediatric clinical genomics research. Genet Med. 2022;24: 410–418.

120. Fox K. The illusion of inclusion - the "all of us" research program and indigenous peoples' DNA. N Engl J Med. 2020;383: 411–413.

121. Kaye J, Whitley EA, Lund D, Morrison M, Teare H, Melham K. Dynamic consent: a patient interface for twenty-first century research networks. Eur J Hum Genet. 2015;23: 141–146.

122. Mc Cartney AM, Anderson J, Liggins L, Hudson ML, Anderson MZ, TeAika B, et al. Balancing openness with Indigenous data sovereignty: An opportunity to leave no one behind in the journey to sequence all of life. Proc Natl Acad Sci U S A. 2022;119. doi:10.1073/pnas.2115860119

123. OECD. Co-ordination and support of international research data networks. Organisation for Economic Co-Operation and Development (OECD); 2017 Dec. doi:10.1787/e92fa89e-en

124. Rider EA, Kurtz S, Slade D, Longmaid HE 3rd, Ho M-J, Pun JK-H, et al. The International Charter for Human Values in Healthcare: an interprofessional global collaboration to enhance values and communication in healthcare. Patient Educ Couns. 2014;96: 273–280.

125. Friedman CP, Rubin JC, Sullivan KJ. Toward an Information Infrastructure for Global Health Improvement. Yearb Med Inform. 2017;26: 16–23.

126. Bodenheimer T, Sinsky C. From triple to quadruple aim: care of the patient requires care of the provider. Ann Fam Med. 2014;12: 573–576.

127. Etheredge LM. A rapid-learning health system. Health Aff . 2007;26: w107–18.

128. Institute of Medicine (US) Roundtable on Evidence-Based Medicine. The Learning Healthcare System: Workshop Summary. Olsen L, Aisner D, McGinnis JM, editors. Washington (DC): National Academies Press (US); 2007.

129. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409: 860–921.

130. Passarge E. Origins of human genetics. A personal perspective. Eur J Hum Genet. 2021;29: 1038–1044.

131. Ginsburg G. Medical genomics: Gather and use genetic data in health care. Nature. 2014;508: 451–453.

132. Davis FD, Williams MS, Stametz RA. Geisinger's effort to realize its potential as a learning health system: A progress report. Learn Health Syst. 2021;5: e10221.

133. Psek WA, Stametz RA, Bailey-Davis LD, Davis D, Darer J, Faucett WA, et al. Operationalizing the Learning Health Care System in an Integrated Delivery System. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2015. p. 6. doi:10.13063/2327-9214.1122

134. Atkins D, Kilbourne AM, Shulkin D. Moving From Discovery to System-Wide Change: The Role of Research in a Learning Health Care System: Experience from Three Decades of Health Systems Research in the Veterans Health Administration. Annu Rev Public Health. 2017;38: 467–487.

135. Enticott J, Braaf S, Johnson A, Jones A, Teede HJ. Leaders' perspectives on learning health systems: a qualitative study. BMC Health Services Research. 2020. doi:10.1186/s12913-020-05924-w

136. Vargas N, Lebrun-Harris LA, Weinberg J, Dievler A, Felix KL. Qualitative Perspective on the Learning Health System: How the Community Health Applied Research Network Paved the Way for Research in Safety-Net Settings. Prog Community Health Partnersh.

2018;12: 329–339.

137. Al-Busaidi ZQ. Qualitative research and its uses in health care. Sultan Qaboos Univ Med J. 2008;8: 11–19.

138. Rieger KL. Discriminating among grounded theory approaches. Nurs Inq. 2019;26: e12261.

139. Moser A, Korstjens I. Series: Practical guidance to qualitative research. Part 3: Sampling, data collection and analysis. Eur J Gen Pract. 2018;24: 9–18.

140. McDonald N, Schoenebeck S, Forte A. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. Proc ACM Hum-Comput Interact. 2019;3: 1–23.

141. Tsai GJ, Chen AT, Garrett LT, Burke W, Bowen DJ, Shirts BH. Exploring relatives' perceptions of participation, ethics, and communication in a patient-driven study for hereditary cancer variant reclassification. J Genet Couns. 2020;29: 857–866.

142. Burla L, Knierim B, Barth J, Liewald K, Duetz M, Abel T. From text to codings: intercoder reliability assessment in qualitative content analysis. Nurs Res. 2008;57: 113–117.

143. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant. 2018;52: 1893–1907.

144. O'Connor C, Joffe H. Intercoder reliability in qualitative research: Debates and practical guidelines. Int J Qual Methods. 2020;19: 160940691989922.

145. Krippendorff K. Computing Krippendorff's Alpha-Reliability. 2011 [cited 17 Nov 2022]. Available: https://repository.upenn.edu/asc_papers/43/

146. Shabankhani B, Charati JY, Shabankhani K, Cherati SK, Bizhan M. Survey of agreement between raters for nominal data using krippendorff's Alpha. [cited 8 Dec 2022]. Available: https://archivepp.com/storage/models/article/Ax8JTUZ2hr0L7lYSZrZnCfnjHXEPcXw4I4idWZIgrqjsVNTOhaW0dRQ6x4T6/survey-of-agreement-between-raters-for-nominal-data-using-krippendorffs-alpha.pdf

147. JGraph. diagrams.net, draw.io. 2021. Available: https://www.diagrams.net/

148. Nordo AH, Levaux HP, Becnel LB, Galvez J, Rao P, Stem K, et al. Use of EHRs data for clinical research: Historical progress and current applications. Learn Health Syst. 2019;3: e10076.

149. Farmer R, Mathur R, Bhaskaran K, Eastwood SV, Chaturvedi N, Smeeth L. Promises and

pitfalls of electronic health record analysis. Diabetologia. 2018;61: 1241–1248.

150. of Health USD, Services H, Others. The feasibility of using electronic health data for research on small populations. Population# 1: Asian-American subpopulations. 2013.

151. Bots SH, Groenwold RHH, Dekkers OM. Using electronic health record data for clinical research: a quick guide. Eur J Endocrinol. 2022;186: E1–E6.

152. Landry LG, Ali N, Williams DR, Rehm HL, Bonham VL. Lack Of Diversity In Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice. Health Aff . 2018;37: 780–785.

153. Amendola LM, Jarvik GP, Leo MC, McLaughlin HM, Akkari Y, Amaral MD, et al. Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. Am J Hum Genet. 2016;99: 247.

154. Amendola LM, Muenzen K, Biesecker LG, Bowling KM, Cooper GM, Dorschner MO, et al. Variant Classification Concordance using the ACMG-AMP Variant Interpretation Guidelines across Nine Genomic Implementation Research Studies. Am J Hum Genet. 2020;107: 932–941.

155. Phillips KA, Deverka PA, Sox HC, Khoury MJ, Sandy LG, Ginsburg GS, et al. Making genomic medicine evidence-based and patient-centered: a structured review and landscape analysis of comparative effectiveness research. Genet Med. 2017;19: 1081–1091.

156. Peterson JF, Roden DM, Orlando LA, Ramirez AH, Mensah GA, Williams MS. Building evidence and measuring clinical outcomes for genomic medicine. Lancet. 2019;394: 604–610.

157. Lu CY, Williams MS, Ginsburg GS, Toh S, Brown JS, Khoury MJ. A proposed approach to accelerate evidence generation for genomic-based technologies in the context of a learning health system. Genet Med. 2018;20: 390–396.

158. Charon JM. Symbolic Interactionism: An Introduction, an Interpretation, an Integration. Prentice-Hall; 1985.

159. Bonham VL, Green ED. The genomics workforce must become more diverse: a strategic imperative. Am J Hum Genet. 2021;108: 3–7.

160. Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev. 2009;19: 212–219.

161. Lappalainen T, MacArthur DG. From variant to function in human disease genetics. Science. 2021;373: 1464–1468.

162. Loos RJF. 15 years of genome-wide association studies and no signs of slowing down. Nat Commun. 2020;11: 5900.

163. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med. 2013;15: 761–771.

164. Denny JC, Crawford DC, Ritchie MD, Bielinski SJ, Basford MA, Bradford Y, et al. Variants Near FOXE1 Are Associated with Hypothyroidism and Other Thyroid Conditions: Using Electronic Medical Records for Genome- and Phenome-wide Studies. The American Journal of Human Genetics. 2011. pp. 529–542. doi:10.1016/j.ajhg.2011.09.008

165. Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Inform Assoc. 2012;19: 212–218.

166. Crosslin DR, The electronic Medical Records and Genomics (eMERGE) Network, McDavid A, Weston N, Nelson SC, Zheng X, et al. Genetic variants associated with the white blood cell count in 13,923 subjects in the eMERGE Network. Human Genetics. 2012. pp. 639–652. doi:10.1007/s00439-011-1103-9

167. Kullo IJ, Ding K, Shameer K, McCarty CA, Jarvik GP, Denny JC, et al. Complement Receptor 1 Gene Variants Are Associated with Erythrocyte Sedimentation Rate. The American Journal of Human Genetics. 2011. pp. 131–138. doi:10.1016/j.ajhg.2011.05.019

168. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. Genet Med. 2002;4: 45–61.

169. Sella G, Barton NH. Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies. Annu Rev Genomics Hum Genet. 2019;20: 461–493.

170. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014;95: 5–23.

171. Gibson G. Rare and common variants: twenty arguments. Nat Rev Genet. 2012;13: 135–145.

172. Gilissen C, Hoischen A, Brunner HG, Veltman JA. Disease gene identification strategies for exome sequencing. Eur J Hum Genet. 2012;20: 490–497.

173. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. 2010;11: 415–425.

174. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nature Reviews Methods Primers. 2021;1: 1–21.

175. Ikegawa S. A short history of the genome-wide association study: where we were and where we are going. Genomics Inform. 2012;10: 220–225.

176. Kitsios GD, Zintzaras E. Genome-wide association studies: hypothesis-“free” or “engaged”? Transl Res. 2009;154: 161–164.

177. Ioannidis JPA, Trikalinos TA, Khoury MJ. Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. Am J Epidemiol. 2006;164: 609–614.

178. Wei W-Q, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome Medicine. 2015. doi:10.1186/s13073-015-0166-y

179. O’Sullivan JW, Ioannidis JPA. Reproducibility in the UK biobank of genome-wide significant signals discovered in earlier genome-wide association studies. Sci Rep. 2021;11: 18625.

180. Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. Int J Med Inform. 2008;77: 291–304.

181. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc. 2013;20: e147–54.

182. Kuijper EJ, Coignard B, Tüll P, ESCMID Study Group for Clostridium difficile, EU Member States, European Centre for Disease Prevention and Control. Emergence of Clostridium difficile-associated disease in North America and Europe. Clin Microbiol Infect. 2006;12 Suppl 6: 2–18.

183. McDonald LC, Killgore GE, Thompson A. An Epidemic, Toxin Gene–Variant Strain of Clostridium difficile. England Journal of …. 2005. Available: https://www.nejm.org/doi/full/10.1056/nejmoa051590

184. O’Connor JR, Johnson S, Gerding DN. Clostridium difficile infection caused by the epidemic BI/NAP1/027 strain. Gastroenterology. 2009. Available: https://www.sciencedirect.com/science/article/pii/S0016508509003606

185. Aas J, Gessert CE, Bakken JS. Recurrent Clostridium difficile Colitis: Case Series Involving 18 Patients Treated with Donor Stool Administered via a Nasogastric Tube. Clin Infect Dis. 2003. Available: https://academic.oup.com/cid/article-abstract/36/5/580/452381

186. Guo B, Harstall C, Louie T, van Zanten SV, Dieleman LA. Systematic review: faecal transplantation for the treatment of Clostridium difficile-associated disease. Alimentary Pharmacology & Therapeutics. 2012. pp. 865–875. doi:10.1111/j.1365-2036.2012.05033.x

187. McDonald LC, Clifford McDonald L, Gerding DN, Johnson S, Bakken JS, Carroll KC, et al. Clinical Practice Guidelines for Clostridium difficile Infection in Adults and Children: 2017 Update by the Infectious Diseases Society of America (IDSA) and Society for Healthcare Epidemiology of America (SHEA). Clinical Infectious Diseases. 2018. pp. 987–994. doi:10.1093/cid/ciy149

188. Crobach MJT, Vernon JJ, Loo VG, Kong LY, Péchiné S, Wilcox MH, et al. Understanding Clostridium difficile Colonization. Clin Microbiol Rev. 2018;31. doi:10.1128/CMR.00021-17

189. Pépin J, Saheb N, Coulombe M-A, Alary M-E, Corriveau M-P, Authier S, et al. Emergence of Fluoroquinolones as the Predominant Risk Factor for Clostridium difficile–Associated Diarrhea: A Cohort Study during an Epidemic in Quebec. Clin Infect Dis. 2005;41: 1254–1260.

190. Lalla F de, de Lalla F, Privitera G, Ortisi G, Rizzardini G, Santoro D, et al. Third generation cephalosporins as a risk factor for Clostridium difficile-associated disease: a four-year survey in a general hospital. Journal of Antimicrobial Chemotherapy. 1989. pp. 623–631. doi:10.1093/jac/23.4.623

191. Bignardi GE. Risk factors for Clostridium difficile infection. J Hosp Infect. 1998;40: 1–15.

192. Fekete T. Concurrent PPIs and antibiotics for incident C. difficile infection were associated with increased risk for recurrent infection. Ann Intern Med. 2010. Available: https://www.acpjournals.org/doi/full/10.7326/0003-4819-153-8-201010190-02012

193. Wurfel MM, Hawn TR. Genetic variants associated with susceptibility to Helicobacter pylori. JAMA: the journal of the American Medical Association. 2013. p. 976.

194. Flores J, Okhuysen PC. Genetics of susceptibility to infection with enteric pathogens. Curr Opin Infect Dis. 2009;22: 471–476.

195. Ananthakrishnan AN, Oxford EC, Nguyen DD, Sauk J, Yajnik V, Xavier RJ. Genetic risk factors for Clostridium difficile infection in ulcerative colitis. Aliment Pharmacol Ther. 2013;38: 522–530.

196. Apewokin S, Lee JY, Goodwin JA, McKelvey KD, Stephens OW, Zhou D, et al. Host genetic susceptibility to Clostridium difficile infections in patients undergoing autologous stem cell transplantation: a genome-wide association study. Support Care Cancer. 2018;26: 3127–3134.

197. Shen J, Mehrotra DV, Dorr MB, Zeng Z, Li J, Xu X, et al. Genetic Association Reveals Protection against Recurrence of Clostridium difficile Infection with Bezlotoxumab Treatment. mSphere. 2020. doi:10.1128/msphere.00232-20

198. Jiang Z-D, DuPont HL, Garey K, Price M, Graham G, Okhuysen P, et al. A common polymorphism in the interleukin 8 gene promoter is associated with Clostridium difficile diarrhea. Am J Gastroenterol. 2006;101: 1112–1116.

199. Garey KW, Jiang Z-D, Ghantoji S, Tam VH, Arora V, Dupont HL. A common polymorphism in the interleukin-8 gene promoter is associated with an increased risk for recurrent Clostridium difficile infection. Clin Infect Dis. 2010;51: 1406–1410.

200. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc. 2016;23: 1046–1052.

201. Carrell D, Denny J. Clostridium Difficile Colitis. PheKB, 2012. Available: https://phekb.org/phenotype/70

202. Carroll KC. Tests for the diagnosis of Clostridium difficile infection: the next generation. Anaerobe. 2011;17: 170–174.

203. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48: 1284–1287.

204. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016;48: 1443–1448.

205. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4: 7.

206. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81: 559–575.

207. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010;26: 2336–2337.

208. Zheng X. HIBAG: an R Package for HLA Genotype Imputation with Attribute Bagging. 2014. Available: https://bioconductor.riken.jp/packages/3.2/bioc/vignettes/HIBAG/inst/doc/HIBAG_Tutorial.pdf

209. Habets THPM, Hepkema BG, Kouprie N, Schnijderberg MCA, van Smaalen TC, Bungener LB, et al. The prevalence of antibodies against the HLA-DRB3 protein in kidney transplantation and the correlation with HLA expression. PLoS One. 2018;13: e0203381.

210. Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, et al. Gene map

of the extended human MHC. Nat Rev Genet. 2004;5: 889–899.

211. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012;22: 1790–1797.

212. Trowsdale J. “Both man & bird & beast”: comparative organization of MHC genes. Immunogenetics. 1995. Available: https://link.springer.com/article/10.1007/BF00188427

213. de Bakker PIW, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet. 2006;38: 1166–1172.

214. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics. 2015;31: 3555–3557.

215. Chaplin DD. Overview of the immune response. J Allergy Clin Immunol. 2010;125: S3–23.

216. Kasahara M, Klein D, Vincek V, Sarapata DE, Klein J. Comparative anatomy of the primate major histocompatibility complex DR subregion: evidence for combinations of DRB genes conserved across species. Genomics. 1992;14: 340–349.

217. Penn DJ, Damjanovich K, Potts WK. MHC heterozygosity confers a selective advantage against multiple-strain infections. Proc Natl Acad Sci U S A. 2002;99: 11260–11264.

218. Dunstan SJ, Hue NT, Han B, Li Z, Tram TTB, Sim KS, et al. Variation at HLA-DRB1 is associated with resistance to enteric fever. Nat Genet. 2014;46: 1333–1336.

219. Horn GT, Bugawan TL, Long CM, Manos MM, Erlich HA. Sequence analysis of HLA class II genes from insulin-dependent diabetic individuals. Hum Immunol. 1988;21: 249–263.

220. Wordsworth P, Pile KD, Buckely JD, Lanchbury JS, Ollier B, Lathrop M, et al. HLA heterozygosity contributes to susceptibility to rheumatoid arthritis. Am J Hum Genet. 1992;51: 585–591.

221. Sospedra M, Muraro PA, Stefanová I, Zhao Y, Chung K, Li Y, et al. Redundancy in Antigen-Presenting Function of the HLA-DR and -DQ Molecules in the Multiple Sclerosis-Associated HLA-DR2 Haplotype. The Journal of Immunology. 2006. pp. 1951–1961. doi:10.4049/jimmunol.176.3.1951

222. Yamamoto-Furusho JK, Rodríguez-Bores L, Granados J. HLA-DRB1 alleles are associated with the clinical course of disease and steroid dependence in Mexican patients with ulcerative colitis. Colorectal Disease. 2010. pp. 1231–1235. doi:10.1111/j.1463-

1318.2009.02025.x

223. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013;45: 1452–1458.

224. Mahdi BM. Role of HLA typing on Crohn's disease pathogenesis. Ann Med Surg (Lond). 2015;4: 248–253.

225. Fogdell A, Hillert J, Sachs C, Olerup O. The multiple sclerosis- and narcolepsy-associated HLA class II haplotype includes the DRB5*0101 allele. Tissue Antigens. 1995;46: 333–336.

226. Catanzaro JR, Strauss JD, Bielecka A, Porto AF, Lobo FM, Urban A, et al. IgA-deficient humans exhibit gut microbiota dysbiosis despite secretion of compensatory IgM. Sci Rep. 2019;9: 13574.

227. Lycke NY, Bemark M. The regulation of gut mucosal IgA B-cell responses: recent developments. Mucosal Immunol. 2017;10: 1361–1374.

228. Kubinak JL, Zac Stephens W, Soto R, Petersen C, Chiaro T, Gogokhia L, et al. MHC variation sculpts individualized microbial communities that control susceptibility to enteric infection. Nature Communications. 2015. doi:10.1038/ncomms9642

229. Khan AA, Yurkovetskiy L, O'Grady K, Pickard JM, de Pooter R, Antonopoulos DA, et al. Polymorphic Immune Mechanisms Regulate Commensal Repertoire. Cell Rep. 2019;29: 541–550.e4.

230. Mielcarz DW, Kasper LH. The gut microbiome in multiple sclerosis. Curr Treat Options Neurol. 2015;17: 344.

231. Kirby TO, Ochoa-Repáraz J. The Gut Microbiome in Multiple Sclerosis: A Potential Therapeutic Avenue. Med Sci (Basel). 2018;6. doi:10.3390/medsci6030069

232. Ventura RE, Iizumi T, Battaglia T, Liu M, Perez-Perez GI, Herbert J, et al. Gut microbiome of treatment-naïve MS patients of different ethnicities early in disease course. Sci Rep. 2019;9: 16396.

233. Seekatz AM, Young VB. Clostridium difficile and the microbiota. J Clin Invest. 2014;124: 4182–4189.

234. Ourth DD. Neutralization of diphtheria toxin by human immunoglobulin classes and subunits. Immunochemistry. 1974;11: 223–225.

235. Wilcox MH, Gerding DN, Poxton IR, Kelly C, Nathan R, Birch T, et al. Bezlotoxumab for Prevention of Recurrent Clostridium difficile Infection. N Engl J Med. 2017;376: 305–

317.

236. Gupta SB, Mehta V, Dubberke ER, Zhao X, Dorr MB, Guris D, et al. Antibodies to Toxin B Are Protective Against Clostridium difficile Infection Recurrence. Clin Infect Dis. 2016;63: 730–734.

237. Rees WD, Steiner TS. Adaptive immune response to Clostridium difficile infection: A perspective for prevention and therapy. Eur J Immunol. 2018;48: 398–406.

238. Riggs AD. X inactivation, differentiation, and DNA methylation. Cytogenet Cell Genet. 1975;14: 9–25.

239. Holliday R, Pugh JE. DNA modification mechanisms and gene activity during development. Science. 1975;187: 226–232.

240. Shoemaker R, Deng J, Wang W, Zhang K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. Genome Res. 2010;20: 883–889.

241. Robertson KD. DNA methylation and human disease. Nat Rev Genet. 2005;6: 597–610.

242. Do C, Dumont ELP, Salas M, Castano A, Mujahed H, Maldonado L, et al. Allele-specific DNA methylation is increased in cancers and its dense mapping in normal plus neoplastic cells increases the yield of disease-associated regulatory SNPs. Genome Biology. 2020. doi:10.1186/s13059-020-02059-3

243. Kular L, Liu Y, Ruhrmann S, Zheleznyakova G, Marabita F, Gomez-Cabrero D, et al. DNA methylation as a mediator of HLA-DRB1*15:01 and a protective variant in multiple sclerosis. Nature Communications. 2018. doi:10.1038/s41467-018-04732-5

244. Do C, Lang CF, Lin J, Darbary H, Krupska I, Gaba A, et al. Mechanisms and Disease Associations of Haplotype-Dependent Allele-Specific DNA Methylation. Am J Hum Genet. 2016;98: 934–955.

245. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A. 2010;107: 21931–21936.

246. Flemming K, Noyes J. Qualitative Evidence Synthesis: Where Are We at? International Journal of Qualitative Methods. 2021;20: 1609406921993276.

247. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.3 (updated February 2022). Cochrane; 2022.

248. Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. ACP J Club. 1995;123: A12–3.

249. Thomas J, Brunton J, Graziosi S (2010). EPPI-Reviewer 4.0: software for research synthesis. EPPI-Centre Software, London: Social Science Research Unit, Institute of Education, University of London.

250. Popay, Roberts, Sowden, Petticrew. Guidance on the conduct of narrative synthesis in systematic reviews. A product from the. Available: https://www.academia.edu/download/39246301/02e7e5231e8f3a6183000000.pdf

251. Weiss CH, Weiss CH. Evaluation: Methods for studying programs and policies. Pearson College Division; 1998.

252. White H. Theory-based systematic reviews. Journal of Development Effectiveness. 2018;10: 17–38.

253. Facey K, Ploug Hansen H, Single A. Qualitative evidence synthesis. In: Facey K, Ploug Hansen H, Single A, editors. Patient Involvement in Health Technology Assessment. Singapore: Adis; 2017. pp. 187–199.

254. Carroll C. Qualitative evidence synthesis to improve implementation of clinical guidelines. BMJ. 2017;356: j80.

255. Carroll C, Booth A, Cooper K. A worked example of“ best fit” framework synthesis: a systematic review of views concerning the taking of some potential chemopreventive agents. BMC Med Res Methodol. 2011;11: 1–9.

256. Carroll C, Booth A, Leaviss J, Rick J. “Best fit” framework synthesis: refining the method. BMC Med Res Methodol. 2013;13: 1–16.

257. Glaser BG. The Constant Comparative Method of Qualitative Analysis. Soc Probl. 1965;12: 436–445.

258. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J Clin Epidemiol. 2009;62: e1–34.

259. Simon MA, Trosman JR, Rapkin B, Rittner SS, Adetoro E, Kirschner MC, et al. Systematic Patient Navigation Strategies to Scale Breast Cancer Disparity Reduction by Improved Cancer Prevention and Care Delivery Processes. JCO Oncol Pract. 2020;16: e1462–e1470.

260. Shaikh AR, Butte AJ, Schully SD, Dalton WS, Khoury MJ, Hesse BW. Collaborative biomedicine in the age of big data: the case of cancer. J Med Internet Res. 2014;16: e101.

261. Wiley LK, Tarczy-Hornoch P, Denny JC, Freimuth RR, Overby CL, Shah N, et al. Harnessing next-generation informatics for personalizing medicine: a report from AMIA’s

2014 Health Policy Invitational Meeting. J Am Med Inform Assoc. 2016;23: 413–419.

262. David SP, Johnson SG, Berger AC, Feero WG, Terry SF, Green LA, et al. Making Personalized Health Care Even More Personalized: Insights From Activities of the IOM Genomics Roundtable. Ann Fam Med. 2015;13: 373–380.

263. Abernethy A, Abrahams E, Barker A, Buetow K, Burkholder R, Dalton WS, et al. Turning the tide against cancer through sustained medical innovation: the pathway to progress. Clin Cancer Res. 2014;20: 1081–1086.

264. Institute of Medicine, Roundtable on Value and Science-Driven Health Care. Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care: Workshop Series Summary. National Academies Press; 2011.

265. Key KD, Lewis EY. Sustainable community engagement in a constantly changing health system. Learn Health Syst. 2018;2. doi:10.1002/lrh2.10053

266. Wallace PJ, Shah ND, Dennen T, Bleicher PA, Crown WH. Optum Labs: building a novel node in the learning health care system. Health Aff . 2014;33: 1187–1194.

267. Williams MS. Early Lessons from the Implementation of Genomic Medicine Programs. Annu Rev Genomics Hum Genet. 2019;20: 389–411.

268. Schwartz MLB, McCormick CZ, Lazzeri AL, Lindbuchler DM, Hallquist MLG, Manickam K, et al. A Model for Genome-First Care: Returning Secondary Genomic Findings to Participants and Their Healthcare Providers in a Large Research Cohort. Am J Hum Genet. 2018;103: 328–337.

269. Potter D, Brothers R, Kolacevski A, Koskimaki JE, McNutt A, Miller RS, et al. Development of CancerLinQ, a Health Information Learning Platform From Multiple Electronic Health Record Systems to Support Improved Quality of Care. JCO Clin Cancer Inform. 2020;4: 929–937.

270. Etheredge LM. Medicare's future: cancer care. Health Aff . 2009;28: 148–159.

271. Etheredge LM. Rapid learning: a breakthrough agenda. Health Aff . 2014;33: 1155–1162.

272. Ginsburg GS, Phillips KA. Precision Medicine: From Science To Value. Health Aff . 2018;37: 694–701.

273. Hirsch BR, Abernethy AP. Leveraging informatics, mobile health technologies and biobanks to treat each patient right. Per Med. 2012;9: 849–857.

274. Hindorff LA, Bonham VL, Ohno-Machado L. Enhancing diversity to reduce health information disparities and build an evidence base for genomic medicine. Per Med.

2018;15: 403–412.

275. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. Health Aff . 2014;33: 1163–1170.

276. McGinnis JM, Fineberg HV, Dzau VJ. Advancing the Learning Health System. N Engl J Med. 2021;385: 1–5.

277. McInnes G, Sharo AG, Koleske ML, Brown JEH, Norstad M, Adhikari AN, et al. Opportunities and challenges for the computational interpretation of rare variation in clinically important genes. Am J Hum Genet. 2021;108: 535–548.

278. Scollen S, Page A, Wilson J. From the Data on Many, Precision Medicine for "One": The Case for Widespread Genomic Data Sharing. Biomed Hub. 2017;2: 104–110.

279. Trifiletti DM, Showalter TN. Big Data and Comparative Effectiveness Research in Radiation Oncology: Synergy and Accelerated Discovery. Front Oncol. 2015;5: 274.

280. Yu PP, Hoffman MA, Hayes DF. Biomarkers and oncology: the path forward to a learning health system. Arch Pathol Lab Med. 2015;139: 451–456.

281. Glasgow RE, Kwan BM, Matlock DD. Realizing the full potential of precision health: The need to include patient-reported health behavior, mental health, social determinants, and patient preferences data. J Clin Transl Sci. 2018;2: 183–185.

282. Chambers DA, Feero WG, Khoury MJ. Convergence of Implementation Science, Precision Medicine, and the Learning Health Care System: A New Model for Biomedical Research. JAMA. 2016;315: 1941–1942.

283. Jones RD, Krenz C, Griffith KA, Spence R, Bradbury AR, De Vries R, et al. Patient Experiences, Trust, and Preferences for Health Data Sharing. JCO Oncol Pract. 2022;18: e339–e350.

284. Jones RD, Krenz C, Gornick M, Griffith KA, Spence R, Bradbury AR, et al. Patient Preferences Regarding Informed Consent Models for Participation in a Learning Health Care System for Oncology. JCO Oncol Pract. 2020;16: e977–e990.

285. Khalifa A, Mason CC, Garvin JH, Williams MS, Del Fiol G, Jackson BR, et al. A qualitative study of prevalent laboratory information systems and data communication patterns for genetic test reporting. Genet Med. 2021;23: 2171–2177.

286. Khalifa A, Mason CC, Garvin JH, Williams MS, Del Fiol G, Jackson BR, et al. A qualitative investigation of biomedical informatics interoperability standards for genetic test reporting: benefits, challenges, and motivations from the testing laboratory's perspective. Genet Med. 2021;23: 2178–2185.

287. Hsu W, Gonzalez NR, Chien A, Pablo Villablanca J, Pajukanta P, Viñuela F, et al. An integrated, ontology-driven approach to constructing observational databases for research. J Biomed Inform. 2015;55: 132–142.

288. Kehl KL, Elmarakeby H, Nishino M, Van Allen EM, Lepisto EM, Hassett MJ, et al. Assessment of Deep Natural Language Processing in Ascertaining Oncologic Outcomes From Radiology Reports. JAMA Oncol. 2019;5: 1421–1429.

289. Preston CG, Wright MW, Madhavrao R, Harrison SM, Goldstein JL, Luo X, et al. ClinGen Variant Curation Interface: a variant classification platform for the application of evidence criteria from ACMG/AMP guidelines. Genome Med. 2022;14: 6.

290. Finlayson SG, Levy M, Reddy S, Rubin DL. Toward rapid learning in cancer treatment selection: An analytical engine for practice-based clinical data. J Biomed Inform. 2016;60: 104–113.

291. Mandl KD, Glauser T, Krantz ID, Avillach P, Bartels A, Beggs AH, et al. The Genomics Research and Innovation Network: creating an interoperable, federated, genomics learning system. Genet Med. 2020;22: 371–380.

292. Yang U-C, Hsiao T-H, Lin C-H, Lee W-J, Lee Y-S, Fann YC. Integrative LHS for precision medicine research: A shared NIH and Taiwan CIMS experience. Learn Health Syst. 2019;3: e10071.

293. Nwaru BI, Friedman C, Halamka J, Sheikh A. Can learning health systems help organisations deliver personalised care? BMC Med. 2017;15: 177.

294. Bubela T, Genuis SK, Janjua NZ, Krajden M, Mittmann N, Podolak K, et al. Medical Information Commons to Support Learning Healthcare Systems: Examples From Canada. J Law Med Ethics. 2019;47: 97–105.

295. Chuong KH, Mack DR, Stintzi A, O'Doherty KC. Human Microbiome and Learning Healthcare Systems: Integrating Research and Precision Medicine for Inflammatory Bowel Disease. OMICS. 2018;22: 119–126.

296. Braithwaite J, Glasziou P, Westbrook J. The three numbers you need to know about healthcare: the 60-30-10 Challenge. BMC Med. 2020;18: 102.

297. Holm S, Ploug T. Big Data and Health Research—The Governance Challenges in a Mixed Data Economy. J Bioeth Inq. 2017;14: 515–525.