Shaping the future of sharing genomic data: Analysis of previously collected clinical requirements

Amira Shymbolatova

Department of Psychology, University of Twente

Faculty of Behavioural Sciences, Management and Social Sciences

PSY 202000384: BSc Thesis PSY

First Supervisor - Dr. Simone Borsci

Second Supervisor - Dr. Valeria Resendez Gomez

July 2nd, 2025

Abstract

This study investigates the alignment of the genomic data-sharing platform requirements between the theoretical expectation and clinical practice. It proposes a theoretical framework on the "work-as-imagined" and "work-as-done" for the evaluation of functional and non-functional requirements derived from a systematic literature review. We evaluated the set of requirements through a mixed-method approach. First, a quantitative survey ($N_{clin} = 30$) assessed the perceived importance and agreement on the 62 system requirements. Second, we conducted a thematic analysis of the requirements mapped through a workshop (N = 36). Results from the survey revealed that 91.9% of requirements were both rated as important and agreed upon, demonstrating strong validation of the theoretical model. However, several technical and visualization-related elements, such as data organization methods and graphical presentation of results presented disagreement. Despite the perceived importance, which indicates the variability in relevance to clinical practice. The qualitative findings provided further in-depth insights by highlighting infrastructural, ethical, and usability barriers, especially regarding federated computing, privacy, and user-centred design. Experts emphasized the lack of intuitive interfaces, challenges in accessing the data, and the need for platforms to better align with clinical workflow. Key inconsistencies, such as usability and implementation, underscore the necessity of adapting the model to actual clinical contexts and institutional restrictions, even if the model was extensively validated. Beyond its empirical findings, this study brings methodological contribution by including an exploratory clustering analysis to capture the requirements derived from unstructured data. It offers a reproducible framework for evaluating stakeholder perspectives toward the software design by the combined use of different analytical approaches, such as quantitative requirements, consensus analysis, and exploratory clustering. Additionally, this research contributes to the theoretical advancement of mental models in platform design and offers actionable guidance for developing clinically meaningful, secure, and interoperable genomic data-sharing systems.

Keywords: genomic data-sharing platform, genomic data exchange, platform requirements, mixed methods, clinician perspectives, mental models

Acknowledgements

I would like to express my deepest gratitude to Dr. Simone Borsci for his guidance throughout the development of this thesis. The insightful feedback and support contributed greatly to the quality and the direction of my work. I would also like to extend my sincere thanks to Dr. Valeria Resendez Gomez, whose assistance went much beyond academic mentoring. Her willingness to answer questions, offer thoughtful advice, and provide emotional encouragement during challenging moments was instrumental in the successful completion of this thesis.

Additionally, I want to express my gratitude to Vincent for his collaborative spirit and valuable contributions during the project. It was gratifying and productive to work with him, and his input greatly enhanced the outcome of this research.

Finally, I would like to put forward my most sincere appreciation for the people that matter in my world: my family and my friends. Without them, who supported me all along my journey and looked upon me with belief, I would not have had the confidence to carry this through to graduation.

Table of Contents

List of Tables

List of Figures

Introduction

Genomic data has made way for new paradigms in the prevention, diagnosis, and treatment of human diseases, being a cornerstone of precision medicine (León & Pastor, 2021). Unprecedented insights into drug response, illness susceptibility, and general health can be gained by analysing a person's genetic makeup (Ginsburg & Phillips, 2018). This allows for developing personalised treatment strategies that optimise therapeutic efficacy while minimising side effects. The paradigm changes call for effective and secure methods of exchanging genomic data so that medical professionals can obtain thorough genomic data for well-informed decision-making (Raza & Hall, 2017). However, the effective exchange of genomic data presents challenges in terms of data privacy and security, interoperability, standardisation, and ethical and legal concerns (Alzu'bi et al., 2014). Most of the existing integrative genomic data-sharing platforms face challenges hindering the seamless exchange of information between the experts (Xue et al., 2023). Despite the continuous challenges, genomic testing is gradually being incorporated into standard clinical care (Raza & Hall, 2017). This change highlights the importance of efficient data sharing among various researchers and healthcare providers (Raza & Hall, 2017).

A broader concept such as healthcare data exchange refers to the process by which various healthcare systems and organisations share electronic medical records and other health-related data (Haque et al., 2023). This includes test results, prescription drugs, treatment plans, and medical histories, in order to provide prompt access to thorough patient data. The exchange of data is an important step toward advancing healthcare delivery and enabling more informed clinical decisions. However, in the context of precision medicine, healthcare data alone is not enough. There is a need to include genomic data as well. Identifying genes and their role in disease development requires complex genomic variations analysis, which demands a multidisciplinary approach (Ma et al., 2024). This involves collaboration among a wide range of clinical experts forming multidisciplinary teams (MDTs), including oncologists, pathologists, geneticists, bioinformaticians, clinical pharmacologists, radiologists, and genomic scientists (Qian et al., 2023). To support these collaborative efforts, efficient platforms for health and genomic data exchange and analysis are essential. These platforms can enhance disease understanding, support personalised treatment strategies, and address challenges related to data accessibility in clinical practice.

However, clinical needs vary across experts. Therefore, it is important to identify and map the distinct requirements of MDTs, such as real-time data access, integration of diverse data sources, and collaborative decision support, as these needs can shape the design and functionality of data-sharing platforms (Credle, 2022). This study aims to understand what a genomic data-sharing platform should provide to enable efficient healthcare data exchange in precision medicine among clinical experts. The genomic data-sharing platform in this thesis refers explicitly to a digital infrastructure that stores, manages and facilitates the exchange of genomic data in clinical settings (Byrd et al., 2020). Including the perspectives of clinical experts, such as MDTs, in the design process is key to ensuring the platform is usable, actionable, and adaptable to the needs of clinicians.

European Union Regulations

Europe is actively evolving its approach to data management, particularly in the healthcare sector, by focusing on making data more accessible while ensuring security and privacy protections (WHO, 2021). This transformation is driven by the increasing awareness of health data's potential to promote precision medicine and enhance healthcare outcomes (Vayena et al., 2017). This approach mainly depends on the availability of comprehensive and integrated health data in order to spot trends, forecast hazards, and improve treatment plans. However, the effective use of precision medicine requires resolving challenges related to data privacy, security, and interpretation, along with ethical concerns and regulatory hurdles (Pandey & Gupta, 2024).

The European landscape is currently influenced by several key regulations and government frameworks that focus on the development of a safe and trusted environment for healthcare data exchange. The European Health Data Space (EHDS) is a fundamental component of the European health global strategy that is regulated by the European Parliament and Council (2022) (European Commission, 2024). In an environment as diverse and complicated as the European Union (EU), the EHDS ecosystem of regulations, standards, practices and technical infrastructures attempts to overcome the technological, legal and ethical challenges of exchanging interoperable health data. In this context, interoperability refers to the ability of various health systems and organisations to work together efficiently by exchanging data and knowledge (European Parliament and Council, 2022). The EHDS aims to promote research and innovation, improve healthcare delivery, and create a standard framework for accessing and sharing health data within the European Union.

While EHDS focuses on regulating the health data infrastructure, other regulatory frameworks focus on how this data can be used in advanced analytical applications, such as artificial intelligence (AI). AI provides opportunities for the improvement of personalised and efficient healthcare (Morley et al., 2022). However, the integration of AI brings attention to issues of data privacy, ethical transparency, and patient safety (Morley et al., 2022). AI governance in Europe places a strong emphasis on ethical, trustworthy and reliable technology advancement (Stix, 2021). Its strategies involve funding AI research, navigating member states' AI strategies and promoting integrated ideas of "trusted AI" and "human-centric AI". This approach entails a wide range of regulations and laws. The EU Artificial Intelligence Act (AI Act) reflects the EU's commitment to ethical AI regulation. The AI Act offers extensive legal guidelines for creating and implementing AI systems in the EU (Van Kolfschooten & Van Oirschot, 2024). The Act classifies AI systems based on their risk and imposes rules on accountability, transparency, and human control, especially for high-risk applications. For instance, AI systems that are used in the healthcare sector to manage health data are classified as high-risk applications (Van Kolfschooten & Van Oirschot, 2024). The requirements for this AI application are coordinated in a way that ensures robustness, explainability, and non-discrimination. In this context, explainability refers to the ability of healthcare professionals to understand the AI decision-making process.

To ensure safety and maximise privacy, the system design for data-sharing platforms complies with European Regulations, such as adherence to the General Data Protection Regulation (GDPR, 2016) and European Data Act (Data Act; European Parliament and Council, 2023). These frameworks provide explicit guidelines for data minimisation, consent, and legitimate data processing. These guidelines guarantee that individuals’ rights are protected while still permitting useful data usage.

The aforementioned legal and regulatory frameworks shape the operational models, such as the client-centred model and federated model, through which health data is managed and exchanged across systems (Li et al., 2024). The first model follows a centralised approach that stores all data in a single location, such as a hospital, making it available to authorised stakeholders and providing a handy source for inquiries. However, it requires a large upfront investment in infrastructure, staff, and data harmonisation, raising challenges related to patient privacy and data security. The second model stores data locally at its source, such as participating institutions or national systems, which enhances data security while analytical results are shared across networks

without exposing raw data. This model is being widely used in European research collaborations and data-sharing programs, especially where cross-border data exchange or privacy-preserving solutions are needed (Raab et al., 2023). The EHDS supports client-centred and federated models, however, the emphasis on interoperability aligns more with federated systems (Raab et al., 2023).

In order to potentially expand the system beyond the European Union and enable global operations, the US Department of Health and Human Services (Office for Civil Rights (OCR) & Office of the Secretary, Department of Health and Human Services, 2024) and the European Health Data Space (EHDS, European Parliament and Council of the European Union, 2024) should work together to discuss data exchange regulations. These organisations represent an ecosystem that addresses both technical and ethical issues related to data exchange in their respective fields (Marcus et al., 2022). These efforts will address the security and privacy of the patient's data while performing actions with the genomic data, like processing, exchanging and storing.

As Europe moves toward a more regulated but accessible health data space, the need for interoperable, high-quality health data continues to increase (EHDS, European Parliament and Council of the European Union, 2024). Artificial intelligence, advanced data analytics, and customised treatment plans all depend on vast amounts of accurate, timely, and diverse data. However, there are several challenges to overcome in order to integrate and leverage such data (Karacic, 2022).

General Challenges in Genomic Data Exchange

Genomic data is still typically gathered and examined independently by different factors: by illness, by institution, and by nation (Global Alliance for Genomics and Health, 2016). Based on the literature and prior studies to improve genetic data management, exchange and analysis, three main challenges should be addressed (Alzu'bi et al., 2014; Ceri & Pinoli, 2020; Xue et al., 2023).

The first challenge is managing genomic data while maintaining compliance with ethical, societal and political issues. Genetic research heavily relies on strict ethical and legal standards to ensure individual data privacy (Balagurunathan & Sethuraman, 2024). From the ethical side, there is a strong need for individuals to get informed consent to store their personal information in the data system. Therefore, the security aspect should focus on the issues related to hack attacks that could potentially leak the private genetic data of the patients (Balagurunathan & Sethuraman, 2024). Encouraging responsible genomic data sharing requires fostering trust between patients,

clinicians, and institutions (Tommel et al., 2023). Patients can be hesitant to share their health data if they fear it being misused in commercial, governmental, or AI-driven contexts, specifically when there is no transparency regarding the use of health data. These concerns are further complicated by inconsistencies in the broader regulatory landscape. Although the GDPR governs data protection in the EU, individual member states can interpret and apply some provisions in different ways (Pormeister, 2018; Molnár-Gábor & Korbel, 2020). The difference in the application of the regulations within the EU member states challenges the process of sharing genomic data across borders. This leads to the limited possibility of creating a unified research collaboration and genomic data-sharing platform (Pormeister, 2018). The fragmentation poses institutional and political barriers to the development of a reliable and collaborative European genomic research ecosystem.

The second challenge is the integration of enormous and diverse health data that is required to pursue precision medicine (Alzu'bi et al., 2014). The different types of data require specific analytical approaches, such as the computational capacity to handle it, the infrastructure to share the data, and data standardisation (Alzu'bi et al., 2014; Ceri & Pinoli, 2020). Stephens et al. (2015) predicted that by 2025, genomics will be one of the biggest databases, which will be 20-40 times bigger than the size of astronomical data. Genomic data itself mainly consists of 4 types: genomic sequences (DNA, RNA, protein) and gene expression profiles (Alzu'bi et al., 2014). However, genetic analysis does not facilitate the goal of precision medicine, which states that individualised treatment is based on genetics, physiology, and environment (Elhussein et al., 2024). In order to reach full utility in the development of precision medicine, it is required to combine it with other data types, such as socioeconomic data and electronic health records (EHRs). Integration of socioeconomic data can enhance the understanding of disparities in genetic disorders by providing information related to individuals' educational and income levels and predisposition to certain racial and ethnic minorities (Khoury et al., 2022). Integration of EHRs supports informed decision-making and personalised treatment plans by providing information on the patient's medical history (Robertson et al., 2024; Mani et al., 2025). However, data has different characteristics that complicate the analysis. First, genomic studies vary in how they format genomic data, such as gene variants and sequences, using different normalisation techniques (Ceri & Pinoli, 2020). An attempt to integrate gene expression information from TCGA (National Cancer Institute, 2019) with the

Genotype-Tissue Expression (GTEx) (Lonsdale et al., 2013) presented a challenge because of the differing normalisation criteria between these datasets (Ceri & Pinoli, 2020). Second, EHR and socioeconomic data differ in their data formats and standards with genomic data. EHR contains structured clinical data, medical history of patients, and unstructured clinical data (clinical notes) (Robertson et al., 2024). Furthermore, socioeconomic data lacks a structured format (Khoury et al., 2022). These barriers hinder the integration process and require applying additional analytical tools to address different aspects of data management to ensure interoperability and data standardisation.

The last challenge relates to organising and structuring large, complex genomic data to make it accessible, specifically clinically usable (Alzu'bi et al., 2014). Existing genomic databases are often designed for researchers rather than clinicians, which makes them difficult to navigate for real-time decision-making (Ashton-Prolla et al., 2015). Therefore, there is a need for a well-designed genomic platform that prioritises clinical workflows by enabling proper storage, searchability and interpretation of the genome sequences, genetic variations and associated disease information (Alzu'bi et al., 2014). The platform should support filtering, prioritisation, and case-specific relevance, allowing clinicians to quickly extract insights that inform personalised treatment plans (Ceri & Pinoli, 2020). However, if these platforms are not tailored to clinical use, there is a risk that the genomic data insights may be misunderstood, which hinders their impact on real-world patient care. Furthermore, beyond the platform's technical features, there is a strong need for clear guidelines and training for the experts to access, interpret and apply genomic data effectively in patient care (Ashton-Prolla et al., 2015). By focusing on usability, learnability, and clinical integration, the genomic data-sharing platform can function as a tool that supports tailored diagnosis and treatment decisions. The following section provides further details on the collaboration-related needs and challenges clinical experts face when accessing, interpreting, and applying genomic data in clinical practice.

Challenges from the Perspective of Clinicians in Accessing and Managing Genomic Data

The exchange of genomic data between the different health providers is crucial for the development of disease understanding and tailored treatments for the patients (Ma et al., 2024; Ma'ruf et al., 2023). However, genomic databases are overwhelming to handle, which directly

influences stakeholders' efficient usage of this platform (Hide, 2005). Clinicians are one of the potential stakeholders who will integrate and access the genomic database to implement it in clinical practices. Bowdin et al. (2016) described potential challenges that clinicians can face in accessing and using genomic data.

First, clinicians should be able to order genomic data by performing basic clinical genetics evaluation and analysing the big data (Bowdin et al., 2016). However, many clinical experts lack education and training in genomic data management, which hinders their ability to analyse genomic reports and causes the burden of managing an enormous volume of genomic data (McLaughlin et al., 2024). According to Bowdin et al. (2016), clinicians should have basic knowledge and the possibility to have additional support from genetic professionals. Thus, it is critical for clinicians to hold specific skills for interpreting and translating genomic sequences into patient treatment.

Second, genomic data is often stored separately from electronic health records (EHRs), making real-time access difficult for clinicians (Balagurunathan & Sethuraman, 2024; Bowdin et al., 2016). Integrating genomic data with EHRs would allow for the complete labelling of human genome variations with ontological terms, which would enhance clinicians' integration of bioinformatics (Balagurunathan & Sethuraman, 2024; Bowdin et al., 2016).

Lastly, universal guidelines are needed for interpreting and applying genomic data in clinical decision-making (Bowdin et al., 2016). Specifically, there is a strong need for practice guidelines that support the integration of genomic data into clinical workflows. Achieving the integration of genomics into clinical workflows requires a multidisciplinary effort, where medical institutions and laboratories collaborate to establish and share best practices for managing, exchanging, processing, and interpreting genomic variants (Bowdin et al., 2016). Credle (2022) highlights several requirements to enable such collaboration, including real-time data access, shared interpretation tools, and harmonised data formats. However, many of these needs remain only partially addressed, highlighting the importance of actively involving clinical experts in developing clinical guidelines and designing genomic data platforms. Doing so will facilitate more accurate and efficient interpretation of genomic data, ultimately improving the quality and completeness of clinical reporting.

Research Aim and Objectives

To address the challenges of healthcare data exchange, it is necessary to identify the key functional and non-functional requirements that would serve the goals of the clinical experts. Functional requirements define the system requirements (Kotonya & Sommerville, 1998), such as facilitating an easier analysis process for clinicians. Non-functional requirements define the system properties and constraints (Kotonya & Sommerville, 1998), like real-time access difficulties and data synchronisation.

Based on the literature, no prior study had focused on measuring the distance between the "work-as-imagined" and "work-as-done" on the set of functional and non-functional requirements of the genomic data platform identified as essential by clinical experts for their efficient work with genomic data management. "Work-as-imagined" (WAI) refers to the expectations, procedures, and mental models of how tasks should be carried out to produce desired results (Hollnagel & Clay-Williams, 2022; Thompson et al., 2023). The mental model extrapolated by Resendez et al. (2025) in the literature is WAI, which will be discussed in the next section. WAI conceptualisation is used to design systems, procedures, and training based on how work is supposed to be done, usually from the perspective of individuals who design, manage, or develop systems. On the contrary, "work-as-done" (WAD) refers to the reality of how work is actually carried out in practice (Hollnagel & Clay-Williams, 2022; Thompson et al., 2023). WAD provides an understanding of work's needs and challenges by researching how cognitive processes are deployed in the workplace. If interventions are based only on imagined tasks rather than actual reality, the gap between WAI and WAD may result in poorly designed systems and higher cognitive burden. Therefore, addressing this gap by integrating WAD insights into the design and ongoing enhancement of work systems is crucial for ensuring that interventions correspond with the limitations and cognitive strategies that employees encounter in actual environments.

To address this gap, this paper aims to: (a) analyse a previously identified set of functional and non-functional requirements to build a genomic data exchange platform, (b) analyse the data of a survey study and a workshop to gather consensus among clinicians and key stakeholders on the importance of the requirements, and (c) qualitatively explore the challenges and barriers they can expect in using such platforms.

Furthermore, in achieving the goal of the present study, we will also attempt to answer the following questions:

1. To what extent do clinicians agree or disagree with the importance of having in the future platforms the functional and non-functional requirements identified in a previous analysis of the literature?
2. What challenges and needs do clinical experts report about their practice when working with genomic data?

This study is being conducted under the PROTECT-CHILD project. A European Union project that aims to improve child healthcare transplant outcomes through federated data sharing and precision medicine solutions (PROTECT-CHILD, 2025). PROTECT-CHILD brings clinicians, researchers, and technical experts together to co-design and implement innovative platforms for secure genomic data exchange by focusing on addressing the needs of paediatric patients. The project contributes to the development of personalised and efficient treatments and ensures better long-term health outcomes for children who face various kinds of rare and complex diseases. The analysis of this study draws upon the data collected through expert surveys and workshops conducted in the context of this project. Specifically, under the PROTECT-CHILD consortium in January 2025, which aimed at identifying and prioritising system requirements for managing genomic data in clinical practice. Additionally, it is important to note that this research was conducted in collaboration with another study (MSc), which had different objectives but was integrated into a single workshop.

Preliminary Work: Define the Requirement of an Ideal Platform for Genomic Clinical Study

The foundation of this study is built upon the unpublished work by Resendez et al. (2025) conducted as part of a PRISMA-based systematic literature review. It aimed to identify the set of functional and non-functional requirements for genomic data-sharing platforms based on the existing literature. In the preliminary work, key challenges from Alzu'bi et al. (2014) were identified, and design factors were categorized after a systematic analysis of current genomic data systems. The findings offer a representation of the features that must be considered for developing an effective, secure, and user-friendly genetic data exchange framework. In a way, this representation of features could be considered the necessary mental model.

Justifying the Mental Model of an Ideal Platform

We understand a mental model as internal representations that are shaped by past experiences and assumptions, helping individuals make sense of the world, form patterns or images of the external

reality, and use these to interpret information and predict future events (Binson et al., 2024; Chen et al., 2008). Several researchers have examined the concept of a mental model in the healthcare context (Chen et al., 2008; Yildirim et al., 2024). A recent study on multimodal healthcare AI for radiology (Yildirim et al., 2024) focused on the integration of artificial intelligent systems in radiology, emphasizing user-centric design. The results of the study presented a mental model that is relevant for the design and integration of AI in the radiology workflow. The mental model included features such as draft report generation, augmented report review, visual search and querying. These features help users understand how AI can support their decision-making process without replacing their expertise. The study provides insights into the usage of the mental model in technological systems, specifically identifying important features that align with user needs and expectations.

In a similar way, Chen et al. (2008) introduced the concept of a systems-informed mental model that helps physicians understand how healthcare systems function and how they can work within these systems. These researchers constructed a cognitive framework for understanding healthcare systems by identifying seven key features of a clinical mental model: purpose (s) or goal(s), boundaries, resources, interactions, outcomes, effectiveness and efficiency, and ability to evolve. Building on previous work, Resendez and colleagues (2025) produced a structured set of requirements for an ideal genomic data-sharing platform through a systematic literature review. A direct, evidence-based mapping can be drawn between the following constructs. The functional requirement for General Data Management is a concrete manifestation of what Chen et al. (2008) describe as 'Resources'. General Data Management and 'Resources' are understood as the sources that provide the foundational infrastructure necessary for robust data-driven research and system functionality. Likewise, the need for Communication and Support tools serves the same function as the' Interactions' feature of Chen et al. (2008). Communication and Support and 'Interactions' are understood as concepts that focus on facilitating the effective exchange of information and collaboration between system components and users. Finally, the critical requirement for Scalability directly reflects the system's 'Ability to Evolve' feature of Chen et al. (2008). In summary, the requirements identified in the literature review function as concrete examples of a recognized mental model. Based on this rationale, the mapped requirements can be understood as a mental model representation for a genomic data-sharing platform.

The approach taken in the preliminary work by Resendez and colleagues (2025) was to systematically analyze the literature to identify what experts repeatedly emphasize as the essential features of genomic data-sharing platforms. This process can be considered as a common mental model available or reported in the literature, built on recurring functionalities like data standardization and visualization. For example, Xia et al. (2014), Suciu et al. (2015), Warner et al. (2018) and other author groups listed in the PRISMA review note that data standardization is important for ensuring consistent analysis across systems. Sauria et al. (2015), Rodchenkov et al. (2020), Reiff et al. (2022), and another author group demonstrate that visualization tools improve the interpretability of complex datasets. Consequently, they are not only technical requirements (Yildirim et al., 2024) but manifestations of collective knowledge derived from observation, experimentation, and error within the domain, which form a mental model, as articulated by Binson et al. (2024). Binson and colleagues (2024) draw a parallel between how machine learning algorithms learn patterns from data and how humans develop mental models through observation and trial and error.

In this study, the Resendez et al. (2025) findings are interpreted as a "work-as-imagined" mental model. This model provides an aggregated understanding of what genomic data-sharing platforms should ideally provide to the potential stakeholders for their understanding of the platform and interaction with it based on patterns identified across prior studies. Understanding this imagined mental model is critical for this study research question because system design must align with clinicians' actual workflows and cognitive expectations.

Functional and Non-Functional Requirements of the Ideal Platform

Through a systematic literature review, the mental model developed by Resendez et al. (2025) identifies a structured set of functional and non-functional requirements for genomic data-sharing platforms. According to the mental model, there are three key functional requirements that the genomic data platform should include. First is general data management, where effective data acquisition, integration, and upload are important for genomic data exchange platforms that require interoperability across electronic health records (EHRs), laboratory results, open-source databases, and hospital datasets. The second functional requirement defined by the mental model (Resendez et al., 2025) is developing advanced data processing and analysis that supports diverse analytical methods such as pathway, network, and statistical modelling. The third and last functional requirement is data visualization and reporting tools, which are essential for making genomic data

interpretable (Resendez et al., 2025). The infrastructure required to integrate the three key functional requirements that are most commonly stated in the genomic research process was made up of non-functional requirements. The model identifies security, usability, scalability, and communication as non-functional needs (Resendez et al., 2025). The list of all the requirements composing the mental model can be found in Appendix A.

Clinical experts consistently emphasize the importance of platforms that streamline data management, including data acquisition, integration, upload, and sharing. Healthcare professionals, as reported by Tommel et al. (2023), highlighted inefficiencies in current systems. They stressed that accessing historical genomic test results could take days or weeks, delaying critical patient care. Healthcare professionals supported their concern by stating that instant retrieval of centralized genomic records was viewed as transformative, enabling immediate clinical action (Tommel et al., 2023). Moreover, scalable technical infrastructure is vital for handling the rapidly expanding volume of genomic data generated in modern healthcare (Office of the National Coordinator for Health Information Technology, 2023). Scalable cloud-based and federated computing architectures allow distributed teams to collaborate efficiently, ensuring data accessibility and reliable performance under growing data demands (Office of the National Coordinator for Health Information Technology, 2023). Additionally, advanced bioinformatics tools and automated analytical methods are essential. These tools ensure rapid, accurate interpretation of genomic data, helping clinicians overcome time and expertise constraints inherent to their workflows (Kaasalainen, 2025).

Given the complexity of genomic data, clinicians strongly advocate for robust visualization and reporting tools. Visualization simplifies complex datasets, enabling clinicians without deep bioinformatics expertise to quickly identify clinically relevant trends and anomalies (Qu et al., 2019). Interactive charts and automated reports facilitate clear communication among healthcare professionals, researchers, and patients, thereby supporting precision medicine and collaborative decision-making (Qu et al., 2019; Tommel et al., 2023). Furthermore, user-friendly interfaces that integrate seamlessly into existing clinical workflows significantly reduce the cognitive burden of interpreting genomic data, promoting clinician adoption and trust (Lau-Min et al., 2022; Tommel et al., 2023). Clinicians also stress the importance of comprehensive user support, including professional training and educational resources, to ensure efficient adoption and effective use of genomic platforms (Malakar et al., 2023). Effective communication features, such as timely

notifications about medical developments, enhance dynamic engagement, collaboration, and patient-centred care (Tommel et al., 2023).

Healthcare professionals identify security, compliance, and patient autonomy as critical to the ethical use of genomic platforms. Clinicians stress that robust data security measures—such as encryption, authentication methods, and strict access controls—are not merely regulatory requirements but also essential to maintaining patient trust (Tommel et al., 2023). Compliance with regulations such as GDPR and HIPAA shapes clinicians' willingness to adopt new platforms due to concerns over legal accountability and patient privacy risks (Tommel et al., 2023). Furthermore, transparent data handling practices and clear patient consent processes are viewed as foundational elements necessary to sustain ethical standards and patient autonomy. However, clinicians stressed that inadequate security compliance can lead to additional administrative burdens, increased hesitation in data sharing, and overall workflow disruptions (Dahlquist et al., 2023; Tommel et al., 2023). Trust emerges as a core requirement: Without transparent, secure, and respectful management of sensitive genomic data, clinicians and patients alike may resist engaging with genomic platforms (Tommel et al., 2023).

From the 'Imagined' to the 'Done': The Research Gap

While Resendez et al.'s. (2025) mental model defines the ideal genomic data exchange platform, it is unclear if these key requirements meet clinical experts' expectations and practical needs in real life. Continuing the research, this study will address the gap mentioned in the preliminary work to map the distance between the mental model based on the literature 'work-as-imagined' and real-world application 'work-as-done' through the collection of quantitative and qualitative information from stakeholders, where clinicians are presenting relevancy in this research. By recognizing this disparity, these findings will indicate the extent to which the 'work-as-imagined' mental model meets clinical expectations. Additionally, guiding future refinements toward ensuring genomic data-sharing platforms are conceptually robust and practically aligned with clinical settings.

Following the study by Resendez et al. (2025), the set of requirements identified and the goal of this study – the next sections will present the results of two actions undertaken in parallel to assess the importance of the requirements. Phase 1 Survey Study – in this part, the data collected from the survey will be reviewed to assess consensual agreement among clinical experts regarding the importance of the functional and non-functional requirements described in the model. The

survey was inspired by the set of requirements identified from the literature review with additional inputs from another collaborator within the PROTECT-CHILD project. The final set of requirements includes categories: genomic data acquisition, genomic data upload, data standardization, file formats, data sharing factors, data quality control, automated data completeness checks, types of analyses in research, reproducibility, use of command-line tools, preferred visualization methods, data export & download, knowledge sharing, data privacy protection, security standards awareness, platform usability (mobile-friendly), multi-language support, platform notifications, access to federated computing, federated computing criteria, data organization in research, participant selection methods and federated computing frameworks.

Phase 2 Workshop Study – this phase reviews the qualitative data collected from the workshop setting to better understand the challenges clinicians face when working with genomic data-sharing platforms.

Phase 1 – Survey Study: Quantitative Analysis of the Survey Data

Methods

Study Design

This study utilised a quantitative and exploratory design to investigate the opinion of experts on the requirements of an ideal genomic data-sharing platform based on the model previously established by Resendez et al. (2025). An online survey was designed to determine which platform features experts believe are necessary to facilitate the secure and effective sharing and analysis of genomic data.

In this phase, we present how we analysed the survey to answer the first research question: "*To what extent do clinicians agree or disagree with the importance of having in the future platforms the functional and non-functional requirements identified in previous work?"*.

Participants

An initial sample of 60 participants (N = 60) completed the survey correctly. Data from seven participants were excluded due to the discrepancies found during the data validation procedure to guarantee the dataset's integrity. The outliers were identified and removed based on the z-score (Hair, 2019). Following several dataset modifications, data from 53 participants was used. This study included experts recruited from three distinct fields, such as clinical, legal, and technical, all of whom had implicit experience in Federated Computing and/or genomic data. The sample

consisted of 30 clinical experts (56.6%), 8 legal experts (15.1%), 14 technical experts (26.4%), and 1 not identified expert (1.9%) who all participated voluntarily. Participants were recruited through professional networks, including the PROTECT-CHILD network and other European projects, as well as through snowball sampling and targeted email lists.

Out of the 53 participants in the sample, 29 (54.7%) identified as male, 23 (43.4%) as female, and 1 (1.9%) preferred not to say their gender identity. The participants' ages ranged from 22 to 75, with a mean of 40.42 (*SD* = 12.21).

This research was approved by the Ethics Committee of the University of Twente's Behavioural, Management and Social Sciences Department and conducted in accordance with its guidelines. Before the participation, informed consent was presented and had to be completed (see Appendix B).

Materials

This research survey was conducted using Qualtrics survey software. Participants were required to complete the questionnaire using a technological device, such as a laptop or smartphone, and an internet connection. Two versions of the survey were used during data collection. The differences between the surveys will be explained later in the procedure section.

The revised survey consisted of 163 items in total. At the start, participants were presented with informed consent and demographic questions. For this research, 77 items were used (questions asked for clinical experts), 62 of which were rated on a 7-point Likert Scale ("extremely important" = 7, "not important at all" = 1), 13 were open questions asking if the participants had something to add and 2 were open-ended questions with predefined answer options. The total number of requirements resulted in 62. The questionnaire sample is included in Appendix C.

In order to make it easier for experts to answer the survey, it was designed to display random subsets of questions depending on the participant's area of expertise. Those with a multidisciplinary background responded to the complete set of questions, while others received roughly half of the items total. Thus, this approach ensured that participants only answered questions aligned with their knowledge domain.

Procedure

The survey study was conducted in four phases, ensuring a structured approach for the survey development, recruitment, and data collection. The process began with the developmental phase, which is followed by the initial recruitment and data collection phase. The last two phases included

the survey revision by incorporating expert feedback and an expanded recruitment phase with the final data collection.

The survey was first developed in November 2024, following a literature review by Resendez et al. (2025) that identified a list of requirements for a genomic data-sharing platform. A draft survey version was created and refined throughout December 2024, and on December 15, 2024, the first version of the survey was finalised and launched.

The second phase – the initial recruitment phase – took place between December 15, 2024, and January 31, 2025, during which the survey was distributed through the email lists of the Europe consortium members and the PROTECT-CHILD consortium members. At this phase, 25 responses were received, including 14 clinical experts, 3 legal experts and 8 technical experts. Based on the expert review and initial analysis of the responses, several modifications were made that enhanced the clarity and ensured that only the participants with the relevant expertise completed the survey.

The third phase of the study on January 31, 2025, included revisions that were implemented to refine the survey further. These modifications included requiring participants to declare their expertise level before proceeding with the survey and in case those without relevant expertise exit the survey. Furthermore, rewording the questions to focus on expectations rather than experience ensured inclusivity for participants without direct experience with Federated Computing. The refined survey included "non-applicable" options for users. Additionally, changes were made to specific survey sections in order to improve clarity, incorporate additional relevant items and resolve minor technical issues, such as fixing HTML visualisation issues (see Appendix D).

Following these revisions, the last phase - expanded recruitment and final data collection - was conducted from February 1 to March 1, 2025. During this phase, the second refined version of the survey was shared outside the initial consortium using snowball sampling and targeted email list outreach to recruit additional experts. By the end of the final data collection, 35 additional responses had been gathered, which included 19 clinical, 5 legal, 10 technical, and 1 not identified expert.

Finally, to ensure the integrity of the dataset, all responses went through the classification and interpretation procedure. The procedure included categorising the data in two ways based on the time the data was collected. Data collected before January 31, 2025, was categorised as responses from the participants with implicit experience in Federated Computing and genomics

data. Data collected after January 31, 2025, include participants who were classified as having insufficient expertise. This classification process ensured that all responses included in the final analysis accurately interpreted experts' opinions within their respective domains.

Data Analysis

Three phases of data analysis (i.e., pre-processing and descriptive analysis) were carried out using the statistical software *R-studio* (see Appendix E).

In the pre-processing phase, the dataset was filtered and cleaned. Incomplete responses, unfinished surveys, and preview submissions were first excluded. Additionally, the pre-processing of the data included inspecting participants who did not reply consciously to the questions by answering all the questions with extreme values (either selecting minimum or maximum value on all of the items). Uniform response patterns may suggest a lack of engagement or incorrect replies, which may potentially harm the validity of the results (Qualtrics, 2022). However, no participants were found to exhibit such a pattern. Furthermore, *z*-score modifications were used for each expert group's key survey variables in order to discover outliers (Venkataanusha et al., 2019). Statistical outliers were excluded from additional analysis if their *z*-scores in any response variable were more or less than 3. Furthermore, Welch's two-sample *t* test analysis was applied to compare the means of items that were subjected to change during the modification of the survey items (West, 2021), specifically an item related to the importance of data access through federated computing. This analysis was performed to assess whether the difference in means was statistically significant, and in case it was significant, the item's responses that were gathered before the modification were excluded. Additionally, as participants received different subsets of questions based on their expertise, the percentage of responses per each item was calculated. Finally, each question was grouped into categories to make the analysis more efficient and comparable when evaluating responses from experts.

In the second phase, different analyses were computed, such as descriptive and comparative statistical evaluations. First, descriptive statistics were computed (Cooksey, 2020), including mean (*M*) and standard deviation (*SD*) for key survey variables such as *" How important is it for you to inspect the quality of data before analysis?"* and *" How important is it for you to employ the following types of analyses in your research? (Epidemiological analysis) "*. Bar plots with presented mean scores and error bars (standard deviation indicators), which show response variability within each requirement category, were used as a quality check to control the perceived

expert's importance of each requirement (Correll & Gleicher, 2014). A cut-off score of 4 on the 7-point Likert scale was used to determine which items were considered to have perceived importance. Values > 4 indicate that the requirement is important, and values ≤ 4 indicate that the requirement is not important. This approach is justified by the 7-point Likert structure, the midpoint of which 4 is a neutral response. Scores above 4 indicate increasing levels of agreement or perceived importance, while scores below or equal to 4 are neutral or indicate disagreement. This interpretation is widely agreed upon in the survey methodology literature (Boone & Boone, 2012; Sullivan & Artino, 2013; Joshi et al., 2015). Expert agreement for each item was evaluated separately using the interquartile range (IQR), with values ≤ 2 interpreted as indicating an agreement and values > 2 interpreted as indicating disagreement on the importance or not importance of the requirement implementation (Bodmer et al., 2019).

The third phase of the analysis examined the expert responses to open-ended questions with a predefined answer option, as well as the opportunity for experts to elaborate further on the answer. Frequency analysis was used to record the distribution of responses and to establish which of the predefined options were selected most often (Kalaian, 2008). This method provided a way to quantify the preferences of experts as well as illustrate any dominant patterns of reasoning. Additional qualitative information was gained from the responses, which included further elaboration. However, the focus remained on the frequency of the selected answers.

All computations, visualisations, and statistical procedures were implemented using R libraries like *dplyr, ggplot2,* and *corrplot* for data processing, visualisation, and statistical modelling. The dataset was cleaned, standardised, and examined, with an emphasis on deriving significant insights from expert comments, which was ensured by the structured analytical technique.

Results

Data Pre-processing and Validation

Following the removal of statistical outliers, ineligible participants and pairwise missing data, a total number of clinical experts, $N_{clin} = 30$ out of a total sample of 33 clinical experts, were included in the analysis (*age range* = 22-73, $M_{age} = 39.13$, $SD_{age} = 13.78$, Female = 63.3%, Male = 36.7%). Three participants were excluded after they were detected as statistical outliers since their *z*-scores were ±3 for one or more items across the survey. A Welch two-sample *t* test conducted on item *"How important is it to you to have access to Federated Computing infrastructure for your*

work?", which was changed during the survey modification, showed no significant difference between the two survey versions, $t(3) = 1.73$, $p = .18$. Therefore, all responses to this item were retained in the analysis.

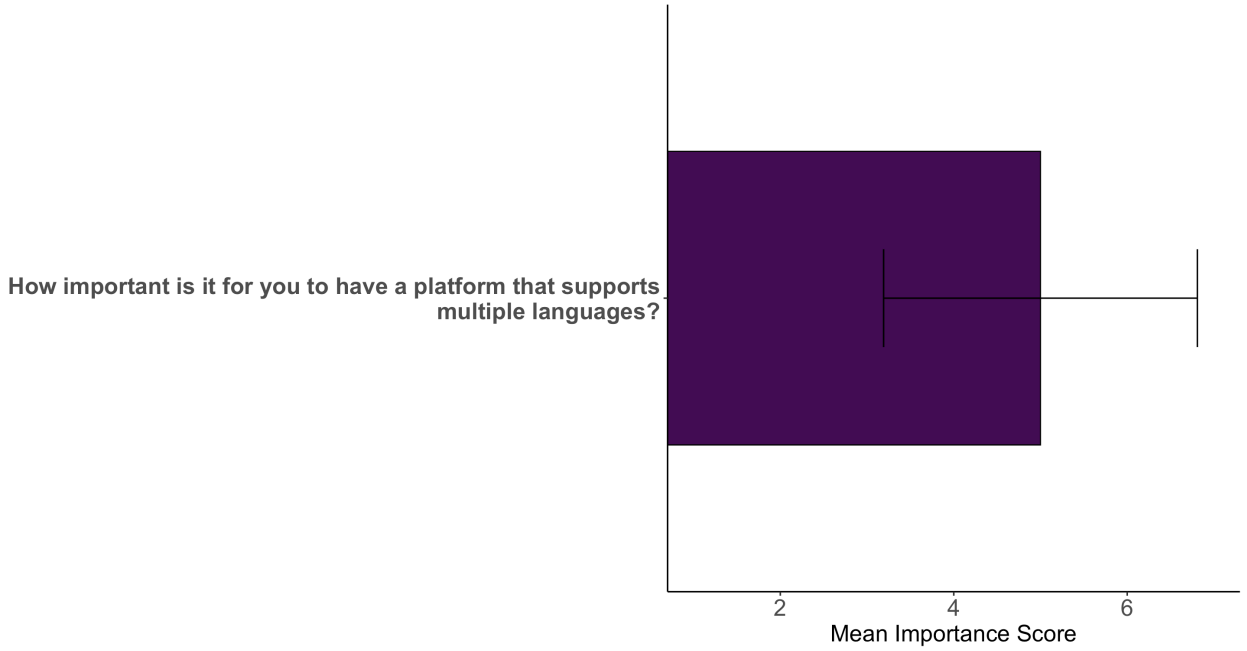
Data Quality Check

The variability of clinical expert responses in terms of *SD* values ranged from 0.65 to 1.81, with an average *SD* equal to 1.28 across all items, which overall can be considered modest variability (see Appendix F).

Five requirement categories showed the most significant divergence in opinions regarding the importance of the requirement items. The "Multi-language Support" category demonstrated the highest overall variability, with its single item showing an *SD* of 1.81, suggesting that clinical experts had different opinions on whether the requirement is important to implement on the platform (see Figure 1).

Figure 1

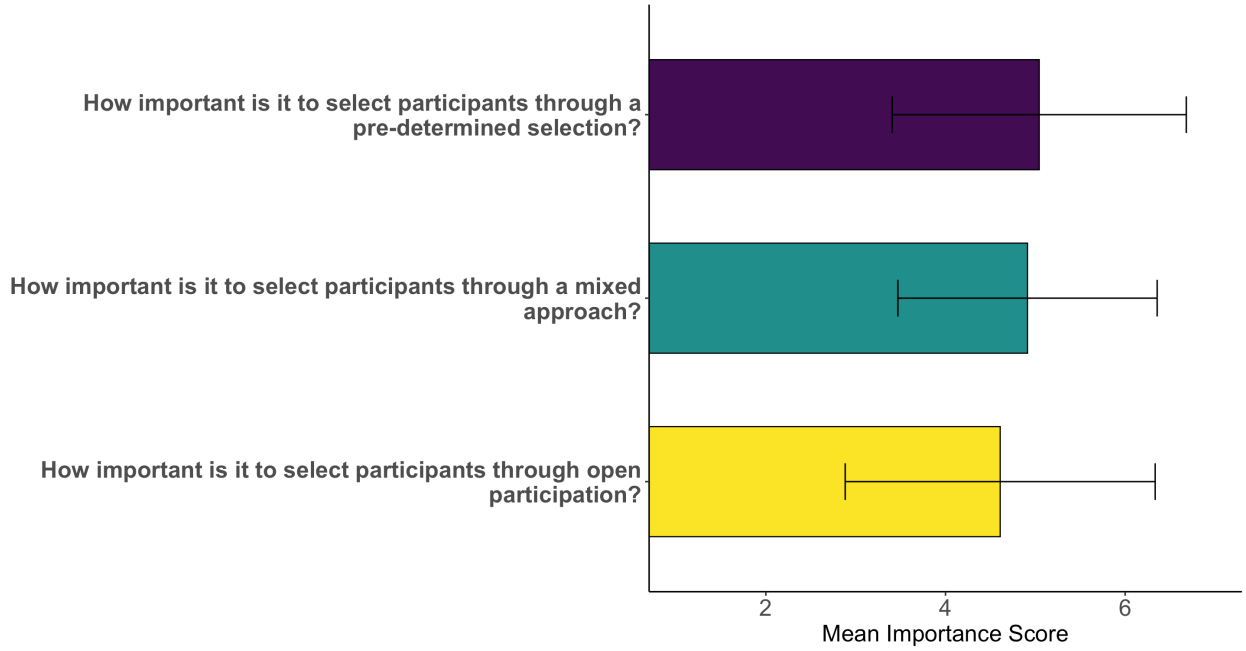
Importance Scores for "Multi-language Support"



As suggested by Figure 2, respondents answered items related to "Participant Selection Methods". Items *SD*s ranged from 1.44 to 1.73, indicating substantial divergence in expert views, particularly for pre-determined (*SD* = 1.73) and open selection strategies (*SD* = 1.64).

Figure 2

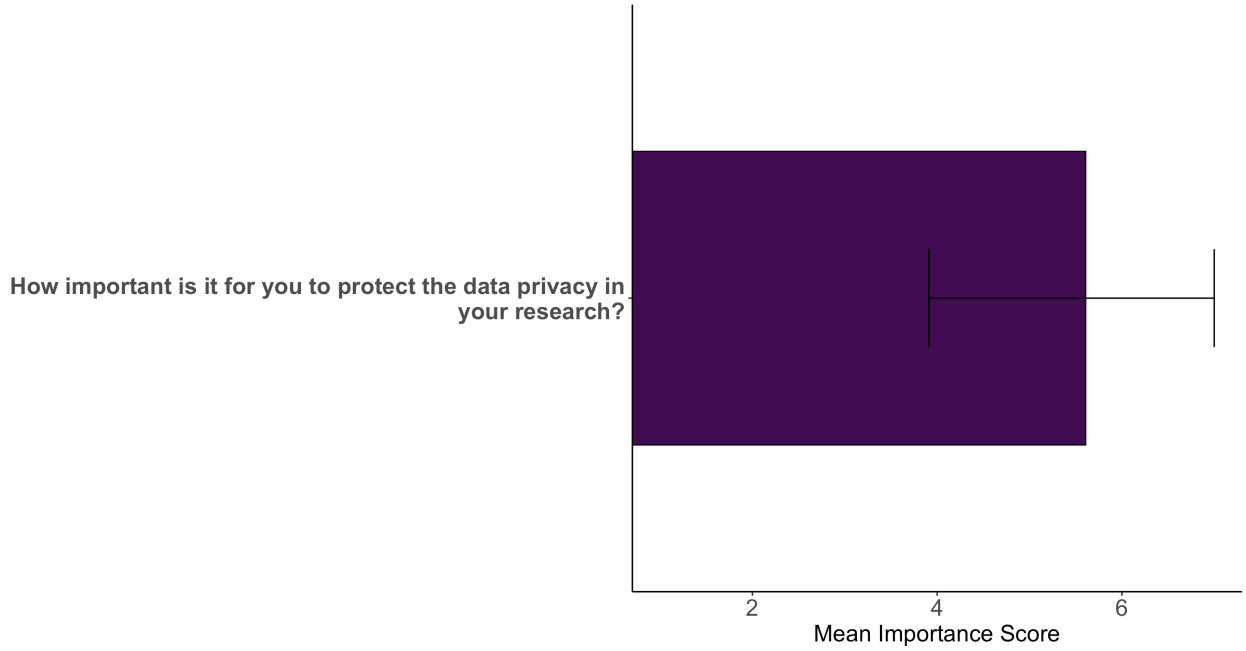
Importance Scores for "Participant Selection Methods" Items

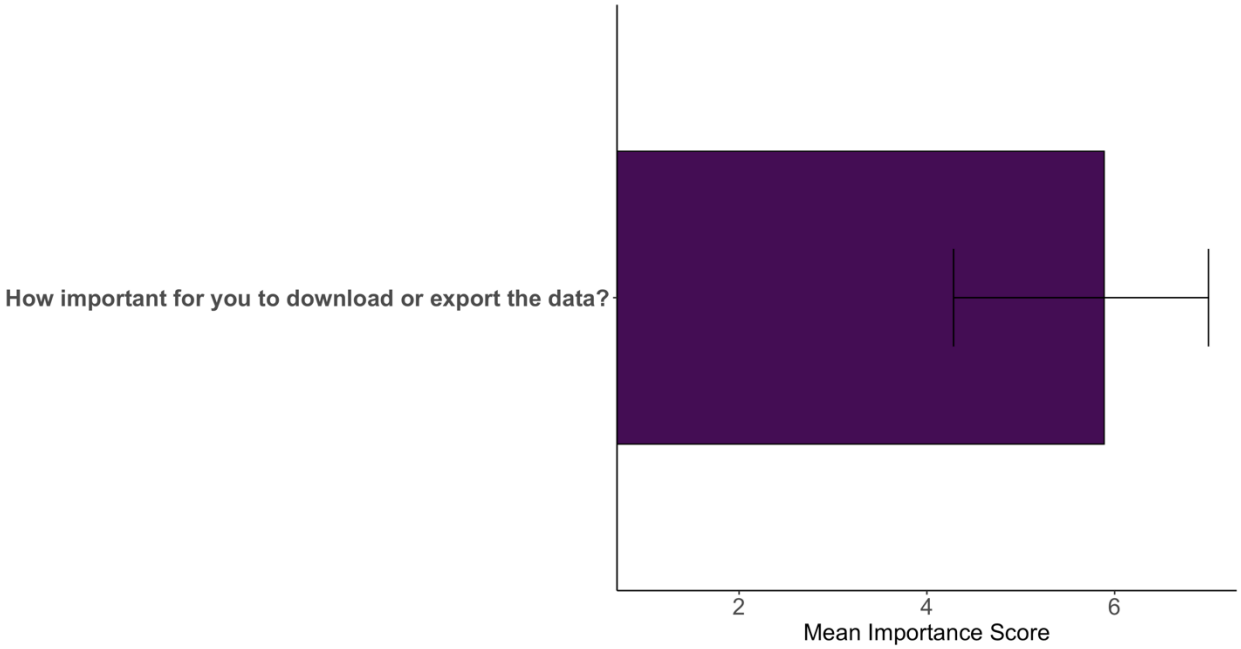


Similarly, experts reflect mixed responses on the importance of the "Data Privacy Protection" item, which showed a high *SD* of 1.70, and the "Data Export and Download" item, which showed a high *SD* of 1.60 (see Figure 3).

Figure 3

Importance Scores for "Data Privacy Protection" and "Data Export and Download"

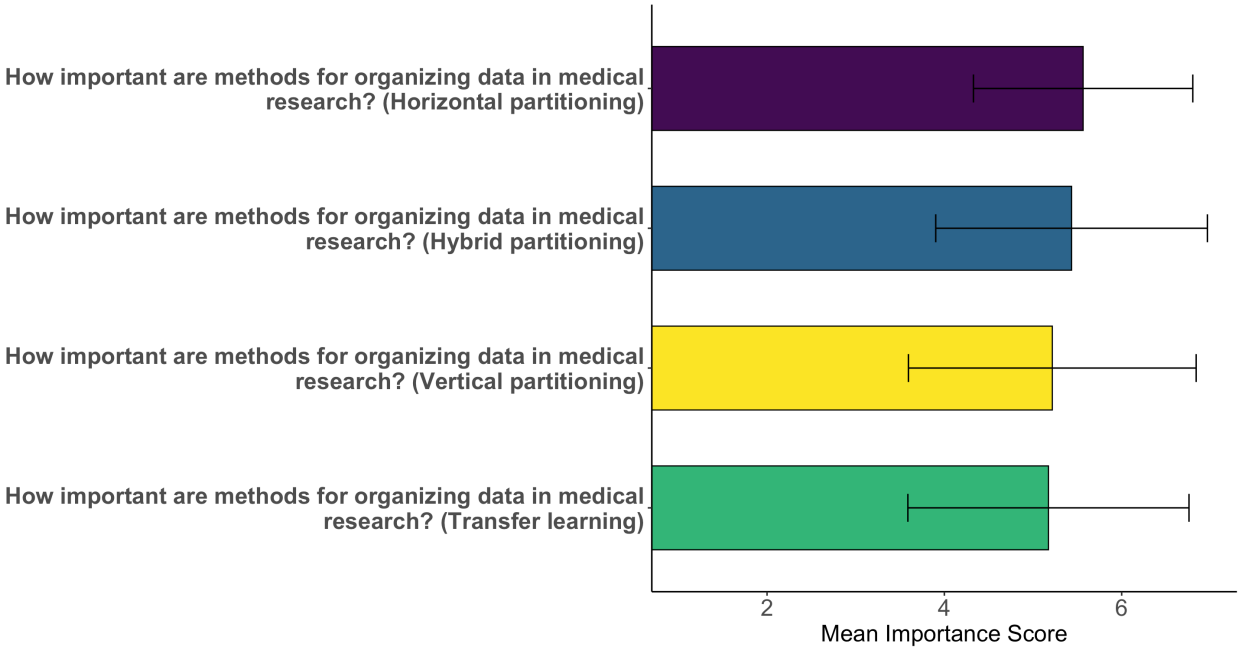




The "Data Organization in Research" category displayed an *SD* range of 1.24 to 1.62, indicating contrasting views, specifically for vertical partitioning (*SD* = 1.62) and transfer learning (*SD* = 1.59) (see Figure 4).

Figure 4

Importance Scores for "Data Organization in Research"



Importance and Agreement Analysis of Platform Requirements

Table 1 presents the results of the importance and agreement analysis for the identified platform requirements (N_{req} = 62) evaluated by clinical experts. The table includes descriptive statistics (mean and standard deviation), interquartile range (IQR), the percentage of respondents who rated each item, and binary decisions on whether the requirement was considered important and agreed upon. These evaluations provide insight into expert consensus regarding the inclusion of specific functional and non-functional requirements within the genomic data-sharing platform.

Table 1

Descriptive Statistics and Expert Consensus on Platform Requirements

Type of Requirement	Requirement Category	Requirements	Mean (*SD*)	IQR	% Respondents	Importance[a]	Agreement[b]
Functional	Genomic Data Acquisition	Acquire genomic data from multiple sources	6.00 (0.97)	1.00	0.56	Yes	Yes
	Genomic Data Upload	Upload your own genomic data	5.72 (0.96)	0.00	0.56	Yes	Yes
	Data Standardization	Use PATRIC model	3.64 (1.12)	0.00	0.34	No	Yes
		Use Genomic Data Model	4.75 (0.87)	1.25	0.38	Yes	Yes
		Use DataSHaPER model	3.91 (1.04)	0.00	0.34	No	Yes
		Use OMOP model	4.69 (0.85)	1.00	0.41	Yes	Yes
		Use FHIR model	4.77 (1.01)	1.00	0.41	Yes	Yes
		Use VCF model	4.92 (1.32)	2.00	0.41	Yes	Yes
		Use PHENOPACKET model	4.50 (0.84)	0.75	0.19	Yes	Yes
	File Formats	Preferred format: VCF	4.85 (1.34)	2.00	0.41	Yes	Yes

Type of Requirement	Requirement Category	Requirements	Mean (*SD*)	IQR	% Respondents	Importance[a]	Agreement[b]
		Preferred format: FAST-Q	4.69 (1.11)	2.00	0.41	Yes	Yes
		Preferred format: BAM	4.86 (1.23)	2.00	0.44	Yes	Yes
		Preferred format: IDAT	5.00 (1.41)	2.00	0.16	Yes	Yes
	Data Sharing Factors	Efficient modality of data sharing	6.06 (0.73)	0.75	0.56	Yes	Yes
		Security modality of data sharing	5.94 (0.80)	1.75	0.56	Yes	Yes
		Scope of the project modality of data sharing	6.06 (0.87)	2.00	0.56	Yes	Yes
	Data Quality Control	Inspect data quality	6.22 (0.65)	1.00	0.56	Yes	Yes
	Automated Data Completeness Checks	Automated checks for data completeness	5.89 (0.76)	1.00	0.56	Yes	Yes
	Types Of Analyses in Research	Use Epidemiological analysis	5.06 (1.51)	2.00	0.56	Yes	Yes
		Use Predictive modelling	5.89 (1.49)	1.00	0.56	Yes	Yes
		Use Statistical analysis	6.00 (1.50)	1.75	0.56	Yes	Yes
		Use Data visualization	5.44 (1.50)	1.00	0.56	Yes	Yes
		Use Exploratory analysis	5.56 (1.38)	1.00	0.56	Yes	Yes
		Use AI modelling	5.50 (1.34)	1.00	0.56	Yes	Yes

Type of Requirement	Requirement Category	Requirements	Mean (*SD*)	IQR	% Respondents	Importance[a]	Agreement[b]
		Use Preventive models	4.80 (1.23)	1.75	0.31	Yes	Yes
	Reproducibility	Importance of reproducibility	5.83 (1.29)	1.75	0.56	Yes	Yes
	Use Of Command-Line Tools	Use command-line tools for analysis	5.00 (1.28)	1.75	0.56	Yes	Yes
	Preferred Visualization Methods	Visual: Graphs	5.94 (1.11)	1.00	0.56	Yes	Yes
		Visual: Charts	5.56 (1.29)	1.00	0.56	Yes	Yes
		Visual: Heat map	5.33 (1.33)	1.75	0.56	Yes	Yes
		Visual: Sequencing	5.22 (1.31)	1.75	0.56	Yes	Yes
		Visual: Networks	5.44 (1.38)	3.00	0.56	Yes	No
	Data Export and Download	Export/download data	5.89 (1.60)	1.00	0.56	Yes	Yes
	Knowledge Sharing	Knowledge sharing with Clinicians	6.22 (1.52)	1.00	0.56	Yes	Yes
		Knowledge sharing with Researchers	6.33 (1.46)	1.00	0.56	Yes	Yes
		Knowledge sharing with Patients	4.72 (1.45)	2.00	0.56	Yes	Yes
		Knowledge sharing with Policymakers	5.00 (1.61)	2.00	0.56	Yes	Yes

Type of Requirement	Requirement Category	Requirements	Mean (*SD*)	IQR	% Respondents	Importance[a]	Agreement[b]
		Knowledge sharing with Ethical board	5.11 (1.64)	1.75	0.56	Yes	Yes
Non-Functional	Data Privacy Protection	Protect data privacy	5.61 (1.70)	1.50	0.72	Yes	Yes
	Security Standards Awareness	Stay updated on security standards	5.57 (1.44)	2.00	0.72	Yes	Yes
	Platform Usability (Mobile-Friendly)	Platform is Mobile-friendly	4.78 (1.57)	2.00	0.72	Yes	Yes
	Multi-Language Support	Platform supports Multi-language	5.00 (1.81)	2.00	0.72	Yes	Yes
	Platform Notifications	Platform sends Notifications	4.91 (1.59)	0.50	0.72	Yes	Yes
	Access To Federated Computing	Access to Federated Computing infrastructure	5.74 (0.96)	1.50	0.72	Yes	Yes
	Federated Computing Criteria	FC criterion: Computational efficiency	6.17 (0.78)	1.00	0.72	Yes	Yes
		FC criterion: Communication efficiency	6.00 (0.74)	1.00	0.72	Yes	Yes
		FC criterion: Model accuracy	6.26 (0.75)	1.00	0.72	Yes	Yes
		FC criterion: Privacy guarantee	6.09 (1.16)	2.00	0.72	Yes	Yes
	Data Organization in Research	Data organization method -	5.57 (1.24)	1.50	0.72	Yes	Yes

Type of Requirement	Requirement Category	Requirements	Mean (*SD*)	IQR	% Respondents	Importance[a]	Agreement[b]
		Horizontal partitioning					
		Data organization method - Vertical partitioning	5.22 (1.62)	2.50	0.72	Yes	No
		Data organization method - Transfer learning	5.17 (1.59)	2.00	0.72	Yes	Yes
		Data organization method - Hybrid partitioning	5.43 (1.53)	2.50	0.72	Yes	No
	Participant Selection Methods	Participant selection: Pre-determined	5.04 (1.64)	1.00	0.72	Yes	Yes
		Participant selection: Open	4.61 (1.73)	2.00	0.72	Yes	Yes
		Participant selection: Mixed	4.91 (1.44)	2.00	0.72	Yes	Yes
	Federated Computing Frameworks	FC framework: TensorFlow	4.62 (1.54)	2.00	0.50	Yes	Yes
		FC framework: PySyft	4.38 (1.36)	1.00	0.50	Yes	Yes
		FC framework: Flower	4.33 (1.50)	0.50	0.47	Yes	Yes
		FC framework: Vantage6	4.50 (1.59)	1.25	0.50	Yes	Yes
		FC framework: Custom-built	4.75 (1.13)	2.00	0.50	Yes	Yes
		FC framework: NVIDIA [c]	–	–	–	–	–
		FC framework: PyTorch	4.67 (1.59)	2.00	0.47	Yes	Yes

[a] Importance: “Yes” indicates a mean score greater than 4; “No” indicates a mean score of 4 or below. [b] Agreement: “Yes” indicates an interquartile range (IQR) of 2.00 or less; “No” indicates an IQR greater than 2.00. [c] The requirement FC framework: NVIDIA was not rated by the clinical experts

Genomic Data Acquisition. In this category one requirement was evaluated. Experts rated the need to acquire genomic data from multiple sources as important (M = 6.00, SD = 0.97) with strong agreement (IQR = 1.00). This suggests a clear consensus on the relevance of including multi-source genomic data acquisition within the platform.

Genomic Data Upload. Similarly, one requirement was assessed in this category. Uploading one's own genomic data was rated important (M = 5.72, SD = 0.96) and agreed upon (IQR = 0.00), indicating unanimous expert support for incorporating this feature.

Data Standardization. In this category, seven requirements were evaluated. Mean importance scores ranged from M = 3.64 (SD = 1.12) for the PATRIC model to M = 4.92 (SD = 1.32) for the VCF model. IQR values ranged from 0.00 to 2.00.

Important and agreed-upon requirements include the Genomic Data Model (M = 4.75, SD = 0.87, IQR = 1.25), OMOP (M = 4.69, SD = 0.85, IQR = 1.00), FHIR (M = 4.77, SD = 1.01, IQR = 1.00), VCF (M = 4.92, SD = 1.32, IQR = 2.00), and PHENOPACKET (M = 4.50, SD = 0.84, IQR = 0.75). The findings suggest a clear consensus on the relevance of including these modern data standards within the platform.

The PATRIC (M = 3.64, SD = 1.12, IQR = 0.00) and DataSHaPER (M = 3.91, SD = 1.04, IQR = 0.00) models were not considered important but unanimously agreed on their limited relevance to be incorporated into the platform.

File Formats. In this category, four requirements were assessed. Mean importance scores ranged from M = 4.69 (SD = 1.11) for FAST-Q to M = 5.00 (SD = 1.41) for IDAT, with IQR values all at 2.00.

All requirements were considered important and agreed upon, including VCF (M = 4.85, SD = 1.34), FAST-Q (M = 4.69, SD = 1.11), BAM (M = 4.86, SD =1.23), and IDAT (M = 5.00, SD = 1.41). The findings indicate strong expert support for offering a range of file format options for data input and export.

Data Sharing Factors. In this category, three requirements were evaluated. Mean scores ranged from *M* = 5.94 (*SD* = 0.80) for security modality to *M* = 6.06 (*SD* = 0.87) for scope of the project modality, and IQR values ranged from 0.75 to 2.00.

All requirements were considered important and agreed upon, including the efficient modality of data sharing (*M* = 6.06, *SD* = 0.73, IQR = 0.75), the security modality of data sharing (*M* = 5.94, *SD* = 0.80, IQR = 1.75), and the scope of the project modality (*M* = 6.06, *SD* = 0.87, IQR = 2.00). These findings indicate that experts find secure, efficient, and scalable data-sharing modalities vital.

Data Quality Control. In this category, one requirement was evaluated. The requirement to inspect data quality was rated as important (*M* = 6.22, *SD* = 0.65) and agreed upon (IQR = 1.00). This reflects a high level of consensus on the critical role of quality assurance in data handling.

Automated Data Completeness Checks. In this category, one requirement was assessed. The need for automated checks for data completeness was considered important (*M* = 5.89, *SD* = 0.76) and agreed upon (IQR = 1.00), indicating consistent expert support for ensuring data completeness through automated tools.

Types of Analyses in Research. In this category, seven requirements were evaluated. Importance scores ranged from *M* = 4.80 (*SD* = 1.23) for preventive models to *M* = 6.00 (*SD* = 1.50) for statistical analysis. IQR values varied between 1.00 and 2.00.

All requirements were considered important and agreed upon, including epidemiological analysis (*M* = 5.06, *SD* = 1.51, IQR = 2.00), predictive modelling (*M* = 5.89, *SD* = 1.49, IQR = 1.00), statistical analysis (*M* = 6.00, *SD* = 1.50, IQR = 1.75), data visualization (*M* = 5.44, *SD* = 1.50, IQR = 1.00), exploratory analysis (*M* = 5.56, *SD* = 1.38, IQR = 1.00), AI modelling (*M* = 5.50, *SD* = 1.34, IQR = 1.00), and preventive models (*M* = 4.80, *SD* = 1.23, IQR = 1.75). The findings indicate broad support for integrating diverse analytical methods into the platform.

Reproducibility. In this category, one requirement was evaluated. The importance of ensuring reproducibility was rated highly (*M* = 5.83, *SD* = 1.29) and agreed upon (IQR = 1.75), indicating expert consensus on its necessity in data exchange platforms.

Use of Command-Line Tools. In this category, one requirement was assessed. The use of command-line tools for analysis was considered important (*M* = 5.00, *SD* = 1.28) and agreed upon (IQR = 1.75), suggesting that this technical proficiency feature is widely endorsed.

Preferred Visualization Methods. In this category, five requirements were evaluated. Importance scores ranged from *M* = 5.22 (*SD* = 1.31) for sequencing to *M* = 5.94 (*SD* = 1.11) for graphs. IQR values ranged from 1.00 to 3.00.

Requirements considered important and agreed upon included graphs (*M* = 5.94, *SD* = 1.11, IQR = 1.00), charts (*M* = 5.56, *SD* = 1.29, IQR = 1.00), heat maps (*M* = 5.33, *SD* = 1.33, IQR = 1.75), and sequencing (*M* = 5.22, *SD* = 1.31, IQR = 1.75). Although the network visualization method was rated as important (*M* = 5.44, *SD* = 1.38), its IQR value equalled 3.00, suggesting disagreement among experts.

Overall, there was consistent support for diverse and accessible visualization formats, with some debate around network-based visuals.

Data Export and Download. In this category, one requirement was evaluated. The ability to export and download data was rated important (*M* = 5.89, *SD* = 1.60) and agreed upon (IQR = 1.00), reflecting a clear need for data portability features.

Knowledge Sharing. In this category, five requirements were assessed. Mean importance scores ranged from *M* = 4.72 (*SD* = 1.45) for patients to *M* = 6.33 (*SD* = 1.46) for researchers, with IQR values between 1.00 and 2.00.

All requirements were considered important and agreed upon, including knowledge sharing with clinicians (*M* = 6.22, *SD* = 1.52, IQR = 1.00), researchers (*M* = 6.33, *SD* = 1.46, IQR = 1.00), patients (*M* = 4.72, *SD* = 1.45, IQR = 2.00), policymakers (*M* = 5.00, *SD* = 1.61, IQR = 2.00), and ethical boards (*M* = 5.11, *SD* = 1.64, IQR = 1.75). The findings present a broad consensus on the importance of knowledge exchange across different stakeholders.

Data Privacy Protection. In this category, one requirement was evaluated. Protecting data privacy was rated as important (*M* = 5.61, *SD* = 1.70) and agreed upon (IQR = 1.50), underlining experts' prioritization of privacy measures.

Security Standards Awareness. In this category, one requirement was assessed. The need to stay updated on security standards was rated important (*M* = 5.57, *SD* = 1.44) and agreed upon (IQR = 2.00), reflecting general alignment on the relevance of evolving data protection measures.

Platform Usability (Mobile-Friendly). In this category, one requirement was evaluated. Experts considered it important that the platform is mobile-friendly (*M* = 4.78, *SD* = 1.57) and agreed on its relevance (IQR = 2.00), indicating consistent support for accessibility across devices.

Multi-Language Support. In this category, one requirement was assessed. Supporting multiple languages was rated important (M = 5.00, SD = 1.81) and agreed upon (IQR = 2.00), suggesting recognition of language inclusivity in platform design.

Platform Notifications. In this category, one requirement was evaluated. The inclusion of platform notifications was rated as important (M = 4.91, SD = 1.59) with strong agreement (IQR = 0.50), reflecting support for user engagement and alert systems.

Access to Federated Computing. In this category, one requirement was assessed. Access to federated computing infrastructure was rated important (M = 5.74, SD = 0.96) and agreed upon (IQR = 1.50), indicating that decentralized computation capabilities are valued.

Federated Computing Criteria. In this category, four requirements were evaluated. Mean importance scores ranged from M = 6.00 (SD = 0.74) for communication efficiency to M = 6.26 (SD = 0.75) for model accuracy, with IQR values between 1.00 and 2.00.

All four criteria were considered important and agreed upon, including computational efficiency (M = 6.17, SD = 0.78, IQR = 1.00), communication efficiency (M = 6.00, SD = 0.74, IQR = 1.00), model accuracy (M = 6.26, SD = 0.75, IQR = 1.00), and privacy guarantee (M = 6.09, SD = 1.16, IQR = 2.00). This indicates expert endorsement for balancing performance and privacy in federated computing systems.

Data Organization in Research. In this category, four requirements were evaluated. Mean importance scores ranged from M = 5.17 (SD = 1.59) for transfer learning to M = 5.57 (SD = 1.24) for horizontal partitioning, with IQR values ranging from 1.50 to 2.50.

The requirements considered important and agreed upon included horizontal partitioning (M = 5.57, SD = 1.24, IQR = 1.50) and transfer learning (M = 5.17, SD = 1.59, IQR = 2.00). The methods of vertical partitioning (M = 5.22, SD = 1.62) and hybrid partitioning (M = 5.43, SD = 1.53) were also rated as important, but experts did not reach agreement. Both methods' IQR values equalled 2.50, reflecting some divergence in opinion on these approaches.

Overall, there was support for diverse data organization methods. However, there is some debate about vertical and hybrid partitioning methods.

Participant Selection Methods. In this category, three requirements were assessed. Importance scores ranged from M = 4.61 (SD = 1.73) for open selection to M = 5.04 (SD = 1.64) for pre-determined selection, with IQR values between 1.00 and 2.00.

All requirements were considered important and agreed upon, including pre-determined (M = 5.04, SD = 1.64, IQR = 1.00), open (M = 4.61, SD = 1.73, IQR = 2.00), and mixed selection methods (M = 4.91, SD = 1.44, IQR = 2.00). Although importance ratings were moderate, the consistent agreement indicates general support for various participant selection approaches in the data exchange platform design.

Federated Computing Frameworks. In this category, six frameworks were evaluated (excluding NVIDIA due to missing data). Importance scores ranged from M = 4.33 (SD = 1.50) for Flower to M = 4.75 (SD = 1.13) for Custom-built solutions. IQR values varied from 0.50 to 2.00.

All frameworks were considered important and agreed upon, including TensorFlow (M = 4.62, SD = 1.54, IQR = 2.00), PySyft (M = 4.38, SD = 1.36, IQR = 1.00), Flower (M = 4.33, SD = 1.50, IQR = 0.50), Vantage6 (M = 4.50, SD = 1.59, IQR = 1.25), Custom-built (M = 4.75, SD = 1.13, IQR = 2.00), and PyTorch (M = 4.67, SD = 1.59, IQR = 2.00). While rated slightly lower in importance score than other categories, these tools were still endorsed for inclusion, reflecting a range of flexible framework preferences.

Frequency Analysis of Open-Ended Questions

Tables 2 and 3 present frequency analysis that was carried out on answers to two predetermined open-ended questions responded to by clinical experts, pinpointing options that were most predominantly selected during the analysis.

Table 2 summarises responses to the question: *"What is the primary focus of your research involving the collaborative use of medical data?"*. As shown in the table, clinicians most frequently selected the creation of AI predictive models (n = 9) as their primary analytical need. Other responses selected extracting statistical insights (n = 5). Finally, a smaller group (n = 5) wrote custom entries, while 13 respondents either skipped the question or left it blank.

Table 3 presents responses to the question: *"What is the primary focus of your research involving the collaborative use of medical data?"*. In this table, responses were more varied. Clinicians frequently noted variability in the data formats they work with across institutions (n = 12), highlighting this as an important challenge for clinical analysis. A smaller group (n = 9) indicated they work with uniform, structured data. These distributions highlight the importance of the platform's flexibility in handling diverse data types for clinical research.

Table 2

Frequency Analysis of the Question: "What is the primary focus of your research involving the collaborative use of medical data?"

Response	Frequency
Developing AI predictive models: Using methods such as deep learning to forecast outcomes	9
Extracting statistical insights: Summarizing data patterns and associations	5
Other / custom entries	5
Missing / No response	13

Table 3

Frequency Analysis of the Question: "Which of the following types of data or processes best describe the data you typically work with?"

Response	Frequency
Data that differs significantly across institutions: E.g., real-world hospital data with variable formats	12
Data that is uniform across institutions: For example, structured clinical trial data	9
Training advanced predictive models: Using techniques like ensemble or transfer learning	2
Other / custom entries	7
Missing / No response	2

Phase 2 – Workshop Data: Qualitative Analysis of Challenges and Opportunities in the Usage of Genomic Platforms

Methods

Study Design

The aim of the second phase was to further extract insights regarding system design requirements based on the stakeholder perspectives, including clinical, legal and technical experts. Additionally,

this phase aims to answer the second research question: *"What challenges and needs do clinical experts report in their practice when working with genomic data?"*.

A workshop with experts was organized to review the requirements and to discuss challenges and needs to handle genetic data for international clinical studies. Using *Atlas.ti*, a thematic analysis was performed on the insights collected during the workshop to find emergent viewpoints in terms of challenges and needs.

Participants

This study comprised a convenience sample of technical, clinical and legal experts from the PROTECT-CHILD consortium who attended one of the consortium meetings in January 2025. The total number of participants consisted of 36 experts representing a range of professional disciplines, including nephrology, paediatrics, computer science, biomedical research, and social sciences, who all participated voluntarily. This is a highly diverse group of experts with experience ranging from 1 year to over 10 years. The age range also varies significantly, with clinical experts spanning from 28 to 60-70 years. All participants represented different European nationalities, and the discussion during the workshop was held in English. Since the participants' conversations were audio recorded, each participant signed an informed consent form to participate at the start of the study (see Appendix G). This research was approved by the Ethics Committee of the University of Twente's Behavioural, Management and Social Sciences Department and conducted in accordance with its guidelines.

Materials

An email was sent prior to the meeting to advertise the workshop. The email included materials for the session, along with the survey link, which was sent to participants prior to the session (see Appendix H). To guarantee a high-quality recording, independent tools were employed to record the workshop on audio and video, such as 6 Zoom H4N Pro Voice/Sound recorder and 2 JVC 4k Camera. For the analysis, only the audio recordings from the primary audio device were used.

The core material utilized during the workshop consisted of an outline that structured various activities and guided participant interaction (see Appendix I). The outline included two primary activities that were intended to extract experts' opinions about the PROTECT-CHILD system. Activity 1 focused on understanding the system's intended usage and identifying important technological enablers to meet clinical objectives. In order to enable structured group conversations, participants were given a printed question sheet with two discussion prompts.

Activity 2 focused on the current challenges as well as possible solutions in using health data platforms. An extra set of written questions that focused on existing practices, new current challenges, and possible solutions were distributed to participants.

Additionally, participants were given a printed version of the summary sheet detailing the key functionalities and attributes of genetic data-sharing platforms in addition to the discussion materials (see Appendix J). It included a high-level description of the functionalities and quality characteristics of usable health data platforms. Additionally, the sheet included four key challenges identified from the literature and an AI-driven clustering model linking functionalities and characteristics to challenges. Participants used this paper as a guide to make sure their conversations reflected current understanding and technological considerations.

All of the materials were created to promote structured but open-ended conversations, enabling experts from different fields to express their opinions on Federated Computing and genetic data-sharing platforms.

Procedure

Prior to the program, participants were divided into six groups of five or six people each. In order to balance the expertise within each group, experts from various fields (such as clinical, technical, ethical, and legal) were paired together to promote more discussions from various angles. All attendees received an email with their assigned group number and a list of other members in their groups. The training took place in two classrooms, with three groups assigned to each. Before the beginning of the study activities, all participants submitted a written informed consent form.

The workshop comprised two structured activities designed to elicit expert insights on genomic data-sharing challenges. The first activity focused on understanding how participants plan to use the system and identifying key technical enablers to support clinical objectives. Each group was asked the two following questions: "*How do you plan to use the system (or how do you think potential users will use it?)",* and *"What are the technical enablers/tools that can support the achievement of clinical objectives?*". Each group then engaged in a 30-minute discussion, after which they delivered a five-minute presentation summarising their findings to the other participants.

The second activity addressed key challenges in health data platforms and aimed to map out relevant solutions and requirements. Each group of participants was assigned to discuss two challenges from a previously defined list: managing a large amount of data (how to handle different

data sources and formats; providing support for data upload and use), protecting sensitive genomic data, managing the complexity of genomic data analysis (analysing complex genomic data), and dealing with legal and social aspects of data sharing (ethical, legal, and social concerns in health data platforms). Afterwards, the participants engaged in a 30-minute discussion about their current preferred approaches, tools, and strategies to cope with these challenges, along with new obstacles and ways to solve them in the future. Each group then delivered a five-minute presentation summarising their findings to the other participants.

Discussions were audio recorded with participants' permission to guarantee data dependability. After the activities had ended, participants were debriefed, during which they were told that the data was going to be transcribed and used for a report that would be shared with them for their feedback. The recordings were transcribed using AmberScript for qualitative analysis. Upon completion of transcriptions, the recordings were deleted.

Data Analysis

A thematic analysis was performed to examine the content of the workshop discussions' by systematically coding, classifying, and interpreting expert ideas using *Atlas.ti*. This analysis included a hybrid approach combining deductive and inductive coding. While inductive coding was utilised to capture emerging themes beyond the basic framework of the conversations, deductive coding was based on pre-defined tasks. Preliminary codes were allocated to relevant textual units after an iterative evaluation of the transcripts. These codes were refined and grouped into higher-order themes that stand for key barriers, challenges, facilitators, and solutions pertaining to the sharing of genetic data.

In order to assess an inter-rater reliability agreement and guarantee the validity of the codes in collaboration with another study (MSc), multiple researchers had independently coded the themes (Belotto, 2018). Discussions and group code improvement was used to settle discrepancies and disputes. In order to make the data more quantitative, frequency counts of the emerging themes had also been gathered.

Additionally, following the qualitative analysis method by Henry et al. (2015), a two-step clustering approach was applied across the coded qualitative data. First, the hierarchical cluster analysis was performed using Ward’s method and binary distance metrics on the transposed binary coding matrix where rows presented interview quotes and columns were codes. This allowed for identifying natural groupings, as codes were grouped depending on their pattern of co-occurrence

across the entire dataset. A dendrogram was generated to visualise the clustering structure and to guide the selection of an appropriate number of clusters. After inspection of merge distances with regards to thematic interpretability, a five-cluster solution was viewed as meaningful. While the dendrogram suggested a cut-off between four and six clusters, the option of five clusters provided the best thematic distinction without losing interpretability. The decision in the number of clusters was supported by the cluster profile plot for k-means, which showed distinguishable patterns in code frequency for five clusters.

Subsequently, a k-means cluster analysis was performed to validate and profile the identified themes. The same matrix was used in the transposed form while setting the number of cluster centres to five, as per the hierarchical analysis, allowing the calculation of average code frequency within each cluster. Multi-line plot showing resulting cluster profile was made to support the interpretation. All analyses and visualisations were carried out in *R-Studio* (see Appendix K). This allowed flexibility in implementing custom clustering, data transformation, and plotting solutions with *cluster*, *ggplot2*, and *reshape2* packages.

Finally, the results of the thematic manual coding, frequency distributions of themes that arose were combined in the final analysis to give a thorough and complete picture of the expert viewpoints on the problems with genetic data-sharing and possible solutions.

Results

The thematic analysis revealed experts' perspectives, needs, and challenges regarding the genomic data-sharing system design requirements. Based on the thematic analysis, 39 unique codes were identified (see Table 4). These codes include functional and non-functional requirements of the system design that cover aspects such as technical infrastructure, data interoperability, user-centred support features, legal and ethical considerations, and broader visions for platform evolution.

Table 4

List of Unique Codes and their Representative Quote

Unique Codes	Example Quote
Current platform shortcomings	"We use the cloud. We need to put the other question there longevity. How we can share that in the cloud because I know that I love the privacy and everything. but our biggest problem is to share, something, I mean in the hospital or the patient should be able to …" (Group 4, Medical Expert)

Unique Codes	Example Quote
Data quality, standardization and formats	"So second important piece is related to data standardization. If data collected cannot talk to each other, that's failed …" (Group 4, Medical Expert) "Number three important piece is related to the data quality. Now you have to clean the data before using it." (Group 4, Medical Expert)
Ethics support infrastructure	"Also, a need, as in other cases to support ethics approval. I mean the data governance layer part of the system, because data ethics are in general big issues with long time and so on, and they are even more complex in genomics is important." (Group 4, Technical Expert)
Legal and regulatory frameworks	"For the future requirements regarding the structure in which we will have a legal and institutional restrictions to access the data." (Group 4, Medical Expert)
Need for system simplicity	"So, if you're if you don't have time to talk with the patient, you don't have time to to fill the form. So it's why the clinicians they go to the open text box and they put some strange abbreviations to to summarize that. So we need an easy collection, an error check and a tool for that that defines the data model that is scalable, that understands different language, different contexts, probably even in the same language or same country. Different doctors will use different abbreviations. So we need something like that because at the end, the searching of data for the clinician mostly is a summary." (Group 4, Technical Expert)
Platform benchmarking	"I expect from this system to be more flexible, more easily extensible. I mean, to add new data that were not considered at the last month. It's easier with such a system than with a registry." (Group 2, Technical Expert)
Stakeholder involvement	"There are need for investment from the Policymaker. So by having this kind of data, it's going to provide evidence for investing more in this area." (Group 6, Big data analytics - Technical Expert)
Technical enablers	"We understand the the standardization, the normalization, the using ontologies, etc. but a clinician has very limited time. And having a good standardization … we need an easy

Unique Codes	Example Quote
	collection, an error check and a tool for that that defines the data model that is scalable, that understands different language, different contexts, probably even in the same language or same country … and in the in the part of emerging issues is about the space and resources, cost, etc. on that because we tend to talk about the cloud. That is something like wait in the in the space. But in the end it's servers and servers every day at home and wasting a lot of energy on that." (Group 5, Lab Scientist - Clinician)
Technical enablers: Advanced analytics and learning tools	"And so, for instance one enabler in this sense would be the quantum computing part of the project that can ensure a more strong data security, which goes in the direction of this. Then the huge number of data and the possibility to overcome differences among center and conduct. A multicenter study brings with it the possibility, in fact, to have a strong, stronger statistical power which will build more data, more centers, and to also develop, develop, sorry, deploy and exploit different boundaries. In addition to statistical techniques, also new deep learning, relatively new deep learning techniques that can be used to find patterns in the patterns." (Group 6, Technical Expert)
Technical enablers: Automation and data collection and processing	"Regarding the technical enablers, uh, the most important thing when we consider is that we need automation. We need to to ease the way the data is, gather, uploaded, etc. because if not, it's just another gizmo that we have to endure on that." (Group 5, Lab Scientist - Clinician)
Technical enablers: Data accessibility and querying	"If I think, uh, of interrogating my own capsule, uh, what I should expect is as much. Granularity as possible in both making the queries and both of obtaining the data. So that if my aim is to make a scientific research, I can have a nice categorization. Of all the variables. Having a significant granularity in the information that we want to ask. If the question is more clinical. What I would ask to the platform, to the capsule is to make a kind of evaluation of the clinical questions. And to get with the answer with. A specific clinical question, for example. I mean. Just like an example, I want to understand what

Unique Codes	Example Quote
	are the risk factors for kidney rejection in a subset of patients receiving kidney transplantation? At the end I would like to say these are the most important risk factors." (Group 2, Pediatrics - Clinical Expert)
Technical enablers: Data security and privacy preservation	"This is also and this is also an enabler because everyone can use it. I mean, if we manage to build something that is privacy preserving is much better to it's much easier to share it, to share data. With this perspective we used to use the data this this is of course the technical and the enabler. So if we have a data structure that is able to preserve privacy and gather most of the information. Is one of the strong enablers, and I find it that these tools are given the challenges that we will discuss later. These tools are still a very important output of the project." (Group 6, AI - Technical Expert)
Technical enablers: Data standardization and harmonization	"So, we first need to standardize the vocabulary and data model vocabulary so we can record. So, data harmonization to me is a enabler." (Group 6, Big data analytics - Technical Expert)
Technical enablers: Interface design and user-friendliness	"And then uh, the other two I think we really need is the real good, uh, mobile dashboard app. For all the users. This is a, a very important tool." (Group 6, Pediatrician - Clinician)
Technical enablers: Interoperability	"Regarding the technical enablers, uh, the most important thing when we consider is that we need … and also that brings to interoperability. We don't need another platform that is provided for a couple of years, then it is discontinued. So the data gets there is in us all because it's in a forgotten format, etc. and for doing that we need a I.T legal particularly to be in the same page to talk together to, to to understand that because the current situation doesn't work, because it's too complex." (Group 5, Lab Scientist - Clinician)
Technical enablers: Scalability and performance	"Regarding the resources that we need regarding the the security, the backup, the maintenance, and all of, uh, necessary service to handle this large volume of data, and this also, uh, this implies some scalability. Scalability issues, uh, that we don't know how to or we put as an issue because

Unique Codes	Example Quote
	we don't know how to handle them. Um, and maybe I like to for the future, we will need a data flow management tool, but, uh, you just solve this, uh, functional requirements and create, like, a unified system for the for the clinical side." (Group 6, Standardization - Technical Expert)
Technical enablers: Support	"But selecting in primary user with a a worldwide standard, the format of the entire medical history of addiction is a technical enabler to support the achievement of this objective is basically, uh, assistant like, uh, some summary dashboard that in a single view, allow clinicians to understand the medical history of a patient and how to get relation, you know, uh, symptoms, medication, uh, and so on. And also to help to to make a decision on public health." (Group 1, Computer Science ML - Technical Expert)
Value of the platform	"We expect that this will lead to a better care, better anticipation of a possible a complication to a to a tailoring, the treatments …" (Group 5, Lab Scientist - Clinician)
Vision of the platform	"We want to create. This kind of platform will be useful to find biomarkers of the information. Maybe it's just like, okay, we know this clinician." (Group 2, Applied ethics – Legal Expert)
Vision of the platform: Clinician-focused dashboard	"You can perform some kind of analysis in the data, and then you aggregate the results and get back to the web page. Right. So in the end, it's like, um, you are not seeing the, uh, the data in each of the hospitals, but you are going to you are doing your analysis. and you are receiving the feedback." (Group 3, Computer Engineer – Technical Expert)
Vision of the platform: Clinincal decision making and benchmarking	"We summarize the into two major groups who make one group is directly related to the clinical outcome because we have, uh, the, uh, any kind of, uh, transplantation there is important for donor receiver matching. And the second thing is when we. Actually having a patient receive the transplant, there is a therapy process, uh, in terms of rejection, as the current state of art, one third of the follow up are not needed. But in the past, we don't know it. In our platform, I included the genomic, uh, this kind of data, it's there is a

Unique Codes	Example Quote
	possibility to customize to personalize the to decide the therapy for follow up. And the third there also an important piece related in the stakeholder." (Group 6, Big data analytics - Technical Expert)
Vision of the platform: Data integration	"Always centralization is cheaper, faster and better. But uh, in comparison with other solutions, potential solutions. Once it is centralized in each one of the persons will have access to different kinds of data." (Group 1, Geneticist – Clinical Expert)
Vision of the platform: Open source and flexibility	"This software has to be open for teaching, for implementing ... Oh if, if now if I change my provider for the clinical history software, I have to restart at the beginning, I have to download all my, my history from decades, all hundreds of thousands of of patients, hundreds of thousands of visits, test analytics, etc., and upload in a new vendor." (Group 5, Lab Scientist - Clinician)
Vision of the platform: Research collaboration	"We want to create this kind of platform will be useful to find biomarkers of the information. Maybe it's just like, okay, we know this clinician. End of these words or maybe not. And also a key. This platform will be useful to create new collaborations between data centers. Because we can see if we do the same things in the same way or not. And we can like share different point of view." (Group 2, Applied ethics – Legal Expert)
Vision of the platform: Scalability, adaptibility, access and management of data	"It's just a matter of scalability with the algorithm and quality of it. More on that data we have. We have. Maybe you don't have to. The data are more accurate." (Group 4, Technical Expert)
Vision of the platform: Support and documentation	"Platforms should help to overcome the complexities of A data analysis. Genomic data is complex, therefore it requires a specific tools and methods to extract meaningful insights." (Group 3, Geneticist – Clinical Expert)
Visualization needs	"I mean, if we speak about results, tables are the best way. So. So you can manage the data easily. Usually. If you have a. Specific clinical. Question, maybe to obtain. A. Graphical result or even a descriptive result, maybe it's better, but this is just something very personal." (Group 2, Pediatrics - Clinical Expert)

Unique Codes	Example Quote
Current procedures of legal, private, and ethical protocols	"I think we can agree that one of the first things that we would say is follow the national regulation. Or. Yeah. Or your local standard procedure on on sharing. Yes. For example, as they were saying before, uh, not writing genomic results in the electronic health records or in my case, not writing, uh, personal information on WhatsApp groups or on emails, that that would be okay." (Group 5, Lab Scientist - Clinician)
Data access	"When it comes to rare diseases, the data is very limited, which makes the analysis even more complex because you need to take into account more elements, and you're not entirely certain of the data quality and reliability of the results. So again, having this platform where you have access to multiple data, uh, is very crucial." (Group 3, Legal Expert)
Data analysis needs	"The data aggregation actually is done manually by the clinicians by checking the different data sources. And this is a problem because there's no the the information is scattered across multiple databases in the clinical sites. So there's a really real need of creating this, um, a unified data model or with a common vocabulary and common data types, uh, that, uh, can create or can solve the problem of, uh, this data creation." (Group 4, Technical Expert)
Data ownership	"Because some countries advocate for, uh, general permission for reuse, the data for research, etc., and other countries argued that the patient must have a button on his record to allow the research, even sometimes for a study to study." (Group 5, Lab Scientist - Clinician)
Data sharing	"There are practices and maybe the fact that everyone has their own procedures and policies in place and of course, have everything that the law requiring them to do, not sharing personally or by email or by, for instance, having password or encryption in the cloud, having authentication mechanisms." (Group 2, Pediatrics - Clinical Expert)
Data storage	"And also, when you are talking about data server uh talking about power consumption storage security also backup." (Group 6, Big data analytics - Technical Expert)

Unique Codes	Example Quote
	"You have dedicated the same server for storage storing a large amount of data." (Group 1, Computer Science ML - Technical Expert)
Financial matters	"Hospitals now need to invest more. Uh, both in terms of personnel but also education programs. Uh, in order to be able to perform such analysis." (Group 3, Legal Expert)
Issues with data handling	"So, the more complex systems become, the harder it is for clinicians to keep up … not only for clinicians. Like for. Yeah. For everyone to keep up on that." (Group 3, Legal Expert)
Patient communication	"That's really difficult now because if you if you talk about like these huge amounts of data that as a clinician, it'll get a lot more difficult to explain to parents what what to make use of the even now talking about like small genetic analysis. Yeah, it's very difficult. And for some patients you need half an hour to explain them." (Group 6, Pediatrician - Clinician)
Privacy and ethical needs	"For us is important to understand that privacy is an advantage but shouldn't be a limit for care because actually the patients are concerned about privacy." (Group 5, Lab Scientist - Clinician)
Privacy, legal and ethics-related challenges	"And then we have been talking about privacy as one major challenge, uh, a hard hurdle for all these different hospitals and around different, uh, countries to work together. So that's why we have to continue to develop the federated learning as an enabler." (Group 6, Big data analytics - Technical Expert)
Usability challenges and needs	"And of course some of the tools are also very important here, like having an easy and reliable user interface where also a clinician could perform this analysis without having to go through a specialized department in order to do that." (Group 3, Legal Expert)

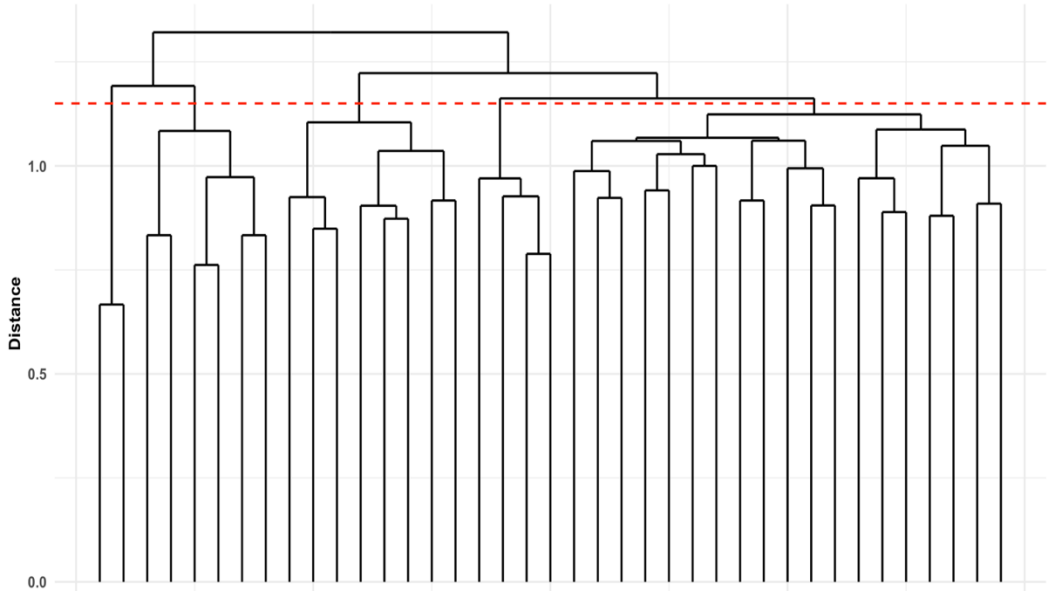
Note. The table includes the column quotes where the verbalization of an expert explains the sense of the unique code.

To examine the relationships between the themes, a cluster analysis was performed using the co-occurrence patterns of the 39 codes across participant quotes. Five main clusters were identified from the findings of Ward's method of hierarchical cluster analysis (see Figure 5). A cut-off

threshold was applied at a height of approximately 1.1 (red horizontal line), resulting in a five-cluster solution.

Figure 5

Hierarchical Clustering

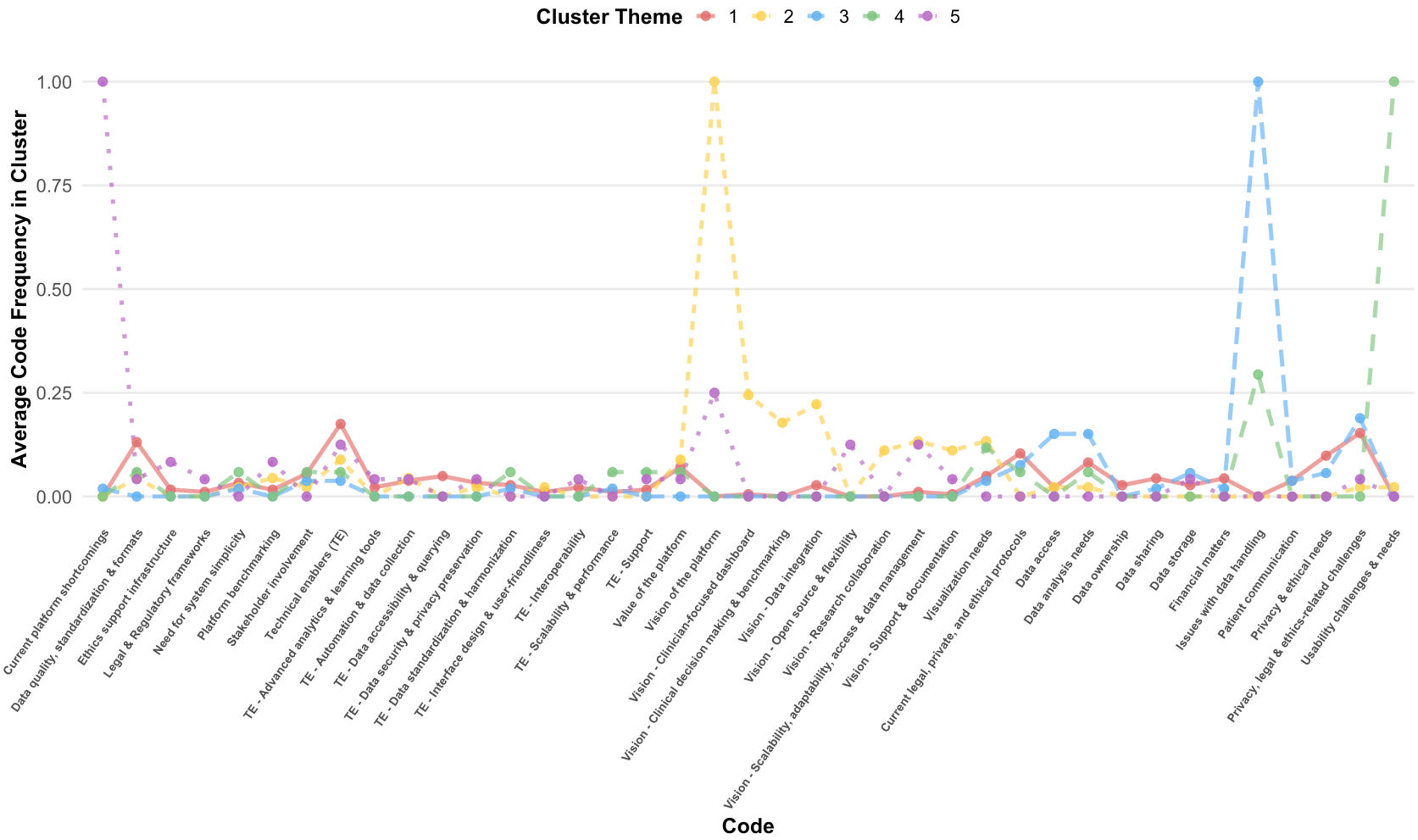


Note. The vertical axis represents the distance (dissimilarity) between clusters.

To validate and further characterise the thematic groupings identified through hierarchical clustering, Figure 6 depicts the cluster profiles generated from the k-means cluster analysis. Each cluster is represented by a separate line that presents the codes' distribution (distribution of the expert's needs). Peaks in the line signify codes that are especially prominent in particular clusters and, hence, central to the cluster's thematic orientation. Thus, the visualization provides an outline of the cluster constitution and the relative prominence of different codes across the clusters. Additionally, the display of these profiles further confirms that there is a unique distribution of codes defining the five clusters. Detailed descriptions of the cluster profiles and associated codes are provided further.

Figure 6

K-means Clustering



Note. The x-axis conveys the set of codes that emerged from qualitative analysis (n = 39), and the y-axis conveys the average code frequency within a cluster. Each coloured line represents one of the five clusters.

Key clusters emerged around the infrastructure integrity and legal foundations, strategic platform vision, ethical and operational barriers to data use, user-centric design and functional limitations, platform limitations and aspirations for openness. The clusters are defined by their most prominent themes, specifically based on the five codes with the highest frequency. Frequency in this context represents the average proportion of quotes within a cluster that were assigned a given code, thereby highlighting the central topics characterizing each cluster (see Table 5).

Table 5

Overview of Cluster Themes

Cluster Theme	Code	% Frequency
Infrastructure Integrity and Legal Foundations	Technical enablers (TE)	0.17
	Privacy, legal and ethics-related challenges	0.15
	Data quality, standardization and formats	0.13
	Current legal, private, and ethical protocols	0.1
	Privacy and ethical needs	0.1
Strategic Platform Vision	Vision of the platform	1.0
	Vision - Clinician-focused dashboard	0.24
	Vision - Data integration	0.22
	Vision - Clinical decision making and benchmarking	0.18
	Vision - Scalability, adaptability, access and data management	0.13
Ethical and Operational Barriers to Data Use	Issues with data handling	1.0
	Privacy, legal and ethics-related challenges	0.19
	Data access	0.15
	Data analysis needs	0.15

Cluster Theme	Code	% Frequency
	Current legal, private, and ethical protocols	0.08
User-Centric Design and Functional Limitations	Usability challenges and needs	1.0
	Issues with data handling	0.29
	Visualization needs	0.12
	Data quality, standardization and formats	0.06
	Need for system simplicity	0.06
Platform Limitations and Aspirations for Openness	Current platform shortcomings	1.0
	Vision of the platform	0.25
	Technical enablers (TE)	0.12
	Vision - Open source and flexibility	0.12
	Vision - Scalability, adaptability, access and data management	0.12

The first cluster is focused on *Infrastructure Integrity and Legal Foundations* (see Figure 7). This includes the need for experts for "*Technical enablers",* and contains discussion about *"Privacy, legal and ethics-related challenges",* as well as the need for *"Data quality, standardization and formats"*. This cluster captures the foundational prerequisites for a trustworthy data-sharing environment. Participants emphasized the importance of technical enablers, such as interoperability standards and system integration, alongside strong legal and ethical safeguards. Topics related to privacy needs, current protocols, and standardization suggest a concern for compliance, security, and quality in the underlying infrastructure. These concerns are directly elaborated on by participants:

> "…the most important thing... is that we need automation. We need to ease the way the data is... uploaded... And also that brings to interoperability. We don't need another platform... So the data gets there... because it's in a forgotten format..." (Group 4, Technical Expert)

"A part important technology enabler is how to make the data collection easy... Second, the important piece is related to data standardization. If data collected cannot talk to each other, that's a fail... The next important step is multi-modality data integration." (Group 4, Medical Expert)

"…privacy as one major challenge, uh, a hard hurdle for all these different hospitals and around different... countries to work together. So that's why we have to continue to develop the federated learning as an enabler..." (Group 4, Medical Expert)

Figure 7

Average Frequency of Each Code in the "Infrastructure Integrity and Legal Foundations" Cluster

Note. Codes are listed on the Y-axis. The X-axis indicates their average frequency within the cluster. Higher points reflect codes more commonly referenced in this cluster.

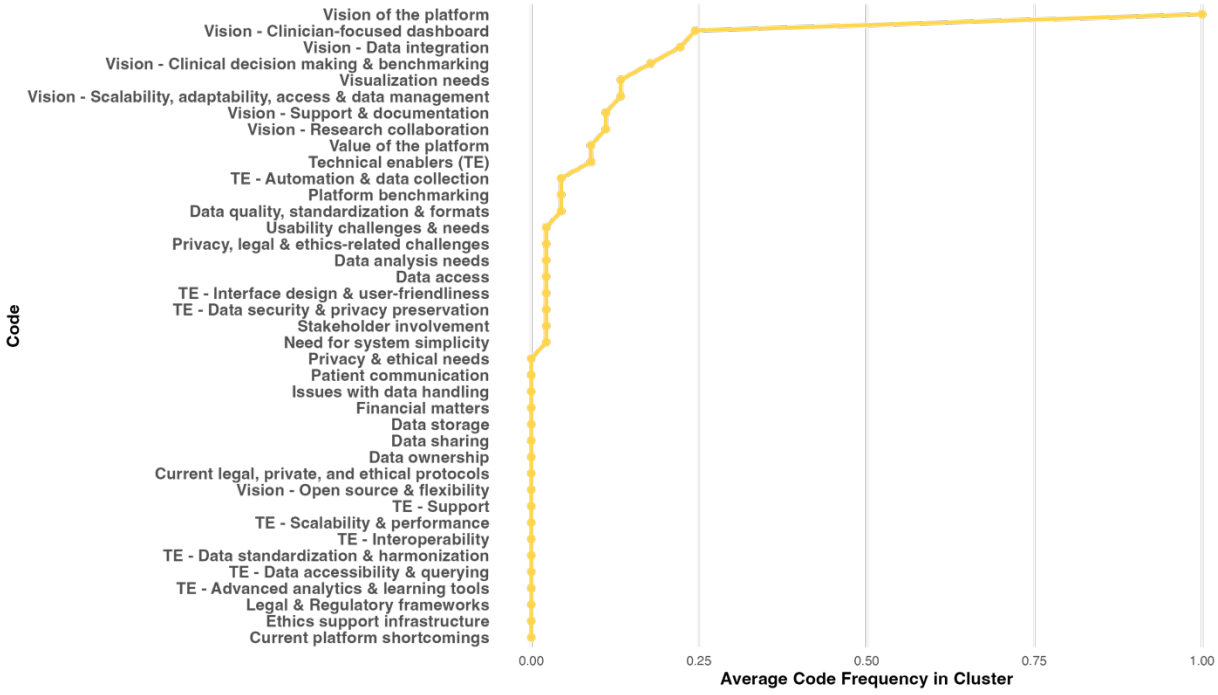
The second cluster, *Strategic Platform Vision*, was composed of topics regarding the "*Vision of the platform"* and contained a discussion about the expert's needs related to *"Vision - Clinician-focused dashboard",* and *"Vision - Data integration"* (see Figure 8). This cluster shared an aspirational and clinician-centered outlook on platform development. Emphasis was placed on future functionalities, including clinical dashboards, benchmarking support, and adaptable,

integrated systems. These needs reflect forward-thinking design considerations that prioritize usability and long-term scalability tailored to the needs of healthcare professionals. Several experts mentioned that:

> "In our platform, I included the genomic... data, there is a possibility to customize... decide the therapy for follow up... By having this kind of data, it's going to provide evidence for investing more in this area... introducing the molecular profile will tremendously help." (Group 6, Big data analytics - Technical Expert)
>
> "The tools... depend a lot on the information we want to take... meaningful ways, like practicing predefined questions... or interrogate the database... presenting the results... depends if the question is clinical or research... tailored for different needs." (Group 1, Geneticist - Clinical expert)

Figure 8

Average Frequency of Each Code in the "Strategic Platform Vision" Cluster



Note. Codes are listed on the Y-axis. The X-axis indicates their average frequency within the cluster. Higher points reflect codes more commonly referenced in this cluster.

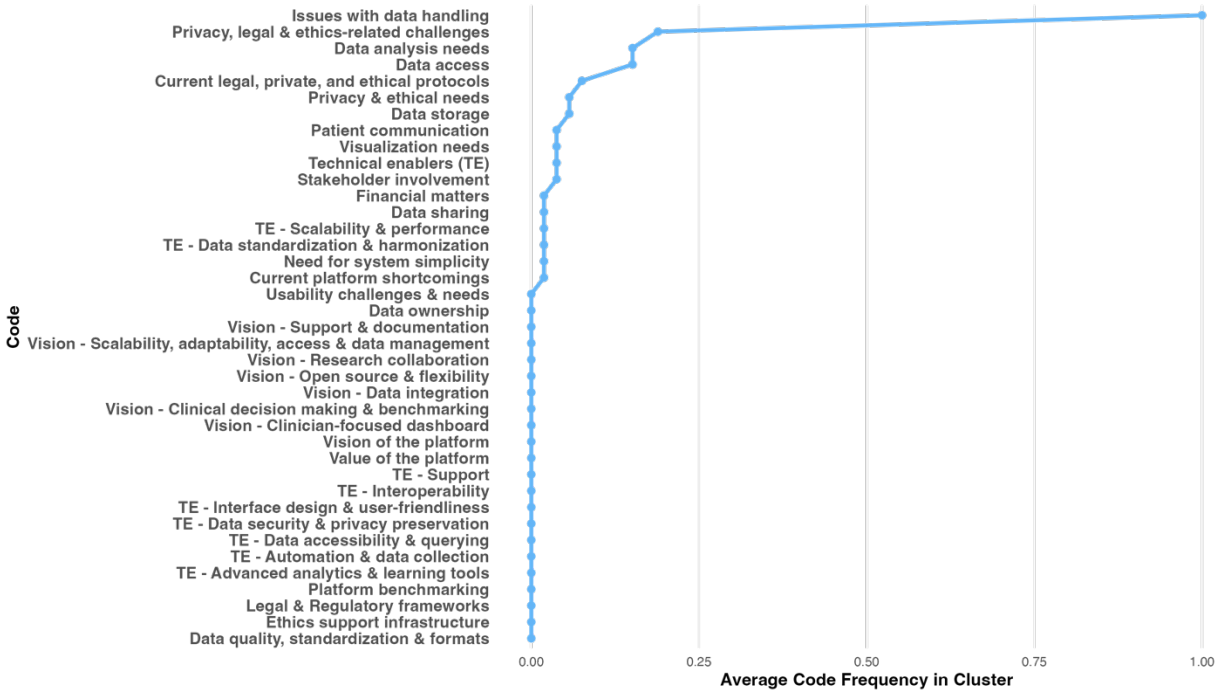
The third cluster, *Ethical and Operational Barriers to Data Use*, includes the need for experts for "*Issues with data handling*", where the discussion centers on the needs related to *"Data analysis*

needs", and *"Privacy, legal and ethics-related challenges"* (see Figure 9). The cluster is marked by a strong presence of concerns surrounding data accessibility, analysis, and ethical management. The co-occurrence of codes related to legal and procedural challenges points to systemic barriers that hinder efficient data use. Additionally, the inclusion of analysis-related needs highlights the tension between ethical safeguards and analytical utility. There were expressed by experts stating that:

> "…handling of large volumes of data... clinicians have very limited time... good standardization means having a form... a lot of screens... at a time, the best case scenario is the double of time allocated for every patient... So this is why clinicians... go to the open text box... different doctors will use different abbreviations... we need something like that..." (Group 5, Lab Scientist - Clinician)
>
> "…one of the first things that we would say is follow the national regulation... or your local standard procedure on sharing... not writing genomic results in the electronic health records... not writing personal information on WhatsApp groups or on emails..." (Group 5, Lab Scientist - Clinician)

Figure 9

Average Frequency of Each Code in the "Ethical and Operational Barriers to Data Use" Cluster



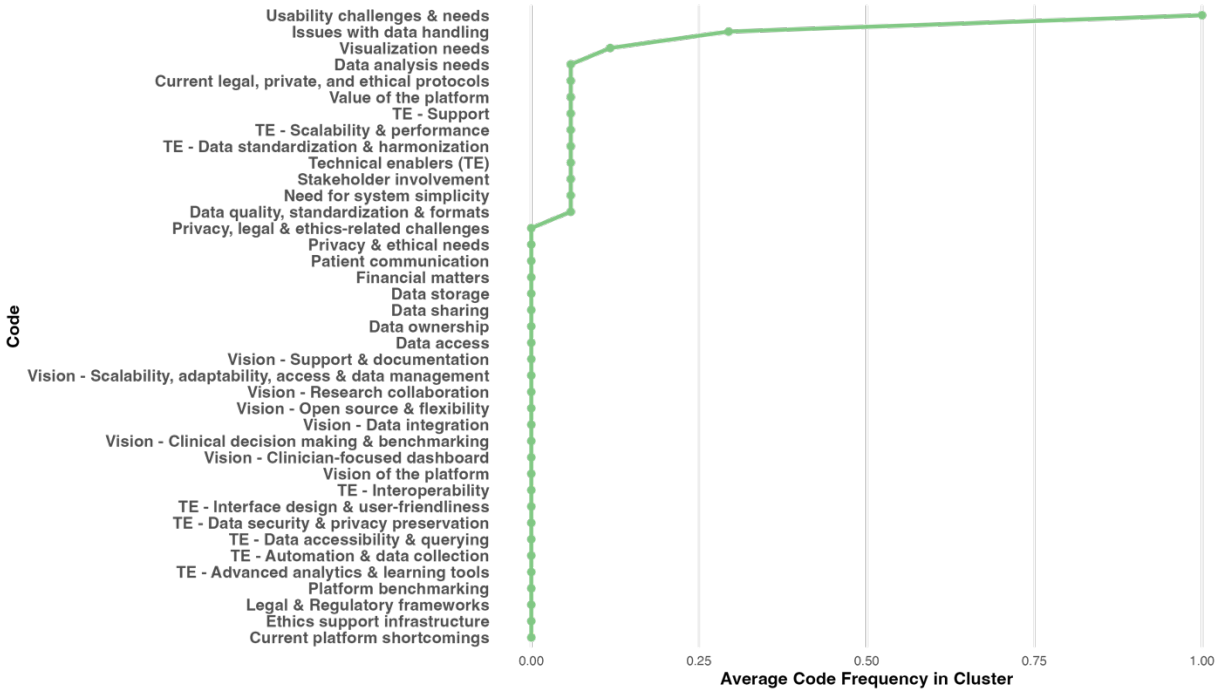
Note. Codes are listed on the Y-axis. The X-axis indicates their average frequency within the cluster. Higher points reflect codes more commonly referenced in this cluster.

The fourth, *User-Centric Design and Functional Limitations*, was supported by the needs of experts for "*Usability challenges and needs",* including the discussion about *"Visualization needs",* and *"Need for system simplicity"* (see Figure 10). The cluster draws attention to usability-related concerns, including challenges with interface design, visualization limitations, and the need for overall system simplicity. The combination of these codes reflects frustration with existing tools and a desire for more intuitive, user-friendly systems that support the practical workflows of clinicians and data professionals:

> "…the procedures are very neuropathic... either in multiple documents or in one huge document... all of these procedures take a lot of time... things that are very relevant... to speed up the process... standardized format for... communications between departments..." (Group 2, Pediatrics - Clinical Expert)
>
> "…if we speak about results, tables are the best way... if you have a specific clinical question... a graphical result or even a descriptive result, maybe it's better, but this is just something very personal." (Group 5, Lab Scientist - Clinician)

Figure 10

Average Frequency of Each Code in the "User-Centric Design and Functional Limitations" Cluster



Note. Codes are listed on the Y-axis. The X-axis indicates their average frequency within the cluster. Higher points reflect codes more commonly referenced in this cluster.

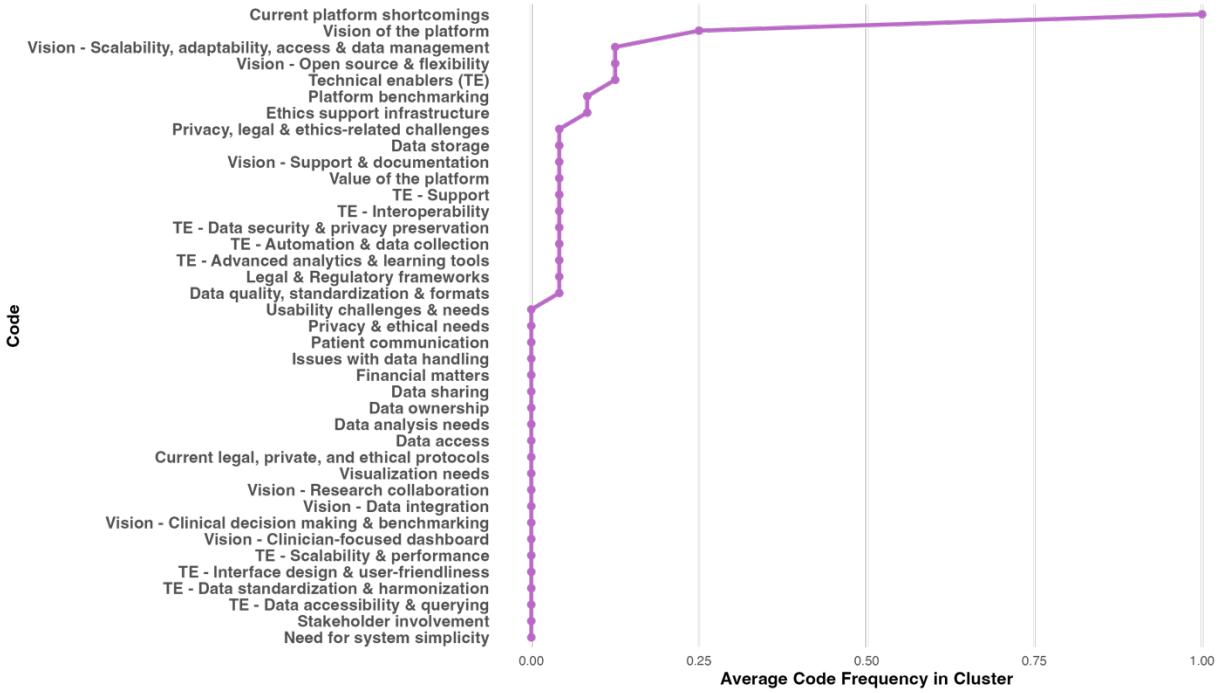
Finally, the last cluster, *Platform Limitations and Aspirations for Openness*, covers experts' discussion related to "*Current platform shortcomings" and the* "*Vision of the platform",* with the need *for "Vision – Scalability, adaptability, access & data management", as well as "Vision – Open source & flexibility"* (see Figure 11). The cluster blends critique of current platforms with forward-looking aspirations. Participants highlighted issues with existing systems while simultaneously advocating for technical enablers and open-source, adaptable architectures. The presence of both shortcomings and visionary codes suggests a recognition that current solutions are inadequate. However, a strong optimism exists regarding future improvements.

> "We want a lot of data... we need really a new technology that allows us to gather... integrate... interpret this data... we need a storage capacity and... data sharing or access, because currently... you misallocate... we expect that this new platform allows this consistency..." (Group 5, Lab Scientist - Clinician)
>
> "…the real good, uh, mobile dashboard app... All the users. This is a... very important tool... the visualization, the presentation... usability... we really need that... Visualization is

helping the expandability... Otherwise... Don't trust it. Can I use it?" (Group 6, Pediatrician - Clinician)

Figure 11

Average Frequency of Each Code in the "Platform Limitations and Aspirations for Openness" Cluster



Note. Codes are listed on the Y-axis. The X-axis indicates their average frequency within the cluster. Higher points reflect codes more commonly referenced in this cluster.

Discussion

This study aimed to examine the alignment between the literature-based expectations regarding the set of functional and non-functional system requirements needed to support genomic data-sharing platforms and the perspectives of clinical experts working with genomic data in real-world settings. Specifically, it sought to explore to what extent clinical experts agree with the importance of functional and non-functional requirements identified in the previous analysis of the literature by Resendez et al. (2025) (the “work-as-imagined” mental model) and to identify the practical challenges and unmet needs clinicians encounter when engaging with such platforms in clinical practice (the "work-as-done" perspective). To address these objectives, the study conducted a two-phase mixed-methods approach that included the data gathered from the survey of the expert opinions on system requirements and data collected from a workshop that explored expert perspectives on the design, usability, and real-world limitations that they can expect in the usage of such platforms. The findings from both phases suggest a substantial degree of alignment between the “work-as-imagined” mental model and the "work-as-done" perspective (Hollnagel & Clay-Williams, 2022; Thompson et al., 2023). However, several crucial divergences emerged.

Main Findings

Phase 1 clarified that the majority of the functional and non-functional requirements proposed by Resendez et al. (2025) were perceived as important and supported by clinical experts' agreement. Specifically, out of the total 62 evaluated requirements, 56 requirements (90.3%) were rated as important and reflected consensus, which indicates a strong validation of the “work-as-imagined” model. The analysis revealed 3 requirements (4.8%), including network visualization method and vertical and hybrid partitioning methods of data organization, that were rated as important, yet they presented a lack of agreement. This reflects divergent perspectives of the clinicians despite the perceived relevance. Additionally, 2 requirements (3.2%), such as use of PATRIC and DataSHaPER models for data standardization reached clinical experts' consensus on unimportance, suggesting a shared perception of its limited clinical value. Notably, no items were considered unimportant and contested, which further supports the overall relevance of the requirement set. Finally, 1 requirement (1.6%) defined as NVIDIA federated computing framework could not be evaluated due to missing data and was therefore excluded from the analysis, which reflects the clinical experts shared limited knowledge of this requirement. These

findings support that while the literature-based system features align with expert expectations, a few aspects may require refinement to address interpretation variability.

Requirements Clinically Validated

The majority of requirements (90.3%) were rated as important and agreed among experts, which indicates strong support for the foundational elements of the literature-based model. These findings support the relevance of the "work-as-imagined" conceptualization (Resendez et al., 2025) and suggest that generally, the imagined design of the genomic data-sharing platform corresponds closely with current clinical expectations. These requirements covered a broad range of functional and non-functional requirements, which reflects shared professional priorities across clinical contexts.

Functional requirements related to the core data processing functions, such as genomic data acquisition, genomic data upload, automated data completeness check and data quality control, received consistent agreement among the clinical experts. Their importance aligns with the prior research that emphasized the significance of consistent and standardized data input for ensuring clinical interpretability and maintaining data integrity in collaborative genomic platforms (Resendez et al., 2025; Tommel et al., 2023). In the workshop discussion, the theme "Infrastructure Integrity and Legal Foundations" highlights the challenge identified by the experts in integrating heterogeneous datasets across institutions and the need for reliable input tools. These findings reflect the shared expert's perspectives on reducing time spent on formatting issues and improving the trustworthiness of downstream analysis. Future platform development should focus on implementing automatic data validation tools, supporting commonly used genomic data formats, and ensuring upload compatibility with various institutional systems. Further studies can focus on exploring how clinical experts interact with these processes.

Requirements related to knowledge sharing and data sharing were rated as important and agreed upon by clinical experts in the survey. According to the defined literature-based model by Resendez et al. (2025), knowledge exchange is one of the key components that ensures platform efficiency. Additionally, according to the workshop outcomes knowledge sharing was raised by the experts under the theme "Strategic Platform Vision", which reflects the need of the experts for a more collaborative environment in platform design. The participants also state that these will help them to ensure the interoperability of the genomic data analysis outcomes. It will allow the exchange of different views, which will further facilitate the accuracy of diagnosing the diseases

and creating treatment plans. Future platform development should explore these collaborative tools that emphasize shared interpretation. As well as further research is needed to evaluate which forms of knowledge sharing support decision-making and reduce interpretive uncertainty in genomic diagnostics.

Furthermore, requirements that were rated as important and agreed upon relate to the types of analyses supported by the platform and the need for reproducibility. These requirements reflect a shared expectation that genomic platforms should not only store the data, but also support robust and consistent analytical workflows. The literature supports that the wide range of genomic analysis has been highlighted as crucial for trust in decision-making workflows and for rapid interpretation of genomic data (Kaasalainen, 2025; Alzu'bi et al., 2014; Ceri & Pinoli, 2020). This helps clinical experts overcome time constraints inherent to their workflows. During the workshop, experts discussed the need for the integration and challenges of different data analysis tools. These discussions were underlined from the theme "Infrastructure Integrity and Legal Foundations" and "Ethical and Operational Barriers to Data Use". Further research can investigate how clinicians engage with analytic outputs and what design elements best support interpretability, trust, and cross-institutional re-use.

Experts also validated requirements related to platform usability as important components to reduce barriers to utilizing genomic data in clinical practice. Requirements such as mobile-friendly access, multilingual support, and notifications received an agreement from the experts on the importance of their integration to the genomic data-sharing platform. The concerns for usability and functional relevance expressed by clinicians parallel the concerns raised by Kaasalainen (2025) and Lau-Min et al. (2022), who observe that technical functionalities must come together with pragmatic design. The findings related to the integration of notifications align with those of Tommel et al. (2023) suggesting that it enhances dynamic engagement with the system. Additionally, the findings from the survey align with the identified theme of "User-Centric Design and Functional Limitations" from the workshop discussion and reflect a demand for platforms that accommodate real-world user needs. The experts noted during the workshop that integrating the mobile dashboard app is an important tool to ensure user-friendliness design. Future research could investigate how different types of notifications, such as reminders, alerts and updates, affect clinician engagement, decision-making efficiency, and alert fatigue. Additionally, usability testing

of multilingual and mobile interfaces in real-world settings would help determine whether these features improve access and reduce workflow disruption.

A strong consensus concerning data privacy and security standards was observed, which is consistent with previous studies highlighting the need for secure data infrastructures in genomic medicine to ensure trust among users and legitimize data transfer within the required legal framework (Balagurunathan & Sethuraman, 2024; Bowdin et al., 2016). Additionally, this theme was actively discussed in the workshop, where clinicians highlighted the need for platform security and data privacy protection measures. These concerns emerged under the themes of "Infrastructure Integrity and Legal Foundations" and "Ethical and Operational Barriers to Data Use". These findings expand on previous research that emphasizes the value of robust data security measures, such as encryption, authentication methods, etc. (Tommel et al., 2023). Though the literature portrays these measures as helpful in maintaining patient trust and privacy, the workshop discussions provided a more nuanced picture. Despite experts' need for privacy and security concepts, they stressed during the workshop that additional protective measures should not hinder their workflow. Experts concurred that clinicians find it more challenging to manage increasingly complex systems, which is supported by the Dahlquist et al. (2023) findings. Consequently, the workshop findings enhance the literature rather than contradict it, implying that although these protective measures are helpful, their applicability mostly depends on how well they integrate with the expert's workflow and cognitive load. Future development of platforms should focus on integrating adaptive security protocols to ensure trust among the users, and that would comply with evolving legal regulations while remaining flexible to support cross-border data use. Additionally, future research should be conducted on integrating specific security and privacy measures that align with experts' expectations and workflow.

Requirements related to federated computing criteria, such as access to federated infrastructure and criteria for implementing distributed models, received strong support for the implementation of these requirements in the system. This aligns with the findings related to the growing recognition of federated and privacy-preserving technologies in genomic data exchange (Office of the National Coordinator for Health Information Technology, 2023). During the workshop, experts stressed the challenge and relevance of federated approaches to enable cross-institutional collaboration. These discussions were underlined from the theme "Ethical and Operational Barriers to Data Use". The consistent agreement on these requirements suggests that

clinical experts accept federated computing as an important element for the implementation of modern platforms. Future development should focus on integrating the federated technologies into the system, and further studies should assess the institutional readiness and technical support which is required for widespread implementation.

These results suggest that the foundational assumption of the mental model developed by Resendez et al. (2025) presents significant validity from the clinical perspective. The similarities between the theoretical expectations and real-world practice provide a strong foundation to proceed with the design and implementation of these requirements for the future genomic data-sharing platform. The Phase 2 findings added valuable insights by revealing how widely agreed-upon requirements can support the clinical practice and satisfy the needs of the clinical experts.

Requirements with Disagreement Despite Importance

Three requirements, data organisation methods - hybrid and vertical partitioning, as well as network-based visualisation were rated as important by clinical experts, however, lacked expert consensus. While their average scores indicate that experts saw value in these requirements, the lack of agreement suggests that they might be context-dependent (applicable only in certain situations or types of hospitals), technically complex (lack the understanding of the requirement meaning), or conceptually ambiguous (wording or idea might be unclear). These issues may affect the consistent interpretation across clinical settings.

Both hybrid and vertical partitioning are advanced techniques used in federated learning systems to determine how data is distributed and processed. These approaches are often discussed in the technical literature (Yu et al., 2024), however, their implementation in real-world clinical practice remains limited. Some experts may likely understand their operational benefits and rated it as important, other experts probably were unfamiliar with the term or uncertain about their relevance to routine workflows. These may explain the disagreement on both requirements. Additionally, during the workshop discussion, these terms were not discussed in depth. The absence of these elaborations may suggest that such features may not be broadly understood by the limited familiarity with the term or prioritised in day-to-day clinical settings. However, taking into consideration the perceived importance of these partitioning strategies, there is a need for further exploration. Future platform development should ensure that such advanced techniques are either well-explained or offered through adaptable interfaces, where clinical experts can engage with the platform according to their knowledge and expertise. Furthermore, studies can also benefit

from clarifying the terminologies in surveys and interviews to reduce the potential misinterpretation of complex technical concepts.

One of the preferred visualisation methods, network-based visualisation, scored highly in importance, but received disagreement among the clinical experts. During the workshop discussion experts indicated the importance of visualisation tools for the interpretation of the genomic data. However, there was no specific reference to network visualisations. This may suggest that either experts lack familiarity with the tool or that they are not commonly applied in clinical workflow. Instead, general comments such as the need for clear and interpretable visualisation interfaces were more prominent, which supports the idea by Qu et al. (2019) that states the relevance of visualisation tools to simplify complex datasets and enable efficient interpretation of the genomic data results. Even though the findings result in a lack of information related to its application, the requirement should not be dismissed as justified by the perceived importance identified in the survey. Future research could explore how and when network-based visualisation tools are helpful in different user groups and whether offering them as optional tools could enhance the usability of genomic platforms without overwhelming less experienced users.

Requirements Considered Not Important but Agreed

The requirements that are identified by the expert's agreement as not important are the standardization models named DataSHaPER and PATRIC. These models are designed to harmonize data and promote a standardization framework in global genomic research (Fortier et al., 2010). However, in this study, clinicians presented a shared perception that the relevance of these models is limited to clinical workflow. This result aligns with broader challenges that highlight the complex adoption of “top-down” data harmonization tools, which are designed by research institutions without consideration of their relevance to clinical workflow (Doiron et al., 2013). Experts may view such models as overly abstract or lacking support for implementation in real-time diagnostic contexts. Furthermore, no specific mention of the DataSHaPER and PATRIC models was made during the workshop discussion, which may indicate unfamiliarity or disinterest, further supporting its low perceived clinical value. While data standardization models are important in the genomic data-sharing platforms, future systems may benefit more from context-sensitive standards (fit specific clinical situations) and interoperable standards (allow different systems to talk to each other smoothly), which will reflect the workflow of clinical experts. Future

platform development should re-evaluate the relevance of implementing the DataSHaPER and PATRIC models.

Together, the quantitative and qualitative findings provide a complementary yet nuanced understanding of the alignment between imagined systems requirements and clinical realities. Quantitative results show a broad agreement on many key requirements, such as core data processing function, data sharing, type of analyses, platform usability, security and privacy measures, and federated computing. This was also observed through qualitative results, where clinicians identified legal and ethical compliance, infrastructure integrity, and usability as the primary concerns. It suggests that, generally, the requirements of the imagined mental model are agreed upon. However, how practitioners see it actually implemented needs to take contextual, ethical, and usability concerns into consideration.

Theoretical and Practical Implications

This study builds upon and furthers the existing theoretical foundation surrounding mental models in genomic data-sharing platform design. As was discussed in the literature review, Resendez's (2025) framework was a "work-as-imagined" mental model, an aggregation of features and assumptions gleaned from systematic literature to describe functional and non-functional requirements of genomic data-sharing platforms. The present study attempted to fill the gap between the theoretical model with respect to real-world clinical needs and practices by empirically studying whether or not and how these imagined requirements align with "work-as-done" in clinical practice.

The findings of this study offer several theoretical and practical contributions to the genomic data-sharing platforms design and evaluation. From a theoretical standpoint, the research findings facilitate the application of "work-as-imagined" and "work-as-done" frameworks (Hollnagel & Clay-Williams, 2022; Thompson et al., 2023) within the context of digital healthcare infrastructures. The application of this approach to the field of platform requirement validation offers a fresh perspective, even if it has usually been utilised to examine disparities in safety-critical domains. The results show that the "work-as-imagined" mental model can be a useful starting point for matching stakeholder expectations with system design. Specifically those mental models that are developed through organised literature synthesis, such as the mental model developed by Resendez et al. (2025).

The high level of agreement among clinical experts on most functional and non-functional requirements was validated by the quantitative and qualitative phases with the “work-as-imagined” model. It provides empirical support for the theoretical proposition that effective system design must align with users’ cognitive expectations and workflows, as previously highlighted in studies by Yildirim et al. (2024) and Chen et al. (2008). Additionally, it supports the theoretical utility of “work-as-imagined” mental models as a source for the anticipation of the stakeholder's needs (Hollnagel & Clay-Williams, 2022; Thompson et al., 2023). However, there are also important improvements that need to be made in areas of divergence, such as network-based visualisation tools, data organisation methods and standardisation methods. Instead of contradicting the theoretical model, these results enhance it by showing how organised, literature-based depictions of ideal platforms can accurately reflect clinical realities when they are regularly assessed against feedback from the real world. Thus, the study advances a more sophisticated theoretical understanding of the role that “work as imagined” models play as design scaffolds in complex multi-user settings. Therefore, this study provides insights that “work-as-imagined” model can meaningfully inform the design of other digital systems in healthcare and beyond, given that pre-implementation models need to be evaluated against actual practice.

On a practical level, the findings explicitly provide guidelines for the genomic data-sharing platform designers. Following the recommendations in Table 1, the implementation of the requirements that received consensus on the importance among clinicians would facilitate clinical workflows in genomic diagnostics and precision medicine.

Moreover, the findings highlight the necessity for federated computing features for privacy, model accuracy, and efficient computation. This confirms the growing timeliness of decentralised architecture for healthcare data exchange, in line with privacy and interoperability demands set forth in various EU regulations, such as the EHDS, GDPR, and AI Act. Expert concerns about usability, ethics, and legal adaptability suggest system designers must take into consideration legal interpretations as well as the linguistic and technological diversity of users across the EU, which is supported by the literature (Pormeister, 2018; Molnár-Gábor & Korbel, 2020). Finally, this study strengthens existing calls in the clinical literature for better training and support for clinicians dealing with genomic data (Ashton-Prolla et al., 2015). Insinuating that adoption of the platform

must be accompanied by institutional policies on education, interdisciplinary collaboration, and streamlining workflows.

Strengths, Limitations and Recommendations for Future Research

This study has several strong points to take into consideration. Firstly, the study utilized a mixed-method approach to explore clinical experts' perspectives on the requirements of the genomic data-sharing platform. To our knowledge, it is the first research study that explores the clinician's opinion on the complete mental model that includes the requirements for the genomic data-sharing platform. It combines a quantitative survey data collection supported with a qualitative workshop interview to obtain additional insights. Furthermore, the diversity in the clinical expert disciplines who participated in the survey and workshop significantly improves the validity and relevance of the findings. Experts ranging from paediatricians to genetics make certain that the requirements and challenges that have been discovered encompass a wide range of practical viewpoints and applications throughout the healthcare environment.

While this study provides important insights into the clinical experts' perspectives on the previously identified set of requirements for the genomic data-sharing platform, several methodological and conceptual limitations of the study must be addressed. These limitations affect the interpretation of the results and provide direction for future research.

For the analysis of the survey study, I adopted a commonly used classification system commonly used to determine consensus on the importance of the requirements, such as threshold for mean to determine importance and IQR to determine agreement (Bodmer et al., 2019). While this method effectively highlights the central tendency and the level of expert agreement on the requirements needed, it also presents an entree for venturing more in-depth exploration of the complexity of expert opinions. For instance, this approach may underrepresent minority viewpoints or overlook the reasoning behind disagreements, especially in cases where requirements were rated as important but revealed divergent views (Nair et al., 2011). These instances of disagreement present an opportunity to identify unresolved issues or conflicts within the professional community. Future research can build on this by systematically collecting and analyzing the qualitative justifications provided by experts alongside quantitative measures. This method can further shed light on the reasons behind agreement or disagreement, capture complex viewpoints, and result in platform requirements that are more inclusive and context sensitive.

In designing the workshop component, participants were provided with a predefined set of requirements to round their discussion. This structured approach was chosen to ground conversations in a shared understanding and ensure the relevancy of the discussed topics. While the predefined list helped to structure and streamline the workshop discussion, it may have influenced participants' ability to challenge the foundational assumptions or propose some new insight. The methodological choice may have brought a framing effect, which is an individual's cognitive bias that influences people's judgements by how information is presented rather than the information itself (Tabesh et al., 2019). This presents an opportunity for future research to build a defined methodological approach by implementing exploratory formats. For instance, hybrid workshops could be conducted where some groups would handle their discussion with the list of requirements and other groups independently from the predefined models to allow for more bottom-up discussion. This method can yield deeper insights by comparing the outcomes between these formats into the emergence of needs and system visions.

The study deliberately focused on clinical experts who shared a common background drawn from European clinical networks. This choice was justified by the study's association with the EU-funded PROTECT-CHILD project. This targeted sampling ensured contextual relevance and consistency with the policy environment in which the platform is intended to be used. Nevertheless, the chosen sampling approach presents an opportunity for future research to explore whether the identified requirements are universally applicable by bringing a broader sample that includes clinicians from a broader range of countries and healthcare settings. Since healthcare systems outside the EU differ in their data-sharing infrastructures, governance models, and clinical workflows. Including in the future studies clinicians from non-European contexts would assist in evaluating the cross-cultural and cross-system generalizability of the findings, and perhaps point to context-specific needs that fall outside the EU framework.

Furthermore, the study explored the experts' perspectives on the identified system requirements, providing valuable insights into design priorities and anticipated challenges and needs. Although this strategy was suitable for the initial phases of requirements validation, it also offers a chance for further study to expand on it by examining the requirements' real-world applications. Incorporating usability testing or observational field studies could help to assess how requirements manifest in practice. Specifically, how it influences cognitive load, workflow integration, and clinical decision-making. These kinds of studies would offer vital proof to close

the gap between the real user experience and the perceived demands, resulting in more practical and efficient genomic data-sharing systems.

Conclusion

This study provided an accurate assessment of the needs and challenges faced by clinicians in the context of genomic data-sharing platforms. The findings validate the application of most of the literature-based requirements while also highlighting significant areas where theoretical presumptions and practical demands diverge. By combining expert consensus with practical insights, the study offers useful guidance for developing genomic data-sharing systems that are both technically sound and clinically meaningful. These contributions support the development of more practical, ethically aware, and clinician-aligned precision medicine systems in the future.

References

Alzu'bi, A., Zhou, L., & Watzlaf, V. (2014). Personal genomic information management and personalized medicine: Challenges, current solutions, and roles of him professionals. *Perspectives in Health Information Management*, *11(Spring)*, 1c

Ashton-Prolla, P., Goldim, J. R., Vairo, F. P. E., Da Silveira Matte, U., & Sequeiros, J. (2015). Genomic analysis in the clinic: benefits and challenges for health care professionals and patients in Brazil. *Journal of Community Genetics*, *6*, 275–283. https://doi.org/10.1007/s12687-015-0238-0

Balagurunathan, Y., & Sethuraman, R. R. (2024). An analysis of ethics-based foundation and regulatory issues for genomic data privacy. *Journal of the Institution of Engineers (India): Series B, 105*, 1097–1107. https://doi.org/10.1007/s40031-024-01058-3

Belotto, M. (2018). Data analysis methods for qualitative research: Managing the challenges of coding, interrater reliability, and thematic analysis. *The Qualitative Report*, *23*(11), 2622–2633. https://doi.org/10.46743/2160-3715/2018.3492

Binson, V. A., Thomas, S., Subramoniam, M., Arun, J., Naveen, S., & Madhu, S. (2024). A review of machine learning algorithms for biomedical applications. *Annals of Biomedical Engineering*, *52*, 1159–1183. https://doi.org/10.1007/s10439-024-03459-3

Bodmer, N. S., Häuselmann, H. J., Frey, D., Aeberli, D., & Bachmann, L. M. (2019). Expert consensus on relevant risk predictors for the occurrence of osteoporotic fractures in specific clinical subgroups – Delphi survey. *BMC Rheumatology*, *3*. https://doi.org/10.1186/s41927-019-0099-y

Boone, H., & Boone, D. (2012). Analyzing likert data. *Journal of Extension*, *50*(2). https://doi.org/10.34068/joe.50.02.48

Bowdin, S., Gilbert, A., Bedoukian, E., Carew, C., Adam, M. P., Belmont, J., Bernhardt, B., Biesecker, L., Bjornsson, H. T., Blitzer, M., D'Alessandro, L. C., Deardorff, M. A., Demmer, L., Elliott, A., Feldman, G. L., Glass, I. A., Herman, G., Hindorff, L., Hisama, F., . . . Krantz, I. D. (2016). Recommendations for the integration of genomics into clinical practice. *Genetics in Medicine*, *18*(11), 1075–1084. https://doi.org/10.1038/gim.2016.17

Byrd, J. B., Greene, A. C., Prasad, D. V., Jiang, X., & Greene, C. S. (2020). Responsible, practical genomic data sharing that accelerates research. *Nature Reviews Genetics*, *21*, 615–629. https://doi.org/10.1038/s41576-020-0257-5

Ceri, S., & Pinoli, P. (2020). Data science for genomic data management: Challenges, resources, experiences. *SN Computer Science*, *1*, 5. https://doi.org/10.1007/s42979-019-0005-0

Chen, D. T., Mills, A. E., & Werhane, P. H. (2008). Tools for tomorrow's health care system: A systems-informed mental model, moral imagination, and physicians' professionalism. *Academic Medicine*, *83*(8), 723–732. https://doi.org/10.1097/acm.0b013e31817ec0d3

Cooksey, R. W. (2020). Descriptive statistics for summarising data. In *Illustrating Statistical Procedures: Finding Meaning in Quantitative Data* (pp. 61–139). https://doi.org/10.1007/978-981-15-2537-7_5

Correll, M., & Gleicher, M. (2014). Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 2142–2151. https://doi.org/10.1109/tvcg.2014.2346298

Credle, K. (2022). *Bridging the gap: How molecular tumor boards thrive with standardized genomic data*. FrameShift Genomics. https://www.frameshift.com/blog/how-molecular-tumor-boards-thrive

Dahlquist, J. M., Nelson, S. C., & Fullerton, S. M. (2023). Cloud-based biomedical data storage and analysis for genomic research: Landscape analysis of data governance in emerging NIH-supported platforms. *Human Genetics and Genomics Advances*, *4*, 100196. https://doi.org/10.1016/j.xhgg.2023.100196

Doiron, D., Burton, P., Marcon, Y., Gaye, A., Wolffenbuttel, B. H. R., Perola, M., Stolk, R. P., Foco, L., Minelli, C., Waldenberger, M., Holle, R., Kvaløy, K., Hillege, H. L., Tassé, A., Ferretti, V., & Fortier, I. (2013). Data harmonization and federated analysis of population-based studies: the BioSHaRE project. *Emerging Themes in Epidemiology*, *10*. https://doi.org/10.1186/1742-7622-10-12

Elhussein, A., Baymuradov, U., Elhadad, N., Natarajan, K., & Gürsoy, G. (2024). A framework for sharing of clinical and genetic data for precision medicine applications. *Nature Medicine*, *30*, 3578–3589. https://doi.org/10.1038/s41591-024-03239-5

European Commission. (2024, May 21). *Global health*. EU Global Health Strategy. https://health.ec.europa.eu/internationalcooperation/global-health_en

European Parliament & Council of the European Union. (2024, March 18). *Regulation of the European Parliament and of the Council on the European Health Data Space (PE 76/2024*

INIT) [PDF]. Council of the European Union. Retrieved June 24, 2025, from https://data.consilium.europa.eu/doc/document/PE-76-2024-INIT/en/pdf

European Parliament & European Council (2022). Proposal for a Regulation of the European Parliament and of the Council on the European Health Data Space, No. Regulation (EU) 2022 (2022). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52022PC0197

European Parliament & European Council (2023). Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on Harmonised Rules on Fair Access to and Use of Data and Amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act) (2016). https://doi.org/10.5040/9781782258674

Fortier, I., Burton, P. R., Robson, P. J., Ferretti, V., Little, J., L'Heureux, F., Deschenes, M., Knoppers, B. M., Doiron, D., Keers, J. C., Linksted, P., Harris, J. R., Lachance, G., Boileau, C., Pedersen, N. L., Hamilton, C. M., Hveem, K., Borugian, M. J., Gallagher, R. P., . . . Hudson, T. J. (2010). Quality, quantity and harmony: The DataSHaPER approach to integrating data across bioclinical studies. *International Journal of Epidemiology*, *39*(5), 1383–1393. https://doi.org/10.1093/ije/dyq139

General Data Protection Regulation (GDPR), No. 2016/679 (2016). https://gdpr-info.eu/

Ginsburg, G. S., & Phillips, K. A. (2018). Precision medicine: From science to value. *Health Affairs*, *37*(5), 694–701. https://doi.org/10.1377/hlthaff.2017.1624

Global Alliance for Genomics and Health. (2016). A federated ecosystem for sharing genomic, clinical data. *Science, 352*(6291), 1278–1280.

Haque, S. N., Kansky, J. P., & Dixon, B. E. (2023). Managing the business of health information exchange: moving towards sustainability. In *Health information exchange* (2nd ed., pp. 113–130). Academic Press. https://doi.org/10.1016/b978-0-323-90802-3.00009-5

Henry, D., Dymnicki, A. B., Mohatt, N., Allen, J., & Kelly, J. G. (2015). Clustering methods with qualitative data: A mixed-methods approach for prevention research with small samples. *Prevention Science*, *16*, 1007–1016. https://doi.org/10.1007/s11121-015-0561-z

Hide, W. (2005). Genome Databases. In *Encyclopedia of Life Sciences*. Wiley. https://doi.org/10.1038/npg.els.0005314

Hollnagel, E., & Clay-Williams, R. (2022). Work-as-Imagined and Work- as-Done. In *Routledge eBooks* (pp. 175–177). https://doi.org/10.4324/9781003109945-52

Joshi, A., Kale, S., Chandel, S., & Pal, D. (2015). Likert scale: explored and explained. *British Journal of Applied Science & Technology*, *7*(4), 396–403. https://doi.org/10.9734/bjast/2015/14975

Kaasalainen, T. (2025, January 22). *Clinical genomics: a key component of modern healthcare*. Euformatics. https://www.euformatics.com/blog-post/clinical-genomics-a-key-component-of-modern-healthcare

Kalaian, S. A. (2008). Frequency distribution. *Encyclopedia of Survey Research Methods*, 293–294. https://doi.org/10.4135/9781412963947.n195

Karacic, J. (2022). Europe, we have a problem! Challenges to health data-sharing in the EU. *In 2022 18th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)* (pp. 47–50). IEEE. https://doi.org/10.1109/WiMob55322.2022.9941532

Khoury, M. J., Bowen, S., Dotson, W. D., Drzymalla, E., Green, R. F., Goldstein, R., Kolor, K., Liburd, L. C., Sperling, L. S., & Bunnell, R. (2022). Health equity in the implementation of genomics and precision medicine: A public health imperative. *Genetics in Medicine*, *24*(8), 1630–1639. https://doi.org/10.1016/j.gim.2022.04.009

Kotonya, G., & Sommerville, I. (1998). Requirements Engineering: Processes and techniques. In *Wiley Publishing eBooks*. http://ci.nii.ac.jp/ncid/BA36137984

Lau-Min, K. S., McKenna, D., Asher, S. B., Bardakjian, T., Wollack, C., Bleznuck, J., Biros, D., Anantharajah, A., Clark, D. F., Condit, C., Ebrahimzadeh, J. E., Long, J. M., Powers, J., Raper, A., Schoenbaum, A., Feldman, M., Steinfeld, L., Tuteja, S., VanZandbergen, C., . . . Nathanson, K. L. (2022). Impact of integrating genomic data into the electronic health record on genetics care delivery. *Genetics in Medicine*, *24*(11), 2338–2350. https://doi.org/10.1016/j.gim.2022.08.009

León, A., & Pastor, Ó. (2021). Enhancing precision medicine: A big data-driven approach for the management of genomic data. *Big Data Research*, *26*. https://doi.org/10.1016/j.bdr.2021.100253

Li, R., Romano, J. D., Chen, Y., & Moore, J. H. (2024). Centralized and federated models for the analysis of clinical data. *Annual Review of Biomedical Data Science*, *7*, 179–199. https://doi.org/10.1146/annurev-biodatasci-122220-115746

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., . . . Moore, H. F. (2013). The genotype-tissue expression (GTEx) project. *Nature Genetics*, *45*(6), 580–585. https://doi.org/10.1038/ng.2653

Ma, A., O'Shea, R., Wedd, L., Wong, C., Jamieson, R. V., & Rankin, N. (2024). What is the power of a genomic multidisciplinary team approach? A systematic review of implementation and sustainability. *Europe Journal of Human Genetics, 32*, 381–391. https://doi.org/10.1038/s41431-024-01555-5

Ma'ruf, M., Irham, L. M., Adikusuma, W., Sarasmita, M. A., Khairi, S., Purwanto, B. D., Chong, R., Mazaya, M., Muhammad, L., & Siswanto, L. M. H. (2023). A genomic and bioinformatic-based approach to identify genetic variants for liver cancer across multiple continents. *Genomics and Informatics*, *21*(4). https://doi.org/10.5808/gi.23067

Malakar, Y., Lacey, J., Twine, N. A., McCrea, R., & Bauer, D. C. (2023). Balancing the safeguarding of privacy and data sharing: perceptions of genomic professionals on patient genomic data ownership in Australia. *European Journal of Human Genetics*, *32*, 506–512. https://doi.org/10.1038/s41431-022-01273-w

Mani, S., Lalani, S. R., & Pammi, M. (2025). Genomics and multiomics in the age of precision medicine. *Pediatric Research*, *97*, 1399-1410. https://doi.org/10.1038/s41390-025-04021-0

Marcus, J. S., Martens, B., Carugati, C., Bucher, A., & Godlovitch, I. (2022). The European Health Data Space. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4300393

McLaughlin, L., Mahon, S. M., & Khemthong, U. (2024). A systematic review of genomic education for nurses and nursing students: Are they sufficient in the era of precision health? *Nursing Outlook*, *72*(5), 102266. https://doi.org/10.1016/j.outlook.2024.102266

Molnár-Gábor, F., & Korbel, J. O. (2020). Genomic data sharing in Europe is stumbling—Could a code of conduct prevent its fall? *EMBO Molecular Medicine*, *12*. https://doi.org/10.15252/emmm.201911421

Morley, J., Murphy, L., Mishra, A., Joshi, I., & Karpathakis, K. (2022). Governing data and artificial intelligence for health care: Developing an international understanding. *JMIR Formative Research*, *6*(1), e31623. https://doi.org/10.2196/31623

Nair, R., Aggarwal, R., & Khanna, D. (2011). Methods of formal consensus in classification/diagnostic criteria and guideline development. *Seminars in Arthritis and Rheumatism*, *41*(2), 95–105. https://doi.org/10.1016/j.semarthrit.2010.12.001

National Cancer Institute. (2019). *The Cancer Genome Atlas Program (TCGA)*. U.S. Department of Health and Human Services. Retrieved June 21, 2025, from https://www.cancer.gov/ccg/research/genome-sequencing/tcga

Office for Civil Rights (OCR), & Office of the Secretary, Department of Health and Human Services. (2024). *HIPAA Privacy Rule to support reproductive health care privacy* (No. RIN 0945-AA20; p. 89). Department of Health and Human Services. https://www.govinfo.gov/content/pkg/FR-2024-04-26/pdf/2024-08503.pdf

Office of the National Coordinator for Health Information Technology. (2023). *Sharing genomic data and information for patient care via application programming interfaces: Sync for Genes Phase 4 final report*. U.S. Department of Health and Human Services. https://www.healthit.gov/sites/default/files/page/2022-06/Sync%20for%20Genes%20Phase%204%20Final%20Report_508.pdf

Pandey, A., & Gupta, S. P. (2024). Personalized medicine: A comprehensive review. *Oriental Journal Of Chemistry*, *40*(4), 933–944. https://doi.org/10.13005/ojc/400403

Pormeister, K. (2018). Genetic research and applicable law: The intra-EU conflict of laws as a regulatory challenge to cross-border genetic research. *Journal of Law and the Biosciences*, *5*(3), 706–723. https://doi.org/10.1093/jlb/lsy023

PROTECT-CHILD Consortium. (2025). *Improving pediatric transplant outcomes with genomic data integration*. PROTECT-CHILD EU. Retrieved June 21, 2025, from https://protect-child.eu

Qian, M., Zhan, Y., Ji, L., & Cheng, Y. (2023). Strategy on precision medicine multidisciplinary team in clinical practice. *Clinical and Translational Discovery*, *3*(4). https://doi.org/10.1002/ctd2.217

Qu, Z., Lau, C. W., Nguyen, Q. V., Zhou, Y., & Catchpoole, D. R. (2019). Visual Analytics of Genomic and Cancer Data: A Systematic review. *Cancer Informatics*, *18*. https://doi.org/10.1177/1176935119835546

Qualtrics. (2022, February 17). *Cleaning survey data: Everything you need to know*. https://www.qualtrics.com/en-gb/experience-management/research/survey-data-cleaning/

Raab, R., Küderle, A., Zakreuskaya, A., Stern, A. D., Klucken, J., Kaissis, G., Rueckert, D., Boll, S., Eils, R., Wagener, H., & Eskofier, B. M. (2023). Federated electronic health records for the European Health Data Space. *The Lancet Digital Health*, *5*(11), e840–e847. https://doi.org/10.1016/s2589-7500(23)00156-5

Raza, S., & Hall, A. (2017). Genomic medicine and data sharing. In *British Medical Bulletin* (Vol. 123, Issue 1, pp. 35–45). Oxford University Press. https://doi.org/10.1093/bmb/ldx024

Reiff, S. B., Schroeder, A. J., Kırlı, K., Cosolo, A., Bakker, C., Mercado, L., Lee, S., Veit, A. D., Balashov, A. K., Vitzthum, C., Ronchetti, W., Pitman, K. M., Johnson, J., Ehmsen, S. R., Kerpedjiev, P., Abdennur, N., Imakaev, M., Öztürk, S. U., Çamoğlu, U., ... Park, P. J. (2022). The 4D nucleome data portal as a resource for searching and visualizing curated nucleomics data. *Nature Communications*, *13*(1), 2365. https://doi.org/10.1038/s41467-022-29697-4

Resendez, V., Yıldırım, F., Gaeta, E., Fico, G., & Borsci, S. (2025). *Towards a common set of interface requirements for genomic data management: A systematic literature review* [Manuscript submitted for publication]. University of Twente.

Robertson, A. J., Mallett, A. J., Stark, Z., & Sullivan, C. (2024). It is in our DNA: bringing electronic health records and genomic data together for precision medicine. *JMIR Bioinformatics and Biotechnology*, *5*, e55632. https://doi.org/10.2196/55632

Rodchenkov, I., Babur, O., Luna, A., Aksoy, B. A., Wong, J. V., Fong, D., Franz, M., Siper, M. C., Cheung, M., Wrana, M., Mistry, H., Mosier, L., Dlin, J., Wen, Q., O'Callaghan, C., Li, W., Elder, G., Smith, P. T., Dallago, C., ... Sander, C. (2020). Pathway commons 2019 update: Integration, analysis and exploration of pathway data. *Nucleic Acids Research*, *48*(D1), D489–D497. Scopus. https://doi.org/10.1093/nar/gkz946

Sauria, M. E. G., Phillips-Cremins, J. E., Corces, V. G., & Taylor, J. (2015). HiFive: A tool suite for easy and efficient HiC and 5C data analysis. In *GENOME BIOLOGY* (Vol. 16). https://doi.org/10.1186/s13059-015-0806-y

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., & Robinson, G. E. (2015). Big data: Astronomical or genomical? *PLoS Biology*, *13*(7). https://doi.org/10.1371/journal.pbio.1002195

Stix, C. (2021). The ghost of AI governance past, present and future: AI governance in the European Union. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3882493

Suciu, R. M., Aydin, E., & Chen, B. E. (2015). GeneDig: A web application for accessing genomic and bioinformatics knowledge. *BMC Bioinformatics*, *16*(1), 67. https://doi.org/10.1186/s12859-015-0497-0

Sullivan, G. M., & Artino, A. R. (2013). Analyzing and interpreting data from Likert-Type scales. *Journal of Graduate Medical Education*, *5*(4), 541–542. https://doi.org/10.4300/jgme-5-4-18

Tabesh, P., Tabesh, P., & Moghaddam, K. (2019). Individual and contextual influences on framing effect: Evidence from the Middle East. *Journal of General Management*, *45*(1), 30–39. https://doi.org/10.1177/0306307019851337

Thompson, N., Shapiro, E., & Ponton, H. (2023). The importance of understanding work-as-done: Implications for research and practice in organizational psychology. *EWOP in Practice*, *17*(2), 116–142. https://doi.org/10.21825/ewopinpractice.89676

Tommel, J., Kenis, D., Lambrechts, N., Brohet, R. M., Swysen, J., Mollen, L., Hoefmans, M. F., Pusparum, M., Evers, A. W. M., Ertaylan, G., Roos, M., Hens, K., & Houwink, E. J. F. (2023). Personal genomes in practice: Exploring citizen and healthcare professionals' perspectives on personalized genomic medicine and personal health data spaces using a Mixed-Methods design. *Genes*, *14*(4). https://doi.org/10.3390/genes14040786

Van Kolfschooten, H., & Van Oirschot, J. (2024). The EU Artificial Intelligence Act (2024): Implications for healthcare. *Health Policy*, *149*, 105152. https://doi.org/10.1016/j.healthpol.2024.105152

Vayena, E., Dzenowagis, J., Brownstein, J. S., & Sheikh, A. (2017). Policy implications of big data in the health sector. *Bulletin of the World Health Organization*, *96*, 66–68. https://doi.org/10.2471/blt.17.197426

Venkataanusha, P., Anuradha, C., Murty, P. S. C., & Chebrolu, S. K. (2019). Detecting outliers in high dimensional data sets using Z-Score methodology. *International Journal of Innovative Technology and Exploring Engineering*, *9*(1), 48–53. https://doi.org/10.35940/ijitee.a3910.119119

Warner, J. L., Prasad, I., Bennett, M., Arniella, M., Beeghly-Fadiel, A., Mandl, K. D., & Alterovitz, G. (2018). SMART cancer navigator: A framework for implementing asco workshop recommendations to enable precision cancer medicine. *JCO Precision Oncology*, *2*, 1–14. https://doi.org/10.1200/PO.17.00292

West, R. M. (2021). Best practice in statistics: Use the Welch t-test when testing the difference between two groups. *Annals of Clinical Biochemistry International Journal of Laboratory Medicine*, *58*(4), 267–269. https://doi.org/10.1177/0004563221992088

World Health Organization. Regional Office for Europe. (2021). *The protection of personal data in health information systems: Principles and processes for public health*. World Health Organization. Regional Office for Europe. https://iris.who.int/handle/10665/341374

Xia, J., Benner, M. J., & Hancock, R. E. W. (2014). NetworkAnalyst. Integrative approaches for protein-protein interaction network analysis and visual exploration. *Nucleic Acids Research*, *42*(W1), W167–W174. Scopus. https://doi.org/10.1093/nar/gku443

Xue, B., Khoroshevskyi, O., Gomez, R. A., & Sheffield, N. C. (2023). Opportunities and challenges in sharing and reusing genomic interval data. *Frontiers in Genetics*, *14*. https://doi.org/10.3389/fgene.2023.1155809

Yildirim, N., Richardson, H., Wetscherek, M. T., Bajwa, J., Jacob, J., Pinnock, M. A., Harris, S., Coelho De Castro, D., Bannur, S., Hyland, S., Ghosh, P., Ranjit, M., Bouzid, K., Schwaighofer, A., Pérez-García, F., Sharma, H., Oktay, O., Lungren, M., Alvarez-Valle, J., ... Thieme, A. (2024). Multimodal healthcare AI: Identifying and designing clinically relevant vision-language applications for radiology. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Article 444, 22 pp.). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642013

Yu, C., Shen, S., Wang, S., Zhang, K., & Zhao, H. (2024). Communication-Efficient Hybrid Federated Learning for E-health with Horizontal and Vertical Data Partitioning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2404.10110

During the preparation of this work, the author used several AI tools. Grammarly was used to assess the grammar of the text and enhance its clarity of the text. Chat GPT was used in order to assist in programming and debugging code in R, receiving feedback on the structure or flow of the text, and for minor revisions for conciseness and clarity of writing. After using this tool/service, the author reviewed and edited the content as needed and take full responsibility for the content of the work.

Appendix A

Excerpt of Functional and Non-Functional Requirements Identified by Resendez et al. (2025) Requirements of PROTECT-CHILD

Table A 1

Excerpt of Summary of Functional Requirements Identified by Resendez et al. (2025)

Functional requirements		
1. General Data Management	**2. Data Processing and Analysis**	**3. Data Visualization and Reporting**
1.1. Data Acquisition	**2.1. Data Pre-processing**	**3.1. Visualization of Data**
…	…	…
…	…	…
…	…	…
1.2. Data Standardization	**2.2. Data Analysis**	**3.2. Generate Reports**
…	…	**3.3. Download Data**
…	…	
…	…	**3.4. Knowledge Dissemination**
1.3. Data Sharing	…	
…	…	**3.5. Citation Buttons**
…	…	
…	…	

Table A 2

Excerpt of Summary of Functional and Non-Functional Requirements Identified by Resendez et al. (2025)

Non Functional Requirments			
1. Communication and Support	**2. UX/UI Features**	**3. Security and Compliance**	**4. Platform technical infrastructure**
…	…	…	…
…	…	…	…

Note. The content presented in this appendix is based on prior work by Resendez et al. (2025) and reflects an excerpt of the original list of functional and non-functional requirements. The full list and original material are not included in this document. For complete access to the original requirement framework, please contact Dr. Valeria Resendez at the University of Twente.

Appendix B
Informed Consent Form for the Survey Participation

Dear Participant,

You are invited to take part in a survey designed to gather insights for the development of a platform aimed at improving medical data accessibility for research purposes, streamlining analysis, and enhancing the replicability and reporting of research findings. This platform will support secure data sharing, and collaborative projects, and incorporate privacy-preserving techniques to protect sensitive information, with applications in fields such as medical and genomic research. In particular, we will discuss or refer to the case of children's genetic data exchange and utilization for clinical research, as an example of a case where data privacy and security should be maximized.

Purpose of the Survey: This survey seeks to gather your preferences and experiences regarding data management, analysis, and privacy-preserving methods, such as Federated Learning, Secure Multi-Party Computation, Homomorphic Encryption, and Differential Privacy. Your feedback will help guide the design of a platform that can support various research needs, including data exploration, dataset uploads, information standardization, analysis, visualization, and secure reporting—all within a protected environment.

Survey Content: In this 15-minute survey, we may ask for your insights on some of the following topics: Awareness and familiarity with privacy-preserving techniques and Federated Computing Needs related to data management, analysis, visualization, and reporting

Network topologies, data partitioning strategies, and client selection in Federated Computing Communication protocols and methods for local computation Preferences for combining Federated Computing with other privacy-preserving techniques

Data Usage: The data you provide will be used solely for research and development purposes to develop a platform within the EU project Protect Child. All responses will be anonymized, and no

personal identifying information will be collected or shared. Your data will be stored securely and treated with confidentiality.

Voluntary Participation: Participation in this survey is entirely voluntary. You may choose to withdraw at any time without consequence. Should you decide to withdraw, your data will be securely deleted, and no further information will be collected.

Benefits and Risks: By participating in this survey, you will help shape the design of a platform that could significantly enhance data-sharing processes in genomic research. There are no foreseeable risks associated with participating in this survey.

Contact Information: If you have any questions about this survey or wish to withdraw your participation, please contact us at v.d.c.resendez@utwente.nl. Thank you for your time and input. Your feedback is essential to the success of this project. Sincerely, Protect Child Team.

Note. This consent form was developed by Resendez et al. (2025) as part of the original study. It is included here in full with permission for transparency.

Appendix C

Excerpt of Survey from Resendez et al. (2025) Requirements of PROTECT-CHILD

Start of Block: Clinical experts 1

Clin1_Q91 **General data management** In genomic research, we have identified ***data management as an important first step***, establishing key processes before moving on to data analysis and visualization (Reska et al., 2021). These processes include **data acquisition**, or gathering information from various sources (Amer-Yahia et al., 2022; Reska et al., 2021); **data standardization**, which ensures uniformity and consistency (Canakoglu et al., 2021; McLeod et al., 2021; Pang et al., 2021); and **data sharing**, enabling access for broader use (Gilbert et al., 2022; McLeod et al., 2021; Ullah et al., 2022). As you respond to this survey, please imagine you are conducting research on thrombosis following liver transplantation. In this scenario, you are exploring genetic or biomolecular data for a pediatric cohort, with the goal of understanding the genomic inflation factors that may influence transplantation outcomes. In the next section, we will ask for your feedback on these processes to understand whether they align with your research needs and which specific features would support your work.

Clin0_Filter Do you have experience with genomic data management?

- I know nothing about genomic data management and do not think I can answer the survey [You will exit the survey] (0)
- I know enough to answer the survey (1)
- I am an expert (2)

Skip To: End of Survey If Do you have experience with genomic data management? = I know nothing about genomic data management and do not think I can answer the survey [You will exit the survey]

Note. This appendix includes only the first item from the original survey designed by Resendez et al. (2025). The complete questionnaire and associated materials are not reproduced here. For full access to the original survey, please contact Dr. Valeria Resendez.

Appendix D

Excerpt of Survey Changes from Resendez et al. (2025)

Based on expert review and personal testing of the survey, several modifications were made:

- Clinician Section: Added "NA" options where appropriate.
- Clinician Section - Q20_R: Fixed an HTML visualization issue and revised wording.
- Clinician Section - Q146: New question added regarding expectations/experience with Federated Computing.

Note. The survey structure and changes discussed in this appendix are excerpts derived from the original materials developed by Resendez et al. (2025). Full access to the survey design and change documentation can be requested from Dr. Valeria Resendez.

Appendix E

R-Studio Code for the Data Analysis of the Survey

```
########################## DATA PREPROCCESSING AND CLEANING ##########
data <- read_excel("data.xlsx")
View(data)
survey_data <- data

# Rename variations of the same column to a standard name
names(survey_data)[names(survey_data) %in% c("Legal_.Q90", "Legal_Q90")] <- "Legal_Q90"

# See the data
summary(survey_data)
##########

#Check colnames before applying the functions
colnames(survey_data)
##########

#Check that the column Progress is numeric, if not make it numeric
str(survey_data$Progress)
#Check for the nas
survey_data$Progress <- as.numeric(survey_data$Progress)

##########

#Note we cannot consider the mean or median time to remove participants since we split the survey in two parts for clinicians. Removing them would lead to no clinician as a group.
#So we remove tests with only preview of the data and also filter those that reply less than 80% of the survey:
data <- survey_data %>%
  filter(!grepl("Preview", Status, ignore.case = TRUE)) %>%
  filter(Progress >= 80) %>%
  filter(Demo_Q117 != 'xxx' | is.na(Demo_Q117)) # Use `!=` to filter out rows where Demo_Q117 is 'xxx'


#####################
metadata_columns <- c(
  "StartDate", "EndDate", "Status", "IPAddress", "Progress",
  "Duration", "Finished", "RecordedDate", "ResponseId", "Recipient",
  "ExternalReference", "LocationLatitude", "LocationLongitude",
"DistributionChannel", "UserLanguage",
  "Q_RecaptchaScore", "Q1"
)
text_columns <- c("Q116","Demo_Q114","Demo_NEWFEBQ167", "Demo_Q110", "Demo_Q112", "Demo_Q111", "Demo_Q123",
      "Demo_Q124",
      "Demo_Q117",
      "Demo_Q115", "Demo_Q144",

"Clin1_Q159","Clin1_Q160","Clin1_Q161","Clin1_Q92","Clin1_Q162",
"Clin1_Q163","Clin1_Q164","Clin2_Q146","Clin1_Q3_R",
"Clin2_Q165","Clin1_Q8_R",
"Clin1_Q8_R_3_TEXT","Clin2_Q94",
"Clin2_Q27_R",

"Clin2_Q27_R_5_TEXT", "Clin2_Q146",
                      "Tech1_Q145",
"Tech1_Q150","Tech1_Q151",
"Tech1_Q152","Tech1_Q153",
"Tech1_Q155","Tech1_Q156","Tech1_Q157","Tech1_Q158","Tech1_Q118",
"Tech1_Q28", "Tech1_Q30",
                      "Legal_Q147",
"Legal_Q166", "Legal_Q90")

yes_no_cols <- c("Tech1_Q28", "Tech1_Q30")

#####################
tech_columns <- c("Tech1_Q6", "Tech1_Q9_1", "Tech1_Q9_2", "Tech1_Q9_3", "Tech1_Q9_4",
                  "Tech1_Q150", "Tech1_Q11_1", "Tech1_Q11_2", "Tech1_Q11_3", "Tech1_Q11_4", "Tech1_Q151",
                  "Tech1_Q13_1", "Tech1_Q13_2", "Tech1_Q13_3", "Tech1_Q152", "Tech1_Q17_1", "Tech1_Q17_2",
                  "Tech1_Q17_3", "Tech1_Q17_4", "Tech1_Q153", "Tech1_Q18", "Tech1_Q19", "Tech1_Q121",
                  "Tech1_Q22", "Tech1_Q24_1", "Tech1_Q24_2", "Tech1_Q24_3", "Tech1_Q155", "Tech1_Q26_1",
                  "Tech1_Q26_2", "Tech1_Q26_3", "Tech1_Q26_4", "Tech1_Q26_5", "Tech1_Q26_6", "Tech1_Q26_7",
                  "Tech1_Q156", "Tech1_Q28", "Tech1_Q29_1",
```

```
"Tech1_Q29_2",        "Tech1_Q29_3",
"Tech1_Q157",
                  "Tech1_Q30",
"Tech1_Q31_1",        "Tech1_Q31_2",
"Tech1_Q31_3",         "Tech1_Q158",
"Tech1_Q143_1",
                  "Tech1_Q143_2",
"Tech1_Q143_3", "Tech1_Q143_4")
# Define columns with text (already
identified)
text_columns_tech <- c("Tech1_Q145",
"Tech1_Q150","Tech1_Q151",
"Tech1_Q152","Tech1_Q153",
"Tech1_Q155","Tech1_Q156","Tech1_Q1
57","Tech1_Q158","Tech1_Q118",
"Tech1_Q28", "Tech1_Q30")


# Define columns related to clinical
experts (from the document, we see
these start with "Clin")
clin_columns   <-   c(   "Clin1_Q1R",
"Clin1_Q2_R",        "Clin1_Q3_R_1",
"Clin1_Q3_R_2","Clin1_Q3_R_3","Clin
1_Q3_R_4","Clin1_Q3_R_5","Clin1_Q3_
R_6",

"Clin1_Q3_R_7","Clin1_Q159","Clin1_
Q4_R_1","Clin1_Q4_R_2","Clin1_Q4_R_
3","Clin1_Q4_R_4",
"Clin1_Q160","Clin1_Q5_R_1","Clin1_
Q5_R_2","Clin1_Q5_R_3"
                  ,"Clin1_Q161"   ,
"Clin1_Q6_R"    ,        "Clin1_Q7_R"
                    "Clin1_Q8_R",
"Clin1_Q8_R_3_TEXT" , "Clin1_Q9_R_1"
                ,"Clin1_Q9_R_2"
,             "Clin1_Q9_R_3"        ,
"Clin1_Q9_R_4"    ,   "Clin1_Q9_R_5",
"Clin1_Q9_R_6"  ,     "Clin1_Q9_R_7"
                  ,"Clin1_Q162"    ,
"Clin1_Q10_R"    ,      "Clin1_Q11_R"
                  "Clin1_Q12_R_1",
"Clin1_Q12_R_2" ,    "Clin1_Q12_R_3"
                  ,"Clin1_Q12_R_4"
,           "Clin1_Q12_R_5"         ,
"Clin1_Q163"    ,     "Clin1_Q13_R",
"Clin1_Q14_R_1" ,    "Clin1_Q14_R_2"
                  ,"Clin1_Q14_R_3"
,               "Clin1_Q14_R_4"     ,
"Clin1_Q14_R_5"   ,     "Clin1_Q164",
"Clin2_Q15_R"   ,      "Clin2_Q16_R"
                  ,"Clin2_Q17_R"
,             "Clin2_Q18_R"         ,
"Clin2_Q19_R"    ,    "Clin2_Q146" ,
"Clin2_Q20_R" ,      "Clin2_Q21_R_1"
                  ,"Clin2_Q21_R_2"
,             "Clin2_Q21_R_3"       ,
"Clin2_Q21_R_4"                     ,
"Clin2_Q22_R_1",     "Clin2_Q22_R_2"
,     "Clin2_Q22_R_3"
                  ,"Clin2_Q22_R_4"
,            "Clin2_Q23_R"         ,
"Clin2_Q24_R"    ,    "Clin2_Q25_R",
"Clin2_Q26_R_1",     "Clin2_Q26_R_2"
                  ,"Clin2_Q26_R_3"
,         "Clin2_Q26_R_4"           ,
"Clin2_Q26_R_5"                     ,
"Clin2_Q26_R_6",
"Clin2_Q26_R_7",      "Clin2_Q165"
                  ,"Clin2_Q27_R"
,     "Clin2_Q27_R_5_TEXT")
#  Define the text columns for
clinical experts (if applicable)
text_columns_clin                   <-
c("Clin1_Q159","Clin1_Q160","Clin1_
Q161","Clin1_Q92","Clin1_Q162",
"Clin1_Q163","Clin1_Q164","Clin2_Q1
46",      "Clin2_Q165","Clin1_Q8_R",
"Clin1_Q8_R_3_TEXT","Clin2_Q94",
"Clin2_Q27_R","Clin2_Q27_R_5_TEXT")

legal_columns   <-   c("Legal_Q147",
"Legal_Q64",          "Legal_Q125",
"Legal_Q126",         "Legal_Q127",
"Legal_Q129", "Legal_Q130",
                    "Legal_Q133",
"Legal_Q134",         "Legal_Q135",
"Legal_Q132",         "Legal_Q103",
"Legal_Q137", "Legal_Q66",
                  "Legal_Q78_1",
"Legal_Q78_2",        "Legal_Q78_3",
"Legal_Q78_4",        "Legal_Q166",
"Legal_Q32_1", "Legal_Q32_2",
                  "Legal_Q32_3",
"Legal_Q32_4")

# Define text columns that should NOT
be converted to numeric
text_columns_legal                  <-
c("Legal_Q147",       "Legal_Q166",
"Legal_Q90","Legal_Q147")  # Modify
if needed

# Convert only the non-text columns
to numeric (leaving the text columns
untouched)
tech_columns_to_convert             <-
tech_columns[!tech_columns     %in%
text_columns_tech]
# Identify which clinical columns are
numeric (excluding text columns)
clin_columns_to_convert             <-
clin_columns[!clin_columns     %in%
text_columns_clin]
legal_columns_to_convert            <-
legal_columns[!legal_columns   %in%
text_columns_legal]

####################################
######
########################## UPDATED
#We will change the values for the
two      groups      we      have:
consortium_experts and other_experts
#  1.  Normalize  Function  (clean
strings)
```

```
normalize_response <- function(x) {
  x <- tolower(trimws(x))
  x <- gsub("–", "-", x)  # fix long
dash to hyphen
  x <- gsub("[[:punct:]]+$", "", x)
# remove trailing punctuation
  x <- gsub(" +", " ", x)  # normalize
spaces
  x <- gsub("exremely", "extremely",
x)
  x      <-      gsub("expereince",
"experience", x)
  # Add more custom replacements as
needed
  return(x)
}
# 2. Convert with fallback
convert_column_to_numeric         <-
function(col, mapping) {
  col                             <-
tolower(trimws(as.character(col)))
  col[col %in% c("na", "n/a", "")] <-
NA
  col[col == "yes"] <- "1"
  col[col == "no"] <- "0"

  numeric_mask                    <-
suppressWarnings(!is.na(as.numeric(
col)))
  result <- rep(NA, length(col))

  result[numeric_mask]            <-
as.numeric(col[numeric_mask])

  # Apply mapping only to non-numeric
and known string values
  mapped <- mapping[col]
  mapped_mask  <-  !numeric_mask  &
!is.na(mapped)

  result[mapped_mask]             <-
mapped[mapped_mask]

  # Show warnings for unmapped string
values
  unmapped   <-   !numeric_mask   &
is.na(mapped) & !is.na(col)
  if (any(unmapped)) {
    message("Unmapped         values
(preserved      as      NA):      ",
paste(unique(col[unmapped]),
collapse = ", "))
  }

  return(result)
}


# Original mapping
mapping <- c(
  "1" = 1,
  "2" = 2,
  "3" = 3,
  "4" = 4,
  "5" = 5,
  "6" = 6,
  "7" = 7,
  "extremely important" = 7,
  "very important" = 6,
  "fairly important" = 5,
  "Fairly important" = 5,
  "moderately important" = 5,
  "somewhat important" = 3,
  "slightly important" = 2,
  "slightly importants" = 2,
  "neutral" = 4,
  "not important at all" = 1,
  "not at all important" = 1,
  "na" = NA,
  "very common" = 7,
  "common" = 6,
  "uncommon" = 4,
  "very uncommon" = 2,
  "not at all common" = 1,
  "always" = 7,
  "very often" = 6,
  "often" = 5,
  "sometimes" = 4,
  "never" = 1,
  "occasionally" = 3,
  "rarely" = 2,
  "not critical at all" = 1,
  "slightly critical" = 2,
  "somewhat critical" = 3,
  "neutral" = 4,
  "moderately critical" = 5,
  "moderately common" = 5,
  "very critical" = 6,
  "extremely critical" = 7,
  "not valuable at all" = 1,
  "slightly valuable" = 2,
  "somewhat valuable" = 3,
  "neutral" = 4,
  "moderately valuable" = 5,
  "very valuable" = 6,
  "extremely valuable" = 7,
  "not relevant at all" = 1,
  "slightly relevant" = 2,
  "somewhat relevant" = 3,
  "moderately relevant" = 5,
  "very relevant" = 6,
  "extremely relevant" = 7,
  "5 - extremely important" = 5,
  "7 - very critical" = 7,
  "1- not at all common" = 1,
  "7 - extremely important" = 7,
  "7 - Exremely important" = 7,
  "7 - extremely common" = 7,
  "1 - not at all important" = 1,
  "1- not relevant at all" = 1,
  "1 - never" = 1,
  "7 - always" = 7,
  "1 - not important at all" = 1,
  "1 – not important at all" = 1,
  "7 - extremely relevant" = 7,
  "7- extremely relevant" = 7,
  "7 - extremely critical" = 7,
  "7 – extremely critical" = 7,
```

```
  "7 - extremely valuable" = 7,
  "1- Not relevant at all" = 1,
  "1- not relevant at all" = 1,
  "1 - not at all common" = 1,
  "7 - exremely important" = 7,
  "1 - not critical at all" = 1,
  "7 - Extremely important" = 7,
  "Neutral" = 4,
  "NA" = NA
)


# Apply the conversion function to the relevant columns in both datasets
# Define exclusion patterns
exclude_columns <- c(text_columns, yes_no_cols, metadata_columns)

tech_cols <- setdiff(grep("^Tech", names(data), value = TRUE), c(text_columns, yes_no_cols))
clin_cols <- setdiff(grep("^Clin", names(data), value = TRUE), text_columns)
legal_cols <- setdiff(grep("^Legal", names(data), value = TRUE), text_columns)

data[tech_cols] <- lapply(data[tech_cols], convert_column_to_numeric, mapping = mapping)
data[clin_cols] <- lapply(data[clin_cols], convert_column_to_numeric, mapping = mapping)
data[legal_cols] <- lapply(data[legal_cols], convert_column_to_numeric, mapping = mapping)

########################
########################

data <- data %>%
  mutate(
    all_sevens_tech = rowSums(data[tech_columns] == 7, na.rm = TRUE) == length(tech_columns),
    all_ones_tech = rowSums(data[tech_columns] == 1, na.rm = TRUE) == length(tech_columns),
    all_four_tech = rowSums(data[tech_columns] == 4, na.rm = TRUE) == length(tech_columns),
    all_sevens_clin = rowSums(data[clin_columns] == 7, na.rm = TRUE) == length(clin_columns),
    all_ones_clin = rowSums(data[clin_columns] == 1, na.rm = TRUE) == length(clin_columns),
    all_four_clin = rowSums(data[clin_columns] == 4, na.rm = TRUE) == length(clin_columns),
    all_sevens_leg = rowSums(data[legal_columns] == 7, na.rm = TRUE) == length(legal_columns),
    all_ones_leg = rowSums(data[legal_columns] == 1, na.rm = TRUE) == length(legal_columns),
    all_four_leg = rowSums(data[legal_columns] == 4, na.rm = TRUE) == length(legal_columns)
  )
##########
##########
##########

# Now we need to check again the progress of each group. how many people we have.
#General
data <- data %>%
  mutate(Group = case_when(
    apply(data[, grep("Tech", names(data))], 1, function(x) any(!is.na(x))) ~ "Technical Expert",
    apply(data[, grep("Clin", names(data))], 1, function(x) any(!is.na(x))) ~ "Clinical Expert",
    apply(data[, grep("Legal", names(data))], 1, function(x) any(!is.na(x))) ~ "Legal Expert",
    TRUE ~ "Other"
  ))

# Summary of progress by group
progress_summary <- data %>%
  group_by(Group) %>%
  summarise(
    count = n(),
    mean_progress = mean(Progress, na.rm = TRUE),
    sd_progress = sd(Progress, na.rm = TRUE),
    min_progress = min(Progress, na.rm = TRUE),
    max_progress = max(Progress, na.rm = TRUE)
  )

# View the progress summary
print(progress_summary)

########################## Outliers
##########################
# Create a copy of the original dataset
```

```
data_with_outliers <- data

# Add a new column to number
participants
data <- data %>%
  mutate(Participant_ID =
paste0("Participant_",
row_number()))  # Unique numbering

########

# Function to remove participants
with outliers based on Z-scores
remove_participants_with_outliers <-
function(df, question_columns,
group_filter) {

  stats <- df %>%
    filter(Group == group_filter)
%>%

summarise(across(all_of(question_co
lumns),
                    list(mean = ~
mean(.x, na.rm = TRUE),
                         sd = ~
sd(.x, na.rm = TRUE)),
                    .names =
"{col}_{fn}"))

  # Compute Z-score table (only for
question columns)
  df_with_z <- df %>%
    filter(Group == group_filter)
%>%

mutate(across(all_of(question_colum
ns),
                  ~ (. -
stats[[paste0(cur_column(),
"_mean")]]) /
stats[[paste0(cur_column(),
"_sd")]],
                  .names =
"z_{col}"))  # Store Z-score for each
question

  # Identify rows with any outlier
(Z-score > 3 or < -3) #Based on Cohen
  outlier_participants <- df_with_z
%>%

filter(if_any(starts_with("z_"), ~
abs(.) > 3)) %>%
    pull(Participant_ID)  # Extract
Participant IDs with outliers

  # Remove outlier participants from
original dataset
  df_cleaned <- df %>%
    filter(!Participant_ID %in%
outlier_participants)  # Remove rows
with outliers
  return(df_cleaned)
}

# Apply the function for each expert
group to remove outlier participants
data_cleaned <- data %>%

remove_participants_with_outliers(t
ech_columns_to_convert, "Technical
Expert") %>%

remove_participants_with_outliers(c
lin_columns_to_convert, "Clinical
Expert") %>%

remove_participants_with_outliers(l
egal_columns_to_convert, "Legal
Expert")

# Print summary after removal
print("Summary of cleaned dataset
(without outliers):")
summary(data_cleaned)
print(paste("Number of participants
removed:", nrow(data) -
nrow(data_cleaned)))

# Save cleaned dataset to Excel
library(writexl)
write_xlsx(data_cleaned,
"cleaned_survey_data_30May.xlsx")

################
##########################

# Load necessary libraries
library(dplyr)
library(readxl)
library(ggplot2)
library(readr)
library(readxl)
library(writexl)


cleaned_survey_data_30May <-
read_excel("cleaned_survey_data_30M
ay.xlsx")
View(cleaned_survey_data_30May)
data_noout<-
cleaned_survey_data_30May

# Define columns related to clinical
experts (from the document, we see
these start with "Clin")
clin_columns <- c( "Clin1_Q1R",
"Clin1_Q2_R", "Clin1_Q3_R_1",
"Clin1_Q3_R_2","Clin1_Q3_R_3","Clin
1_Q3_R_4","Clin1_Q3_R_5","Clin1_Q3_
R_6",

"Clin1_Q3_R_7","Clin1_Q159","Clin1_
Q4_R_1","Clin1_Q4_R_2","Clin1_Q4_R_
3","Clin1_Q4_R_4",
```

```
"Clin1_Q160","Clin1_Q5_R_1","Clin1_
Q5_R_2","Clin1_Q5_R_3"
                    ,"Clin1_Q161"   ,
"Clin1_Q6_R"    ,       "Clin1_Q7_R"
                  ,"Clin1_Q8_R",
"Clin1_Q8_R_3_TEXT" , "Clin1_Q9_R_1"
                  ,"Clin1_Q9_R_2"
              "Clin1_Q9_R_3"       ,
"Clin1_Q9_R_4"    ,   "Clin1_Q9_R_5",
"Clin1_Q9_R_6"  ,     "Clin1_Q9_R_7"
                  ,"Clin1_Q162"   ,
"Clin1_Q10_R"   ,      "Clin1_Q11_R"
                   "Clin1_Q12_R_1",
"Clin1_Q12_R_2" ,    "Clin1_Q12_R_3"
                  ,"Clin1_Q12_R_4"
            "Clin1_Q12_R_5"       ,
"Clin1_Q163"    ,     "Clin1_Q13_R",
"Clin1_Q14_R_1" ,    "Clin1_Q14_R_2"
                  ,"Clin1_Q14_R_3"
                "Clin1_Q14_R_4"   ,
"Clin1_Q14_R_5"   ,     "Clin1_Q164",
"Clin2_Q15_R"   ,      "Clin2_Q16_R"
                  ,"Clin2_Q17_R"
               "Clin2_Q18_R"      ,
"Clin2_Q19_R"    ,    "Clin2_Q146" ,
"Clin2_Q20_R" ,      "Clin2_Q21_R_1"
                  ,"Clin2_Q21_R_2"
              "Clin2_Q21_R_3"     ,
"Clin2_Q21_R_4"                   ,
"Clin2_Q22_R_1",     "Clin2_Q22_R_2"
,     "Clin2_Q22_R_3"
                  ,"Clin2_Q22_R_4"
              "Clin2_Q23_R"       ,
"Clin2_Q24_R"    ,    "Clin2_Q25_R",
"Clin2_Q26_R_1",     "Clin2_Q26_R_2"
                  ,"Clin2_Q26_R_3"
         "Clin2_Q26_R_4"          ,
"Clin2_Q26_R_5"                   ,
"Clin2_Q26_R_6",
"Clin2_Q26_R_7",       "Clin2_Q165"
                  ,"Clin2_Q27_R"
,     "Clin2_Q27_R_5_TEXT")
#  Define  the  text  columns  for
clinical experts (if applicable)
text_columns_clin                 <-
c("Clin1_Q159","Clin1_Q160","Clin1_
Q161","Clin1_Q92","Clin1_Q162",
"Clin1_Q163","Clin1_Q164","Clin2_Q1
46",      "Clin2_Q165","Clin1_Q8_R",
"Clin1_Q8_R_3_TEXT","Clin2_Q94",
"Clin2_Q27_R","Clin2_Q27_R_5_TEXT")


# Identify which clinical columns are
numeric (excluding text columns)
clin_columns_to_convert           <-
clin_columns[!clin_columns      %in%
text_columns_clin]

##########################  T  test
for      the      ClinicianSection-
Q20_R:FixedanHTMLvisualizationissue
andrevised wording.

# Convert EndDate column to Date type
if it is not already in Date format
data_noout$EndDate                <-
as.Date(data_noout$EndDate,
format="%Y-%m-%d")

# Split the data into two groups:
before and after January 31, 2025
group_before <- data_noout %>%
  filter(EndDate < as.Date("2025-01-
31")) %>%   # Group before Jan 31,
2025
  select(Clin2_Q20_R)  # Adjust this
to select the desired column

group_after <- data_noout %>%
  filter(EndDate  >=  as.Date("2025-
01-31")) %>%  # Group after Jan 31,
2025
  select(Clin2_Q20_R)

# Perform the t-test to compare the
responses from the two groups
t_test_result                     <-
t.test(group_before$Clin2_Q20_R,
group_after$Clin2_Q20_R)

# View the results of the t-test
print(t_test_result)

##########################
###########
#Clinical experts

#   Identify   columns   related   to
Clinicians
clin_columns     <-     grep("^Clin",
names(data), value = TRUE)
# Define columns related to clinical
experts (from the document, we see
these start with "Clin")
clin_columns   <-   c(   "Clin1_Q1R",
"Clin1_Q2_R",         "Clin1_Q3_R_1",
"Clin1_Q3_R_2","Clin1_Q3_R_3","Clin
1_Q3_R_4","Clin1_Q3_R_5","Clin1_Q3_
R_6",

"Clin1_Q3_R_7","Clin1_Q159","Clin1_
Q4_R_1","Clin1_Q4_R_2","Clin1_Q4_R_
3","Clin1_Q4_R_4",
"Clin1_Q160","Clin1_Q5_R_1","Clin1_
Q5_R_2","Clin1_Q5_R_3"
                    ,"Clin1_Q161"   ,
"Clin1_Q6_R"    ,       "Clin1_Q7_R"
                  ,"Clin1_Q8_R",
"Clin1_Q8_R_3_TEXT" , "Clin1_Q9_R_1"
                  ,"Clin1_Q9_R_2"
              "Clin1_Q9_R_3"       ,
"Clin1_Q9_R_4"    ,   "Clin1_Q9_R_5",
"Clin1_Q9_R_6"  ,     "Clin1_Q9_R_7"
                  ,"Clin1_Q162"   ,
"Clin1_Q10_R"   ,      "Clin1_Q11_R"
                   "Clin1_Q12_R_1",
"Clin1_Q12_R_2" ,    "Clin1_Q12_R_3"
```

```
                    ,"Clin1_Q12_R_4"
             "Clin1_Q12_R_5"          ,
"Clin1_Q163"    ,     "Clin1_Q13_R",
"Clin1_Q14_R_1" ,    "Clin1_Q14_R_2"
                    ,"Clin1_Q14_R_3"
                  "Clin1_Q14_R_4"     ,
"Clin1_Q14_R_5"   ,     "Clin1_Q164",
"Clin2_Q15_R"   ,      "Clin2_Q16_R"
                    ,"Clin2_Q17_R"
                "Clin2_Q18_R"         ,
"Clin2_Q19_R"     ,     "Clin2_Q146" ,
"Clin2_Q20_R" ,       "Clin2_Q21_R_1"
                    ,"Clin2_Q21_R_2"
               "Clin2_Q21_R_3"        ,
"Clin2_Q21_R_4"                       ,
"Clin2_Q22_R_1",     "Clin2_Q22_R_2"
,     "Clin2_Q22_R_3"
                    ,"Clin2_Q22_R_4"
                "Clin2_Q23_R"         ,
"Clin2_Q24_R"     ,     "Clin2_Q25_R",
"Clin2_Q26_R_1",      "Clin2_Q26_R_2"
                    ,"Clin2_Q26_R_3"
,         "Clin2_Q26_R_4"             ,
"Clin2_Q26_R_5"                       ,
"Clin2_Q26_R_6",
"Clin2_Q26_R_7",       "Clin2_Q165"
                    ,"Clin2_Q27_R"
,     "Clin2_Q27_R_5_TEXT")

# Define the text columns for
clinical experts (if applicable)
text_columns_clin                     <-
c("Clin1_Q159","Clin1_Q160","Clin1_
Q161","Clin1_Q92","Clin1_Q162",
"Clin1_Q163","Clin1_Q164","Clin1_Q3
_R",       "Clin2_Q165","Clin1_Q8_R",
"Clin2_Q94","Clin2_Q27_R_5_TEXT")


# Map the column names to their
respective questions for clinical
experts
question_mapping_clin <- c(
  "Clin1_Q1R" = "How important is it
for you to acquire genomic data from
multiple sources?",
  "Clin1_Q2_R" = "How important is it
for you to upload your own collected
genomic data into a system?",
  "Clin1_Q3_R_1" = "How relevant are
the   following   data   models   for
standardizing the data you typically
work with? (PATRIC)",
  "Clin1_Q3_R_2" = "How relevant are
the   following   data   models   for
standardizing the data you typically
work with? (Genomic Data Model)",
  "Clin1_Q3_R_3" = "How relevant are
the   following   data   models   for
standardizing the data you typically
work with? (DataSHaPER)",
  "Clin1_Q3_R_4" = "How relevant are
the   following   data   models   for
standardizing the data you typically
work with? (OMOP)",
  "Clin1_Q3_R_5" = "How relevant are
the   following   data   models   for
standardizing the data you typically
work with? (FHIR)",
  "Clin1_Q3_R_6" = "How relevant are
the   following   data   models   for
standardizing the data you typically
work with? (VCF)",
  "Clin1_Q3_R_7" = "How relevant are
the   following   data   models   for
standardizing the data you typically
work with? (PHENOPACKET)",
  "Clin1_Q4_R_1"   =   "What   file
formats   are   important   for   your
workflows? (VCF)",
  "Clin1_Q4_R_2"   =   "What   file
formats   are   important   for   your
workflows? (FAST-Q)",
  "Clin1_Q4_R_3"   =   "What   file
formats   are   important   for   your
workflows? (BAM)",
  "Clin1_Q4_R_4"   =   "What   file
formats   are   important   for   your
workflows? ((IDAT)",
  "Clin1_Q5_R_1" = "How important
are the following factors in your
decision     to     share     data?
(Efficiency)",
  "Clin1_Q5_R_2" = "How important
are the following factors in your
decision to share data? (Security)",
  "Clin1_Q5_R_3" = "How important
are the following factors in your
decision to share data? (Scope of the
project)",             #Clin1_Q5_R_3
(Interquartile range higher than 2)
-   mostly   for   members   in   the
consortium group
  "Clin1_Q6_R" = "How important is it
for you to inspect the quality of
data before analysis?",
  "Clin1_Q7_R" = "How important is it
for you to have automated checks for
data completeness?",
  "Clin1_Q8_R_1"   =   "What   is   the
primary   focus   of   your   research
involving the collaborative use of
medical   data?   (Developing   AI
predictive models)",
  "Clin1_Q8_R_2"   =   "What   is   the
primary   focus   of   your   research
involving the collaborative use of
medical      data?      (Extracting
statistical insights)",
  "Clin1_Q9_R_1" = "How important is
it for you to employ the following
types of analyses in your research?
(Epidemiological         analysis)",
#Interquartile range higher than 2
(mostly creates discrepancy in the
whole dataset)
```

```
  "Clin1_Q9_R_2" = "How important is it for you to employ the following types of analyses in your research? (Predictive modelling)",
  "Clin1_Q9_R_3" = "How important is it for you to employ the following types of analyses in your research? (Statistical analysis)",
  "Clin1_Q9_R_4" = "How important is it for you to employ the following types of analyses in your research? (Data visualization)",
  "Clin1_Q9_R_5" = "How important is it for you to employ the following types of analyses in your research? (Exploratory data analysis)",
  "Clin1_Q9_R_6" = "How important is it for you to employ the following types of analyses in your research? (AI modelling)",
  "Clin1_Q9_R_7" = "How important is it for you to employ the following types of analyses in your research? (Preventive models)",
  "Clin1_Q10_R" = "How important is reproducibility in your work?",
  "Clin1_Q11_R" = "How important is it for you to have command-line tools for analysis?",
  "Clin1_Q12_R_1" = "How important are the following visualizations? (Graphs)",
  "Clin1_Q12_R_2" = "How important are the following visualizations? (Charts)",
  "Clin1_Q12_R_3" = "How important are the following visualizations? (Heat map)",
  "Clin1_Q12_R_4" = "How important are the following visualizations? (Sequencing)", #Interquartile range higher than 2 (mostly for other experts outside the consortium group)
  "Clin1_Q12_R_5" = "How important are the following visualizations? (Networks)", #Interquartile range was 3 (mostly for those in the consortium group and the whole dataset)
  "Clin1_Q13_R" = "How important for you to download or export the data?",
  "Clin1_Q14_R_1" = "How important is it for you to share the knowledge produced through the platform with (Clinicians)",
  "Clin1_Q14_R_2" = "How important is it for you to share the knowledge produced through the platform with (Researchers)",
  "Clin1_Q14_R_3" = "How important is it for you to share the knowledge produced through the platform with (Patients)", #Interquartile range higher than 2 (mostly for other experts group and thus the whole dataset)
  "Clin1_Q14_R_4" = "How important is it for you to share the knowledge produced through the platform with (Policy makers)", #Interquartile range higher than 2 for those in the consortium (mostly for experts in the consortium group and thus the whole dataset)
  "Clin1_Q14_R_5" = "How important is it for you to share the knowledge produced through the platform with (Ethical board)",
  "Clin2_Q15_R" = "How important is it for you to protect the data privacy in your research?",
  "Clin2_Q16_R" = "How important is it to stay informed about the latest standards in security for genomic data management?", #Interquartile range higher than 2 (only on the whole dataset)
  "Clin2_Q17_R" = "How important is it for you that the platform is mobile-friendly?", #Interquartile range higher than 2 (mostly for those in the other expert group and thus the whole dataset)
  "Clin2_Q18_R" = "How important is it for you to have a platform that supports multiple languages?",
  "Clin2_Q19_R" = "How important do you consider notifications in the platform?",
  "Clin2_Q20_R" = "How important is it to you to have access to Federated Computing infrastructure for your work?",
  "Clin2_Q21_R_1" = "What criteria are most important when designing Federated Computing systems? (Computational efficiency)",
  "Clin2_Q21_R_2" = "What criteria are most important when designing Federated Computing systems? (Communcation efficiency)",
  "Clin2_Q21_R_3" = "What criteria are most important when designing Federated Computing systems? (Model accuracy)",
  "Clin2_Q21_R_4" = "What criteria are most important when designing Federated Computing systems? (Privacy guarantee)",
  "Clin2_Q22_R_1" = "How important are methods for organizing data in medical research? (Horizontal partitioning)", #Interquartile range higher than 2 (mostly for those in the consortium group but does not appear in the whole dataset)
```

```
  "Clin2_Q22_R_2" = "How important are methods for organizing data in medical research? (Vertical partitioning)", #Interquartile range higher than 2 (mostly for other experts outside the consoritum thus it also appears in the whole dataset)
  "Clin2_Q22_R_3" = "How important are methods for organizing data in medical research? (Transfer learning)",
  "Clin2_Q22_R_4" = "How important are methods for organizing data in medical research? (Hybrid partitioning)", #Interquartile range higher than 2 (appears only in the whole dataset)
  "Clin2_Q23_R" = "How important is it to select participants through a pre-determined selection?",
  "Clin2_Q24_R" = "How important is it to select participants through open participation?", #Interquartile range higher than 2 (mostly for those outside the consortium)
  "Clin2_Q25_R" = "How important is it to select participants through a mixed approach?", #Interquartile range higher than 2 (appears only in the whole dataset)
  "Clin2_Q26_R_1" = "Which Federated Computing frameworks are important for your work? (TensorFlow)", #higher than 2 (only for other experts outside the consortium it does not appear in the whole dataset)
  "Clin2_Q26_R_2" = "Which Federated Computing frameworks are important for your work? (PySyft)",
  "Clin2_Q26_R_3" = "Which Federated Computing frameworks are important for your work? (Flower)",
  "Clin2_Q26_R_4" = "Which Federated Computing frameworks are important for your work? (Vantage6)",
  "Clin2_Q26_R_5" = "Which Federated Computing frameworks are important for your work? (Custom-built frameworks)",
  "Clin2_Q26_R_6" = "Which Federated Computing frameworks are important for your work? (NVIDIA)",
  "Clin2_Q26_R_7" = "Which Federated Computing frameworks are important for your work? (PyTorch)" #Interquartile range higher than 2 for those in the consortium group (appears only on the consortium)
  )

# Define the mapping of column names to question categories for clinical experts
category_mapping_clin <- c(
  "Clin1_Q1R" = "Genomic data acquisition",
  "Clin1_Q2_R" = "Genomic data upload",
  "Clin1_Q3_R_1" = "Data standardization",
  "Clin1_Q3_R_2" = "Data standardization",
  "Clin1_Q3_R_3" = "Data standardization",
  "Clin1_Q3_R_4" = "Data standardization",
  "Clin1_Q3_R_5" = "Data standardization",
  "Clin1_Q3_R_6" = "Data standardization",
  "Clin1_Q3_R_7" = "Data standardization",
  "Clin1_Q4_R_1" = "File formats",
  "Clin1_Q4_R_2" = "File formats",
  "Clin1_Q4_R_3" = "File formats",
  "Clin1_Q4_R_4" = "File formats",
  "Clin1_Q5_R_1" = "Data sharing factors",
  "Clin1_Q5_R_2" = "Data sharing factors",
  "Clin1_Q5_R_3" = "Data sharing factors",
  "Clin1_Q6_R" = "Data quality control",
  "Clin1_Q7_R" = "Automated data completeness checks",
  "Clin1_Q8_R_1" = "Research focus",
  "Clin1_Q8_R_2" = "Research focus",
  "Clin1_Q9_R_1" = "Types of analyses in research",
  "Clin1_Q9_R_2" = "Types of analyses in research",
  "Clin1_Q9_R_3" = "Types of analyses in research",
  "Clin1_Q9_R_4" = "Types of analyses in research",
  "Clin1_Q9_R_5" = "Types of analyses in research",
  "Clin1_Q9_R_6" = "Types of analyses in research",
  "Clin1_Q9_R_7" = "Types of analyses in research",
  "Clin1_Q10_R" = "Reproducibility",
  "Clin1_Q11_R" = "Use of command-line tools",
  "Clin1_Q12_R_1" = "Preferred visualization methods",
  "Clin1_Q12_R_2" = "Preferred visualization methods",
  "Clin1_Q12_R_3" = "Preferred visualization methods",
  "Clin1_Q12_R_4" = "Preferred visualization methods",
  "Clin1_Q12_R_5" = "Preferred visualization methods",
  "Clin1_Q13_R" = "Data export & download",
```

```
  "Clin1_Q14_R_1"    =    "Knowledge
sharing",
  "Clin1_Q14_R_2"    =    "Knowledge
sharing",
  "Clin1_Q14_R_3"    =    "Knowledge
sharing",
  "Clin1_Q14_R_4"    =    "Knowledge
sharing",
  "Clin1_Q14_R_5"    =    "Knowledge
sharing",
  "Clin2_Q15_R"   =   "Data   privacy
protection",
  "Clin2_Q16_R"     =     "Security
standards awareness",
  "Clin2_Q17_R"     =     "Platform
usability (mobile-friendly)",
  "Clin2_Q18_R"   =   "Multi-language
support",
  "Clin2_Q19_R"     =     "Platform
notifications",
  "Clin2_Q20_R"    =    "Access    to
federated computing",
  "Clin2_Q21_R_1"    =    "Federated
computing criteria",
  "Clin2_Q21_R_2"    =    "Federated
computing criteria",
  "Clin2_Q21_R_3"    =    "Federated
computing criteria",
  "Clin2_Q21_R_4"    =    "Federated
computing criteria",
  "Clin2_Q22_R_1"       =       "Data
organization in research",
  "Clin2_Q22_R_2"       =       "Data
organization in research",
  "Clin2_Q22_R_3"       =       "Data
organization in research",
  "Clin2_Q22_R_4"       =       "Data
organization in research",
  "Clin2_Q23_R"    =    "Participant
selection methods",
  "Clin2_Q24_R"    =    "Participant
selection methods",
  "Clin2_Q25_R"    =    "Participant
selection methods",
  "Clin2_Q26_R_1"    =    "Federated
computing frameworks",
  "Clin2_Q26_R_2"    =    "Federated
computing frameworks",
  "Clin2_Q26_R_3"    =    "Federated
computing frameworks",
  "Clin2_Q26_R_4"    =    "Federated
computing frameworks",
  "Clin2_Q26_R_5"    =    "Federated
computing frameworks",
  "Clin2_Q26_R_6"    =    "Federated
computing frameworks",
  "Clin2_Q26_R_7"    =    "Federated
computing frameworks"
)


# Compute the mean and standard
deviation of responses for each
numeric Clin-related question for
Clinical Experts
agreement_analysis_clin          <-
data_noout %>%
  filter(Group == "Clinical Expert")
%>%

summarise(across(all_of(clin_column
s_to_convert),
                   list(mean   =   ~
mean(.x, na.rm = TRUE),
                        sd    =    ~
sd(.x, na.rm = TRUE),

mean_percent = ~ mean((.x - 1) / 6,
na.rm = TRUE) * 100,  # Transform to
percentage (1 -> 0%, 7 -> 100%)
                        sd_percent =
~ sd((.x - 1) / 6, na.rm = TRUE) *
100  # Transform to percentage (1 ->
0%, 7 -> 100%)
                   ),
                   .names          =
"{col}_{fn}"))  # This will include
the column name + function name in
the output

#   Reshape   the   data   correctly,
ensuring only two parts in the name
agreement_analysis_clin          <-
tidyr::pivot_longer(
  agreement_analysis_clin,
  cols = everything(),
  names_to      =      c("question",
".value"),
  names_pattern                    =
"^(.*)_(mean|sd|mean_percent|sd_per
cent)$"   # Add pattern to include
mean_percent and sd_percent
)


# Add question text & category to the
data
agreement_analysis_clin          <-
agreement_analysis_clin %>%
  mutate(category = recode(question,
!!!category_mapping_clin)) %>%
  mutate(question_text             =
recode(question,
!!!question_mapping_clin))


########################## AGREEMENT
ANALYSIS AMONG CLINICAL EXPERTS

#Include IQR range for this:
# Calculate the IQR for each numeric
Clin-related question for Clinical
Experts
iqr_analysis_clin <- data_noout %>%
  filter(Group == "Clinical Expert")
%>%
```

```
summarise(across(all_of(clin_columns_to_convert),
                 list(iqr = ~ IQR(.x, na.rm = TRUE)),  # Calculate IQR
                 .names = "{col}_iqr"))   # Name the columns accordingly

iqr_analysis_clin <- tidyr::pivot_longer(
  iqr_analysis_clin,
  cols = everything(),
  names_to = c("question", ".value"),
  names_pattern = "^(.*)_(iqr)$"  # Capture question name and value type
)

# View the IQR results for each question
print(iqr_analysis_clin)

############## Export dataset RAW

# Display the agreement analysis with question names and their results
print("Agreement analysis for Clinical experts (with question names and results):")
print(agreement_analysis_clin)

############## Export dataset for clinicians TABLE
# Step 1: Join mean/sd table with IQR
agreement_clin_full <- agreement_analysis_clin %>%
  left_join(iqr_analysis_clin, by = "question")

# Step 2: Add Agreement* and Importance*
agreement_clin_full <- agreement_clin_full %>%
  mutate(
    Agreement = ifelse(iqr <= 2, "Yes", "No"),
    Importance = ifelse(mean > 4, "Yes", "No")
  )

# Step 3: Compute % of respondents
clin_total <- data_noout %>% filter(Group == "Clinical Expert")

percent_response_clin <- sapply(clin_columns_to_convert, function(q) {
  valid <- sum(!is.na(clin_total[[q]]))
  total <- nrow(clin_total)
  round(100 * valid / total, 1)
})

agreement_clin_full$percent_respondents <- percent_response_clin[agreement_clin_full$question]
agreement_clin_full$percent_respondents <- percent_response_clin[agreement_clin_full$question] / 100  # as fraction
agreement_clin_full <- agreement_clin_full %>%
  mutate(
    `Mean %` = round(mean_percent / 100, 4),
    `SD %` = round(sd_percent / 100, 4)
  )

# Step 4: Rename and reorder columns for final output
final_table_clin <- agreement_clin_full %>%
  select(
    `Question Code` = question,
    `Survey Item` = question_text,
    `Category` = category,
    `Mean` = mean,
    `SD` = sd,
    `Mean %`,
    `SD %`,
    `IQR` = iqr,
    `% of respondents` = percent_respondents,
    `Agreement*` = Agreement,
    `Importance*` = Importance
  )

# Step 5: Save or print
print(final_table_clin)
write_xlsx(final_table_clin, "summary_agreement_clinical_table_14May.xlsx")

##########################

########################## PLOTS


library(ggplot2)
library(viridis)
library(stringr)
library(dplyr)

# Ensure folder exists
dir.create("plots_Clin_ZAKONCHILA", showWarnings = FALSE)

# Wrap long labels
agreement_analysis_clin <- agreement_analysis_clin %>%
  mutate(wrapped_question = str_wrap(question_text, width = 60))

# Split by category
```

```
plots_data_clin                    <-
split(agreement_analysis_clin,
agreement_analysis_clin$category)

# Loop to generate enhanced plots
using human-readable question names
for (cat in names(plots_data_clin))
{

  q                                <-
ggplot(plots_data_clin[[cat]], aes(x
= reorder(wrapped_question, mean), y
= mean, fill = wrapped_question)) +
    geom_col(width = 0.6, color =
"black") +
    geom_errorbar(aes(
      ymin = pmax(mean - sd, 1),
      ymax = pmin(mean + sd, 7)
    ),
    width = 0.2, color = "black",
linewidth = 0.6) +
    coord_flip(ylim = c(1, 7))+
    labs(
      title = cat,
      x = NULL,
      y = "Mean Importance Score"
    ) +
    scale_fill_viridis_d(option    =
"D") +
    theme_classic(base_size = 14)+
    theme(
      panel.background             =
element_rect(fill = "white", color =
NA),
      plot.background              =
element_rect(fill = "white", color =
NA),
      axis.title.x                 =
element_text(size = 20),
      axis.text = element_text(size
= 20),
      axis.text.y                  =
element_text(face = "bold", size =
20),
      plot.title = element_text(size
= 20, face = "bold", hjust = 0.5),
      legend.position = "none"
    )

  # Save to file
  n_items                          <-
nrow(plots_data_clin[[cat]])
  ggsave(
    filename                       =
paste0("plots_Clin_ZAKONCHILA/Agree
ment_Analysis_",  gsub("  ",  "_",
cat), ".png"),
    plot = q,
    width = 12,
    height = max(5, n_items * 0.6),
# Height grows with items
    dpi = 300,
    bg = "white"
  )

  # Optional preview
  print(q)
}

##########################
```

Appendix F

The Variability of Clinical Expert Responses Across All Items

Figure F 1

Importance Scores for "Genomic Data Acquisition"

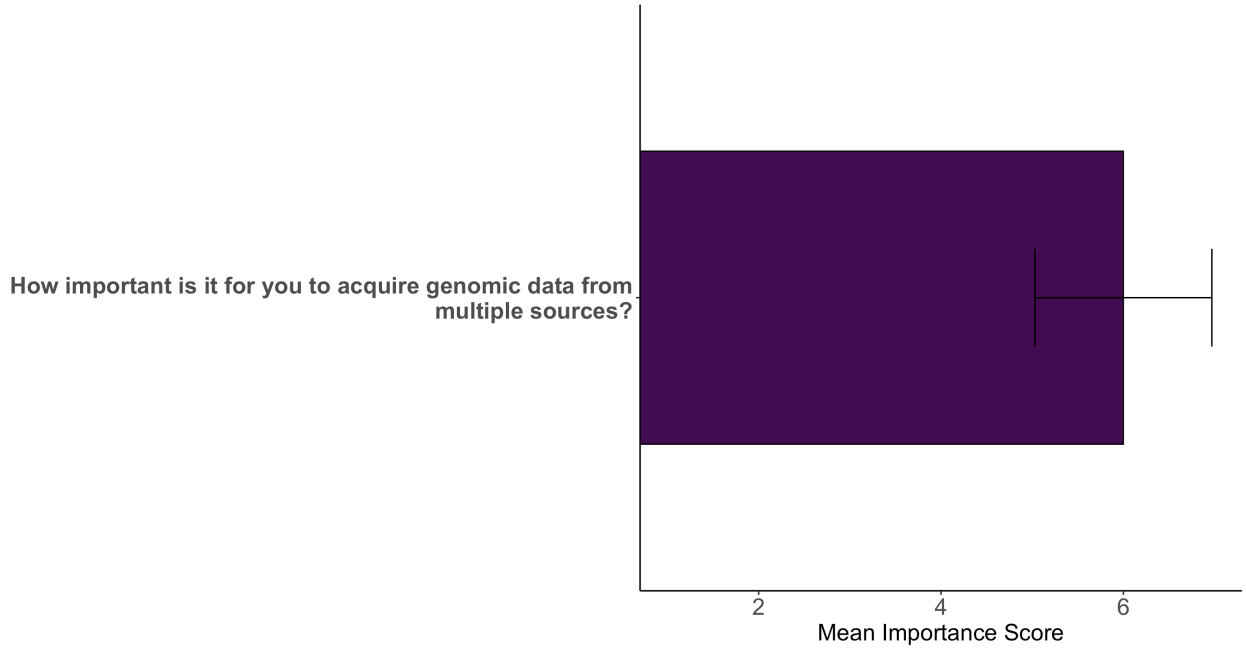


Figure F 2

Importance Scores for "Genomic Data Upload"

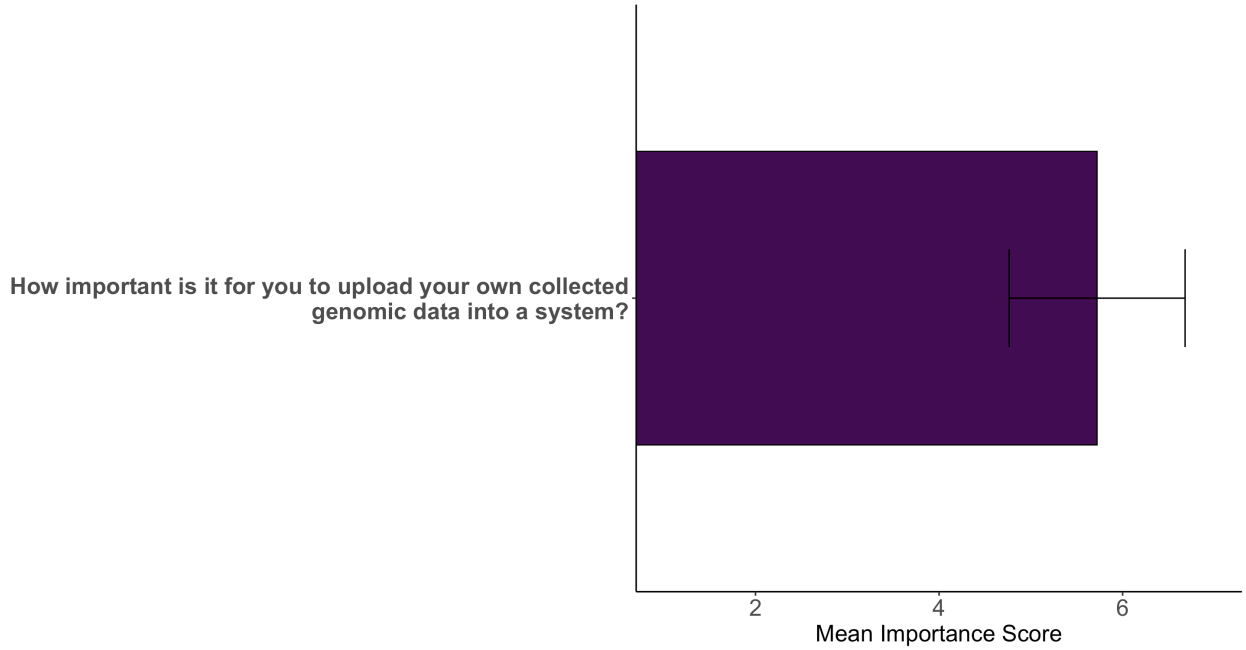


Figure F 3

Importance Scores for "Data Standardization"

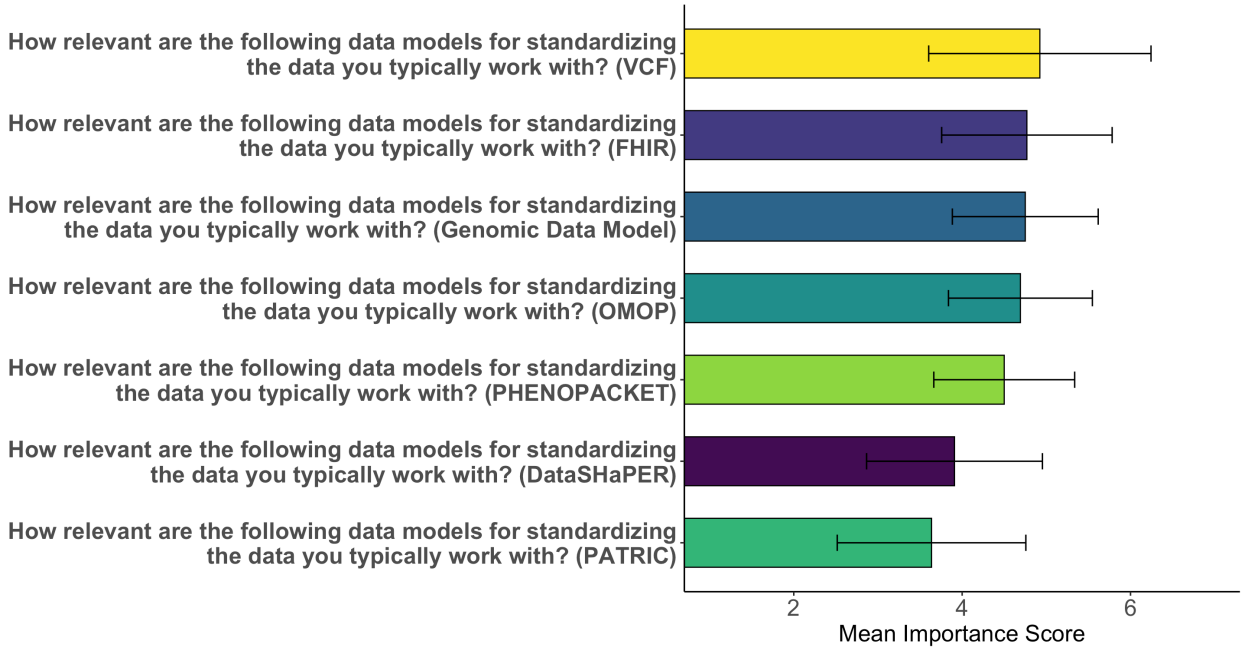


Figure F 4

Importance Scores for "File Formats"

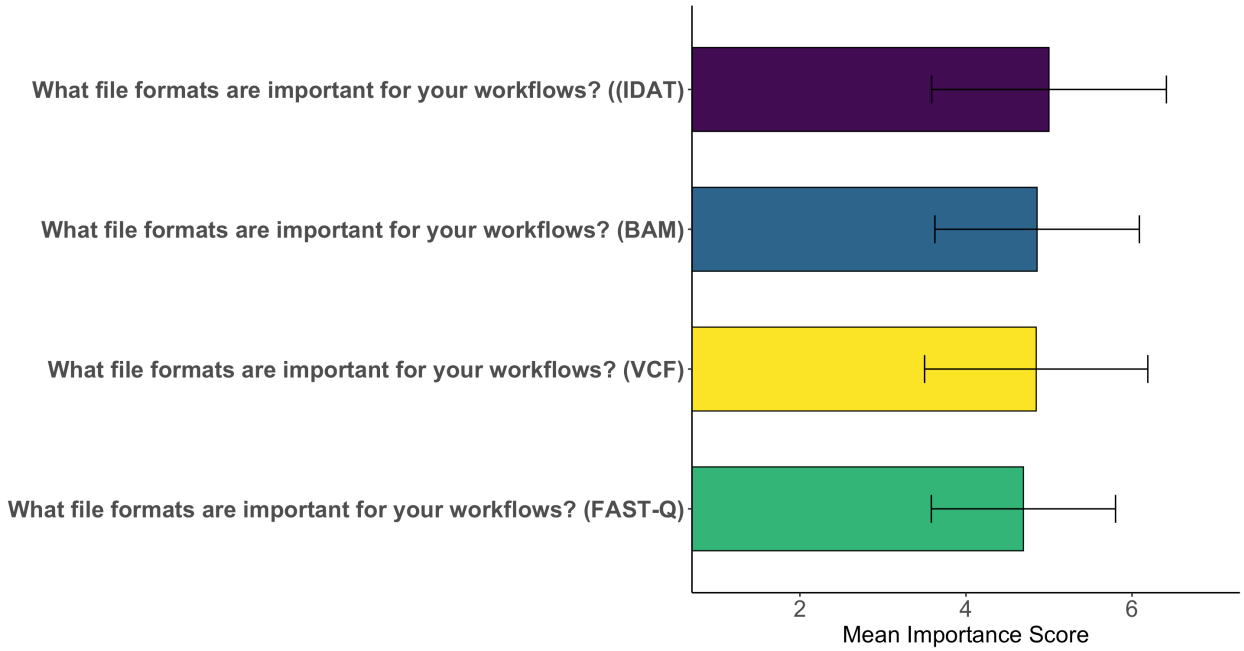


Figure F 5

Importance Scores for "Data Sharing Factors"

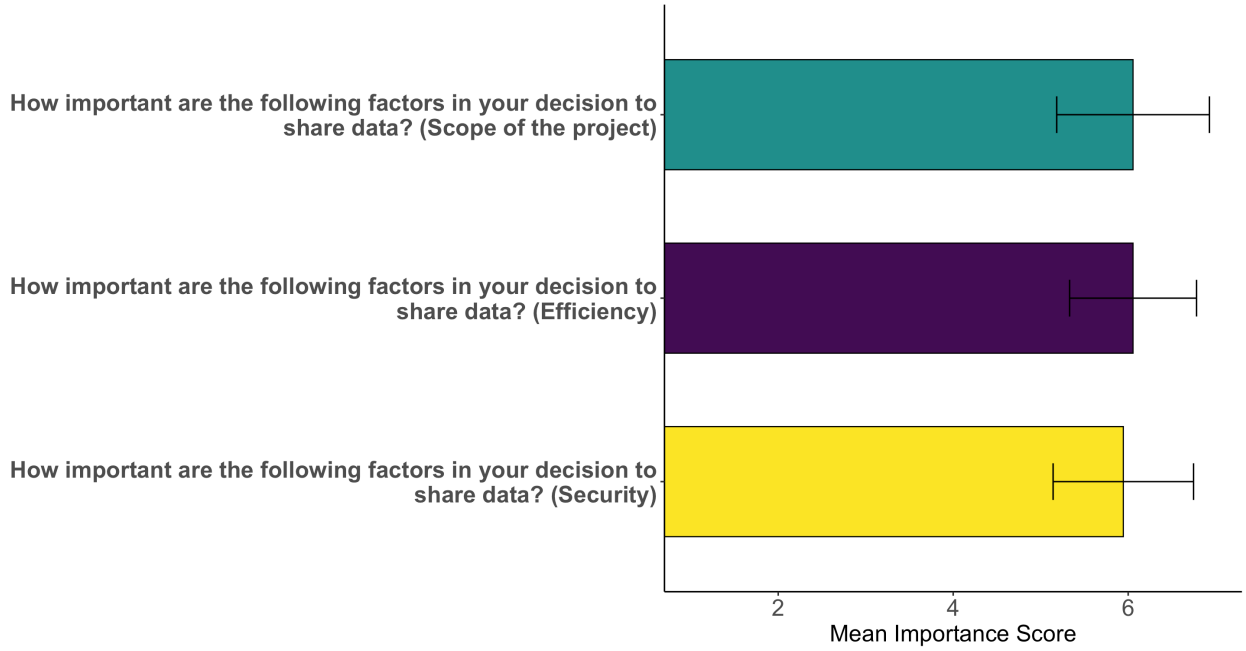


Figure F 6

Importance Scores for "Data Quality Control"

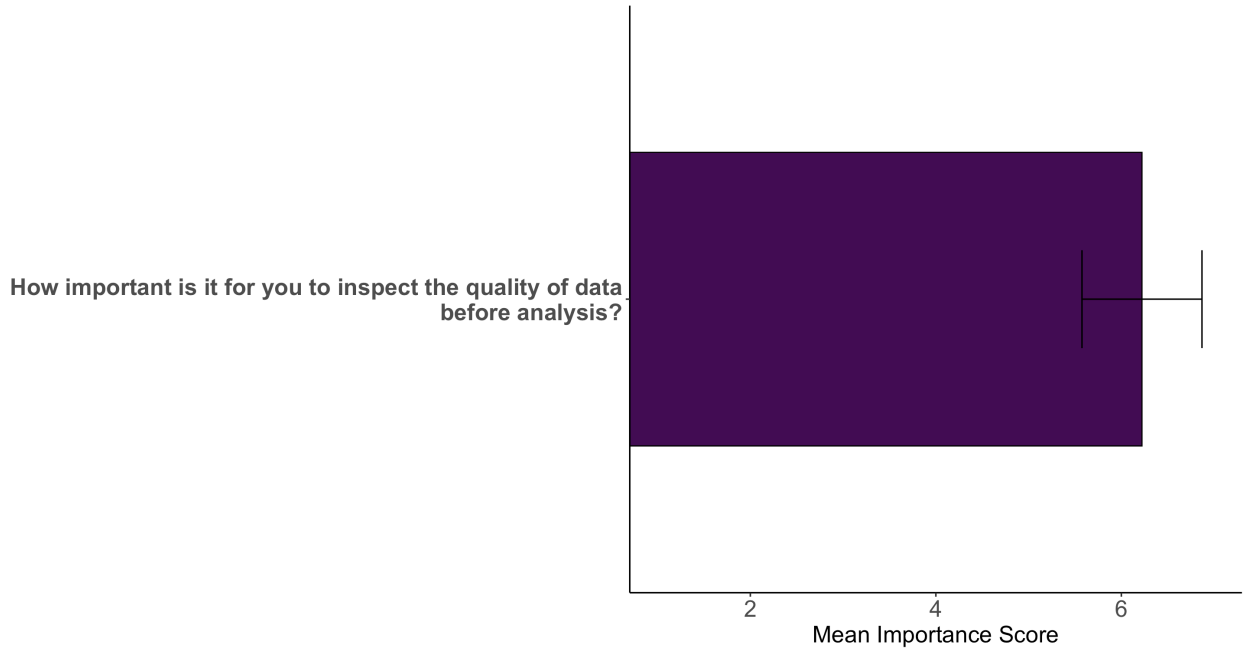


Figure F 7

Importance Scores for "Automated Data Completeness Checks"

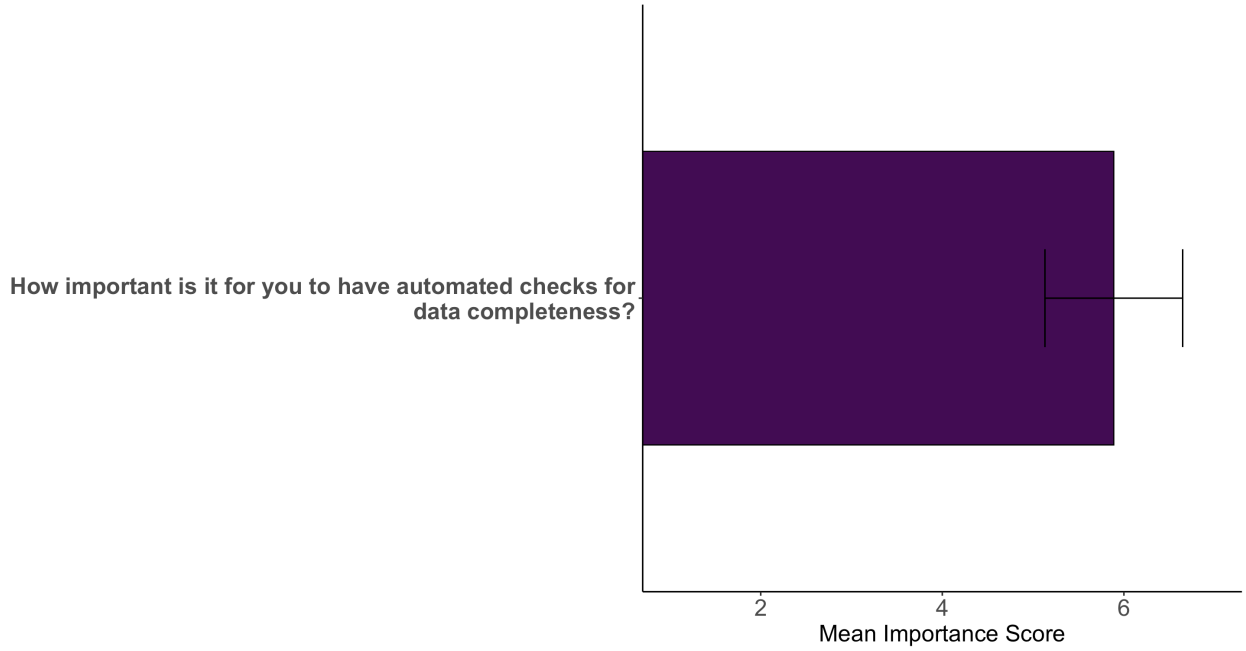


Figure F 8

Importance Scores for "Types Of Analyses in Research"

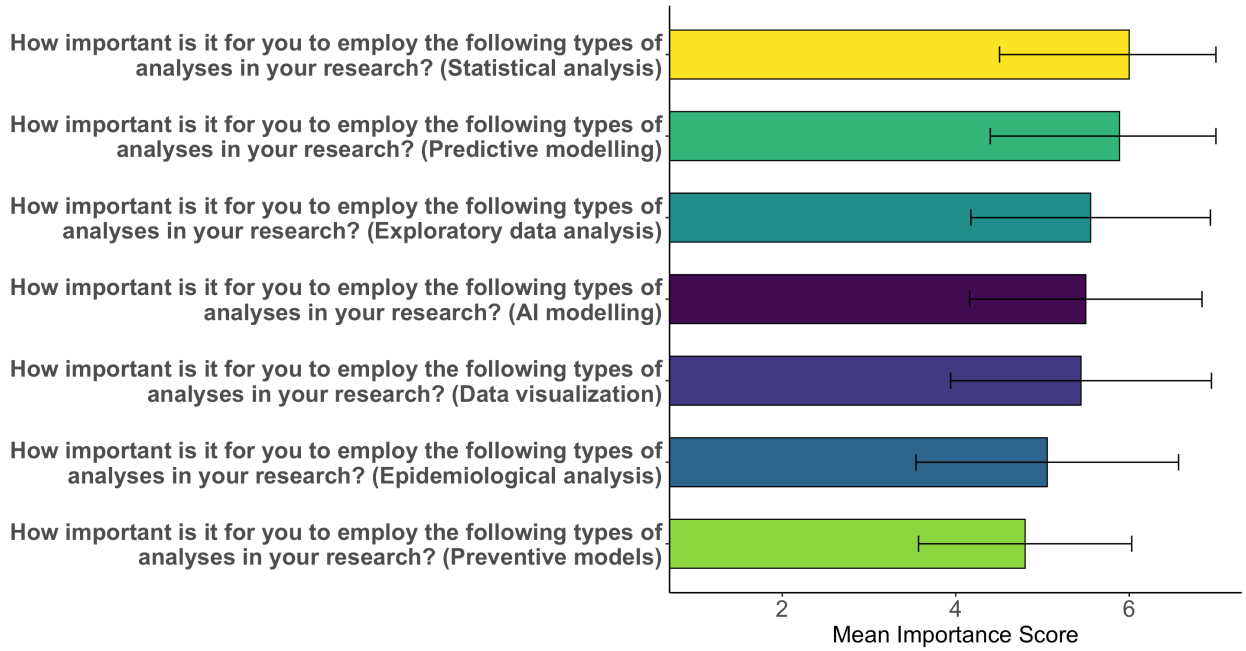


Figure F 9

Importance Scores for "Reproducibility"

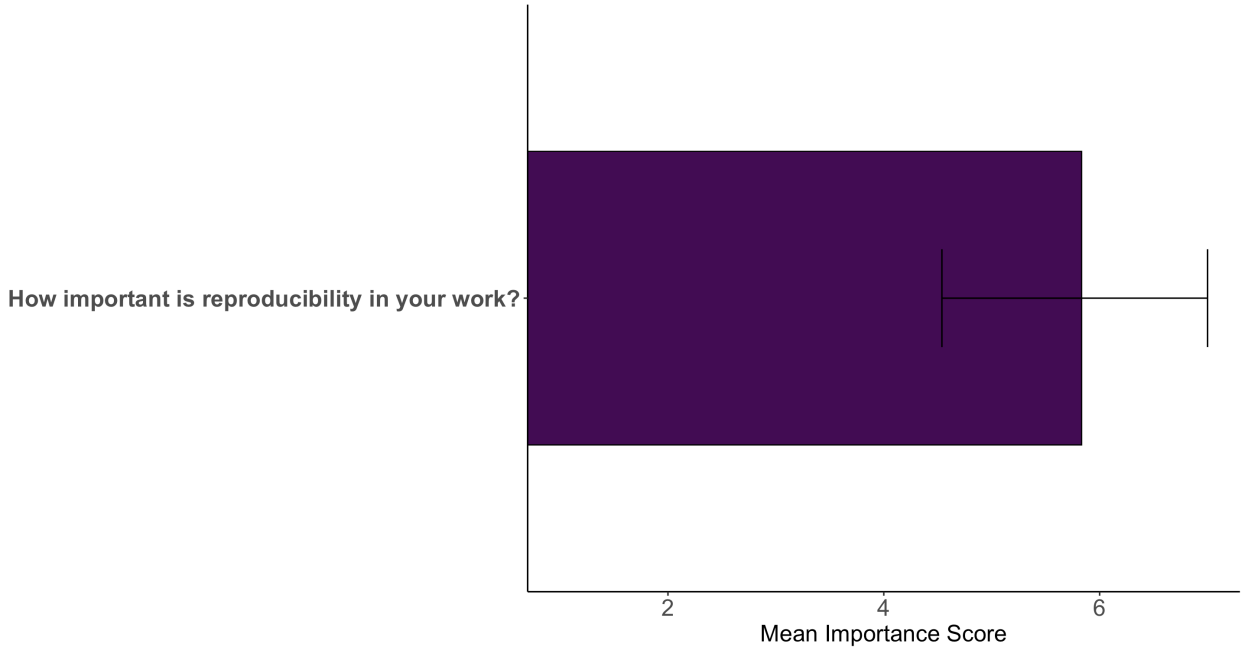


Figure F 10

Importance Scores for "Use Of Command-Line Tools"

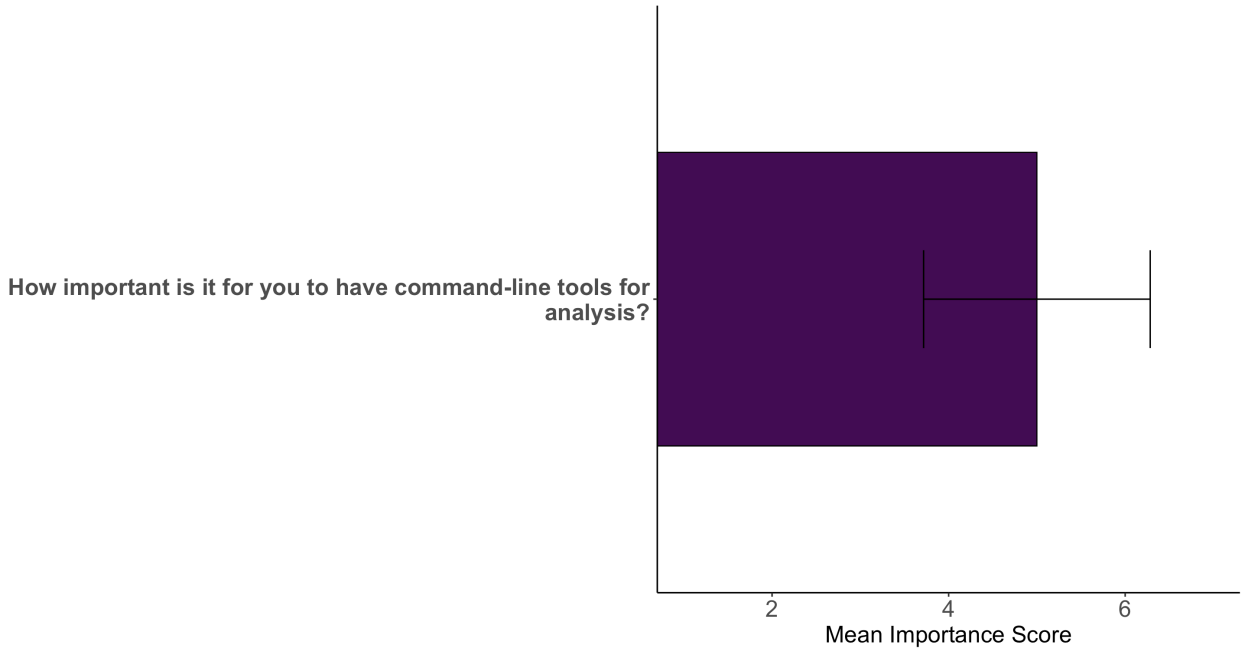


Figure F 11

Importance Scores for "Preferred Visualization Methods"

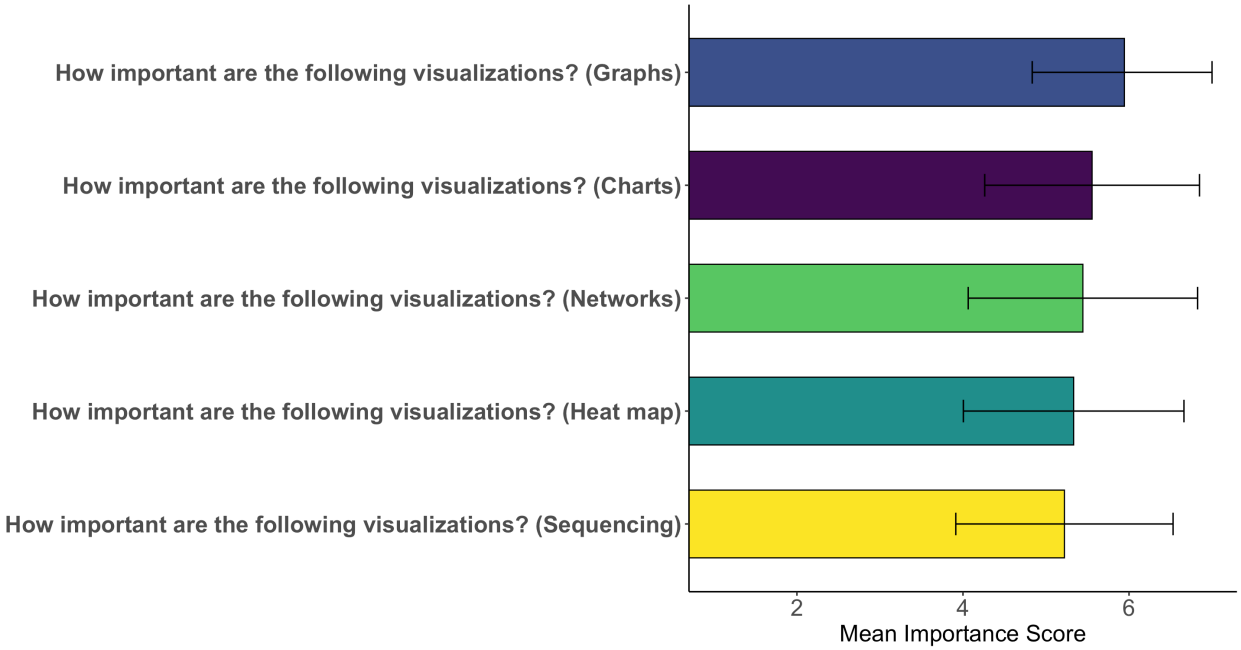


Figure F 12

Importance Scores for "Data Export & Download"

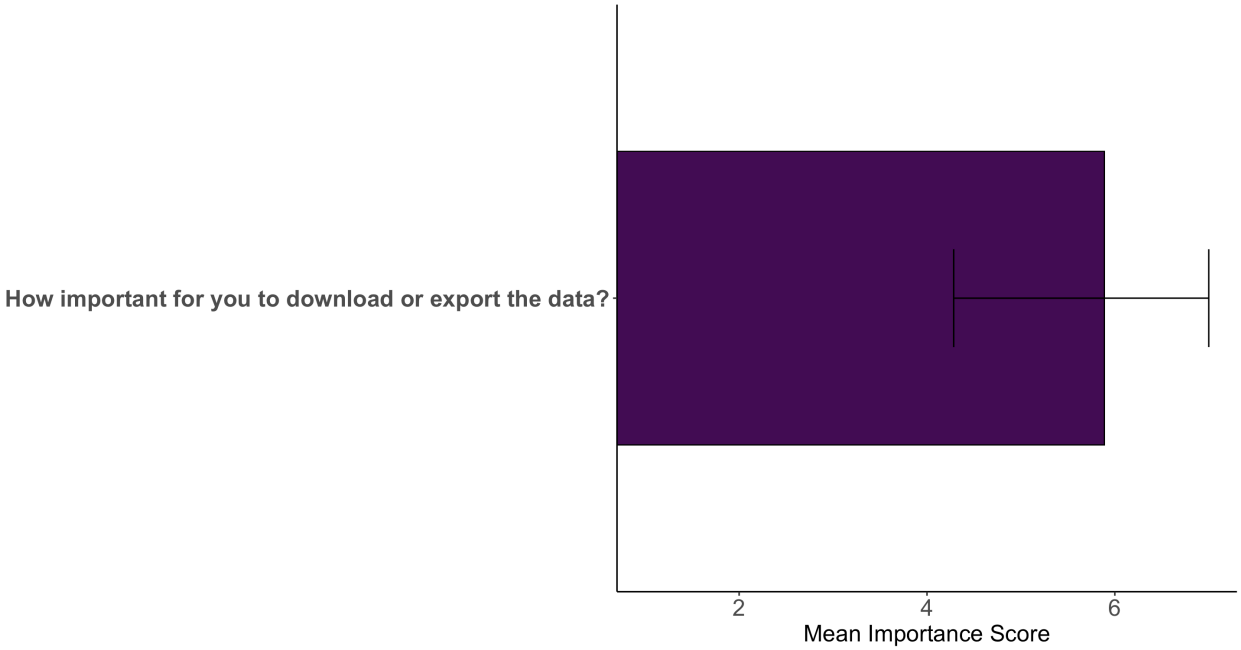


Figure F 13

Importance Scores for "Knowledge Sharing"

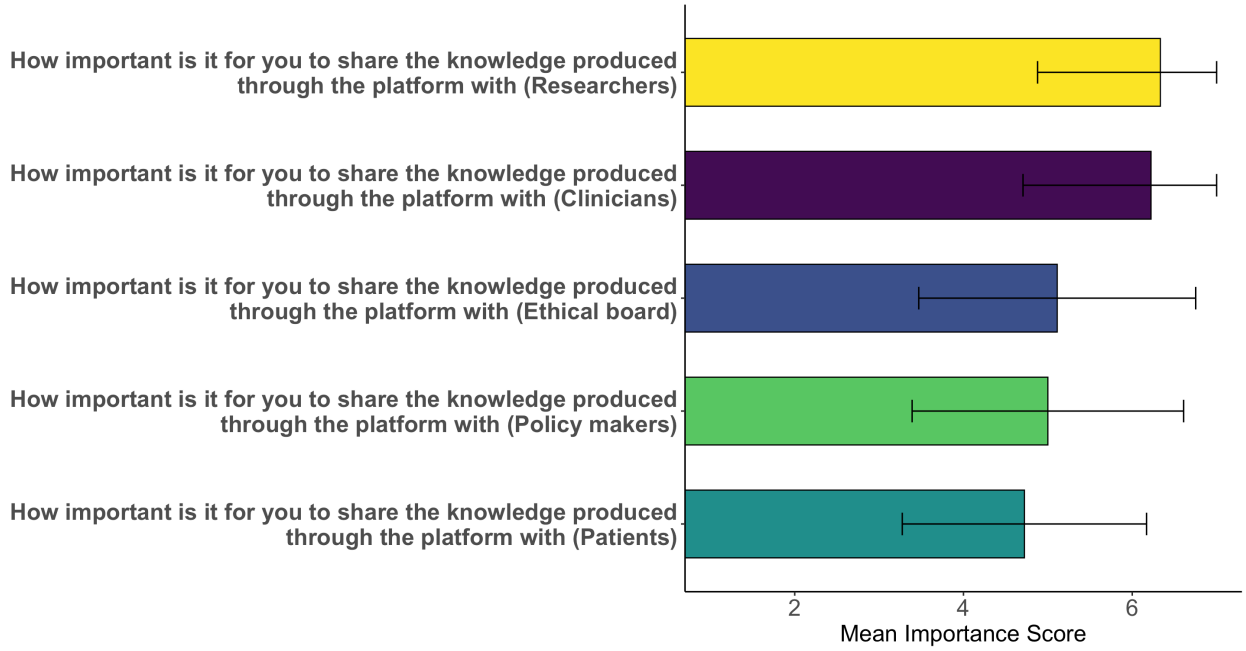


Figure F 14

Importance Scores for "Data Privacy Protection"

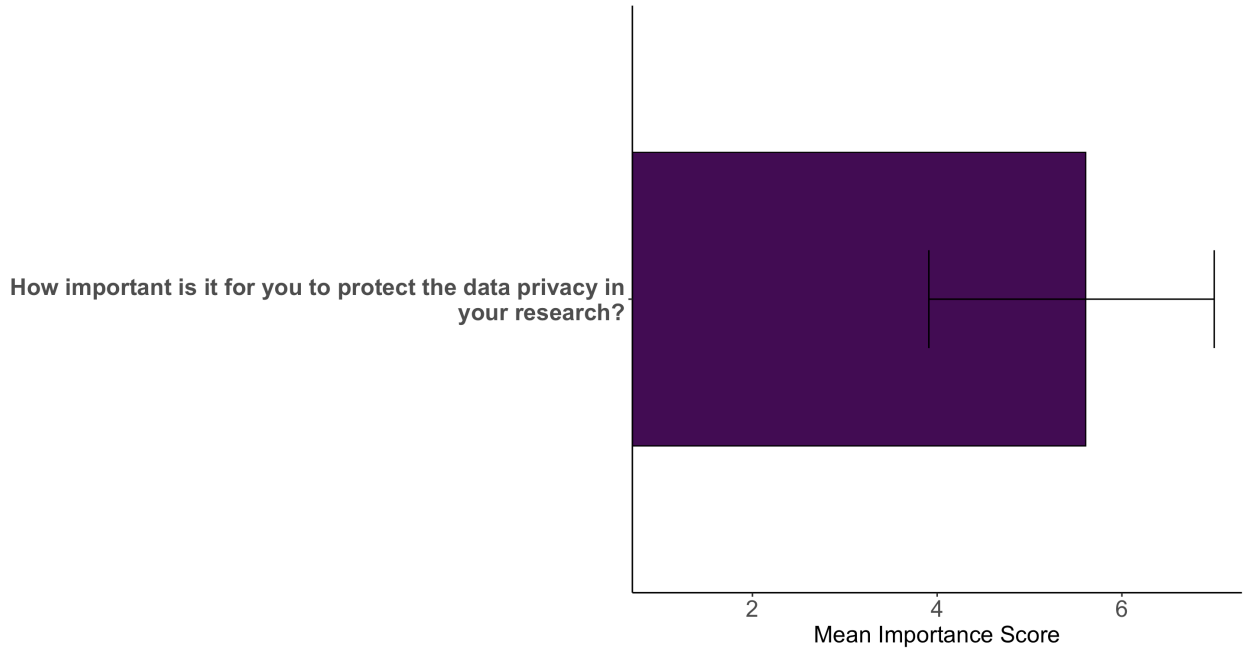


Figure F 15

Importance Scores for "Security Standards Awareness"

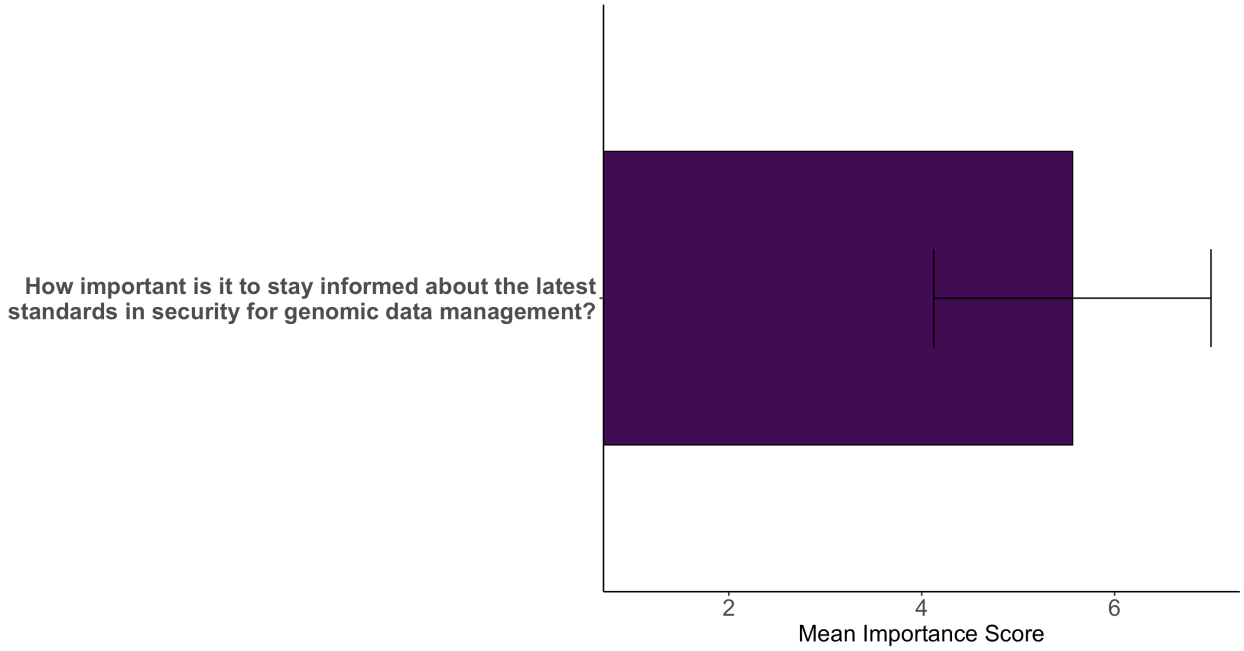


Figure F 16

Importance Scores for "Platform Usability (Mobile-Friendly)"

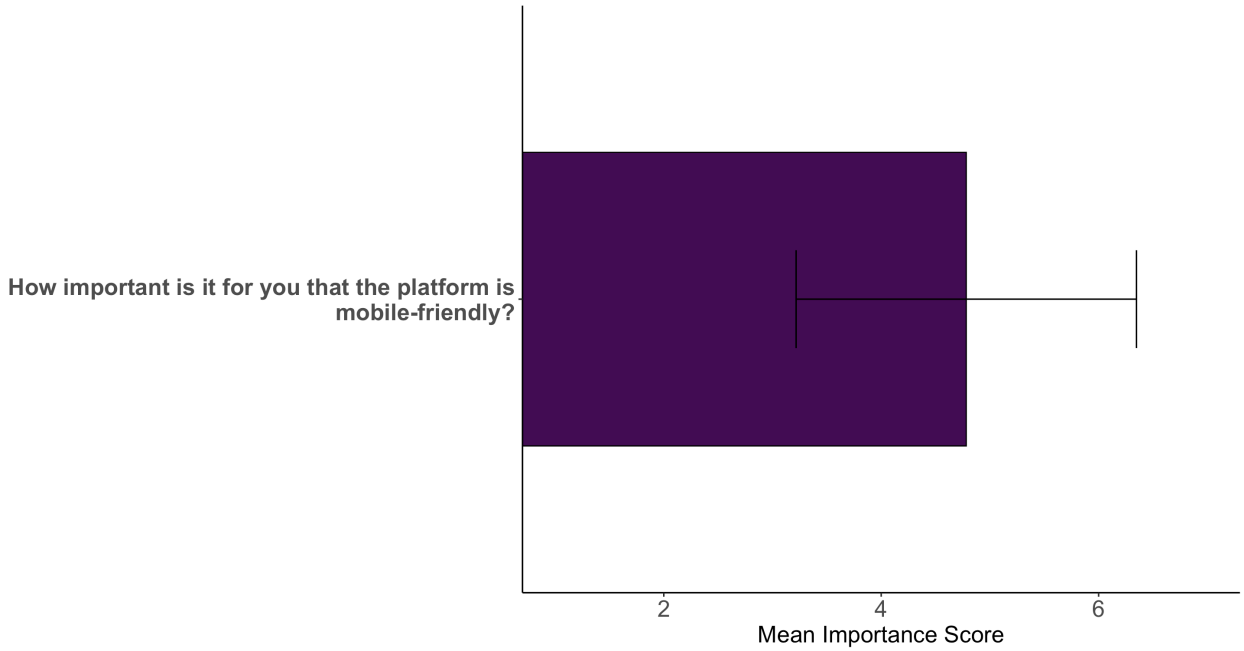


Figure F 17

Importance Scores for "Multi-Language Support"

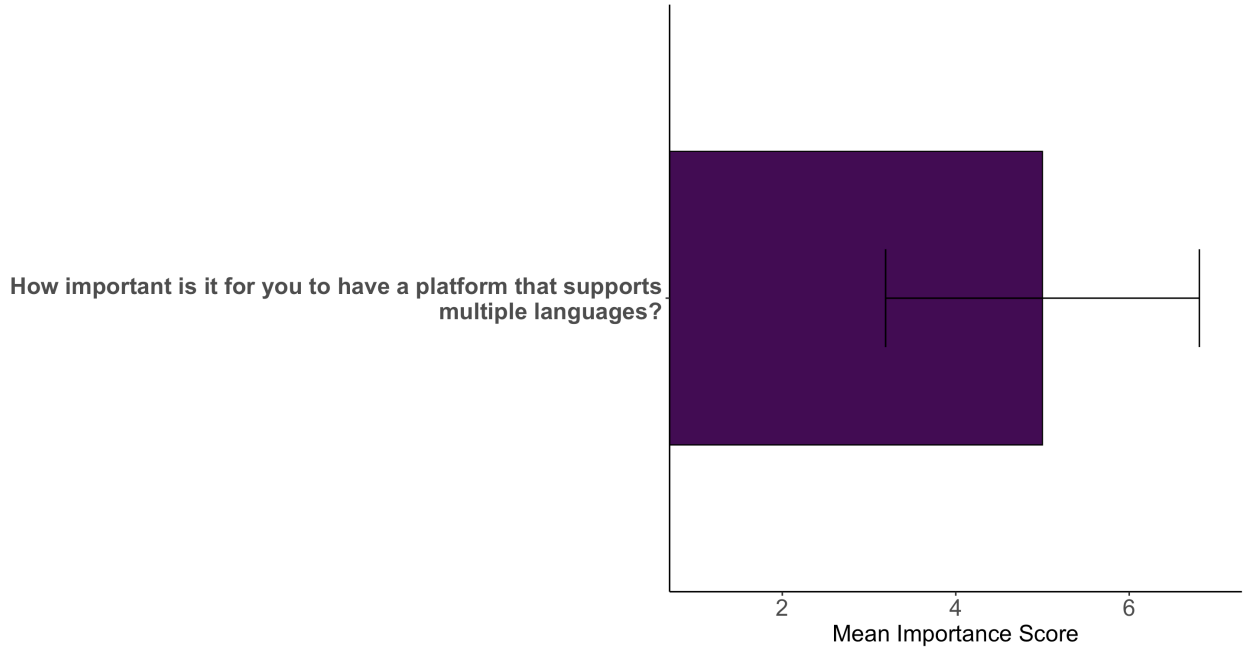


Figure F 18

Importance Scores for "Platform Notifications"

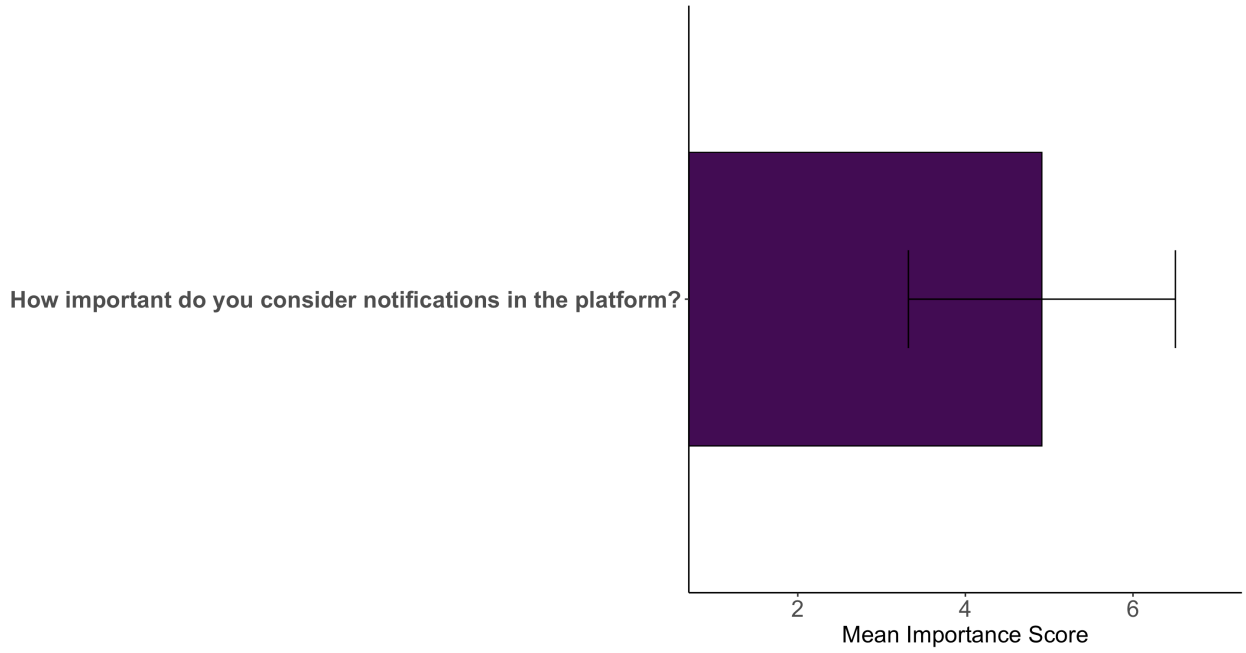


Figure F 19

Importance Scores for "Access to Federated Computing"

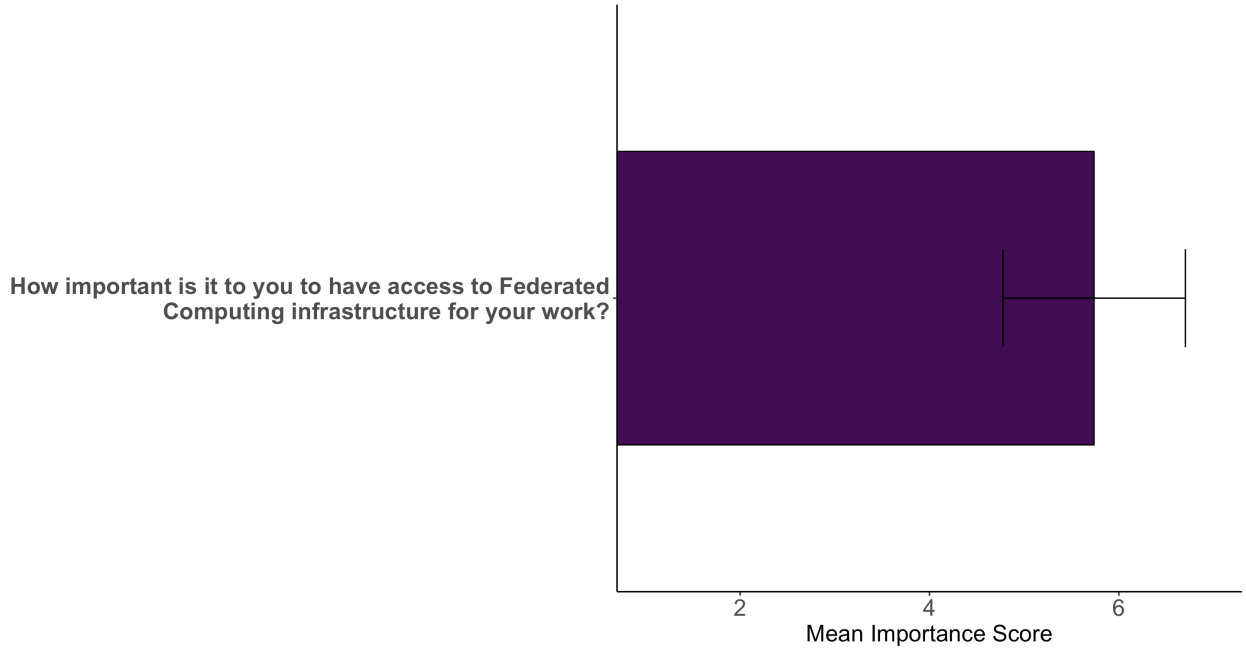


Figure F 20

Importance Scores for "Federated Computing Criteria"

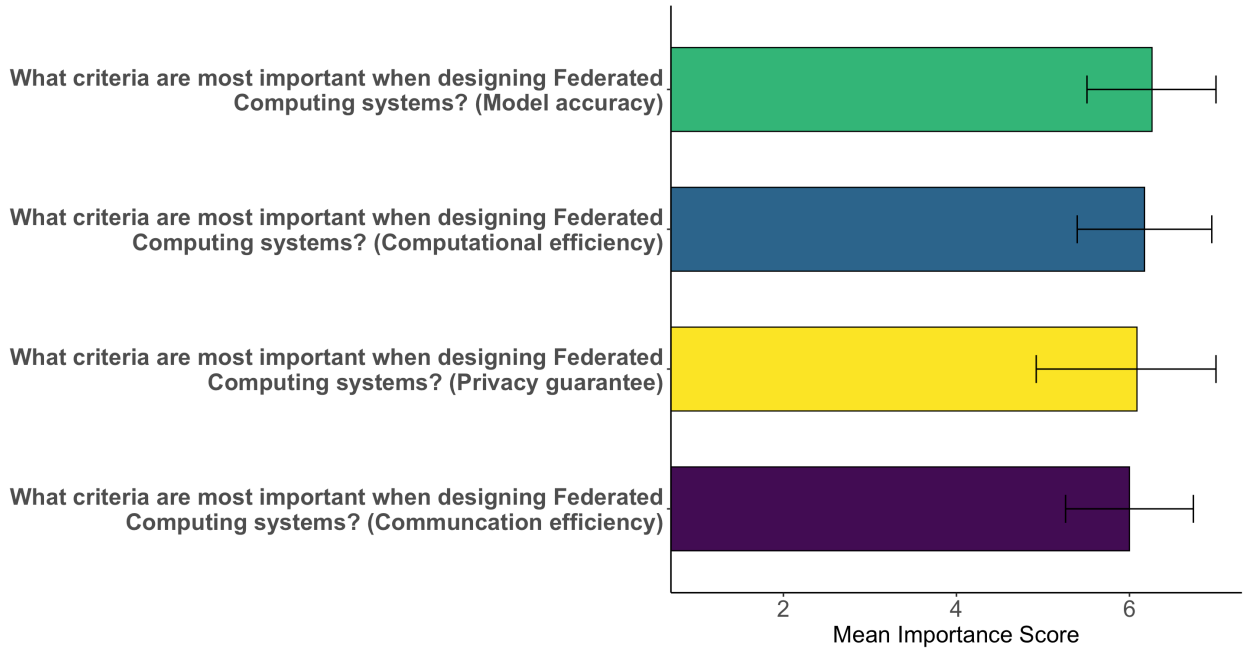


Figure F 21

Importance Scores for "Data Organization in Research"

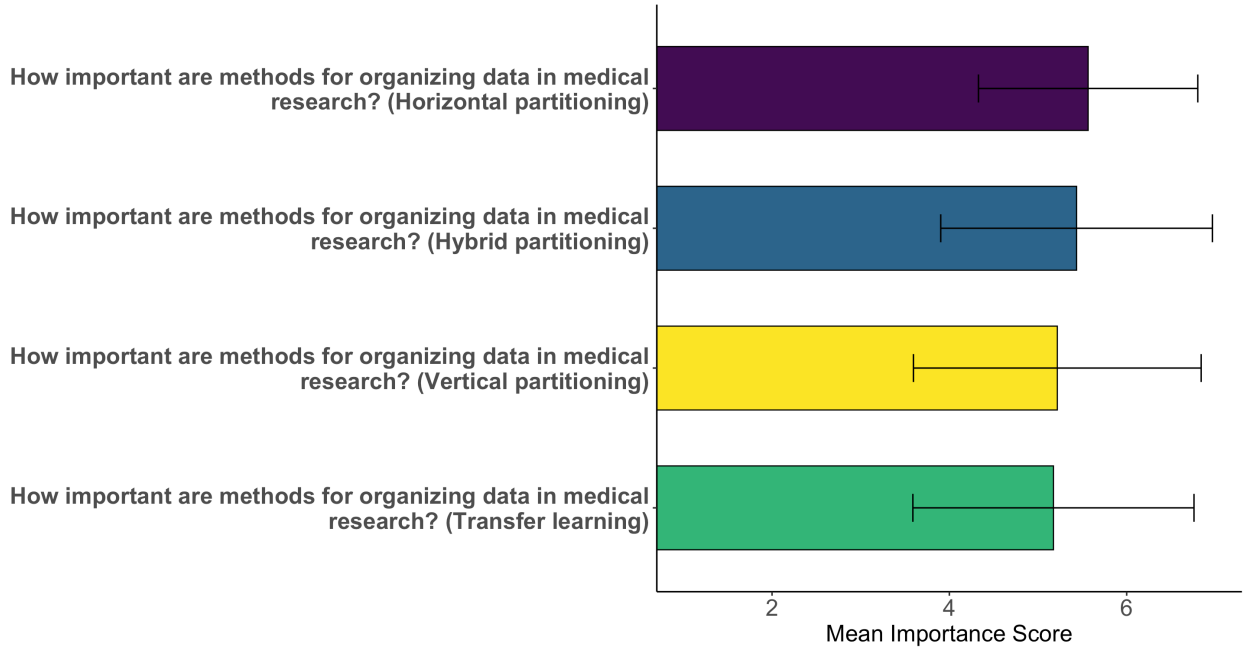


Figure F 22

Importance Scores for "Participant Selection Methods"

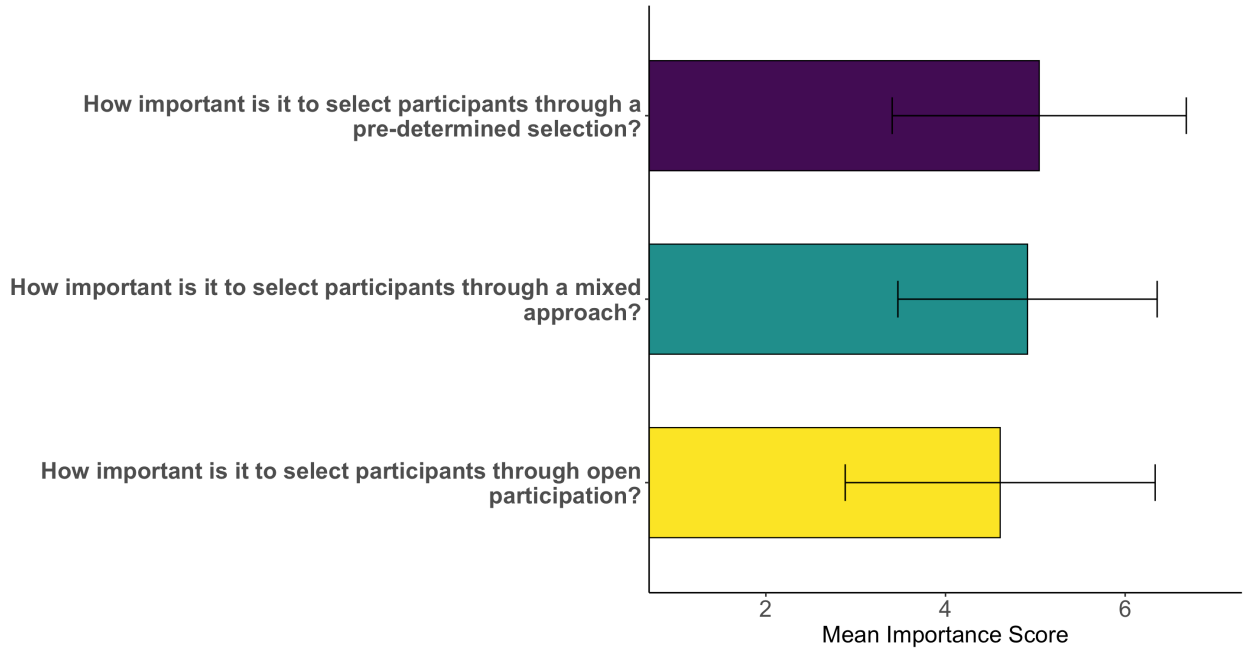
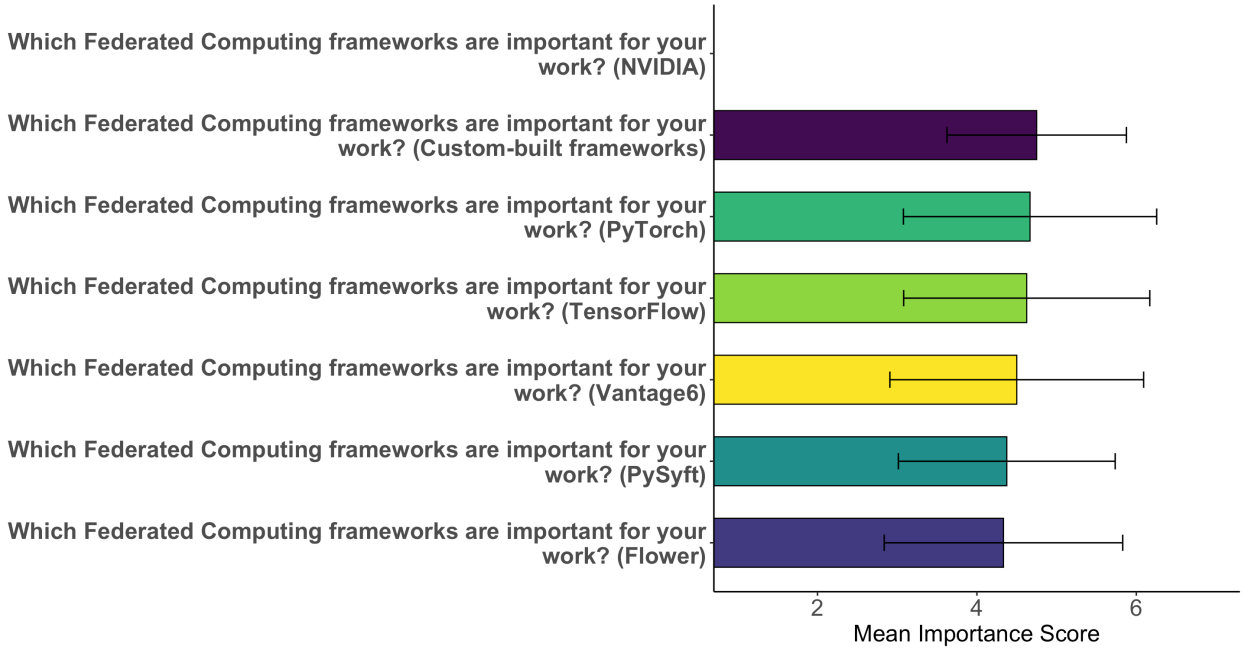


Figure F 23

Importance Scores for "Federated Computing Frameworks"


Which Federated Computing frameworks are important for your work? (NVIDIA)
Which Federated Computing frameworks are important for your work? (Custom-built frameworks)
Which Federated Computing frameworks are important for your work? (PyTorch)
Which Federated Computing frameworks are important for your work? (TensorFlow)
Which Federated Computing frameworks are important for your work? (Vantage6)
Which Federated Computing frameworks are important for your work? (PySyft)
Which Federated Computing frameworks are important for your work? (Flower)
2
4
6
Mean Importance Score

Appendix G

Informed Consent Form for the Workshop Participation

Informed consent

Social science approach towards PROTECT-CHILD

Dear Participant,

You are invited to participate in a workshop aimed at discussing challenges in creating federated health data platforms. This workshop is part of a research initiative focused on improving data sharing, analysis, and privacy protection in health data research, with an emphasis on Federated Learning and other privacy-preserving techniques. The session will involve discussions to identify barriers and potential solutions, helping to shape the future of secure, efficient, and privacy-preserving health data management.

Purpose of the Workshop

The goal of this workshop is to collaboratively identify key challenges and practical solutions related to handling large volumes of data, ensuring data privacy, facilitating data analysis, and managing legal and social aspects. Your insights will contribute to the development of innovative solutions that improve data accessibility, collaboration, and compliance with privacy regulations, particularly in fields like medical and genomic research.

Workshop Activities

During the workshop, you will be asked to:

- Discuss the challenges you face in health data environments.
- Share insights on strategies or solutions to overcome these challenges.
- Collaborate with fellow participants to propose potential tools and approaches.

The session will include interactive discussions and written activities. Your contributions will play a crucial role in shaping the design and development of future federated platforms.

Data Usage

The information you provide will be anonymized and used solely for research and development purposes. No personal identifying information will be collected or shared. All data will be securely stored and handled with strict confidentiality.

Voluntary Participation

Participation in this workshop is entirely voluntary. You may withdraw at any time, and your data will be securely deleted. There will be no consequences for choosing to withdraw.

Benefits and Risks

By participating, you will contribute to advancements in federated data platforms, benefiting research and clinical practice by enabling secure, collaborative, and efficient data use. There are no foreseeable risks associated with this workshop.

Contact Information

If you have any questions about the workshop or wish to withdraw your participation, please contact us at **v.d.c.resendez@utwente.nl**. If you have concerns about your rights as a participant or would like to discuss the study with someone independent of the research team, you can reach the Secretary of the Ethics Committee for the Humanities & Social Sciences domain at the Faculty of Behavioural, Management, and Social Sciences, University of Twente, viaethicscommittee-hss@utwente.nl

Thank you for your valuable time and input. Your participation is crucial to the success of this initiative.

Sincerely,

WP 2 Coordination Team -PROTECT CHILD

University of Twente

Consent Form for

Social science approach towards the PROTECT-CHILD

Please tick the appropriate boxes	Yes	No
Taking part in the study		
I have read and understood the study information dated 15/01/2025. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	□	□
I can refuse to answer questions and withdraw from the study at any time without having to give a reason.	□	□
I understand that taking part in the study involves: 1. Discussing challenges encountered in dealing with genomic data. 2. Sharing insights and my approach to working with genomic data. 3. An audio and video recording of these discussions will be destroyed after transcription.	□	□
Risks associated with participating in the study		
I understand that taking part in the study does not involve any foreseeable risks.	□	□
Use of the information in the study		
I understand that the information I provide will be used to create a report for PROTECT-CHILD, academic articles, and other secondary purposes, such as website blog posts.	□	□
I understand that any personal information collected about me, such as my voice or work activities, that could identify me will remain confidential and not be shared outside the study team.	□	□
I agree that my information can be quoted in research outputs.	□	□
I agree to share a joint copyright of the notes written during the workshop with the researchers of this initiative.	□	□
I agree to be audio and video recorded.	□	□
Future use and reuse of the information by others		

I give permission for the audio and video recordings I provide during this study to be stored in an encrypted Surfspot folder until they are transcribed, after which the recording will be permanently deleted.	□	□
I give permission for the anonymized transcription of the recording to be retained and made available for future research and learning purposes.	□	□
I give the researchers permission to keep my contact information and to contact me for future research projects.	□	□

Signatures

_____ _____ _____

Name of participant Signature Date

I have accurately share the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting.

_____ _____

_____

Researcher name Signature

Date

Study contact details for further information: Valeria Resendez: v.d.c.resendez@utwente.nl

Note. This workshop consent form was authored by Resendez et al. (2025) and is reproduced here in full for reference purposes.

Appendix H

Excerpt of Email from Resendez et al. (2025)

Dear PROTECT CHILD Team,

We're excited to have you join us for the upcoming workshop in Rome. To ensure a productive and engaging session, we kindly ask for your help with two preparatory steps:

1. **Review the attached document**

   Attached to this email, you'll find a document outlining key high-level requirements (functionalities and characteristics) for the future PROTECT CHILD platform. Please take a moment to review it, as it will form the basis of one of our workshop activities.

2. **Complete the survey in this link:**

Completing this survey will help you engage more deeply with the material and provide valuable input for our discussions. Your insights are key to shape the direction of the project.

Bring printed materials

To ensure smooth participation during the activities, print the attached document and bring it with you to the workshop.

Your preparation will be key to make this session a success, and we truly appreciate your time and effort. If you have any question or need support, please don't hesitate to get in touch.

We look forward to meeting you in Rome!

Warm regards,

WP2 Coordinator

Note. This email was originally written by Dr. Valeria Resendez and is presented here in full with appropriate acknowledgment.

Appendix I

Excerpt of Activities from Resendez et al. (2025)

Activity 1

PART 1 – QUESTIONS YOU MUST ANSWER IN GROUP

When you are thinking about a system like PROTECT CHILD for better integration and use of health-related real-world and research data, including genomics, to improve clinical outcomes ...

Question 1: How do you plan to use the system (or how do you think potential users will use it)?

Question 2: What are the technical enablers/tools that can support the achievement of the clinical objectives?

Activity 2

If you have to use data from health platforms like the one you envisage it will be PROTECT CHILD:

1. CURRENT PRACTICE: How do you usually deal with such challenges? i.e. functions you have used/processes you have adopted, shortcuts you have adopted.
2. EMERGING ISSUES AND SOLUTIONS: Are these challenges relevant for the future? Are there other challenges and issues that we are missing or that you see on the horizon? What solutions do you see for these emerging issues/challenges?

Note. The activities included in this appendix were originally developed and conducted by Resendez et al. (2025). They are reproduced here in full for contextual reference.

Appendix J

Excerpt of Participants Summary Sheet from Resendez et al. (2025)

Overview of challenges for health data management platforms

Challenge 1: Handling large volumes of data

With data coming from various sources and in diverse formats, managing and organizing this information can be incredibly complex. Platforms should support users in data upload and utilization.

Challenge 2: Ensuring privacy and security of the data

Ensuring the security and privacy of genomic data is an important concern when dealing with sensitive, highly personal information.

Challenge 3: Complexity of Data Analysis

Platforms should help to overcome the complexities of data analysis**.** Genomic data is complex; therefore, it requires specific tools and methods to extract meaningful insights.

Challenge 4: Managing legal and social aspects

Managing the legal, social, and ethical issues associated with healthcare information. These issues require careful navigation to ensure ethical standards are upheld while balancing research progress. Platforms should be trustworthy, explain how data is managed, and enable for instance the possibility to recall and hide data.

Note. The summary presented here is a partial excerpt from participant documentation developed by Resendez et al. (2025). Images and full detailed materials have been removed. For complete access to the original participant documents and challenge sheets, please contact Dr. Valeria Resendez.

Appendix K

R-Studio Code for the Data Analysis of the Workshop Data

```
# 1. Install / load needed packages
if (!require(readxl)) install.packages("readxl")
if (!require(dendextend)) install.packages("dendextend")
library(readxl)
library(dendextend)
library(dplyr)
library(tidyr)
library(ggplot2)
library(reshape2)

# 2. Read in your coding matrix
#    (adjust the path to wherever you saved the .xlsx file)
df <- read_excel("Cleaned_Coding_Matrix_Numeric.xlsx")

code_labels <- c(
  "Current platform shortcomings",
  "Data quality, standardization & formats",
  "Ethics support infrastructure",
  "Legal & Regulatory frameworks",
  "Need for system simplicity",
  "Platform benchmarking",
  "Stakeholder involvement",
  "Technical enablers (TE)",
  "TE - Advanced analytics & learning tools",
  "TE - Automation & data collection",
  "TE - Data accessibility & querying",
  "TE - Data security & privacy preservation",
  "TE - Data standardization & harmonization",
  "TE - Interface design & user-friendliness",
  "TE - Interoperability",
  "TE - Scalability & performance",
  "TE - Support",
  "Value of the platform",
  "Vision of the platform",
  "Vision - Clinician-focused dashboard",
  "Vision - Clinical decision making & benchmarking",
  "Vision - Data integration",
  "Vision - Open source & flexibility",
  "Vision - Research collaboration",
  "Vision - Scalability, adaptability, access & data management",
  "Vision - Support & documentation",
  "Visualization needs",
  "Current legal, private, and ethical protocols",
  "Data access",
  "Data analysis needs",
  "Data ownership",
  "Data sharing",
  "Data storage",
  "Financial matters",
  "Issues with data handling",
  "Patient communication",
  "Privacy & ethical needs",
  "Privacy, legal & ethics-related challenges",
  "Usability challenges & needs"
)

# 3. Prepare the binary matrix of codes × quotes
#    Drop the first “Quote” column, convert to matrix, then transpose:
code_mat <- t(as.matrix(df[ , -1]))
rownames(code_mat) <- code_labels
# give your rows proper code names

# 4. Compute the pairwise binary distance between codes
#    (this treats each code vector of 0/1 across quotes)
dist_mat <- dist(code_mat, method = "binary")

# 5. Perform hierarchical clustering using Ward’s method
#   ward.D2 implements the classical “minimum variance” Ward
hc <- hclust(dist_mat, method = "ward.D2")


file_path <- "Cleaned_Coding_Matrix_Numeric.xlsx"
code_matrix <- read_excel(file_path)


rownames(code_matrix) <- code_matrix[[1]]
code_matrix <- code_matrix[,-1]

# Transpose the matrix so rows = codes, columns = quotes
code_matrix_t <- t(code_matrix)

# Cluster quotes instead of codes (K-means)
set.seed(123)
km <- kmeans(code_matrix, centers = 5)

# Add cluster info
```

```
code_matrix$Cluster <- as.factor(km$cluster)
code_matrix$Quote <- rownames(code_matrix)

# Melt for plotting
long_df <- melt(code_matrix, id.vars = c("Quote", "Cluster"))

# Calculate mean frequency of each code per cluster
# Ensure code_labels is the correct length and order!
agg <- aggregate(value ~ variable + Cluster, data = long_df, FUN = mean)
agg$variable <- factor(agg$variable, levels = unique(agg$variable), labels = code_labels)


########## Hierarchical Clustering of Codes (Ward's Method)
install.packages("ggdendro")
library(ggdendro)

# Convert hclust to dendrogram
dend <- as.dendrogram(hc)

# Convert dendrogram to ggplot-compatible object
dend_data <- dendro_data(dend, type = "rectangle")

# Plot with ggplot2
library(ggplot2)
ggplot(segment(dend_data)) +
  geom_segment(aes(x = x, y = y, xend = xend, yend = yend), size = 1) +
  geom_hline(yintercept = 1.15, color = "red", linetype = "dashed", size = 1) +  # <--- This adds a red line
  scale_y_continuous(expand = c(0.05, 0)) +
  labs(
    y = "Distance",
    x = NULL
  ) +
  theme_minimal(base_size = 18) +
  theme(
    plot.title = element_text(size = 20, face = "bold"),
    axis.text.x = element_text(size = 12, face = "bold"),
    axis.text.y = element_text(size = 14, face = "bold"),
    axis.title = element_text(size = 16, face = "bold")
  ) +
  geom_text(
    data = label(dend_data),
    aes(x = x, y = y - 0.03, label = label),   # y offset to avoid overlap
    angle = 90, hjust = 1, size = 4, fontface = "bold"
  )



########## --- K-MEANS CLUSTERING ---

# Custom color palette (optional)
custom_palette <- c(
  "#E57373", # red
  "#FFD54F", # yellow
  "#64B5F6", # blue
  "#81C784", # green
  "#BA68C8"  # purple
)

ggplot(agg, aes(x = variable, y = value, group = Cluster, color = Cluster)) +
  geom_line(aes(linetype = Cluster), size = 1.5, alpha = 0.7) +        # Thicker, semi-transparent lines
  geom_point(size = 3, alpha = 0.85) +                                 # Larger points
  scale_color_manual(values = custom_palette) +                        # Custom colors
  theme_minimal(base_size = 17) +                                      # Larger base font
  labs(
    x = "Code",
    y = "Average Code Frequency in Cluster",
    color = "Cluster Theme",
    linetype = "Cluster Theme"
  ) +
  theme(
    axis.text.x = element_text(angle = 55, hjust = 1, vjust = 1, size = 9, face = "bold"),
    axis.text.y = element_text(size = 14),
    axis.title = element_text(size = 16, face = "bold"),
    legend.position = "top",
    legend.title = element_text(size = 16, face = "bold"),
    legend.text = element_text(size = 14),
    plot.title = element_text(size = 20, face = "bold"),
    plot.subtitle = element_text(size = 15, face = "italic"),
    panel.grid.minor = element_blank(),
    panel.grid.major.x = element_blank()
  )
```

```
##########  --- K-MEANS CLUSTERING --- SEPARATE

# Use the same palette as before
custom_palette <- c(
  "#E57373", # Cluster 1: red
  "#FFD54F", # Cluster 2: yellow
  "#64B5F6", # Cluster 3: blue
  "#81C784", # Cluster 4: green
  "#BA68C8"  # Cluster 5: purple
)
names(custom_palette)             <- as.character(1:5)

library(ggplot2)

# Ensure 'Cluster' is character, not factor, for safe color matching
agg$Cluster                       <- as.character(agg$Cluster)

for(cl in unique(agg$Cluster)) {
  subdf <- subset(agg, Cluster == cl)
  # Order variable factor by value, so line goes from lowest to highest frequency
  subdf <- subdf[order(subdf$value), ] # bottom-to-top (lowest value at top)
  subdf$variable                  <- factor(subdf$variable,   levels   = subdf$variable)

  p   <-   ggplot(subdf,   aes(y   = variable, x = value, group = 1)) +
    geom_line(color                = custom_palette[cl], size = 1.7) +
    geom_point(color               = custom_palette[cl], size = 3) +
    theme_minimal(base_size = 17) +
    labs(
      title = paste("K-Means Cluster Profile – Cluster", cl),
      y = "Code",
      x = "Average Code Frequency in Cluster"
    ) +
    theme(
      axis.text.y                  = element_text(size   =   15,   face   = "bold"),
      axis.text.x                  = element_text(size = 12),
      axis.title = element_text(size = 16, face = "bold"),
      plot.title = element_text(size = 20, face = "bold"),
      plot.subtitle                = element_text(size   =   15,   face   = "italic"),
      panel.grid.minor             = element_blank(),
      panel.grid.major.y           = element_blank()
    )
  print(p)

  # Save each plot (optional)
  ggsave(
    filename                       = paste0("cluster_profile_",       cl, ".png"),
    plot = p,
    width = 16, height = 9, dpi = 200
  )
}

########## Frequency analysis

library(dplyr)

top_codes <- agg %>%
  group_by(Cluster) %>%
  arrange(desc(value)) %>%
  slice_head(n = 5) %>%
  ungroup()


#  Optionally,  rename  columns  for clarity
top_codes <- top_codes %>%
  mutate(Frequency   =   round(value, 2)) %>%
  rename(Code = variable) %>%
  select(Cluster, Code, Frequency)

# Print results per cluster
for              (cl              in unique(top_codes$Cluster)) {
  cat("Cluster", cl, "\n")
  print(top_codes %>% filter(Cluster == cl) %>% select(Code, Frequency))
  cat("\n")
}
```