This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement Nº 101137423

D2.1: High-level requirements and user stories

Deliverable No.	D2.1.1	**Due Date**	30/11/2025
Description	This deliverable outlines the stakeholder-driven process used to define, co-design, and validate the high-level functional and non-functional requirements for the PROTECT-CHILD infrastructure. It reports on the activities conducted in collaboration with stakeholders, which resulted in a set of user stories capturing the high-level requirements.		
Type	Report	**Dissemination Level**	PU
Work Package No.	WP2	**Work Package Title**	Codesign and multi-stakeholder requirements
Version	v1.0	**Status**	Final

Authors

Name and surname	Partner name	e-mail
Simone Borsci	UTW	s.borsci@utwente.nl
Funda Yildirim	UTW	f.yildirim@utwente.nl
Valeria Resendez	UTW	v.d.c.resendez@utwente.nl
Elena Martinelli	MME	martinel.ele@virgilio.it
Franco Mercalli	MME	f.mercalli@multimedengineers.com
Eugenio Gaeta	UPM	eugenio.gaeta@upm.es
Maria Fernanda Cabrera	UPM	chiqui@lst.tfo.upm.es
Aya Mahboub	UPM	amahboub@lst.tfo.upm.es

Document History

Version	Date	Changes	Authors
v1.0	04-12-25	Final review and final version	Aya Mahboub
v0.8	25-11-25	Refining user requirements and outlining specifications for the next iteration cycle	Simone Borsci, Funda Yildirim, Valeria Resendez, Elena Martinelli, Franco Mercalli, Eugenio Gaeta, María Fernanda Cabrera
v0.7	03-11-25	Integration of additional literature for the user stories	Simone Borsci, Funda Yildirim, Valeria Resendez, Elena Martinelli, Franco Mercalli, Eugenio Gaeta, María Fernanda Cabrera
v0.6	24-10-25	Refinement of user stories based on input from geneticists and legal experts.	Simone Borsci, Funda Yildirim, Valeria Resendez, Elena Martinelli, Franco Mercalli, Eugenio Gaeta, María Fernanda Cabrera
v0.5	08-07-25	Integration of insights from the workshop in Palermo	Simone Borsci, Funda Yildirim, Valeria Resendez, Elena Martinelli, Franco Mercalli, Eugenio Gaeta, Maria Fernanda Cabrera
v0.4	15-05-25	Integration of prior activities, including literature review, survey, and workshop findings.	Simone Borsci, Funda Yildirim, Valeria Resendez, Elena Martinelli, Franco Mercalli, Eugenio Gaeta, Maria Fernanda Cabrera
v0.3	09-05-25	Refinement of user stories informed by outcomes from previous projects.	Simone Borsci, Funda Yildirim, Valeria Resendez, Elena

			Martinelli, Franco Mercalli, Eugenio Gaeta
v0.2	14-04-25	Further development of user stories based on consultation with project management and experts.	Simone Borsci, Funda Yildirim, Valeria Resendez, Elena Martinelli, Franco Mercalli, Eugenio Gaeta
v0.1	02-04-25	Development of user stories based on literature reviews.	Simone Borsci, Funda Yildirim, Valeria Resendez

Key data

Keywords	Requirements (functional and non-functional requirements), user stories, workshops, a survey, and a literature review
Lead Editor	Simone Borsci, Funda Yildirim, Valeria Resendez
Internal Reviewer(s)	Elena Martinelli, Franco Mercalli

Abstract

The PROTECT-CHILD project aims to develop a federated data infrastructure framework that integrates clinical and genomic data to improve the use of secondary data, particularly for pediatric liver and kidney transplantation. To achieve this, it is necessary to address challenges in data management and data exchange, including the management of complex genomic data, privacy, legal and ethical requirements, and the integration of data from diverse sources. To ensure that the platform addresses key challenges and remains usable for its intended users, we engaged several key stakeholders (e.g., clinical experts and healthcare providers, IT and technical teams, and data governance and compliance experts).

The present document describes the process of iteratively defining and evaluating high-level functional and non-functional requirements and the process of transforming such requirements into User Stories to inform the future PROTECT-CHILD platform. Such a process defines an initial architecture (from the perspective of the key stakeholders) and investigates how data and metadata architecture should be organized in terms of requirements, and how to enable interoperability. Specifically, the present document reports the outcomes of four main activities:

1. **Literature review**: Systematic definition of an initial list of requirements.
2. **Survey with key stakeholders** – Revision and extension of requirements in a user-centered modality.
3. **Focus group with key Stakeholders** – Co-design and co-exploration of requirements and expected challenges in usage and potential solutions.
4. **Generation and Embedding of User Stories** - Refinement of requirements to feed user stories, expert analysis of the stories, and embedding of stories in a low-level interactive Prototype on Figma (see http://tiny.cc/m6vj001) to enable visualization, interaction, and discussion with end-users and stakeholders of the functionalities and the requirements to feed the data model.
5. **Workshop with data users** - Refinement of user stories, testing a prototype based on a prior project.

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation, or both.

Table of Contents

Index of Figures

Index of Tables

Summary of the deliverable

The PROTECT-CHILD project aims to build a federated digital system that securely manages and analyzes pediatric clinical and genomic data for transplantation across the EU. This deliverable describes several activities undertaken to define high-level user requirements, both functional and non-functional, that will inform the platform's design, development, and deployment. The deliverable is connected to the Project Objectives as indicated in Table 1.

Table 1. Deliverable context

PROJECT ITEM IN THE DoA	RELATIONSHIP
Project Objectives	Stakeholders' involvement, co-design, and definition of key requirements from the perspective of stakeholders, including data and metadata utilization, (multi-standard) interoperability, and overall expected architecture of functionalities and relationships among functionalities of the future Protect-Child system.
Exploitable Results	Results are going to be used iteratively to improve requirements, define functionalities, and identify issues to ensure data exchange and interoperability.
Workplan	The next 6 months of activities will consolidate the current User Stories to implement final requirements and indications about the architecture
Milestones	None
Deliverables	High-Level Requirements and User Stories
Risks	Non foreseeable

Background and objectives

Rapid advancements in genomic technologies have opened new frontiers in transplant research. However, the full potential of these technologies is limited by fragmented data ecosystems and infrastructural, ethical, and regulatory barriers. The project addresses these gaps by defining requirements for a platform that balances technical innovation with legal, ethical, and usability standards. To support this goal, a systematic literature review, a survey with experts, and two stakeholder workshops were conducted to map requirements, challenges, and design priorities.

Literature Review (Systematic)

A PRISMA-based review of 53 studies showed three core functional requirements, data management, processing/analysis, and visualization/reporting, and four major non-functional requirements, communication/support, UI/UX, technical infrastructure, and data security/compliance. These insights guided the formulation of initial platform capabilities and informed the next phases of consultation.

Survey with key stakeholders

A survey involving 52 clinical, technical, and legal experts evaluated the importance of 70 requirements. Results confirmed high consensus on priorities such as privacy-

preserving techniques, federated learning support, data acquisition standardization, and intuitive user interface (UI) features. Domain-specific priorities emerged: clinicians emphasized data acquisition from multiple sources, automated data quality checks, predictive/statistical modelling, and visualization, while legal experts focused on the need for standardized EU guidelines to ensure GDPR compliance, privacy-by-design, and transparency in federated environments. Technical respondents highlighted the importance of privacy-preserving techniques, horizontal data partitioning, federated learning/analytics, and reproducibility.

Focus group with key stakeholders

A focus group with 36 stakeholders across disciplines validated these findings and emphasized:

- A federated, scalable infrastructure supporting local analysis and cross-site collaboration.
- Automation for data extraction and standardization;
  Privacy-preserving computation and modular governance models.
- Legal and ethical harmonization of workflows to resolve consent, data ownership, and approval of inconsistencies.
- The need for user-centered design and multilingual, intuitive interfaces.

Generation and embedding of User Stories

Initial User Stories have been developed and categorized by stakeholder type and requirement source (literature, survey, workshops, etc.). The Stories were reviewed with experts in the consortium and then transformed into an initial low-fidelity prototype using Figma. The Stories were further validated in a follow-up workshop using the prototype with different stakeholders. In this workshop, we further explored key needs. Moreover, the workshop also served to raise awareness about future data handling and sharing practices, as well as the infrastructure. These stories provide an overview of the real-world needs and show how they can be aligned with the EU regulatory frameworks (e.g., GDPR, EHDS). Through this activity we also support PROTECT-CHILD's goal to deliver a compliant, adaptable, and effective digital platform.

Definition of functional and non-functional requirements through a workshop

Within the framework of the PROTECT-CHILD initiative, a participatory workshop was organized in Palermo to collect stakeholder perspectives on the platform's intended features and overall direction. The workshop opened with a demonstration of a mock-up version, offering participants an opportunity to explore its layout and envisioned functionalities. To further support discussion, we employed Wooclap for interactive polling, which enabled attendees to share their views on key priorities, needs, and potential challenges. This process allowed us to further specify the functional and non-functional requirements envisioned in the project.

1. Introduction

Advancements in health and genetic data analysis technologies, such as clinical biomarkers, imaging tools, and genetic approaches including gene testing for known variants, genome-wide association studies (GWAS), and next-generation sequencing (NGS) technologies, have the potential to enhance transplant outcomes for liver (Li 2021) and kidney diseases (Duan et al. 2022). Integrating genetic information with clinical, demographic, and other health data can contribute to better clinical decision-making, ultimately enhancing the quality of life for both patients and their donors (e.g., Caliskan et al. 2022). Yet, capitalizing on the opportunities presented by large-scale, heterogeneous health data remains contingent upon efficiently collecting, managing, and analyzing the data (Canakoglu et al. 2021; Fahr, Buchanan, and Wordsworth 2019). Achieving effective data management requires addressing several critical challenges (Alzu'bi, Zhou, and Watzlaf 2014). Among these challenges, a primary concern is the ability to handle diverse and large-scale datasets, which requires robust data pre-processing and standardization. Health and genomic datasets are inherently complex, making their analysis particularly challenging (Alzu'bi et al. 2014). Simplifying these analyses is essential to improve accessibility and usability (Alzu'bi and Zhou 2017; Reska et al. 2021). Additionally, as data is collected from diverse sources, ensuring interoperability, easy navigation, and effective analytical capabilities becomes crucial (Alzu'bi and Zhou 2017; Reska et al. 2021). Beyond technical challenges, managing health and genomic data also involves legal, social, and ethical considerations. Protecting patient privacy, complying with national and international regulations, and implementing access controls are essential components of responsible data management (Alzu'bi and Zhou 2017; Reska et al. 2021). In highly regulated environments such as the European Union, data security and privacy of health and genomic data must align with frameworks such as the General Data Protection Regulation (GDPR; European Parliament and European Council 2016) and the European Health Data Space (European Parliament and European Council 2022). Integrating these regulatory requirements into a platform design ensures secure and compliant data handling.

One promising approach to addressing these challenges is the development of a reference framework for integrating research and real-world clinical and genomic data on child transplants across multiple EU countries. This effort is supported by High-Performance Computing infrastructures and a federation of interoperable data and computing resources. Additionally, dedicated platforms for health and genomic data management play a crucial role in mitigating risks such as data misuse, loss, and theft. Moreover, these platforms can enhance transparency and ensure compliance with the FAIR (Findable, Accessible, Interoperable, and Reusable) principles (Forsdick et al. 2023). However, despite many institutions and research communities having established best practices for effective data management, translating these theoretical frameworks into practical implementations remains a challenge (Forsdick et al. 2023). Bridging this gap requires close collaboration among researchers, clinical experts, technical specialists, legal professionals, and other key stakeholders.

An important first step in this process is evaluating the literature, which provides insights into the functional and non-functional requirements of digital ecosystems for managing health and genomic data:

- **Functional requirements:** the platform's ability to provide functionalities that serve (stated or implied) the needs of the end-users. These functional requirements represent what the system can and must do (Glinz 2007; Rawashdeh and Matalkah 2006; Sommerville 2016).
- **Non-functional requirements:** the qualities of the platform that support the different functionalities to be utilized appropriately (e.g., accessibility, usability, etc.) and features that enable certain functionalities. These non-functional requirements also define the constraints of the software architecture. Non-functional requirements often apply to the system as a whole rather than individual system features or services (e.g., security, access control, availability, etc.) (Cleland-Huang et al. 2006; Glinz 2007; Sommerville 2016; Taraborelli 2003).

For instance, integrating healthcare databases may require non-functional requirements such as efficient data ingestion pipelines. Similarly, enabling non-technical users to analyze data may involve functional requirements such as drag-and-drop interfaces or prebuilt analysis templates, enhancing usability, accessibility, and platform adoption (Davis-Turak et al. 2017; Kho et al. 2013; McLeod et al. 2021; Pang et al. 2021).

Building on the literature review, we employed four additional methods to identify and refine requirements for a federated platform to support health and genomic clinical studies:

1. **Literature review** – Systematic definition of an initial list of requirements
2. **Survey with key stakeholders** – Revision and extension of requirements in a user-centered approach to define high-level requirements
3. **Focus group with key Stakeholders** – Co-design and co-exploration of requirements and expected challenges in usage and potential solutions.
4. **Generation and Embedding of User Stories** – Refinement of requirements to feed user stories, expert analysis of the stories, and embedding of stories in a low-level interactive Prototype on Figma (see http://tiny.cc/m6vj001) to enable visualization, interaction, and discussion with end-users.
5. **Refinement of functional and non-functional requirements through a workshop** – Revision of the low-level interactive prototype with data users.

This document provides an overview of the user needs and of the functionalities required for the Protect-Child ecosystem from two main perspectives, in line with the Protect-Child User journey:

- Data holder (data provider) perspective: to make high-quality data available, accessible, and understandable for use in secure data processing environments at each data holder's premises (the Protect-Child Capsules)
- Data user perspective: to allow the use of data available in federated secure data repositories (the Protect-Child Capsules), facilitating data search and data access by means of a unique data use permit for all relevant data holders.

2. Literature review: Initial Requirements

Process

As an initial step toward informing the development of federated infrastructure frameworks for a clinical and genomic data management platform, we conducted a PRISMA systematic literature review. This review aimed to identify both functional and non-functional requirements that practitioners, such as clinicians and data analysts, consider essential for supporting responsible data interoperability and analysis. We searched four major databases, including Scopus, PubMed, Web of Science, and Google Scholar, for peer-reviewed English-language publications from the last 10 years. Through thematic analysis, we extracted and ranked the most frequently mentioned requirements in the literature. From 544 initial records, 372 were screened, and 53 studies met the inclusion criteria for in-depth analysis. Our findings highlight three core categories of functional requirements: (1) general data management (including acquisition, standardization, and interoperability), (2) data processing and analysis workflows, and (3) data visualization and reporting tools. We also identified four key non-functional design considerations: (1) communication and user support, (2) the platform's technical infrastructure, (3) user interface and experience, and (4) security and compliance mechanisms. In addition to mapping these needs, the review sheds light on unresolved challenges that must be addressed to create a reliable and user-centered platform. These insights provided a base for engaging clinical, technical, and legal experts in discussions about priorities for data standardization, data exchange, and data analysis. The review was registered on OSF (https://shorturl.at/pAXDG).

Overview of key results

We first identified from the literature 69 key requirements (See Appendix 1 and 2 for an overview).

Functional requirements

General data management

In clinical and genomic research, general data management seems to be the first key step, involving key functionalities prior to the analysis and visualization of the data (Reska et al. 2021). These functionalities include:

- **Data acquisition:** Collecting patient data from various sources, such as electronic health records, lab results, open-source databases, and proprietary data from different hospitals (Amer-Yahia et al. 2022; Reska et al. 2021). Here, ensuring access and generating data interoperability is essential.
- **Data standardization:** Ensuring that the collected data is consistent and formatted uniformly, which is crucial for accurate interpretation, comparison, and interoperability (Canakoglu et al. 2021; McLeod et al. 2021; Ullah et al. 2022).
- **Data interoperability:** Allowing users to access open-source data while also permitting them to apply for access to proprietary datasets (Gilbert et al. 2022; McLeod et al. 2021; Ullah et al. 2022).

Data processing and analysis

The second step is data processing and analysis (Reska et al. 2021), including functionalities from data pre-processing and data analysis:

- **Data pre-processing:** Allows users to set standards and parameters to inspect the quality, consistency, and suitability of the data prior to subsequent computational tasks (Gill et al. 2023; Pavia et al. 2023; Reiff et al. 2022).
- **Aggregated data analysis:** Serves the users to apply statistical, computational, and machine learning techniques to process, interpret, and derive meaningful insights from the data (N. T. Li et al. 2022; Ochoa et al. 2021; Reiff et al. 2022).

Data visualization and reporting

The third step of genomic research involves exploring insights through data visualization and generating outputs for reporting. This step incorporates two key functionalities:

- **Aggregated data visualization:** Enables users to represent the aggregated data through graphical or interactive visualization to facilitate understanding, analysis, and the discovery of insights. After the processing of the data, complex information is transformed into visual formats like networks, scatter plots, genome browsers, heatmaps, pie charts, histograms, trend charts, and dashboards, making it easier to identify patterns, trends, and relationships within the data (Cerami et al. 2012; McLeod et al. 2021; Ochoa et al. 2021).
- **Reporting**: Allows users to generate files that support the sharing and publication of the data analysis, ensuring that findings can be communicated, also allowing users to cite findings and disseminate knowledge (L. Li et al. 2022; McLeod et al. 2021; Pavia et al. 2023; Yukselen et al. 2020).

Non-functional requirements

Communication and Support

Communication and support are a unique intersection of functional and non-functional requirements, encompassing both user interactions with the platform and the platform's autonomous functionalities. Some features include:

- **Feedback:** allowing users to provide feedback by leaving comments or rating their favorite datasets constitutes a functional requirement, directly enabling user interaction with the platform (Ware, Satyamoorthy, and Paul 2024).
- **Documentation:** which refers to various resources to help users understand, navigate, and use a platform effectively (Sauria et al. 2015). This includes tutorials, user guides, instructional manuals, and technical documentation.
- **Notifications**: refer to alerts or messages that inform users (such as participants, data managers, or researchers) about new or important updates related to data or research.
- **Communication with patients:** aiming to allow direct interaction with data providers (Gilbert et al. 2022; Suciu, Aydin, and Chen 2015).

UX/UI Features

The UI/UX category emphasizes characteristics that enhance platform usability and interactivity, especially for users without technical expertise (Pearce, Nikiforova, and Roy 2019; Raudvere et al. 2019). Key features that further contribute to a more intuitive and accessible user experience across diverse user groups and devices include:

- **Search functionality** to locate data using filters and keywords efficiently (Raudvere et al. 2019).
- **Interactive visual tools** like region highlighting (Campbell et al. 2021).
- **Zooming** (Suciu et al. 2015).
- **Hover effects** (Pearce et al. 2019).
- **Dropdown menus** (Pavia et al. 2023).
- **Drag-and-drop capabilities** (Raveendran et al. 2022).
- **Mobile responsiveness** (Duong et al. 2020).
- **Multilingual support** (Bhuvaneshwar et al. 2016).

Security and compliance

Security and compliance refer to the set of technical, procedural, and ethical measures implemented to ensure that data is protected from unauthorized access, misuse, or breaches, while also aligning with legal, regulatory, and ethical standards. Requirements regarding security include:

- **Privacy-protection measures: which** refers to measures designed to protect personal and sensitive data by removing or masking identifiers that could be used to trace the data back to individuals. Common methods include data anonymization, de-identification, and encryption (Kuzmenkov et al. 2021; L. Li et al. 2022)**.**
- **User registration:** to enable access to secured areas of the platform and allow the platform to track user activity, enforce permissions, and offer personalized services while maintaining accountability (Ochoa et al. 2021; Reiff et al. 2022).
- **User authentication:** often involves multi-factor authentication (e.g., passwords plus verification codes) and dynamic access control to ensure that only authorized users can interact with sensitive data (Ramirez et al. 2010; Sante et al. 2014; Verbruggen and Menschaert 2019).

Platform technical infrastructure

The platform architecture literature highlights three key technical priorities:

- **Federated infrastructure:** To support decentralized storage and secure, collaborative analysis by interoperable data (McLeod et al. 2021; Mohr et al. 2018; Wolf et al. 2015).
- **Scalability and adaptability:** Essential for managing growing data volumes and evolving user needs (Ochoa et al. 2023; Raveendran et al. 2022).
- **Integrating Application Programming Interfaces (APIs):** To enable data integration and system flexibility.

A summary of the connection between functional and non-functional requirements, as relevant from the literature, is summarized in Figures 1 and 2. An overview of the level

of agreement among the reviewed sources regarding the importance of these requirements is provided in Appendix 1, which lists the functional and non-functional requirements mentioned by the authors of each selected article.

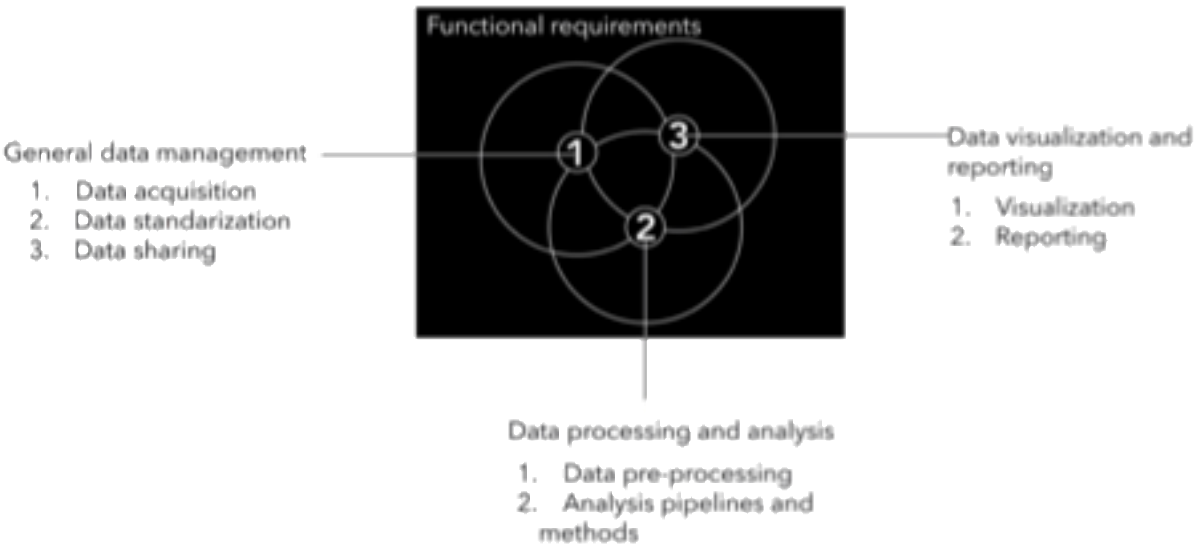

Figure 1. Functional requirements overview

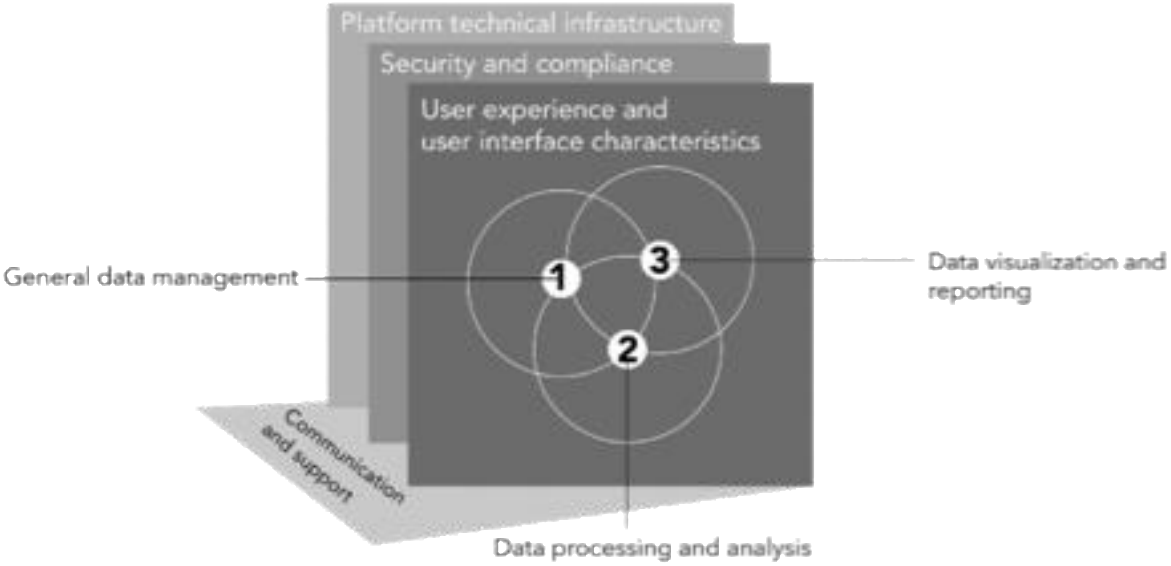

Figure 2. Overview of functional and non-functional requirements – Architecture of functionalities

3. Survey with key stakeholders

Process

We designed and conducted a survey study involving 52 experts across three primary areas of expertise: clinical (n=30), legal (n=8), and technical (n=14). Participants were recruited via email invitations sent to members of the PROTECT-CHILD consortium and other members related to the project.

Materials

The survey was developed based on the requirements list (see the requirements list above). The questionnaire consisted of a combination of Likert scale items and multiple-choice questions about the importance of the requirements from different stakeholders' perspectives.

Data analysis

The analysis followed several steps. First, we preprocessed the data by grouping responses based on the participants' domains of expertise. Outliers were removed. Mean and interquartile range for each requirement were used to establish importance and agreement (consensus) regarding the importance of the requirements. Agreement and importance were used to establish if agreement was achieved regarding the importance or lack of importance of the requirements, and if further investigation was needed.

A summary of the current results is presented in three tables in Appendix 3 for the different groups of stakeholders:

- **Table 1** concerns IT and technical teams (Biomedical engineers, Computer scientists, Cybersecurity experts, Database management)
- **Table 2** concerns Clinical Experts & Healthcare providers (Pediatricians, Hepatologists, Geneticists, and Lab scientists)
- **Table 3** concerns Data governance & compliance (Data standardization experts, Applied ethics, Ethics & AI, Law, and informatics (DPO office), Data managers)

Overview of key survey results

Technical experts

Several requirements emerged with consensus and importance, serving to prioritize the platform's development (Appendix 3). These included privacy-preserving techniques, horizontal data partitioning, transfer learning, and support for federated learning and analytics. Participants emphasized the need for reproducibility, standardized ontologies, and strong privacy guarantees. These requirements highlight a shared commitment to transparency and responsible innovation. Pre-determined governance models were also favored, suggesting that clear roles and responsibilities are seen as essential for effective collaboration.

Network architecture preferences leaned toward peer-to-peer topologies, which were both well-understood and considered highly relevant. In contrast, hierarchical, centralized, and hybrid models sparked more divided opinions, indicating a need for further testing or clarification. Similarly, while techniques like vertical partitioning, homomorphic encryption, and secure multi-party computation were recognized as important, the lack of agreement among participants suggests that these areas require additional discussion or evidence before they can be confidently integrated. Some technologies and approaches, such as gRPC, custom protocols, and open or hybrid governance models, received lower ratings in both importance and consensus. Likewise, several frameworks, including PySyft, custom-built solutions, and NVIDIA FLARE, were not widely endorsed, either due to limited familiarity or perceived lack of added value.

Overall, the feedback reveals a clear stakeholder preference for practical, privacy-focused, and interoperable solutions, grounded in consensus and aligned with the real-world demands of cross-institutional collaboration.

Clinical experts and healthcare providers

Clinical experts and healthcare providers highlighted a wide range of priorities for the PROTECT-CHILD platform, with particular emphasis on data acquisition, analysis, and communication. There was strong agreement on the need to support the acquisition of genomic data from multiple sources and to enable users to upload their own genomic data, reflecting a demand for flexibility and inclusiveness in data integration. Several interoperability standards, such as the Genomic Data Model, OMOP, FHIR, and VCF formats, were also rated as important and achieved sufficient consensus, making them suitable for adoption in the platform. These preferences underline the value of standardization in facilitating data exchange across institutions.

Data quality and reliability emerged as top priorities, with respondents strongly supporting the inclusion of automated checks, quality inspections, and reproducibility safeguards before any analysis takes place. In terms of analytical functionality, experts called for support for predictive modelling, statistical analysis, exploratory data analysis, and AI-based approaches, demonstrating a clear interest in advanced yet reliable computational tools. The ability to share knowledge, not only with clinicians and researchers, but also with patients and ethical boards, was also relevant, emphasizing the importance of communicating the findings and enhancing collaboration in rare disease research.

Visualization tools also scored highly, particularly charts, graphs, heat maps, and sequencing visualizations, which are key to interpreting and communicating complex genomic data. Moreover, participants stressed the need for data export functionalities, federated computing access, and privacy guarantees, ensuring that the system balances utility with security and ethical considerations.

Several items, such as mobile and multilingual platform support or sharing information with policymakers, were considered important but did not reach consensus. This suggests that there is room for further dialogue or phased implementation. Similarly, while some data models (e.g., PATRIC, DataSHaPER) and file formats (e.g., IDAT) were familiar to a few participants, they did not garner broad support and may require additional validation or training.

Overall, the results suggest that clinicians are looking for a platform that is secure, serves to easily navigate metadata, and can support reliable aggregated data analysis. The final goal of clinicians would be to foster collaboration across the healthcare ecosystem and share the findings to advance healthcare.

Data governance and compliance experts

Data governance and compliance experts placed strong emphasis on the importance of clear, standardized frameworks for federated computing across Europe. There was broad agreement on the need to develop EU-wide best practices to ensure compliance with data sovereignty laws, particularly in support of GDPR principles such as data minimization and purpose limitation. Several respondents also highlighted the importance of conducting risk assessments before allowing data access and the critical need to retain data on-premises to align with regulatory requirements. Additionally, experts saw value in using federated computing to enable efficient data interoperability for rare diseases and to support cross-border health research that improves patient care across the EU.

While most requirements were rated as important, not all reached consensus. Concepts such as harmonizing governance across jurisdictions, ensuring transparency, and aligning with privacy-by-design principles were widely supported in theory but had more varied responses, suggesting that implementation of pathways remained unclear. Similarly, while many experts recognized the importance of mitigating risks, such as data breaches, lack of auditability, and accountability challenges, the perceived urgency and approach differed among participants.

In sum, the results reflect a shared understanding of the regulatory and ethical stakes involved in federated computing. Experts advocate for standardized EU guidance and safeguards, especially when it comes to data protection, privacy, and cross-border collaboration, but acknowledge the need for continued discussion to reach alignment on how best to implement them.

A workshop with stakeholders (Section 4) and future activities of co-design serve to further clarify the requirements that should be implemented and in what form.

4. Stakeholder workshop: focus group

Process

To extend and complement the previous activities, a focus group was conducted to explore the challenges and opportunities associated with the requirements identified in the literature as important to develop platforms such as the one we are targeting for PROTECT-CHILD. The session brought together 36 experts from the PROTECT-CHILD consortium ranging from 28 to 70 years. These experts represent the previously mapped expertise: clinical, technical, and legal. To stimulate interdisciplinary discussion, participants were mixed by expertise so that each group included at least one technical expert, one clinician, and one legal expert. The focus group was held in English and aimed to encourage a structured and open environment for dialogue.

The focus group was organized around two core activities designed to elicit stakeholder perspectives about the requirements. In the first activity, participants discussed how they envisioned using the platform in practice and identified the technological enablers required to meet clinical objectives. The second activity focused on the practical challenges of working with health data platforms, encouraging participants to suggest some solutions to tackle those challenges. Each group was provided with guiding questions developed from prior insights, along with a summary of potential platform requirements and common issues identified in the literature (see Appendices 1 and 2), to support and inform their discussion. For instance, we asked to reflect on possible technical enablers to address the identified challenges. The timing for each of the activities was around 30 minutes. After each activity participants were encouraged to summarize their conversation in 5 minutes.

A thematic analysis of the recorded conversations was conducted by a team of three experts, who discussed and categorized the quotes until consensus was reached and Krippendorff's alpha exceeded 0.6 (see Appendix 4 for an overview of the discussed codes). The resulting themes were subsequently grouped into four overarching topics, informed by both k-means clustering and expert evaluation. Insights derived from the thematic analysis are discussed below. Building on the requirements identified in the literature, the survey, insights from the workshop, and previous projects, a visual version of the platform requirements was subsequently developed (see http://tiny.cc/m6vj001).

Overview of results

Vision and value of the platform

Participants envisioned a platform that integrates as much data as possible in a single application. In short, a platform that is robust and scalable. This vision was mostly represented by a geneticist and a technical expert on the focus group. The geneticist explained:

"So, um, my vision is to have a tool or, uh, application or platform that could be able to integrate in one single file. If you think in a file, for example, the clinical information, the genomic information and all the variables, the health variables, important health variables in all the life of one single person. So, in other words, um, you can sum up, or you can summarize in one single file."

The technical expert asked for clarification:

"All the health history or the health?"

To which the geneticist responded:

"Yes, the history of one single person. So probably the transplantation is the excuse or it's the use case, but could be perfectly a good example of, uh, doing the same with any other disorder or any other medical problem in any part of Europe or the world."

This dialogue illustrates the desire for a holistic platform capable of integrating diverse health-related data steams into a unified format. A pediatrician expanded on this idea, stressing the transition from research to clinical application:

"*Maybe now we are using the tool for research, but in the future, we want to use it for clinical practice. Yeah. So maybe in the future the advantage will be to store a huge amount of data that now is impossible to store long time and to collect but also share if the patient moves around now. Now, at present, you cannot share that patient's data between two words in the same hospital or just in the same hospital. So, it's the fact that you can move around Europe and the data will be there and then of course, if the data are shareable, you can have access to a care that is not national but cross-national.*"

This vision comes alongside the need of having a lot of data integrated in one place. One pediatrician shares the perspective sharing:

"*Our clinical point of view, it could be useful to have a lot of data or a lot of patients and to use this data clinical and ethical and to do in future a lot of studies to look for possible correlations between clinical data, clinical and genomic*".

Ideally, the ultimate goal of the platform would be to go beyond the clinical context and integrate different sources of information. Importantly a laboratory scientist from one of the centers shared the vision of traceability within the platform:

"This is about who has access. I mean not all the doctors a certain type of data and you have to have a traceability on that. You have to be able to, to trace that to normally to everything."

Aligned with this vision, but pointing in a slightly different direction a pediatrician suggested that he wants to be able to access as much data as possible, at least from his center. He stated:

"If we interrogate the overall set of capsules. We could obtain only aggregate results. [..], 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 in both making the queries and obtaining the data".

In addition to data integration, participants strongly emphasized the need for open-source and flexible solutions. A technical expert explained:

> *"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."*

Similarly, a laboratory scientist from one of the centers highlighted the importance of an open-source application to ensure adaptability and autonomy:

> *"For me it's important that the software is open source [...] I'm not against the companies to, to, to have proprietary software, but this software has to be open for, uh, for teaching, for, uh, in implementing, uh, tailoring to different, uh, because if not, you are trapped in a, in a, in subcontracting with a company and then you cannot get out because."*

Such perspectives show stakeholders' concerns regarding vendor lock-in and the need for systems that can be adapted to local contexts without external dependency.

Scalability and automation also emerged as a critical theme in participants' vision of the platform. *A pediatrician echoed this perspective sharing:*

"The most important thing when we consider is that we need automation. We need to ease the way the data is, gather, uploaded, etc."

While a technical expert envisioned:

> *"A tool for that that defines the data model that is scalable, that understands different languages, different contexts, probably even in the same language or same country."*

Another expert emphasized the role of scalability in ensuring quality:

> *"It's just a matter of scalability with the algorithm and quality of it."*

These perspectives reveal a strong consensus among stakeholders: the envisioned platform must not only integrate diverse data types but also guarantee traceability, remain open-source and adaptable, support automation, and scale across contexts and geographies to meet the evolving demands of both research and clinical care. Further we moved to challenges related to technical and legal challenges. In the next section we shared the perspectives regarding legal, ethical and regulatory needs.

Legal, ethical, and regulatory needs

To make the vision of the platform actionable, there are challenges that must be address. Managing patient data across different institutions is particularly complex because each organization has its own processes and standards. As one legal expert explained, the problem lies not only in the fragmentation of procedures but also in their overwhelming and impractical presentation:

"*a two-fold problem that either all of the procedures and requirements and everything that you need to do is either scattered in multiple documents or in one huge document, and*

either case, no one reads it practically. And of course, uh, the fact that all of these procedures take a lot of time, uh, that could be used in other ways".

This illustrates how bureaucratic hurdles, while designed to ensure compliance can slow down research and create barriers to collaboration. Clinical experts also echoed these concerns, pointing out the lack of communication between clinical, technical, and legal stakeholders. Without a common language or shared understanding, progress becomes fragmented and inefficient. One laboratory scientist from one of the centers noted:

"*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. and legal particularly need to be in the same page, to talk together to, to understand that because the current situation doesn't work, because it's too complex.*"

The complexity of current systems has also led to workarounds that undermine compliance, reflecting a misalignment between regulations and real-world practices. The conversation with the expert continued by expressing:

"*When we are talking too complex is that usually clinicians know when they have to, to share data, anything that is more complex that simply going to WhatsApp to call a friend and send in the images and a PDF etc.*"

This remark highlights the tension between the urgency of clinical needs and the weight of regulatory processes. If rules become too cumbersome, they risk being bypassed altogether, defeating their intended purpose. Underlying these frustrations is a broader debate about how privacy should serve patients. While experts agree on its importance, many feel that current interpretations prioritize abstract legal categories over patient outcomes. One hepatologist expressed this unease by stating:

"*Surely privacy is an important field, but I don't understand why, uh, a mutation is most important than, uh, to have, uh, an infectious disease or, uh, and, uh, and a tumor. This seems a very odd legal point*".

In this view, privacy should not be a barrier to care or research but rather a tool that protects patients while enabling better health outcomes. Particularly, when it is about are diseases or mutations. As another pediatrician put it, customizable access permissions are essential because:

"*Privacy is an advantage but shouldn't be a limit for care.*"

Together, these reflections reveal a critical tension: the balance between protecting data and enabling its use for patient benefit and scientific progress. Addressing these challenges requires not only technical solutions but also institutional collaboration, education to bridge knowledge gaps, regulatory clarity, and a shift toward frameworks that prioritize patient outcomes without undermining trust.

Data lifecycle needsIn addition to these legal and regulatory complexities, the management and usage of data itself present structural challenges. A central issue identified across discussions was the lack of standardization across different stages of the research process, from data collection and management to storage and analysis. A technical expert illustrated this by pointing to the diversity of incompatible systems in use:

"I think the problem is that we are all the hospitals have different software."

A pediatrician further detailed the consequences of this fragmentation:

"So, you start with a software for the medical record. Then you add a software for the laboratory analysis and other software for the imaging analysis belongs to different companies. So, they don't talk."

These reflections show the complexities (e.g., technical and procedural) to effective data management. Complex regulations and miscommunication among stakeholders can limit efficiency, while the lack of interoperable systems prevents data from being standardized and connected across settings.

Another challenge connected to the differences in storage is the lack of time that makes clinical experts express the data in different formats. A laboratory scientist from one of the centers points out:

"So, for us it's important to understand the need of time. We understand the standardization, the normalization, the using ontologies, etc. but a clinician has very limited time and having a good standardization means having a form, having a form that you have to take a lot of things, a lot of screens. [...] This is why the clinicians they go to the open text box, and they put some strange abbreviations to summarize that."

This shows how standardization, although valuable for data quality and interoperability, risks creating additional burdens that may compromise clinical practice if not carefully implemented.

Data analysis presents another layer of complexity. Experts emphasized the need for tools capable of managing both high complexity and low volume of data. As one technical expert reflected:

"This is huge data. Right? So, they were discussing, like, a quarter of a terabyte, like in informatic terms. That's a lot of data, like huge. Spacing is also a problem. [...] The machines that do the analysis are, like, really, really expensive."

Clinicians echoed this concern, linking storage challenges with the risks of data fragmentation:

"We need a storage capacity, and we need to allow the data sharing or access, because currently when you have huge amounts of data from one patient, it's very easy to miss

because there are different platforms in different places, you mis-allocate, etc. So, we expect that this new platform allows this consistency, and we don't lose anything."

The challenge is further amplified when dealing with genomic data, which not only involves large volumes but also requires highly specialized approaches. A technical expert illustrated the scale of the problem:

"Another issue is, uh, the size of data. I mean, we were hearing [geneticist] speaking about a quarter of terabyte capacity. So, if a researcher would like to manage 200% or 400%, this is 500TB."

A geneticist added that size alone is not the only difficulty, but the analysis as well:

"Genomic data is complex; therefore, it requires specific tools and methods to extract meaningful insights."

Overall, these perspectives highlight that challenges in data management are interconnected. Fragmented systems, the difficulty of standardizing processes without overburdening clinicians, the need for reliable large-scale storage, and the complexities of genomic analysis require solutions that combine technological advances with practical workflow considerations. Following these discussions, experts offered insights into potential solutions, which are detailed in the sections below.

Technical enablers and infrastructure

A central theme emerging from the discussions was the need for technical and procedural enablers to manage, analyze, and integrate complex health and genomic data. Experts emphasized that effective data management requires a combination of automation, standardization, interoperability, and advanced analytics, while also considering usability, legal compliance, and clinical workflows. Automation emerged as a key enabler to reduce manual burden and increase efficiency in data management and analysis. A geneticist noted that many current procedures are still performed manually:

"And so, this is, uh, this is something that is already done in a human way. Not automated or anything."

Technical experts highlighted that automating analyses can save researchers significant time and streamline workflows:

"The automation of an analysis brought in by the project is very important... to contribute to saving the time of the research."

This automation can occur throughout different stages of the process. For instance, in the overall architecture, it can serve to ensure privacy. A technical expert shared:

"You can do anything on the system because once the system has been started, there is no way to access it. You just have to trust at the moment you start the system. This capability reduces reliance on the provider."

Beyond the overall security, automation can also be applied for the legal and regulatory challenges. This point is closely linked to the standardization of procedures. Experts highlighted that more automated and harmonized processes could reduce administrative burden while ensuring compliance. A legal expert suggested automating ethical approvals across multiple centers:

"All of these ten applications, identical approval forms, data protection impact assessments or whatever, but you combine it in one and then every center just looks at this one."

Standardized procedures would allow researchers to avoid repeating identical applications across different hospitals:

"Instead of having to fill in all of these things separately [...] you combine it in one, and then every center just looks at this one. This could speed up the process while ensuring privacy and security."

Automation also plays a role in making the analysis more efficient. One technical expert described using natural language processing (NLP) to simplify and accelerate the identification of patient cohorts:

"We are still developing this and creating a cohort for the patients. It's really difficult. [...] One is based on natural language. So, suppose I want a cohort of female patients with this type of sarcoma. From the natural text, the prompt we create tries to obtain all the data that matches that criteria."

Building a robust platform includes standardization, which a clinical expert emphasized requires several steps, including normalization and the use of ontologies.

"*We understand the standardization, the normalization, the using ontologies, etc."*

He further emphasized the importance of scalability through the standardization of the data:

"A tool that defines the data model that is scalable, that understands different languages, different contexts… because at the end, the setting of data for the clinician mostly is a summary."

In this point, discussing the standardization and interoperability of data become crucial for advancing the tool. A technical expert emphasized that data collected from different sources must be able to *"talk to each other*", otherwise the system fails. He also pointed to the need for integrating multimodal data, combining imaging, clinical, and genomic datasets, to unlock richer insights. In this context, standardization not only supports integration but also enables advanced analytics. Technical experts highlighted the potential of algorithms, deep learning, and large language models to detect hidden patterns and generate new knowledge:

"I mean, you have 500TB of data which means you can use algorithms to find some patterns."

Another technical expert added:

"I would like to use this data for new algorithms to solve the same problem without maybe extracting features from other algorithms, using new techniques."

Beyond analytics, quantum computing was mentioned to strengthen data security while supporting multi-center studies with greater statistical power. Scalability and robustness were also emphasized as key requirements, with one expert pointing to the importance of interoperability with external systems:

"Having an API or ways to communicate with the system."

Finally, participants stressed that technology alone is not enough. As one geneticist noted, a crucial enabler lies in education:

"One of the main methods to overcome the complexity of such data is to educate clinicians and other pertinent people."

This highlights the importance of building not only technical capacity but also user understanding to ensure meaningful adoption of the platform.

User experience, communication, and resources

Usability emerged as both a challenge and a key requirement across interviews. Participants repeatedly stressed the lack of intuitive and efficient user interfaces, which currently prevent clinicians from interacting with complex systems without relying on technical intermediaries. As one pediatrician put it,

"Friendly is quite important"

A technical expert echoed this perspective, emphasizing the importance of design:

"The visualization, the presentation... usability. Yeah, so we really need that."

Another expert added that clinicians should: "*not [have] to go into codes*", suggesting that the interface should be intuitive and self-explaining for clinicians to understand what happened.

While participants found it difficult to fully articulate what a user-friendly interface should look like, many agreed that visualization is central to usability. One technical expert explained:

"To make [results] interpretable [and] understandable by users and in terms of dashboard presentation, have a visualization so that the users can validate. And so that's our description."

At the same time, visualization preferences were shown to vary. A pediatrician stressed that the "*best*" format depends on the use case:

"If we speak about results, tables are the best way... If you have a specific clinical question, maybe a graphical or descriptive result is better. But this is just something very personal."

These reflections highlight the importance of customizable and adaptive visual tools that can accommodate both research- and clinic-oriented needs. Moreover, these requirements reveal both shared needs, such as simplicity, autonomy, and effective visualization, but also nuanced differences in how usability is perceived.

Importantly, user requirements should not be treated as static. Instead, they had evolved iteratively through the literature review, survey, focus group and workshop, and were continuously refined in collaboration with the management team. Insights from earlier projects, along with ongoing consultations across the consortium, further enriched this process.

After the reflection of these comments, we continue structuring and the process, and we encourage a third activity with participants. In the following section, we present the resulting high-level user requirements from the discussion. However, these requirements will continue to be refined as we advance the platform's development and engagement with users and domain experts continues.

5. Workshop for refinement of requirements

As part of the PROTECT-CHILD project, we conducted a workshop during a consortium meeting to gather feedback from stakeholders on the platform's functionality. This workshop built directly on the outcomes of the previous focus group phase, in which participants discussed and articulated areas of agreement and disagreement regarding the initial set of requirements. The outputs of the first workshop served as the basis for this second session, providing a structured overview of requirements that required further elaboration, for example, the identification of technical enablers for processing and using the data.

To address these points, we organized a single plenary session with 27 participants (aged 23–70). The session began with a brief overview of the functional requirements identified in earlier activities and supported by previous projects (Figure 3), followed by a walk-through of a mock-up version of the platform in which the identified requirements were embedded (see http://tiny.cc/m6vj001). This mock-up was used to map potential technical enablers and illustrate how these requirements could be operationalized, helping participants to visualize the platform's structure and potential applications.

Following the demonstration, we used interactive polling via Wooclap to collect input on priorities, needs, and concerns regarding previously identified requirements which still needed to be agreed or clarified. The session was also recorded to capture participants' verbal feedback and spontaneous remarks. This approach enabled active, real-time engagement and provided both quantitative and qualitative insights to guide the ongoing refinement of the platform's design and functionality.

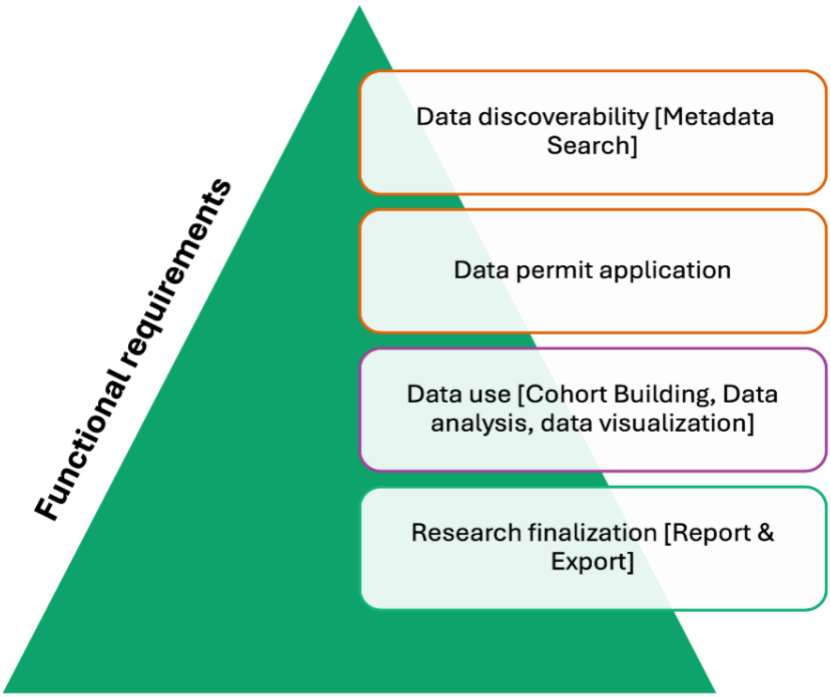

Figure 3. Overview of functional requirements for Protect-Child

High-level functional requirements

For each component, the requirements (and associated user stories) that emerged in previous phases were embedded in a visual, interactive prototype. Any eventual issues or concerns regarding the functionalities presented by each component were captured and discussed with the participants.

Data discoverability [Metadata search]

From the feedback gathered from Wooclap we noticed a strong majority (88%) agreed with the way data discoverability was proposed. A step-by-step approach (i.e., a structured, numeric way), where users explore the data variables from each center. The most valued features for metadata navigation included filtering options, a step-by-step interaction structure, the ability to select multiple transplant types, and the option to download CSV files.

One comment from such discussion was to include a map to display participating hospitals across the network. Some participants raised concerns about accessibility and the complexity of navigating large volumes of variables. One participant recommended exploring Athena.ohdsi.org as a model for effective metadata presentation. Additionally, one clinical expert voiced a preference for logging in before exploring metadata, likely reflecting cultural differences in privacy expectations.

Data permit application

Following data discovery, the proposed process allows users to log in and apply for data access. All participants (100%) in this process expressed support for both displaying data permit statuses and integrating a cohort-building tool driven by natural language processing (NLP).

Data use

Once participants have been granted data access, they can begin creating variables and conducting analyses. During this process, technical experts suggested providing synthetic data for data managers. Clinical experts also expressed a preference for accessing individual-level data for qualitative patient evaluations and the option to download aggregated data. The visualizations presented after the analysis were well-received, with 88% agreement on the proposed setup.

Research finalization

Once users complete the analysis, they have an option to generate an automatic report based on the workspace summary and all the data provided in the previous process. In this step, users agreed with the generation of an automated report they can download from the interface.

Additional discussion on non-functional requirements

Data quality

During the workshop, some experts also voice their preferences and priorities. Participants identified key priorities for metadata exploration and data quality assessment. Regarding data quality, participants prioritized indicators such as completeness (e.g., missing data), consistency across records, reliability of data sources, and harmonization with standards like OMOP and FHIR. Notably, 8 out of 9 participants agreed that it would be helpful to display how data quality is assessed. It was also suggested that the current data quality indicators could support evaluating completeness and reliability during and after cohort creation.

Data governance

Current data governance plans for the project were discussed. The discussion concluded in each hospital managing data access through internal governance structures for now. Participants highlighted the need for technical flexibility to align with evolving regulatory implementations across various sites and countries. Particularly concerning different governance models such as the implementation of a Health Data Access Body.

Progress and Challenges in Federated Infrastructure such as Protect-Child

The workshop highlighted several partially satisfied requirements where progress has been made, but further refinement is still needed. While stakeholders recognized advances in interoperability and automation, they noted that many functionalities remain only partially implemented, requiring continued investment and development. Experts further emphasized that it is important to understand the specific limitations of a federated approac**h**. For example, certain statistical methods (e.g., Monte Carlo simulations) are difficult to implement when data is not available at the individual level. Genomics experts expressed a preference for centralized data solutions, citing efficiency and analytical depth, even though EU regulatory frameworks strongly favor federated, privacy-preserving models. These tensions illustrate the trade-offs that must be negotiated between scientific goals, regulatory compliance, and practical feasibility.

Despite these challenges, insights from the literature review, survey, focus groups, and workshop converge on a clear set of requirements for a federated, privacy-preserving approach to pediatric transplant research. This includes the acquisition of data from multiple sources, standardization and translation between data models (e.g., FHIR and OMOP, VCF files), systematic quality checks, and reproducible analysis pipelines. Together, these elements create a basis for easy, reliable, and reproducible pipelines of advanced analytics. Across groups, there was high agreement on the need for automation (during diverse steps of the process such as ingestion, quality control, and reporting), strong interoperability via APIs, and usable, low-code interfaces that let clinicians explore and validate results without “going into code.” Legal experts highlighted the importance of GDPR-compliant safeguards, on-

premises control, and clear data accountability. Clinicians stressed complementary requirements: actionable clinical insights, traceable data access, and the ability to view information both at the row level within their own Centers of Expertise (CoEs) and in aggregated form across the federation.

At the same time, structural challenges in hospital systems add complexity. Fragmented hospital systems and inconsistent approval processes generate sociotechnical challenges, leading to workarounds that jeopardize legal compliance. Legal requirements are often disparate or difficult to manage in practice. Under time pressure, clinicians often revert to free-text entries and inconsistent formats, which further complicate standardization.

From a technical perspective, the envisioned platform must be capable of handling very large, complex genomic datasets while also integrating multimodal data (clinical, imaging, and genomics) at scale. Equally important, usability was suggested as a key factor for adoption. Stakeholders repeatedly called for “friendly” tools, guided metadata discovery, NLP-assisted cohort building, customizable visualizations, and automated reporting, supported by training and education to ensure that diverse user groups can engage with the system.

Overall, these findings show the potential and challenges of developing a federated infrastructure for pediatric transplant research. Stakeholders prioritized robust data stewardship, automation, interoperability, usability, privacy safeguards, and accountability. However, these goals must be balanced against the practical constraints of federated infrastructures, such as decreased computational efficiency, persistent site-to-site heterogeneity, and fragmented legal and ethical frameworks. Therefore, a federated approach should be seen as an evolving socio-technical system requiring continuous alignment between technological infrastructure, regulatory frameworks, and the realities of clinicians and researchers, rather than a definitive endpoint.

6. User stories

Process

Building on the requirements from literature, survey, focus group, and workshop, as well as utilizing insights and documentation from previous European projects (e.g., *IDEA4RC*), we defined a user journey (Figure 4) and initial user stories. These sources provided a basis for identifying user needs, expectations, current and expected process of usage and expected issues. By translating insights into actionable user stories, we were able to define an initial architecture of functionalities and their (expected) relationships.

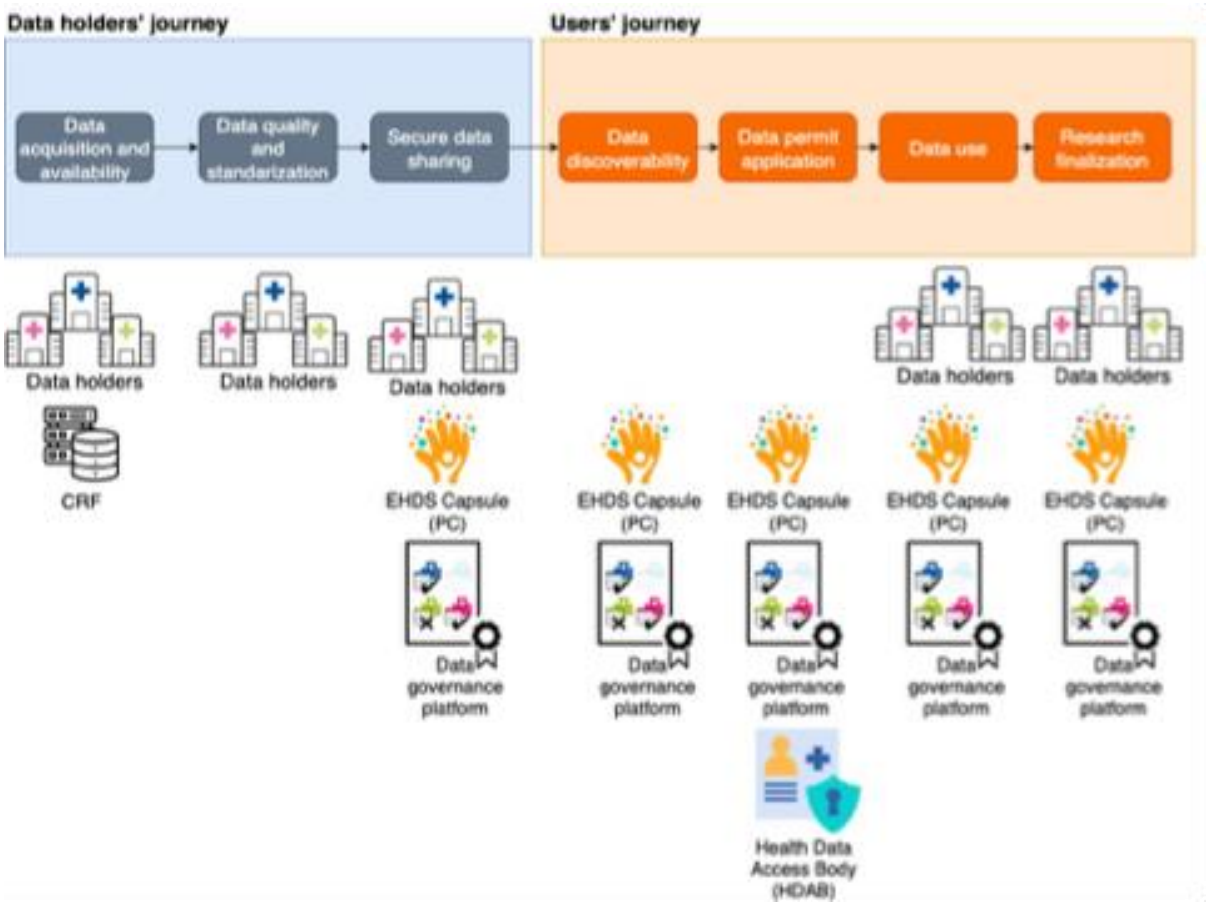

Figure 4. Overview of use journey and infrastructure for Protect-Child

Each user story was structured using the standardized format in Table 4. User stories highlighted in blue represent needs and requirements that were confirmed or validated by co-design activities, while those in orange are stories that require further validation by the technical team to determine their feasibility and the alignment with the project scope.

Table 2. Template of User Story

US #x. y (Responsible Partners) – Descriptive title
As a <*actor addressed by the User Story*>
I want to <a *need expressed by the actor*>
So that <*the reason for having the above need addressed, i.e., the value expected from its satisfaction*>
Initial acceptance criteria ● <*conditions that must be met to consider the User Story correctly implemented*>
Additional notes ● <*information that, although not part of the User Story proper, has been nevertheless elicited during its formulation, and can be useful to report, for the benefit of further project activities*>

In the following sections, we briefly explain each of the steps conducted for the definition of these user stories. Later, we present the general user stories, followed by the user stories identified for the main target group (e.g., Clinical experts and healthcare providers). Moreover, we complemented with source of information the user story **L** for Logic, **LR** for Literature Review, **P** for Previous work (e.g., IDEA4RC, other EU projects), **FG** for Focus group, **W** for Workshop, **E** for Expert insights (formal and informal), **S** for Survey, **R** for Regulations (e.g., EHDS). A glossary of the terms used in the user stories are further detailed in Appendix 5.

Users

In our analysis, we identified three key stakeholder groups involved in the use, management, and governance of health and genomic data: data holders, data managers, and data users (e.g., pediatricians, geneticists, researchers). A more detailed overview of all potential users that could be involved is shown in Figure 5 and further address in Appendix 6. Each group plays a distinct role in ensuring that data is collected, processed, and used responsibly and effectively. The following sections describe these user types in more detail, including their responsibilities, challenges, and relevance within the broader health data ecosystem.

Data holders

Data holders are institutions or individuals that generate, collect, or store health-related data. They act as custodians of data resources and are responsible for granting access to authorized users under defined legal and ethical frameworks. According to Staunton et al. (2024) a data holder is *"any natural or legal person, entity, or body in the health or care sector, or performing research in relation to these sectors."* Under the European Health Data Space (EHDS) framework, when electronic health records are access from a data holder within a single Member State, applicants can apply directly to that data holder under Article 49 (European Parliament and European Council 2022). Data holders thus play a central role in maintaining data quality, ensuring compliance with data protection regulations, and managing access requests. They are also accountable for safeguarding sensitive patient information and ensuring that any data interoperability aligns with legal and ethical requirements.

Data managers

Data managers are responsible for ensuring that the data meets the quality standards necessary for reliable research and clinical use. Jetten et al. (2021) describe data managers as professionals who handle the management of data objects and metadata, providing the technical and administrative infrastructure required to gather and curate high-quality data. They coordinate the development of data collection plans, assign responsibilities, and manage the creation of data collection instruments. Serving as a bridge between technical data systems and the scientific or clinical teams that depend on them, data managers ensure that information flows efficiently and accurately across the research process. As Agrawal et al. (2023) note, an appointed data manager often oversees the transfer, verification, validation, and repackaging of data, ensuring it is securely stored in designated archives. Their expertise spans areas such as data preservation, information science, archiving, and regulatory compliance (Jetten et al. 2021).

Data users

Data users are individuals or organizations that request and utilize data for secondary purposes such as research, innovation, or policy development. Under the EHDS framework, any natural or legal person can apply to access electronic health data, provided they meet ethical and regulatory requirements (European Parliament and European Council 2022). Data users are expected to evaluate the ethical implications of their work, especially when operating across jurisdictions where different ethical standards may apply (Staunton et al. 2024). Data users must also ensure that their research outcomes are shared with the Health Data Access Body (HDAB), promoting transparency and the responsible reuse of data. This category includes a diverse range of professionals, such as clinical experts and healthcare providers with different expertise (e.g., pediatrician, hepatologist, lab scientists), geneticists, bioinformatics, researchers, among others. See figure 5 for an overview of potential data users of PROTECT-CHILD.

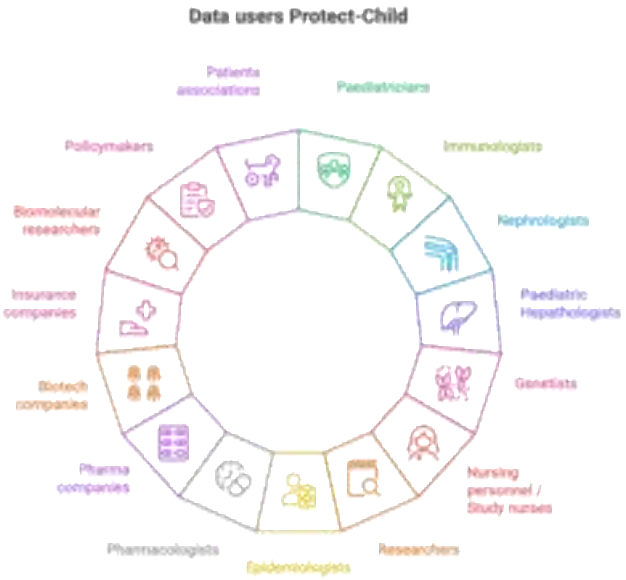

Figure 5. Overview of potential data users for Protect-Child

General user stories

Data holder perspective

Table 3. Registration

0. Partners involved: SERMAS-LA PAZ, UNIPD, UKE, ISMETT, BIOMERIS **Expertise:** Data holders (Managers of the platform) Registration
As a Protect child data holder
I want to enable the registration of user profiles, who are authorized to access to data, within my organization
So that only verified individuals from our institution can securely access the Protect Child platform.
Initial acceptance criteria [P, L]: • [P]: A registration form is available via a secure web page. • [L]: Users must provide required information (e.g., name, email, password, institution, role). • [L]: Passwords must meet security requirements (e.g., minimum length, special characters). • [P, L]: Users receive access credentials and verification email to confirm registration. • [P]: User data is stored in the appropriate directory agreed with other partners. • [L]: Successful verification grants access to the platform.
Additional notes [L, E]: • [L, E]: The platform should comply with GDPR and data security standards. • [L, E]: Multi-factor authentication (MFA) may be required to ensure security. • [L]: Users should be able to update their profile information after logging in. • [L]: Administrators should have access to manage user accounts if necessary. • [E, P]: The registration and login flow must be implemented using secure OAuth2-based authentication, managed by CERTH, to ensure federated and scalable identity verification across partner institutions. • [E]: SPEs must apply security-by-design and security-by-default principles in line with ENISA and ISO 27001/27701 standards, including encryption, access control, and key management. • [E]: All users and administrators accessing health data shall authenticate using MFA and recognized federated identity frameworks (eIDAS, GA4GH Passports). Each identity must be distinct, non-transferable, and consistently verified while active.
Information that supports the user story [P, E]: • [P]: Previous projects such as IDEA4RC project. • [E]: Meeting with coordination team and CERTH. • [E]: Requirements mapped by T2.4 and defined by T2.2.

Table 4. Simple file-based deployment (Excel, CSV, TSV)

1. Partners involved: SERMAS-LA PAZ, UNIPD, UKE, ISMETT, BIOMERIS **Expertise:** Data holders (Managers of the platform) Simple file-based deployment (Excel, CSV, TSV)
As a Protect child data holder
I want to install and set up the data capsule
So that I can enable and manage access to the Protect Child platform and data sources.
Initial acceptance criteria [L, E]: • [E]: The hospital IT team can place validated data (e.g., in .csv, .xlsx, or .tsv format) into a designated folder on the system. • [E]: A predefined script, provided with the capsule, is executed to inject data from the folder into the platform. • [E]: The script logs success or failure, with detailed output for debugging if needed.

• [E]: The capsule automatically processes the data and builds the core project common data model database, generates an OMOP-based database for secondary use and analytics. • [L]: The script includes logs for errors, successful ingestion, and file versioning.
Additional notes:
Information that supports the user story [E]: • [E]: Conversations with technical experts • [P]: Documentation made for other project tasks. • [E]: Interview with BIOMERIS.

Table 5. Semantic-standard-driven deployment (FHIR, OMOP)

2. Partners involved: SERMAS-LA PAZ, UNIPD, UKE, ISMETT **Expertise:** Data holders (Managers of the platform) Semantic-standard-driven deployment (FHIR, OMOP)
As a Protect child data holder
I want to reuse my existing data available in the hospital (CoE),
So that I can contribute to the capsule without redundant reformatting.
Initial acceptance criteria [L, E]: • [E]: The center verifies whether the FHIR or OMOP data matches the implementation guide defined by Protect-Child. • [E]: If not, the center is responsible for adapting its data to meet the expected FHIR IG. • [E]: Once compatibility is confirmed, the FHIR is converted to OMOP and is injected into the capsule without converting from CSV. • [L, E]: Once compatibility of the OMOP data is also confirmed, it can also be injected in the capsule. • [L]: The capsule builds on OMOP data model databases to ensure compatibility across CoEs.
Additional notes:
Information that supports the user story [E]: • [E]: Conversations with technical experts. • [P]: Documentation made for other project tasks.

Table 6. Capsule versioning

3. Partners involved: SERMAS-LA PAZ, UNIPD, UKE, ISMETT, UPM **Expertise:** Data holders (Managers of the platform) Capsule versioning
As a Protect child data holder
I want to create the latest version of the capsule whenever new data needs to be added,
So that each capsule reflects a specific, immutable dataset release, and I can track changes over time.
Initial acceptance criteria [L, E, P]: • [L]: Once a capsule is built and deployed, it is immutable, no further data can be added or modified within it. • [E]: To add new data, the old capsule must be deleted, and a new capsule must be created with the updated dataset. • [L]: The new capsule is assigned a version number or release ID following the agreed convention (e.g., Release_1, Release_2). • [L, E]: The modifications of the capsule are registered in a logbook. • [E]: Each release is uniquely identifiable so that downstream users and tools can trace the exact dataset it contains. • [E]: After implementation, it is expected updates in the data capsule to be carried out annually, in alignment with the requirements of the European Health Data Space regulation. • [L]: The central infrastructure references the latest capsule release by default, unless otherwise specified. • [E]: Data centers are responsible for locally maintaining a copy of each previous capsule version if needed for traceability or auditing.

• [P]: Once the script is re-run, the latest version will be available for the Protect Child orchestrator.
Additional notes [R]: • [R]: The European Health Data Space (EHDS) anticipates implementing data versioning approximately on an annual basis.
Information that supports the user story [E, LR, P]: • [E]: Conversations with technical experts and documentation made for other project tasks. • [P]: Previous projects such as IDEA4RC. • [P]: Osmond, M., Hartley, T., Dyment, D. A., Kernohan, K. D., Brudno, M., Buske, O. J., ... & Mostafavi, S. (2022). Outcome of over 1500 matches through the Matchmaker Exchange for rare disease gene discovery: The 2-year experience of Care4Rare Canada. *Genetics in Medicine*, *24*(1), 100-108. • R]: European Parliament & Council of the European Union. (2025, February 11). *Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847*. Official Journal of the European Union, L 63, 5 March 2025.

Table 7. Integration of data sources through a common data model

4. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD, DEUSTO **Expertise:** Data managers (Managers of the platform) Integration of data sources through a common data model
As a Data Manager
I want the platform to support the secure integration and harmonization of multi-source clinical data (including electronic health records, registries such as PETER and ModSimTher, and laboratory data) into a common data model,
so that I can efficiently share standardized data within the Protect-Child ecosystem.
Initial acceptance criteria [LR, E, P]: • [E, LR, P]: The platform shall enable the secure integration of data including electronic health records (EHRs), registries such as Peter, ModSimTher, and laboratory data) into a common data model. • [E]: The platform allows the successful integration of the extract, transform and load (ETL) workflows validated for each Center of Expertise (CoE). • [E]: The unstructured data should be adapted to the common data model through Natural Language Processing (NLP). • [E]: The data is converted to OMOP data model. • [L]: The platform enables secure interoperability across diverse EHR systems and between hospitals, labs, and research institutions.
Additional notes: • The requirements of this user story are related to User Story 2 and User Story 5.
Information that supports the user story [P, W, LR]: • [P]: Previous projects such as IDEA4RC. • [W]: *“Whatever you use, and this is a term in HPO terms called hypoglycemia... there is a word called hypoglycemia with a code and a standardization... These ontologies are not only clinical signs or symptoms, but also biochemical, x-ray, and neurophysiological findings.”* **— Geneticist** • [W]: *“We have different tools with different electronic software."* **— Hepatologist** • [W]: *"Because probably if in Italy you put 'mis'... by mistake... you need to interpret the context that this particular doctor is putting 'mis' as an algorithm."***— Lab scientist** • [W]: *"For doing that [automation], you need to transform what is now a box… a general form. That takes at least ten to fifteen minutes, even if it's just clicking."***— Lab scientist** • [LR]: Gill, I. S., Griffiths, E. J., Dooley, D., Cameron, R., Savić Kallesøe, S., John, N. S., ... & Hsiao, W. W. (2023). The DataHarmonizer: a tool for faster data harmonization, validation, aggregation, and analysis of pathogen genomics contextual information. *Microbial Genomics*, *9*(1), 000908.

- [LR]: Danahey, K., Borden, B. A., Furner, B., Yukman, P., Hussain, S., Saner, D., ... & O'Donnell, P. H. (2017). Simplifying the use of pharmacogenomics in clinical practice: building the genomic prescribing system. *Journal of biomedical informatics*, *75*, 110-121

Table 8. NLP Tool for converting to common data model

5. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD, DEUSTO **Expertise: Managers of Protect-Child Platform (Potentially BIOMERIS)** **NLP Tool for converting to common data model**
As a Data Manager
I want the platform to help me extract and convert the unstructured data that comes from local hospitals or a research center to the common data model of Protect child
so that I can share the data for the Protect Child ecosystem
Initial acceptance criteria [LR, E, P]: • [E, LR, P]: The platform enforces unstructured data to match predefined health and genomic data standards to ensure interoperability with the agreed data models (e.g., OMOP). • [E, P]: The NLP standardization process of unstructured data should consider the languages of the text field (e.g., Spanish, Italian, German). • [E]: The platform integrates context-aware intelligence to interpret ambiguous or region-specific terms (e.g., "MIS" used differently across countries). • [L, LR]: The model should translate phenotype descriptions into standardized ontology codes and maps them correctly across specialties. • [P, E]: The platform also provides a report of the NLP model with quality indicators (e.g., confidence scoring) for Data Users to validate the extraction of the information needed. • [E]: The model should also provide standard metrics (e.g., precision, recall, F1-score) to data managers to evaluate the models before deployment. • [E, W]: The conversion to standardized formats can be applied not just to genomic data (already well-defined) but can be extended to diverse clinical documents such as pre-transplant evaluations, surgical reports, discharge summaries, histopathology reports, outpatient notes, laboratory results, immunological assessments, and imaging reports. • [E]: Users can run the NLP tool locally on their machine to automatically extract and map unstructured data to the common data model, without requiring full platform execution.
Additional notes [E]: • [E]: The extraction of data should be within the Secure Processing environments and performed using open-source language models (e.g., Mistral, Medgemma) complying with GDPR standards such as Article 6 and EHDS Article 50. • [E]: In accordance with legal requirements, the CoE must maintain a record of all processing activities and an audit trail documenting dataset access, algorithm execution, and data exports to ensure accountability (Article 30, GDPR). • [L]: This aspect still requires additional investigation to be conducted in the iterative phase of T2.1
Information that supports the user story [P, W, LR, E]: • [P]: Previous projects such as IDEA4RC. • [E]: Requirements mapped by T2.4. • [W]: *"Whatever you use, and this is a term in HPO terms called hypoglycemia... there is a word called hypoglycemia with a code and a standardization... These ontologies are not only clinical signs or symptoms, but also biochemical, x-ray, and neurophysiological findings."* **— Geneticist** • [W]: *"We have different tools with different electronic software."* **— Hepatologist** • [W]: *"Because probably if in Italy you put 'mis'... by mistake... you need to interpret the context that this particular doctor is putting 'mis' as an algorithm."***— Lab scientist** • [W]: *"For doing that [automation], you need to transform what is now a box… a general form. That takes at least ten to fifteen minutes, even if it's just clicking."***— Lab scientist** • [LR]: Gill, I. S., Griffiths, E. J., Dooley, D., Cameron, R., Savić Kallesøe, S., John, N. S., ... & Hsiao, W. W. (2023). The DataHarmonizer: a tool for faster data harmonization, validation, aggregation, and analysis of pathogen genomics contextual information. *Microbial Genomics*, *9*(1), 000908.

• [LR]: Danahey, K., Borden, B. A., Furner, B., Yukman, P., Hussain, S., Saner, D., ... & O'Donnell, P. H. (2017). Simplifying the use of pharmacogenomics in clinical practice: building the genomic prescribing system. *Journal of biomedical informatics*, *75*, 110-121

Table 9. Data quality checks

6. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD, DEUSTO **Expertise:** Data managers (Managers of the platform) Data quality checks
As a Data Manager
I want to run quality checks before the data injection across one or more data sources available
So that I can check for any errors in the data and fix them
Initial acceptance criteria [E]: • [E]: Automated validation and error-checking mechanisms detect inconsistencies or mismatches in terminology and suggest corrections. • [E]: Quality is defined as: (1) completeness, (2) conformance, (3) plausibility. • [P]: Only variables above an agreed quality score (threshold) are made available to be transferred to the Protect Child Capsule.
Additional notes:
Information that supports the user story [P, E, LR, FG]: • [P]: Previous projects such as IDEA4RC. • [E]: Conversational with statistician from La Paz. • [LR]: Martínez-García, M., & Hernández-Lemus, E. (2022). Data integration challenges for machine learning in precision medicine. *Frontiers in medicine*, *8*, 784455. • [LR]: Lau, J. W., Lehnert, E., Sethi, A., Malhotra, R., Kaushik, G., Onder, Z., ... & Davis-Dusenbery, B. (2017). The Cancer Genomics Cloud: collaborative, reproducible, and democratized—a new paradigm in large-scale computational research. *Cancer research*, *77*(21), e3-e6. • [LR]: Murtagh, M. J., Turner, A., Minion, J. T., Fay, M., & Burton, P. R. (2016). International data sharing in practice: innovative technologies meet old governance. *Biopreservation and biobanking*, *14*(3), 231-240 • [FG]: *Number two important piece is related to the data quality. Now you have to clean the data before using it.* **- Big data expert** • [E]: Insights provided during consortium meeting in Palermo by DEUSTO.

Table 10. Metadata sharing

7. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD, DEUSTO **Expertise:** Data holders (Managers of the platform) Metadata sharing
As a Protect child data holder
I want to make the metadata from the data sources available to data users to describe the information in the capsule
So that I can make data users able to search whether the information available in the capsule is relevant for their research.
Initial acceptance criteria [L, P, E]: • [L, E]: Metadata is structured and standardized across all relevant data sources. • [P]: Data Users can search and filter metadata to assess relevance before requesting access to raw or aggregated data. • [P]: The metadata presentation includes key descriptors (e.g., data type, number of patients, collection times, disease area, level of aggregation).
Additional notes:
Information that supports the user story [P, FG, E]: • [P]: Previous projects such as IDEA4RC. • [FG]: *"It is very important for clinical experts to have this [aggregated] information in one single screen"* **— Geneticist**

- [FG]: "*I think that [the problem] is the low number of patients in this, if we interrogate the overall set of capsules, we could obtain only aggregated results.*" **— Pediatrician**
- [FG]: "You *can, uh, integrate data from a clinical investigation*" **— Pediatrician**
- [FG]: "*There are there are some big registries but they're not so big and they're not so uh, holistic as in terms of aggregating as they are more specifically for a certain pathology*" **- Pediatrician**
- [E]: Conversations with technical experts, managers of the projects and documentation for other tasks related to the project (E.g., Clinica Project, High level data and material mapping).

Table 11.Managing data access

8. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data holders (Managers of the platform) Managing data access
As a Protect child data holder
I want the platform to help me evaluate the status of the submitted data permit
so that I can understand whether additional information is needed or what steps are required to complete the data permit application process.
Initial acceptance criteria [L, P]: • [P]: The platform allows me to track the different statuses of the requests. For instance, the platform contains a dashboard displaying ongoing, pending, and completed data access applications. • [L]: The platform allows me different filtering options (e.g., by status, dataset, institution). • [P]: The platform allows me to see and manage the uploaded documents (e.g., ethical approval, data manager process, etc.). • [P]: The platform allows me to leave internal notes and access the logs of communication between requests. • [E]: The platform notifies if after 2 months there are still aspects of the data permit under evaluation (e.g., revision of the approval of the study protocol). • [E, LR]: The platform allows me to assign actions to distinct roles to users in my organization (e.g., legal, ethics, clinical). • [P]: The platform allows me to define specific Data Governance Policies outlining what data can be shared and under what conditions.
Additional notes: • [E]: In line with GDPR Article 35, SPE operators must assign a Data Protection Officer (DPO) and assess the impact on data protection for all high-risk processing activities. • [E]: User will only access sensitive data only when authorized through valid permits issued by the Governance layer. • [E, R]: This process should align with the requirements set by the European Health Data Space (EHDS), including the establishment of a centralized Health Data Access Service (HDAS). • [LR]: This user story is also justified based on best practices identified in the literature. • [L]: This aspect still requires additional investigation to be conducted in the iterative phase of T2.1
Information that supports the user story [P, FG, E]: • [P]: Previous projects such as IDEA4RC. • [E]: Conversations with technical experts, managers of the projects and documentation for other tasks related to the project (E.g., Clinica Project, High level data and material mapping). • [E]: Requirements mapped by T2.4 and identified by T2.2. • [LR]: Quinn, P., Ellyne, E., & Yao, C. (2024). Will the GDPR restrain health data access bodies under the European Health Data Space (EHDS)? *Computer Law & Security Review*, *54*, 105993. • [LR]: Pavlenko, E., Strech, D., & Langhof, H. (2020). Implementation of data access and use procedures in clinical data warehouses. A systematic review of literature and publicly available policies. *BMC medical informatics and decision making*, *20*(1), 157. • [LR]: Wood, A., Denholm, R., Hollings, S., Cooper, J., Ip, S., Walker, V., ... & Sudlow, C. (2021). Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. *bmj*, *373*. • [R]: European Parliament & Council of the European Union. (2025). *Regulation (EU) 2025/327 of 11 February 2025 on the European Health Data Space and*

amending Directive 2011/24/EU (OJ L 63, 5 March 2025). Official Journal of the European Union.

Table 12. Define actors and roles for data permit management

9. Partners involved: SERMAS-LA PAZ, UNIPD, UKE, ISMETT, CERTH **Expertise:** Legal representatives (from Data holders) Define actors and roles for data permit management
As a Protect-Child Data Holder
I want to define the actors involved in governance and assign authorization levels,
so that different authorized users can revise and control the workflows for the data permit and ensure that only authorized personnel manage and review applications.
Initial acceptance criteria [P]: ∉ [P, LR]: The system allows data holders to define actors in my organization (legal staff, data managers, reviewers, etc.). ∉ [P]: Each actor can be assigned an authorization level to control different aspects of the workflow. ∉ [P]: The platform has an authentication system that verifies the credentials against authorized roles.
Additional notes:
Information that supports the user story [P]: ∉ [P]: Previous projects such as IDEA4RC. ∉ [LR]: Quinn, P., Ellyne, E., & Yao, C. (2024). Will the GDPR restrain health data access bodies under the European Health Data Space (EHDS)? *Computer Law & Security Review*, *54*, 105993. ∉ [LR]: Pavlenko, E., Strech, D., & Langhof, H. (2020). Implementation of data access and use procedures in clinical data warehouses. A systematic review of literature and publicly available policies. *BMC medical informatics and decision making*, *20*(1), 157. ∉ [LR]: Wood, A., Denholm, R., Hollings, S., Cooper, J., Ip, S., Walker, V., ... & Sudlow, C. (2021). Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. *bmj*, *373*.

Table 13. Data permit rules and documentation

10. Partners involved: SERMAS-LA PAZ, UNIPD, UKE, ISMETT, CERTH **Expertise:** Data holders (Managers of the platform) Data permit rules and documentation
As a Protect Child Data Holder
I want to specify the rules for accessing the data as well as the documentation and templates required for a data permit application (e.g., study protocol, consent form templates),
so that data users know exactly what is needed to complete their application and can submit it using the correct formats.

Initial acceptance criteria [P]:
• [E, P]: The Data Holder can list all required documents and information for a Data Permit (e.g., study protocol, ethics approval, data protection plan).
• [P]: Governance interface allows specifying required application information:
  ○ Name, institution, position, department/address, contact details
  ○ Brief CV
  ○ Type of requesting institution (predefined list)
  ○ Short study description and other requested info
• [P]: The Data Holder can upload templates or examples linked to each document type. Each document in the application interface is linked to a template (editable or linked to a web app).

- [L, E]: Data Users can view the requested information and download templates directly from the Data Governance Platform.
- [L, E]: The platform distinguishes between mandatory and optional documents.
- [E]: The Data Holder can update or revise requirements, with version history visible to Data Users.
- [P]: System allows uploading conditions for data use (PDF or web links).
- [P]: Conditions are linked to automatic checks when Data Users submit an application.
- [L]: The requirements are associated with specific datasets or categories of data access.
- [L, E]: Only authorized Data Holders can edit or manage the requirements.
- [P]: Data Users' applications can be automatically checked for completeness against these defined requirements.
- [P]: Configuration interface allows saving incomplete setups and resuming later.
- [P]: Once all steps are complete, the configuration can be confirmed and saved.

Additional notes [P]:

- [LR]: This user story is informed by the requirements outlined in the European Health Data Space (EHDS), as well as by insights from prior projects and best practices identified through a literature review.
- [L, E]: Recognizing that institutions and countries may have distinct regulatory and operational requirements, this user's story highlights the adaptability of the governance platform allowing customization of requirements for accessing proprietary hospital data.

Information that supports the user story [P, E]:

- ∉ [P]: Previous projects such as IDEA4RC.
- ∉ [E]: Data governance flows.
- ∉ [E]: Discussions with data governance teams (e.g., CERTH, UDGA).
- ∉ [LR]: European Parliament & Council of the European Union. (2025). Regulation (EU) 2025/327 of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU (OJ L 63, 5 March 2025). Official Journal of the European Union.
- ∉ [LR]: Quinn, P., Ellyne, E., & Yao, C. (2024). Will the GDPR restrain health data access bodies under the European Health Data Space (EHDS)? Computer Law & Security Review, 54, 105993.
- ∉ [LR]: Pavlenko, E., Strech, D., & Langhof, H. (2020). Implementation of data access and use procedures in clinical data warehouses. A systematic review of literature and publicly available policies. BMC medical informatics and decision making, 20(1), 157.
- ∉ [LR]: Wood, A., Denholm, R., Hollings, S., Cooper, J., Ip, S., Walker, V., ... & Sudlow, C. (2021). Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. bmj, 373.
- ∉ [LR]: Prasser, F., Spengler, H., Bild, R., Eicher, J., & Kuhn, K. A. (2019). Privacy-enhancing ETL-processes for biomedical data. International journal of medical informatics, 126, 72-81.
- ∉ [LR]: Ienca, M., Ferretti, A., Hurst, S., Puhan, M., Lovis, C., & Vayena, E. (2018). Considerations for ethics review of big data health research: A scoping review. *PloS one*, *13*(10), e0204937.

Table 14. Data permit access oversight

11. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data manager (Data holder) Data permit access oversight
As a Protect child data holder
I want the platform to help me evaluate access to the data and enforce role-based, time-limited access controls
so that I can track usage and ensure data is only used within the agreements previously established

Initial acceptance criteria [P, L]: ● [P, L]: The platform contains a dashboard with the users who have accessed the data, and the data permit information. ● [E]: The platform allows me to see any comments attached to the data permit application. ● [L]: The platform reminds me once an access is about to expire to either remind the data user to update their data permit or revoke the data access. ● [L]: The platform provides a compliance report so that I can download it for audits or internal reviews. ● [L]: The data holder must give access only to the data variables included in the data access permit and have mechanisms of control to revoke access when necessary.
Additional notes: ● [E]: All data processing activities (e.g., under the capsule) should be GDPR compliant (e.g., article 6 and 9) with a specific consent of the patient to share the data for secondary purposes. Additionally, the data shared with data users should be pseudoanonymized or aggregated.
Information that supports the user story [FG, LR, P]: ● [FG]: *"If we need the ethical approval, if you want to, if you want to do, a clinical study or something for the standard practice. Uh, maybe the problem with privacy is, uh, to get access to some data. "***—Geneticist** ● [FG]: "*You need ethical permission and legal permission and many other things*" **— Pediatricians** ● [FG]: "*And complete ethical support, to convince our ethical committee to access*" **— Pediatrician** ● [FG]:*"The main issue that we identified was that the procedures are very bureaucratic [...] all of these procedures take a lot of time that could be used in other ways."* **— Lawyer** ● [FG]: "*An LLM like a ChatGPT thing where you can just go and ask it: What do I need to do if I want to do a clinical trial on this? And it just shows you that*" **— Lawyer** ● [LR]: Rodchenkov, I., Babur, O., Luna, A., Aksoy, B. A., Wong, J. V., Fong, D., ... & Sander, C. (2020). Pathway Commons 2019 Update: integration, analysis, and exploration of pathway data. *Nucleic acids research*, *48*(D1), D489-D497. ● [P]: Previous projects such as IDEA4RC. ● [E]: Data governance workflows. ● [E]: Requirements mapped in T2.4

Table 15. Auditing

12. Partners involved: SERMAS-LA PAZ, UNIPD, UKE, ISMETT, CERTH **Expertise:** Data holders (Managers of the platform) Auditing
As a Protect child data holder
I want to audit data access and data processing activities for my Secure Processing Environment (data capsule)
so that I can comply with governance provisions and legal obligations as mandated by the GDPR and the European Health Data Space (EHDS) Regulation.
Initial acceptance criteria [P]: ∉ [P]: A data usage tracking and auditing dashboard is available (T7.2). ∉ [P]: Audit results are produced to meet the Centres of Expertise (CoEs) obligations under the GDPR, including the facilitation of data subject rights under Chapter III.
Additional notes [P, R]: ∉ [P, R]: This user story is linked to logging requirements expressed in Article 50(1)(e) of the proposed EHDS Regulation. [W]: This user story may evolve once a Health Data Access Body (HDAB) is formally established. ∉ [E]: All data processing activities must comply with the GDPR, particularly Articles 6 and 9, and ensure explicit user consent for allowing data access.

- [E]: In accordance with Article 30 (GDPR), each CoE must maintain a record of processing activities and an audit trail documenting dataset access, algorithm execution, and data exports to ensure accountability.
- [E, P]: Secure Processing Environments (SPEs) must maintain immutable, trusted logs for at least one year, to support internal and external audits.

Information that supports the user story [P]:

- [P]: Previous projects such as IDEA4RC.
- [P]: Previous project such as TEDHAS: Schlünder, I., Bodi, A., Hanché, V., Peculis, R., Rizzuto, L., Rovite, V., Selecka, I., … Schneider, K. (2025, May 27). D7.1 Guideline on how to use data in a secure processing environment (Deliverable D7.1). TEHDAS2 – Second Joint Action Towards the European Health Data Space. https://tehdas.eu/wp-content/uploads/2025/07/d7.1-guideline-on-how-to-use-data-in-a-secure-processing-environment.pdf
- [R]: European Parliament & Council of the European Union. (2016, April 27). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L 119, 1–88. https://eur-lex.europa.eu/eli/reg/2016/679/oj
- [R]: European Parliament & Council of the European Union. (2025, February 11). Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847. Official Journal of the European Union, L 63, 5 March 2025.
- [E]: Requirements mapped by T2.4 and defined by T2.2.

Table 16. Backoffice dashboard for signing agreements PROTECT-CHILD [GDPR compliance]

13. Partners involved: SERMAS-LA PAZ, UNIPD, UKE, ISMETT, CERTH **Expertise:** Legal representatives (from Data holders) Backoffice dashboard for signing agreements PROTECT-CHILD [GDPR compliance]
As a PROTECT-CHILD legal representative
I want to examine and process the requests for digital signature of study specific PROTECT-CHILD Agreements related to the usage of my PROTECT-CHILD Platform component
So that I can process personal data in my PROTECT-CHILD component in compliance with the GDPR and the upcoming EHDS Regulation
Initial acceptance criteria [P]: • [P]: A Dashboard is provided where the list of pending study specific PROTECT-CHILD agreements (related to User Story 29) is shown. • [P]: Only Agreements related to processing occurring in the relevant PROTECT-CHILD component are shown. • [P]: The PROTECT-CHILD Platform component legal representative can digitally sign the pending Agreements. • [P]: The PROTECT-CHILD Platform component legal representative can check the status and history of all pending and signed study specific PROTECT-CHILD agreement.
Additional notes [P]: - [E]: This is a potential but non-mandatory user story. It requires further clarification from data holders regarding the process for accepting digitally signed documents, including which software is recognized and under what conditions. - [E]: Data shared across borders within the federated network should rely on permits recognized by Health Data Access Bodies (HDABs) in line with EHDS Article 50. - [L]: This aspect still requires additional investigation to be conducted in the iterative phase of T2.1
Information that supports the user story [P]: • [P]: Previous projects such as IDEA4RC. • [E]: Requirements mapped by T2.4 and defined by T2.2.

Table 17. Data permit processing overview

14. Partners involved: SERMAS-LA PAZ, UNIPD, UKE, ISMETT, CERTH **Expertise:** Legal representatives (from Data holders) Data permit processing overview
As a Protect Child Legal Representative
I want to provide an overview of the data processing activities performed by my CoE, including those carried out by CoE researchers as well as by any external users who have submitted a data permit application
So that I can respond to a Supervisory Authority's request

Initial acceptance criteria [P]:
• [P]: A Dashboard is provided from where the legal representative(s) will have access to all documentation and information necessary in order to satisfy the request, such as:
  ○ Anonymization procedures
  ○ Access granting procedures
  ○ Technical and organizational measures
  ○ DPIAs
  ○ Data access applications and supporting documents
• [P]: Data permits granted

Additional notes [E]: • [E]: In accordance with legal requirements, the CoE must maintain a record of all processing activities and an audit trail documenting dataset access, algorithm execution, and data exports to ensure accountability (Article 30, GDPR). • [E]: All data processing activities (e.g., under the capsule) should be GDPR compliant (e.g., article 6 and 9) with a specific consent of the patient to share the data for secondary purposes. Additionally, the data shared with data users should be pseudoanonymized or aggregated. • [E]: In line with GDPR Article 35, SPE operators must assign a Data Protection Officer (DPO) and assess the impact on data protection for all high-risk processing activities.
Information that supports the user story [P, E]: ∉ [P]: Previous projects such as IDEA4RC. ∉ [E]: Requirements mapped by T2.4 and defined by T2.2.

Table 18. Data permit granting

15. Partners involved: SERMAS-LA PAZ, UNIPD, UKE, ISMETT, CERTH **Expertise:** Legal representatives (from Data holder) Data permit granting
As a CoE legal representative
I want to examine and process Data Access Applications that involve data held by my CoE
So that I can approve it, reject it with a motivation, or require minor additional information before a Data Permit is granted.
Initial acceptance criteria [P]: • [P]: A Dashboard is provided where the list of pending Data Access Applications per date of filing is shown and can be processed. • [P]: Only Data Access Applications in scope of the relevant CoE responsibility (i.e., requested for cohorts that include data from that CoE) are shown. • [P]: A Data Permit will be granted to the requesting Data User (related to User Story 29 and 30) only when all CoEs whose responsibility the Data Access Application is in the scope of.

- [P, LR]: The data holders must accept or reject access to their dataset. The expected time for the approval is within 2 months. This can be extended due to the complexity of project. However, it is important to communicate to the researcher.
- [LR]: Potentially, the approval of the data permit can also be accepted or rejected through the Health Data Access Body (HDAB). This approval is also expected to happen within 2 months. However, the complexity of the project might influence this timing.
- [P]: The CoE legal representative can digitally sign her/his deliberation
  - In case of approval, the CoE legal representative can digitally sign the PROTECT-CHILD data permit, generated as per User Story 29 and attached to the Data Access Application as per User Story 30)
  - In case of rejection, the legal representative should be able to describe the reason, for the data user to consider.
- [P]: The CoE legal representative can check the status, comments and history of all pending and examined Data Access Applications.

Additional notes [P, R]:

- [R]: The above obligation is provided in Art. 43 of the EHDS Regulation, and it prescribes obligations to Health Data Access Bodies. In the context of PROTECT-CHILD, it is the PROTECT-CHILD federated platform that could be considered as a Data Access Body. However, as it is not intended for the platform to become a legal entity, this obligation shall, most likely, fall onto the personnel responsible for the Data Permit procedures. In the upcoming years, the Health Data Access Body.
- [P, R, E]: Data holders must foresee making the pseudoanonymized datasets available to the Health Data Access Body to allow data access with the corresponding entities when necessary.
- [E]: The data shared should comply with the principles of data minimization and purpose limitation principles.

Information that supports the user story [P, R]:

- ∉ [P]: Previous projects such as IDEA4RC.
- ∉ [R]: Frequently Asked Questions on European Health Data Space .
- ∉ [R]: European Parliament & Council of the European Union. (2025, February 11). *Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847*. Official Journal of the European Union, L 63, 5 March 2025.

Table 19. Data permit revocation

16. Partners involved: SERMAS-LA PAZ, UNIPD, UKE, ISMETT, CERTH **Expertise:** Legal representatives (from Data holders) Data permit revocation
As legal representative
I want to examine compliance of Data Permit with its granting conditions and relevant legislation
So that I can revoke a granted Data Permit not complied to
Initial acceptance criteria [P, L]: • [P]: A Dashboard is provided to the CoE legal representative. In this dashboard legal representatives can: (1) Monitor Data User's compliance with the conditions established by the CoE as well as national and international legislation, (2) Evaluate the information that comes from auditing process (US 11) (Audit data access in CoEs Data capsule), (3) Request Data Users for additional information. • [P]: In this dashboard, legal representatives can terminate the data processing operations if non-compliance is identified. Potentially, they may notify the Data User of the issue, providing about 2 months for corrective action, and offer an opportunity to present their perspective. If non-compliance persists, legal representatives have the possibility to revoke the Data Permit,

cease any relevant data processing operation and, where appropriate, exclude the Data User from any access to the data for a period up to 5 years. • [P]: Inform the Data User of its decisions and the reasons upon which it was made.
Additional notes:
Information that supports the user story [P, R, E]: ∉ [P]: Previous projects such as IDEA4RC. ∉ [E]: Data governance workflows.

User stories focused on data users

The following user stories describe, from the perspective of Protect-Child data users, the key functionalities the platform must support throughout the full research workflow, from registration and login, to metadata exploration, cohort creation, data analysis, visualization, and reporting. Each user story captures what users need to achieve and why, while the accompanying acceptance criteria outline how the system should behave in practice. To structure these requirements, we distinguish between functional requirements, which define the concrete actions and operations the platform must enable, and non-functional requirements, which describe the quality attributes and conditions (e.g., security, usability, compliance, scalability) that ensure those functions can be delivered safely.

Functional high-level requirements

Registration and Login

Table 20. Registration

17. Partners involved: SERMAS-LA PAZ, UNIPD, UKE, ISMETT, CERTH **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Registration
As a Protect child data user
I want to register on the Protect Child platform
So that I can securely access its resources and services.
Initial acceptance criteria [L, P, E]: • [P]: A secure web page is available where I can fill out a registration form. • [L]: The form requires me to provide my full name, institutional email, affiliated institution, role, and a strong password that meets security standards. • [L]: After submitting the form, I receive a verification email with a confirmation link. • [L]: I can confirm my email address by clicking the link in the verification email. • [L, E]: Once my email is verified, I am prompted (or required) to set up multi-factor authentication (MFA) for enhanced security. • [L, P]: I can log in at any time to update my profile information or change my password. Including auto-expiration after inactivity and the ability for users to view and revoke active sessions. • [L]: The platform ensures credentials and tokens are never stored or transmitted in plaintext, using strong hashing algorithms and secure token-based authentication (e.g., OAuth 2.0). • [L]: The platform ensures all backend services mutually authenticate and encrypt communication before exchanging any sensitive data (e.g., via service mesh or mTLS).

Additional notes [E]: • [E]: SPEs must apply security-by-design and security-by-default principles in line with ENISA and ISO 27001/27701 standards, including encryption, access control, and key management. • [E]: All users and administrators accessing health data shall authenticate using MFA and recognized federated identity frameworks (eIDAS, GA4GH Passports). Each identity must be distinct, non-transferable, and consistently verified while active.
Information that supports the user story [P, E]: • [P]: Previous projects such as IDEA4RC. • [E]: Requirements mapped by T2.4 and defined by T2.2.

Table 21. Login

18. Partners involved: SERMAS-LA PAZ, UNIPD, UKE, ISMETT, CERTH **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Login
As a Protect child data user
I want to log in to Protect Child
So that I can access my account and use its features.
Initial acceptance criteria [L, P]: • [P, E]: A login page is provided for users to enter their credentials (from their organizations). • [L]: Users can only access their account after successful authentication. • [E, P]: The platform requires Multi-Factor Authentication (MFA) for all data holders and enforces role-based access control (RBAC), ensuring users only see and interact with the data and functions aligned with their permissions. • [L, P]: The platform provides a secure interface with HTTPS enforced to ensure all login credentials and session data are transmitted securely. • [L]: If incorrect credentials are entered, an error message is displayed. • [L]: Users remain logged in for a session unless they log out or their session expires. • [L]: The platform prevents multiple failed login attempts. • [E]: The platform verifies user context (device, location, behavior) at login and during high-risk actions. • [L]: The platform logs and monitors all login attempts, including successful and failed attempts, with timestamp, IP address, and device fingerprint. • [L]: The platform alerts users and administrators in case of suspicious login attempts, such as from a new device or an unexpected location. • [L]: The platform provides session management features, including auto-expiration after inactivity and the ability for users to view and revoke active sessions. • [L]: The platform ensures credentials and tokens are never stored or transmitted in plaintext, using strong hashing algorithms and secure token-based authentication (e.g., OAuth 2.0). • [L]: The platform ensures all backend services mutually authenticate and encrypt communication before exchanging any sensitive data (e.g., via service mesh or mTLS).
Additional notes [E]: • [E]: SPEs must apply security-by-design and security-by-default principles in line with ENISA and ISO 27001/27701 standards, including encryption, access control, and key management. • [E]: All users and administrators accessing health data shall authenticate using MFA and recognized federated identity frameworks (eIDAS, GA4GH Passports). Each identity must be distinct, non-transferable, and consistently verified while active.
Information that supports the user story [P, E]: • [P]: Previous projects such as IDEA4RC and regulations EHDS. • [E]: Conversations with technical experts and documentation made for other project tasks. • [E]: Workshop Implementing Zero Trust with Istio (20/06/25). • [E]: Requirements mapped by T2.4 and defined by T2.2.

General data management

Data discoverability [Metadata Search]

Table 22. Informed metadata discovery

19. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Informed metadata discovery
As a Protect child data user
I want to see a summary of the platform's goals and purpose before exploring the metadata or logging into the system
so that I can understand what the platform offers and decide if it is relevant to my work.
Initial acceptance criteria [L, P]: ● [P]: The platform provides a summary of how the platform works. ● [P]: The landing page includes two buttons: a) One to explore the metadata of the available databases. b) One to log in and access my personalized workspace and gather information.
Additional notes:
Information that supports the user story [W, P]: ● [P]: Previous projects such as IDEA4RC. ● [W]: Validation during workshop, this user story was accepted.

Table 23. Metadata discovery (1 by Advanced search)

20. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Metadata discovery (1 by Advanced search)
As a Protect child data user
I want to start the metadata search by selecting a specific transplant use case (e.g., liver, kidney)
So that I can explore variables metadata-level information relevant only to my area of interest.
Initial acceptance criteria [P]: ● [P]: The metadata search interface presents a list of available transplant types of datasets with the number of patients for each dataset. ● [P]: Data Users can select between one or multiple transplant types. ● [P]: After selecting a transplant type, I can proceed to the next step by clicking the "Continue" button. ● [P]: A visual progress bar (with steps like "Transplant Type," "Variables Pre-selection," "Availability per Center," and "Detail Analysis") guides me through the metadata exploration workflow. ● [W]: A map showing the centers available is shown after selecting the transplant type to further evaluate the locations I will retrieve data from. ● [P]: The transplant type selection is accessible without logging into the platform.
Additional notes [E]: ● [E]: Each SPE shall deploy a GA4GH Beacon v2 endpoint to enable federated discovery of genomic and phenotypic variants derived from VCF/BAM metadata.
Information that supports the user story [FG, W, P]: ● [FG]: *"There are 2 to 2 ways. The basic one that is composed of predefined questions in order to access the local and uh, center uh information you need, usually the outcomes, uh, the treatments and so on. Um, and the, and the other would be, and these would be predefined questions. That is, it. Okay. And the other would be more sophisticated, uh, tools to interrogate the database, uh, in an open way so that these are the two main uses, as far as I know of others."* **- *Pediatrician*** ● [P]: Previous projects such as IDEA4RC. ● [W]: This user story was discussed during the workshop and was accepted by clinical experts and healthcare providers. ● [W]: A data manager suggested including the map and clinical experts agreed.

• [E]: Requirements mapped by T2.4.

Table 24. Data discovery (Select metadata variables)

21. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Data discovery (Select metadata variables)
As a Protect child data user
I want to search, select, and pre-filter relevant clinical and research variables (e.g., patient demographics, histology, treatment response, variants) from the metadata
So that I can define the scope of my project and ensure I am only requesting the data I need
Initial acceptance criteria [P]: • [P, W]: A searchable and filterable list of variables is available, categorized by topic (e.g., Patient, Hospital Data, Genetic Test Expression). • [P]: Each variable includes a short description, data type (e.g., Label, Number), and example values for clarity. • [P]: Data Users can select or deselect variables using checkboxes, including a bulk-select option. • [P]: Selected variables persist through to the next steps in the workflow (Availability per Center, Detail Analysis), without revealing patient-level counts. • [P]: Selected variables can be exported as a CSV file for documentation or external review. • [P]: A progress indicator at the bottom reflects that the user is in *Step 2: Variables Pre-selection* of the research setup process. • [L, E]: Data Users only see metadata for each variable. The patient-level information, including how many patients have the selected variables, is not available at this stage.
Additional notes [E]: • [E]: Each SPE must implement a GA4GH Beacon v2 endpoint to support federated discovery of genomic and phenotypic variants based on VCF/BAM metadata. • [P, W, LR]: In line with ethical and data protection best practices (e.g., TEHDAS2) the display of subject-level demographic or clinical data should only be permitted for populations exceeding a minimum threshold (e.g., $n > 10$) to prevent re-identification risks. • [W]: The search functionality emerged as one of the most valued features among Data Users.
Information that supports the user story [P, W]: ∉ [P]: Previous projects such as IDEA4RC. ∉ [P]: Previous projects such as TEDHAS: Andacheh, N., Bergdahl, M., Bränd Persson, K., Dolanski-Aghamanoukjan, L., Ferent, L., Gustafsson, A., Jansson, G., Jönsson, C., Leander Knudsen, A., Magnusson, J., Mäkinen, M., Niemeyer, A., Näsänen-Gilmore, P., Peolsson, M., Piironen, V., & Rautiainen, M-R. (2025, September 16). *M5.2 Draft Guideline for Health Data Access Bodies on minimum categories and limitations on the reuse of health data: Reflections and recommendations for health data access bodies on allowed purposes and prohibited use according to the EHDS* (Version 0.5) [Draft guideline]. Joint Action Towards the European Health Data Space (TEHDAS2). https://www.tehdas.eu/wp-content/uploads/2025/09/draft-guideline-for-health-data-access-bodies-on-minimum-categories-and-limitations-on-the-reuse-of-health-data.pdf ∉ [W]: Workshop in Palermo. ∉ Jakob, C. E. M., Mache, S., Kiessling, C., Gruhl, F., & Valeri, F. (2020). Design and evaluation of a data anonymization pipeline to share individual-level health data from a national COVID-19 registry. *BMC Medical Informatics and Decision Making, 20*(1), 275. https://doi.org/10.1186/s12911-020-01258-5 ∉ [E]: Requirements mapped by T2.4.

Table 25. Data discovery (Verify data availability across clinical centers)

22. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Data discovery (Verify data availability across clinical centers)

As a Protect child data user
I want to verify which clinical centers have data in the data capsule for the variables I selected, across different outcomes (Infection, Drug-related toxicity, Rejection, Organ failure)
So that I can decide which centers to include in my analysis based on data completeness and relevance
Initial acceptance criteria [P]: ● [P]: Availability is displayed per center (e.g., SERMAS, INETUM, ISMEET, UKE, UNIPD), with timeframes for data collection clearly shown. ● [P]: Visual indicators (✓ = available, ✕ = not available, ⭕ = unknown/missing) are used to signal availability by variable and by phase (D, P, R). ● [P]: Data Users can select/deselect centers for inclusion based on the availability overview. ● [P]: Data Users can filter or collapse groups to manage long lists of variables. ● [P]: Variables are grouped (e.g., Patient Information) and displayed with descriptions. ● [P]: Variable availability can be filtered and searched to focus on specific data types. ● [P]: A downloadable CSV summary reflects selected variables and their availability across centers. ● [P]: A progress indicator confirms that this is *Step 3: Availability per Center* in the research setup process.
Additional notes:
Information that supports the user story [P, W]: ● [P]: Previous projects such as IDEA4RC. ● [W]: User story discussed during workshop and accepted by users.

Project metadata management

Table 26. Project metadata management (Workspace creation)

23. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Project metadata management (Workspace creation)
As a Protect child data user
I want to create a new research project
So that I can start working on my research question defining objectives, variables, and analysis in one place
Initial acceptance criteria [P]: ● [P]: After the selection of variables relevant for my study, I can create a "New workspace", where a button is visible. This workspace acts as a container for my study's variables, objectives, and subsequent workflow steps. ● [P]: Clicking the "New workspace" leads to a form or setup page for a new workspace where I can enter a name and a brief project description. ● [P]: The workspace form also asks for a definition of research objectives (e.g., identifying genetic determinants of the underlying disease that led to transplantation and developing polygenic risk scores (PRS)). ● [P]: A visual confirmation (e.g., redirect) confirms successful creation of the workspace. ● [P]: After creation, the workspace becomes accessible in "My Workspace" and links to all subsequent project phases (e.g., Metadata Search, Data Access).
Additional notes:
Information that supports the user story [P,W]: ● [P]: Previous projects such as IDEA4RC. ● [W]: User story discussed during workshop and accepted by users.

Table 27. Project metadata management (Workspace management)

24. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Project metadata management (Workspace management)
As a Protect child data user
I want an overview of all my workspace (i.e., research projects)
So that I can quickly find the research project I want to work on
Initial acceptance criteria [P]: ● [P]: Workspaces are organized in a table format. ● [P]: The list displays workspace name, last modification date, and status. ● [P]: Each workspace has an "Open" button. ● [P]: Clicking the "Open" button takes me to the workspace environment. ● [P]: Each workspace row has a delete icon (trash bin). ● [P]: Clicking it prompts a confirmation before deletion. ● [P]: A search bar is available at the top of the workspace list. ● [P]: The results dynamically filter as I type. ● [P]: In the workspace form, users can edit the project description and objectives later if their research focus evolves.
Additional notes:
Information that supports the user story [P,W]: ● [P]: Previous projects such as IDEA4RC. ● [W]: User story discussed during workshop and accepted by users.

Table 28. Project metadata management (Track project status in workspace)

25. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Project metadata management (Track project status in workspace)
As a Protect child data user
I want to navigate the progress of my research project across separate phases (e.g., metadata search, data access, analysis, and results)
So that I can easily understand what has been completed, what is still pending, and where to take action next
Initial acceptance criteria [P]: ● [P]: Each project workspace clearly shows the current status (e.g., Completed, Submitted, Pending) of each project phase. ● [P]: A visual indicator (e.g., progress line with checkmarks and icons) helps track progress through Metadata Search, Data Access, Data Analysis, and Result Report. ● [P]: The last modification date is visible to understand the most recent activity. ● [P]: Status labels (e.g., "Submitted", "Completed") are color-coded for quick identification.
Additional notes:
Information that supports the user story [P, W]: ● [P]: Previous projects such as IDEA4RC. ● [W]: User story discussed during workshop and accepted by users.

Data permit application (Data access)

Table 29. Data permit application (Redirect to data access permits after workspace creation)

26. Partners involved: SERMAS-LA Paz, INETUM, ISMETT, UKE, CERTH. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Data permit application (Redirect to data access permits after workspace creation)

As a Protect child data user
I want to be redirected to the Data Permit Tool to submit a data access request
So that I can submit a data permit application with all the requested information to the centres I have selected for using their data
Initial acceptance criteria [P, E, L]: ● [P]: After selecting metadata variables and creating a workspace, Data Users are prompted to initiate a data permit request. ● [L, E]: The platform auto-verifies whether the requested data type requires ethical approval based on the metadata. ● [L, E]: The platform provides a standardized process for Data Users from different institutions and countries (e.g., Spain, Italy, Germany) to apply for ethical approval and data access efficiently. ● [P]: The platform shows a step-by-step guide outlining the required documents and information before approval. For instance, Data Users should fill out a structured online form specifying: Purpose of the request (e.g., clinical study, research, standard care), data type and sources required, justification for data access, and compliance details (e.g., GDPR, local regulations). Further relevant documents are annexed to the Data Access Application. Examples include the (1) Study protocol, (2) Study-specific data management plan, (3) Study-specific DPIA, (4) Ethical Approval(s). ● [L, P]: The platform should allow Data Users to upload supporting documentation (e.g., patient consent forms, institutional approvals).
Additional notes [L, P, W]: ● [L]: Assess how this process may be optimized to improve efficiency for clinicians within the scope of the current project. ● [L]: This user story is connected to US 10 (from the data holder perspective). ● [E]: Data users can exercise their right of access, rectification, erasure, restriction and objection through this interface, responses should be within the legally required timeframe. ● [P]: The Data Access application may include information in line with art. 45(2) of the proposed EHDS Regulation (e.g. description of the requested data, explanation of the intended use, estimation of the period during which the data is needed, etc.) ● [P]: Art. 45(6) of the proposed EHDS regulation states that the Commission will set out a template for the Data Access Application ● [W]: The data governance method might change in the future. Therefore, the data permit application process should be adaptable for different data governance models.
Information that supports the user story [P, FG, LR, W]: ● [P]: Previous projects such as IDEA4RC. ● [FG]: *"If we need the ethical approval, if you want to, if you want to do, a clinical study or something for the standard practice. Uh, maybe the problem with privacy is, uh, to get access to some data. "***—Geneticist** ● [FG]: *"You need ethical permission and legal permission and many other things"* **— Pediatricians** ● [FG]: "*And complete ethical support, to convince our ethical committee to access"* **— Pediatrician** ● [FG]: *"The main issue that we identified was that the procedures are very bureaucratic [...] all of these procedures take a lot of time that could be used in other ways.""* **— Lawyer** ● [FG]: *"An LLM like a ChatGPT thing where you can just go and ask it: What do I need to do if I want to do a clinical trial on this? And it just shows you that"* **— Lawyer** ● [LR]: Rodchenkov, I., Babur, O., Luna, A., Aksoy, B. A., Wong, J. V., Fong, D., ... & Sander, C. (2020). Pathway Commons 2019 Update: integration, analysis, and exploration of pathway data. *Nucleic acids research*, *48*(D1), D489-D497. ● [W]: "We don't know whether the hospital has a plan for dealing with the requirements of EHDS [...] yes, we need this process to be adaptable" – Pediatrician ● [E]: Requirements mapped by T2.4 and defined by T2.2.

Table 30. Register or retrieve profile from Data governance Platform

27. Partners involved: SERMAS-LA Paz, INETUM, ISMETT, UKE, CERTH.

Expertise: Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics]
Register or retrieve profile from Data governance Platform
As a Protect child data user
I want to register or retrieve my profile on Protect-Child and confirm my privacy statement,
so that I can access the Data governance platform being identity with basic information

Initial acceptance criteria [P, E, L]:
• [P]: Data Users can access to the data permit application platform through Protect-Child via a web interface.
  ○ If already registered, the system retrieves existing profile data.
  ○ If not registered, user completes registration including:
  ■ Name, institution, department, contact email
• [P]: Type of requesting institution/user (predefined list).
• [P]: User must confirm the privacy statement via a checkbox.
• [P]: User ID and profile details are saved in the system.

Additional notes:
Information that supports the user story [P, FG, LR, W]: • [P]: Previous projects such as IDEA4RC.

Table 31. Submission of Data Permit Application to Data governance Platform

28. Partners involved: SERMAS-LA Paz, INETUM, ISMETT, UKE, CERTH.
Expertise: Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics]
Submission of Data Permit Application to Data governance Platform
As a Protect child data user
I want to complete and submit my data access application to the relevant CoEs and update it if requested,
so that my application is processed, reviewed, and tracked through the Protect child workflow.

Initial acceptance criteria [L, P]:
• [P]: The platform will retrieve the list of relevant Data Holders from Data Discovery.
• [P]: Data Users can fill in application templates including:
  ○ [P]: Brief CV of PI, study background, objectives, hypothesis, statistical analyses, data needed, expected results, dissemination plan
  ○ [P]: Any additional legal information requested.
• [P]: Data Users can fill in required documents using provided templates.
• [P]: The platform allows downloading templates and reusing documents across multiple Data Holders if applicable.
• [P]: The platform presents the necessary ethical committee forms, which Data Users can complete and submit.
• [P]: Data Users can review data access policies and accept conditions for each selected Data Holder.
• [P]: The platform indicates if the user is not allowed to apply for a specific Data Holder.
• [P]: Data Users can review completeness of the application before submission.
• [P]: Data Users can save the application and submit it via Protect-Child.
• [P]: The Data Access Application is sent to all CoEs that contribute data to the selected cohort and must deliberate on it.
• [P]: Where required, an Agreement shall be signed between the Data User and the CoEs that will be allowing access to the data.
• [P]: The result of the submission can be an approval, a rejection with a motivation, a request for additional information.
• [P]: The Data Access Application is assigned to the Data Holder's reviewers.
• [P]: The Data Access Application status is updated and linked to the Data Holder's monitoring dashboard.

• [P]: Data Users can access and modify the application if reviewers request changes. • [P]: Modified application can be saved and resubmitted. • [P]: Reviewers are notified of resubmission, and the application status is updated accordingly.
Additional notes: • [L]: This user story is connected to US 10 (from the data holder perspective). • [E]: Data users can exercise their right of access, rectification, erasure, restriction and objection through this interface, responses should be within the legally required timeframe. • [P]: The Data Access application may include information in line with art. 45(2) of the proposed EHDS Regulation (e.g. description of the requested data, explanation of the intended use, estimation of the period during which the data is needed, etc.) • [P]: Art. 45(6) of the proposed EHDS regulation states that the Commission will set out a template for the Data Access Application
Information that supports the user story [FG, P]: • [P]: Previous projects such as IDEA4RC.

Table 32. Notification data permit application

29. Partners involved: SERMAS-LA Paz, INETUM, ISMETT, UKE, CERTH. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Notification data permit application
As a Protect child data user
I want to be notified of the result of my data permit application
so that I continue and meet any additional requirements solicited by the legal representatives of the CoE and obtain a Data permit.

Initial acceptance criteria [P, E, L]:
• **[P]:** The Data Governance platform allows users to track the status of their data access applications. For instance, it provides a dashboard displaying ongoing, pending, and completed requests.
• **[P]:** Data Users receive notifications whenever the status of their application changes (e.g., from Pending to Granted).
• **[P]:** The result of a submission can be:
  ○ **[P]: Approved**: If required, the platform allows Data Users to sign an agreement with the CoEs.
  ■ **[P]:** Once the agreement is submitted, the platform provides access credentials or a secure link to the requested dataset, if applicable.
  ○ **[P, L]: Rejection**: The Data Governance platform allows users to make corrections, ask for clarification or resubmit the application.
  ○ **[P, L]: Request for additional information**: Data Users can revise and edit their application to include the requested information or documents.
• The result of a resubmission can be an approval, a rejection with a motivation, a request for additional information.
• **[P]:** The platform maintains logs of all data application processes.
• [**P]:** Any additional information submitted is included in the previously submitted Data Access Application (related to User Story 11).

Additional notes [L, P, W]: • [L]: This user story is connected to US 10 (from the data holder perspective). • [E]: Data users can exercise their right of access, rectification, erasure, restriction and objection through this interface, responses should be within the legally required timeframe. • [P]: The Data Access application may include information in line with art. 45(2) of the proposed EHDS Regulation (e.g. description of the requested data, explanation of the intended use, estimation of the period during which the data is needed, etc.)

• [P]: Art. 45(6) of the proposed EHDS regulation states that the Commission will set out a template for the Data Access Application • [W]: The data governance method might change in the future. Therefore, the data permit application process should be adaptable for different data governance models.
Information that supports the user story [P, FG, LR, W]: • [P]: Previous projects such as IDEA4RC. • [FG]: *"If we need the ethical approval, if you want to, if you want to do, a clinical study or something for the standard practice. Uh, maybe the problem with privacy is, uh, to get access to some data. "***—Geneticist** • [FG]: *"You need ethical permission and legal permission and many other things"* **— Pediatricians** • [FG]: "*And complete ethical support, to convince our ethical committee to access"* **— Pediatrician** • [FG]: *"The main issue that we identified was that the procedures are very bureaucratic [...] all of these procedures take a lot of time that could be used in other ways.""* **— Lawyer** • [FG]: *"An LLM like a ChatGPT thing where you can just go and ask it: What do I need to do if I want to do a clinical trial on this? And it just shows you that"* **— Lawyer** • [LR]: Rodchenkov, I., Babur, O., Luna, A., Aksoy, B. A., Wong, J. V., Fong, D., ... & Sander, C. (2020). Pathway Commons 2019 Update: integration, analysis, and exploration of pathway data. *Nucleic acids research*, *48*(D1), D489-D497. • [W]: "We don't know whether the hospital has a plan for dealing with the requirements of EHDS [...] yes, we need this process to be adaptable" – Pediatrician • [E]: Requirements mapped by T2.4 and defined by T2.2.

Table 33. Data permit signature

30. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD, CERTH **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Data permit signature
As a Data User
I want to generate and digitally sign the PROTECT-CHILD Data Permit
so that I can fulfill the formal requirements for data agreements with all relevant actors involved in obtaining a Data Permit and accelerate the resolution of legal and data governance issues.
Initial acceptance criteria [P]: • [P]: The platform provides a service for electronic drafting, signature and recording of a study specific PROTECT-CHILD. • [P]: The Data User can digitally sign the generated study specific PROTECT-CHILD Agreement. • [P]: The generated study specific PROTECT-CHILD agreement can be shared with other actors that must sign it (related to user story 28, 29). • [P]: In case that an agreement is required, the agreement must satisfy the requirements of the GDPR.
Additional notes [P]: • [P]: Art. 51(2) of the proposed EHDS regulation states that the Commission will establish a template for the joint controllers' arrangement to be signed among the Data User and the Health Data Access Body. • [E]: This is a potential but non-mandatory user story. It requires further clarification from data holders regarding the process for accepting digitally signed documents, including which software is recognized and under what conditions.

• [P]: The agreement shall be designed to address both GDPR requirements, as well as those provided in the upcoming EHDS Regulation and shall be accordingly updated to address any further updates and modifications in the legal framework. • [E]: For later stages of the project, the process should also simulate, as closely as possible, a central governing body that verifies the signatures of all Centers of Excellence (CoEs). • [L]: This aspect still requires additional investigation to be conducted in the iterative phase of T2.1
Information that supports the user story [P, E]: • [P]: Previous projects such as IDEA4RC. • [E]: Interview with CERTH experts. • [E]: Discussions with data governance teams (e.g., CERTH, UDGA).

Table 34. Digital consent management

31. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD, CERTH **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Digital consent management
As a data manager and clinical expert
I want the platform to provide a secure, auditable, and potentially blockchain-enabled patient's consent management system
so that I can access and share patient data in compliance with ethical, legal, and institutional requirements, without adding unnecessary administrative burdens.
Initial acceptance criteria [E, FG]: • [FG, E]: Patients (or legal guardians) can provide consent at a detailed level (e.g., allow genomic data access but restrict blood type information). • [E]: Patients can update, withdraw, or modify consent only before the data is aggregated. • [E]: Changes to the consent are traceable. • [E]: Once the data is aggregated patients (or legal guardians) cannot change their consent. • [E]: All consent actions (granting, withdrawal, modification) are recorded using blockchain/smart contracts. • [E]: A secure audit trail logs who accessed or modified consent and when, ensuring accountability. • [L, R, E]: Compliance is enforced with national, institutional, and international regulations (e.g., GDPR, hospital-specific SOPs). • [L]: Ethical approvals are integrated into workflows to prevent unauthorized use.
Additional notes: ∉ [E]: Blockchain technology could be implemented as a pilot to evaluate feasibility, auditability, and compliance, but it is not mandatory. ∉ [E]: Needs to be discussed with stakeholders, ethics committees, and legal advisors before implementation. ∉ [E]: This is a potential but not mandatory user story. ∉ [LR]: Based on a review of existing data governance frameworks, the recommendation is that consent should not be required for secondary data use. Nevertheless, Member States with more stringent data protection regimes, such as Germany, may impose additional consent requirements. ∉ [L]: This aspect still requires additional investigation to be conducted in the iterative phase of T2.1
Information that supports the use of story [FG, E, LR]: • [FG]: *"One of the first things that we would say is follow the national regulation."* — ***Geneticist***

- [FG]: *"You anonymize the data and then, which are the data we don't want to share and anonymize to have only the one we need to understand."* **— *Clinical Expert***
- [FG]: *"We have standard operational procedures that state how this should be done [...] When you start working in the hospital, they make you read and sign the standard operational procedures"* **— *Geneticist***
- [FG]: *"Privacy should be an advantage, but not a limit [...] So we agree that there should be customizable access permissions in data sharing capabilities with user registration, authentication and traceability, just not to limit patient care."* **— *Pediatrician***
- [FG]: *"All patients accept to understand. You know, for their child, they want to collaborate."* **— *Geneticist***
- [FG]: *"Complete legal and ethical support is necessary to convince the ethics committee."* **— *Pediatrician***
- [FG]: *"I mean not all the doctors a certain type of data and you have to have a traceability on that."* - ***Lab scientist***
- [FG]: *"The patients could administrate this [their data], this is my power, this is my property"* - ***Lab scientist***
- [E]: Discussions with data governance teams (e.g., CERTH, UDGA).
- [LR]: Jones, L. A., Nelder, J. R., Fryer, J. M., Alsop, P. H., Geary, M. R., Prince, M., & Cardinal, R. N. (2022). Public opinion on sharing data from health services for clinical and research purposes without explicit consent: An anonymous online survey in the UK. *BMJ Open, 12*(5), e057579. https://doi.org/10.1136/bmjopen-2021-057579
- [LR]: Cengiz, N., Kabanda, S. M., & Moodley, K. (2024). Cross-border data sharing through the lens of research ethics committee members in sub-Saharan Africa. *PLOS ONE, 19*(4), e0303828. https://doi.org/10.1371/journal.pone.0303828
- [LR]: Scheibner, J., Ienca, M., Kechagia, S., Troncoso-Pastoriza, J. R., Raisaro, J. L., Hubaux, J.-P., Fellay, J., & Vayena, E. (2020). Data protection and ethics requirements for multisite research with health data: A comparative examination of legislative governance frameworks and the role of data protection technologies. *Journal of Law and the Biosciences, 7*(1), lsaa010. https://doi.org/10.1093/jlb/lsaa010

Table 35. Data permit application (Support application process)

32. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD, CERTH **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Data permit application (Support application process)
As a Protect child data user
I want the data governance platform to help me apply for data access
so that I can understand what type of information and documents are needed.
Initial acceptance criteria [L, E, P]: ● [E]: To facilitate the process, the platform has resources (e.g., LLM) that allow users to know the requirements for ethical procedures beforehand. ● [E]: The platform provides templates for submission.
Additional notes [E]: ● [L]: This is a nice to have user story derived from the need for clarification on certain procedures or questions in the data permit process. ● [E]: CERTH will need structured information provided in full paragraphs rather than forms, along with clear instructions outlining the process to deliver this tool. ● [L]: This aspect still requires additional investigation to be conducted in the iterative phase of T2.1

Information that supports the user story [FG, LR, E]: ● [FG]: *"If we need the ethical approval, if you want to, if you want to do, a clinical study or something for the standard practice. Uh, maybe the problem with privacy is, uh, to get access to some data. "*—**Geneticist** ● [FG]:"*You need ethical permission and legal permission and many other things"* — **Pediatricians** ● [FG]:"*And complete ethical support, to convince our ethical committee to access"* — **Pediatrician** ● [FG]: *"The main issue that we identified was that the procedures are very bureaucratic [...] all of these procedures take a lot of time that could be used in other ways.""* — **Lawyer** ● [FG]: *"An LLM like a ChatGPT thing where you can just go and ask it: What do I need to do if I want to do a clinical trial on this? And it just shows you that"* — **Lawyer** ● [LR]: Rodchenkov, I., Babur, O., Luna, A., Aksoy, B. A., Wong, J. V., Fong, D., ... & Sander, C. (2020). Pathway Commons 2019 Update: integration, analysis, and exploration of pathway data. *Nucleic acids research*, *48*(D1), D489-D497. ● [E]: Conversations with the management team of Protect-Child.

Data use

Data pre-processing

Table 36. Select and manage cohorts for data analysis

33. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Select and manage cohorts for data analysis
As a Protect child data user
I want to select or build a patient cohort using structured and NLP-based tools within my data analysis workflow
So that I can define a relevant subset of patients to include in my research and proceed to analysis using high-quality, well-prepared data
Initial acceptance criteria [P]: ● [P]: Users can access Cohort Selection as the first step in the 5-step analysis pipeline (Cohort Selection → Data Quality → Data Preparation → Tool Selection → Results) ● [P]: Users can search existing cohorts, create new ones via the NLP Cohort Builder, or manage saved cohorts through the Cohort Manager ● [P]: The cohort list displays metadata such as ID, Name, Creation Date, Update Date, Status, and available Actions ● [P]: Users can proceed only when a valid cohort is selected ● [P]: A progress indicator at the bottom shows users where they are in the overall process
Additional notes: ● [P, W, LR]: In line with ethical and data protection best practices (e.g., TEHDAS2) the display of subject-level demographic or clinical data should only be permitted for populations exceeding a minimum threshold (e.g., $n > 10$) to prevent re-identification risks.
Information that supports the user story [P]: ● [P]: Previous projects such as IDEA4RC.

Table 37. Verify cohort building

34. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Verify cohort building
As a Protect child data user
I want to define a patient cohort using specific criteria (e.g., gender, birthdate) via the Cohort Manager tool

So that I can filter only the relevant patients for my research question and proceed with focused, high-quality data analysis
Initial acceptance criteria [P]: ● [P]: Users can add filters using AND/OR logic for patient parameters such as gender, age, or conditions. ● [P]: Each filter includes a search parameter name and a value field with a date or categorical input. ● [P]: Users can freely add or remove constraints using the interface controls. ● [P]: A “Create New Cohort” button finalizes the selection and links the cohort to the active analysis. ● [P]: Redirect back to the Data Analysis workflow with the selected cohort preloaded in Step 1: Cohort Selection ● [P]: All cohort definitions are saved for future review, reuse, or audit.
Additional notes: ● [P, W, LR]: In line with ethical and data protection best practices (e.g., TEHDAS2) the display of subject-level demographic or clinical data should only be permitted for populations exceeding a minimum threshold (e.g., *n > 10*) to prevent re-identification risks. ● [L]: This aspect still requires additional investigation to be conducted in the iterative phase of T2.1
Information that supports the user story [P]: ● [P]: Previous projects such as IDEA4RC.

Table 38. Verify data quality

35. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise**: Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Verify data quality
As a Protect child data user
I want to revise the data quality of the data I’ve selected
So that prior to the data analysis I can ensure the data meets the requirements needed for my research project.
Initial acceptance criteria [P]: ● [L]: Users can view summary dashboards displaying key data quality metrics, including missing values, duplicates, out-of-range values, and sequencing quality for genomic data. ● [L]: Users can interactively filter or highlight records based on quality flags (e.g., incomplete fields, low-quality genomic reads, inconsistent values). ● [L]: Visualizations include histograms, boxplots, heatmaps, coverage plots (for genomic data), and PCA/clustering plots to detect outliers or inconsistencies. ● [L]: Users can annotate or flag data points for review, correction, or exclusion from analysis. ● [L]: The platform provides alerts or notifications when data fails predefined quality thresholds. ● [L]: Quality checks and revisions are saved, and previous QC decisions can be reviewed for audit or reproducibility purposes.
Additional notes: ● [L]: This user story is related to User Story 2. ● [L[: To be further revise with clinical experts including geneticists. ● [L]: This aspect still requires additional investigation to be conducted in the iterative phase of T2.1
Information that supports the user story [S]: ● [S]: How important is for you inspecting data quality? (87,04%). ● [FG]: *“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” -Pediatrician*

• [FG]: "*Once we have the standard we have the data collection the number. Number two important piece is related to the data quality. Now you have to clean the data before using it" - Pediatrician*

Data analysis

Table 39. Prepare datasets by selecting cohorts, centers, and variables

36. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Prepare datasets by selecting cohorts, centers, and variables
As a Protect child data user
I want to select cohorts, centers, and variables to define the dataset for analysis
So that I can ensure the input data aligns with my study design and includes the relevant patient subgroups from the selected sources
Initial acceptance criteria [P]: • [P]: Users can select from a list of participating centers and predefined cohorts (e.g., "Patients with liver transplant > 6 years old"). • [P]: Variables can be selected or created from a dropdown menu for inclusion in the analysis. • [P]: A summary view displays selected elements and prepares summary statistics to help assess cohort balance and completeness. • [P]: Users can preview differences across cohorts and centers using toggles or checkboxes. • [P]: Clicking Continue takes users to Step 4: Analytic Tool Selection. • [P]: Navigation status at the bottom highlights that the user is in Step 3: Data Preparation.
Additional notes: • This is a pre-validation stage, giving users a chance to refine their cohort or variable selection before choosing analytic methods. • [P, W, LR]: In line with ethical and data protection best practices (e.g., TEHDAS2) the display of subject-level demographic or clinical data should only be permitted for populations exceeding a minimum threshold (e.g., $n > 10$) to prevent re-identification risks.
Information that supports the user story [P]: • [P]: Previous projects such as IDEA4RC.

Table 40. Create a new derived variable for analysis

37. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Create a new derived variable for analysis
As a Protect child data user
I want to create a new variable by processing existing ones (e.g., combining, transforming, or categorizing them)
So that I can tailor the dataset to reflect clinically meaningful indicators specific to my research question

Initial acceptance criteria [P, E, L]:
• [P]: Users can open a form to create a new variable by selecting:
  o Variable category (e.g., demographic, clinical outcome)
  o One or more source variables
  o Groupings can be created manually by specifying value ranges (e.g., 12–16) and assigning category names (e.g., "Adolescent")
  o Processing method (e.g., arithmetic operation, categorization, transformation)
  o Output type/definition (e.g., continuous, binary, categorical)
  o Name and description of the new variable
• [P]: The created variable appears in the list of selected variables and can be used in subsequent steps.
• [P]: Variable definitions are stored in the workspace for transparency and reproducibility.

• [P]: Users can cancel or save the new variable, with validation to avoid duplicates or invalid inputs. • [P]: Derived variables are traceable in audit logs or metadata exports. • [P]: A preview or confirmation of how the values are grouped is available before finalizing. • [E]: By hovering over the information icon, users can view a brief explanation that helps them understand the type of variable they are working with (e.g., categorical, continuous) and the relevance of specific analyses (e.g., Kaplan analysis for time-to-event data analysis). • [L, E]: I want the platform to include multiple imputations to address missing data in clinical datasets.
Additional notes: • [L]: This aspect still requires additional investigation to be conducted in the iterative phase of T2.1. • [P, W, LR]: In line with ethical and data protection best practices (e.g., TEHDAS2) the display of subject-level demographic or clinical data should only be permitted for populations exceeding a minimum threshold (e.g., $n > 10$) to prevent re-identification risks.
Information that supports the user story [P, E]: • [E]: Conversation with statistician from La Paz. • [P]: Previous projects such as IDEA4RC.

Table 41. Select and execute an analytic tool for data analysis

38. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Select and execute an analytic tool for data analysis
As a Protect child data user
I want to choose and execute an appropriate analytic tool from a list of available algorithms
So that I can apply the correct statistical or machine learning method to answer my research question using the prepared dataset
Initial acceptance criteria [P]: • [P]: Users can click Select methods to choose from a list of preconfigured analytic tools (e.g., survival analysis, clustering, regression models). • [P]: Previously used tools appear in a searchable, sortable list with metadata (e.g., ID, Name, Creation Date, Update Date). • [P]: Each analytic tool includes an Action option (e.g., run, view, delete, duplicate). • [P]: Upon selecting a method, users are guided through its parameter configuration and execution. • [P]: When a method is selected, an information box appears explaining what the method does and how it works. • [P]: Progress is visually tracked in the workflow (Step 4 of 5: Analytic Tool Selection). • [P]: Results from each tool are stored and linked to the workspace for later retrieval or reporting.
Additional notes [E]: • [E]: Any algorithm operating in an SPE must be pre-approved for compliance to guarantee it avoids re-identification, bias, or unlawful profiling, per AI Act requirements.
Information that supports the user story [P]: • [P]: Previous projects such as IDEA4RC. • [E]: Requirements mapped by T2.4 and defined by T2.2.

Table 42. Define and launch a new data analysis

39. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Define and launch a new data analysis
As a Protect child data user

I want to create a new data analysis within my workspace by entering a clear name and description of the objective
So that I can document my research analysis process (e.g., risk scores, immunosuppressant outcome correlations) and manage multiple analysis versions in a structured way
Initial acceptance criteria [P]: ● [P]: Users can create a new analysis by entering a Name and a Brief Description and clicking Create Analysis. ● [P]: Each created analysis is stored and visible in the workspace with metadata including ID, creation date, update date, and name. ● [P]: Users can later edit or delete analysis entries. ● [P]: Analyses are automatically linked to the workspace where they were created. ● [P]: The system supports managing multiple analyses for one project, enabling iterative or comparative research. ● [C]: The platform offers an additional tab in the analysis procedure, where I can see the code in R or python to share my analysis procedure with other experts and statisticians. ● [E]: The additional tab is something similar to R Commander, depending on the user accessing the data. ● [L, C]: The platform includes advanced analytics pipelines capable of handling multimodal datasets (e.g., clinical records, imaging, genomics, lab test results).
Additional notes:
Information that supports the user story [FG, P, E]: ● [FG]: *"We have that the next important step is multimodality data integration. How do you integrate all kinds of data from molecular to imaging to the traditional clinical one" - Big data expert* ● [FG]: *"This is multimodality. So, the, the thing is. That. We would expect to have a lot of data coming from, from different exams."* - Standardization expert ● [P]: Previous projects such as IDEA4RC. ● [E]: Conversation with statistician from La Paz and further refinement with Eugenio.

Table 43. Data analysis

40. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Data analysis
As a Protect child data user
I want to select the algorithm predefined for statistical analysis
so that I can conduct an analysis easily.
Initial acceptance criteria [C, L, P, E]: ● [P]: I want the platform to have pre-defined statistical analysis in the platform so that I can decide which to use. ● [P]: I want the platform to include an analysis for descriptive analysis, including means, medians, standard deviations, IQR frequencies, and proportions to understand demographics and general trends for the patients. ● [P, E]: I want the platform to include a crosstabulation analysis (frequency tables of categorical variables), for instance, to visualize the associations between categories and outcomes. ● [P]: I want the platform to include the Kaplan-Meier analysis to estimate the probability of survival over time. ● [P, E]: I want the platform to include a chi-squared test or Fisher's Exact test to test the independence between categorical variables. ● [P]: I want the platform to include a generalized linear model, such as logistic regression and Poisson regression, to model univariate and binary outcomes or continuous outcomes with different distributions. ● [P, E]: I want the platform to include a CoxPH (COX Proportional Hazards Model) to model how covariates influence survival time without assuming a specific baseline hazard. ● [P]: I want the platform to include a Log-rank test to compare the survival distributions between groups.

- [L]: I want the platform to include t-tests to compare means between two groups for continuous clinical outcomes.
- [E]: I want the platform to exploratory factor analysis (EFA) and Principal Component Analysis (PCA) for the analysis of the genome.
- [L]: I want the platform to include ANOVAs and MANOVAs to compare means across three or more groups.
- [L]: I want the platform to include Mixed Effects Models to account for repeated measures or hierarchical structure in the data.
- [L]: I want the platform to include Propensity Score Matching to reduce confounding in observational comparisons.
- [E, L]: I want the platform to include Nagelkerke's R^2, Harrell's C-statistic, ROC curves and AUC calculations to evaluate the performance of classification models or diagnostic tests.
- [S, C, L]: The platform has integrated some Machine Learning algorithms (i.e., lasso, random forest, gradient boosting) that allow for further identification of patient outcomes.
- [L]: The platform provides model outputs in an interpretable format, such as risk scores, confidence intervals, and visualizations (e.g., survival curves, ROC curves, decision trees, random forest).
- [E, L]: Traceability and reproducibility features are built not only on the logs, but also in code (e.g., R) that I can share with others, to ensure that any modifications or corrections to patient data are easily traced and changed.
- [E, L]: Every statistical test in the interface includes a hover feature that displays information about its purpose and requirements, helping users easily decide which test is most appropriate for their analysis.
- [E]: The platform has data dashboards that allow clinical experts to visualize aggregated patient data and track disease progression over time.

Additional notes [L, E]:

- [L]: The platform should ensure secure access control, allowing only authorized clinical experts and researchers to analyze sensitive patient data while complying with ethical and regulatory guidelines.
- [E]: For federated learning with Vantage6, participating nodes must expose datasets with an identical structure. Therefore, a crucial requirement is that each dataset includes the same features/columns, regardless of missing values. Missing data will not break analysis but can reduce model accuracy. Data harmonization needs to be completed prior to capsule creation. Any transformations (e.g., normalization, encoding) must happen *before* deployment, as Vantage6 assumes pre-processed data.
- [E]: Any algorithm operating in an SPE must be pre-approved for compliance to guarantee it avoids re-identification, bias, or unlawful profiling, per AI Act requirements.
- [E, R]: In line with EHDS Article 73, all data transfers and algorithm run must be recorded and remain auditable by competent authorities for at least 10 years after the project ends.
- [L]: This aspect still requires additional investigation to be conducted in the iterative phase of T2.1

Information that supports the user story [FG, S, E, P]:

- [E]: Conversation with statistician from La Paz.
- [P]: Previous projects such as IDEA4RC.
- [E]: Requirements mapped by T2.4 and defined by T2.2.
- [E]: Clinical protocol documentation.
- [R]: European Parliament & Council of the European Union. (2025, February 11). *Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847*. Official Journal of the European Union, L 63, 5 March 2025.
- [FG]: *"You can classify the users, for example, based on patient characteristics or outcomes to improve your activity as a physician"* **— Pediatrician**
- [FG]: *"Comparing clinical activities across hospitals is important, especially in cases like children transplantation, where variability in management is high."* **— Pediatrician**
- [FG]: *"Data analysis is difficult because we work with a small number of patients, but we generate massive amounts of data that require specialized tools."* **— Pediatrician**
- [FG]: "*It would be quite useful to make more systematic diagnostic process"* **— Pediatrician**
- [FG]: "*It is not easy to find experts in analyzing this data"* **— Pediatrician**
- [FG]: "*We need systematic diagnostic support, helping us validate clinical decisions based on data-driven insights."* **— Pediatrician**

- [S]: How important is it for you to employ the following types of analyses in your research? (Epidemiological analysis): (68%)
- [S]: How important is it for you to employ the following types of analyses in your research? (Predictive modelling) (81%)
- [S]: How important is it for you to employ the following types of analyses in your research? (Statistical analysis) (83%)

Table 44. Predefined Genomic Variant Analysis for Clinical Experts

41. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise: Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics]** Predefined Genomic Variant Analysis for Clinical Experts
As a Protect child data user (mainly Geneticists)
I want the platform to provide me with pre-defined variant analysis
so that I can extract actionable insights from genomic and related datasets to support diagnosis, personalized treatment, and long-term transplant success in pediatric patients.
Initial acceptance criteria [E, L]: • [E]: The platform should also support Hardy–Weinberg equilibrium testing and a Marker exclusion based on QC results for the quality control of the genome. • [E]: Uploaded VCF and BAM files shall be verified against GA4GH and Ensembl schemas by the system before any annotation occurs. • [E]: The platform shall enable federated aggregation of annotated variant statistics (e.g., allele frequencies, functional impact) across multiple SPEs while keeping individual VCF/BAM data local. • [E, W]: The platform should support whole-genome sequencing (WGS) and Copy Number Variations (CNVs) analysis to detect associations between of SNPs (Single Nucleotide Polymorphisms) with outcomes. • [E]: The platform should support access to FASTQ files. • [E]: The platform could include Linear Discriminant Analysis (LDA) for classifying clinical episodes based on methylation profiles. • [L]: The platform provides dashboards showing interpretable outputs (significant SNPs, enriched pathways, CNVs). • [L]: The platform allows validation of sex and sample identity via SNP array data. • [E, L]: The platform supports modular plugins such as ClinVar, gnomAD, dbNSFP, CADD aiming to reproduce VarSeq clinical annotations and enrichment functionalities. • [E, L]: The platform ensures reproducibility, version control, and audit trails.
Additional notes [L, W]: • [W]: Methylomic analysis is run through PCA. • [E]: Any algorithm operating in an SPE must be pre-approved for compliance to guarantee it avoids re-identification, bias, or unlawful profiling, per AI Act requirements. • [E, R]: In line with EHDS Article 73, all data transfers and algorithm run must be recorded and remain auditable by competent authorities for at least 10 years after the project ends. • [E]: Pipelines should be reproductible across nodes. • [L]: This aspect still requires additional investigation to be conducted in the iterative phase of T2.1
Information that supports the user story [W, E]: • [FG]: *"Platforms should help overcome the complexities of genomic data analysis, extracting meaningful insights efficiently."* **— Geneticist** • [FG]: *"Clinical experts don't usually look at raw data; they rely on processed statistics and visualizations to make decisions."* **— Geneticist** • [W]: Presentation by Dr. Lapunzina. • [E]: Requirements mapped by T2. • [E]: Clinical protocol documentation. • [R]: European Parliament & Council of the European Union. (2025, February 11). *Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the*

European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847. Official Journal of the European Union, L 63, 5 March 2025.

Table 45. Methylation (Epigenetic) analysis module

42. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Methylation (Epigenetic) analysis module
As a Protect child data user (mainly Geneticists)
I want the platform to perform and visualize predefined epigenetic analyses (e.g., PCA, GED, Horvath clock)
So that I can assess epigenetic differences across patient groups and extract predictive or age-related biomarkers relevant to pediatric transplantation.
Initial acceptance criteria [E, L]: • [E]: The platform can compare two or more sample groups for methylation differences. • [L]: Supports multiple control types. • [L]: Includes calculation of epigenetic age (Horvath clock). • [E]: Integrates SVM and AI-based predictive modeling. • [E]: Includes a searchable index of CpG island codes. • [E]: Accepts IDAT files and long-read files for input. • [E]: Supports display of legacy epigenetic signatures. • [E]: This module shall generate epigenome-wide methylation matrices and episignature detection results. • [E]: The system shall support reference-based and unsupervised methylation clustering, enabling identification of disease- or transplant-related CpG patterns. • [E]: Federated aggregation of methylation summary data (e.g., mean β-values by gene or CpG island) via the Federated Computing Protocol while keeping patient-level data local. • [E]: Provides a user-friendly visual interface (GUI) instead of command-line (R) operations.
Additional notes [E, R]: • [E]: To be implemented more visual than R-based analysis interface tailored for clinicians. • [E]: Must align with the usability expectations discussed with Genomic Experts within the Protect-Child team. • [E]: Any algorithm operating in an SPE must be pre-approved for compliance to guarantee it avoids re-identification, bias, or unlawful profiling, per AI Act requirements. • [E, R]: In line with EHDS Article 73, all data transfers and algorithm run must be recorded and remain auditable by competent authorities for at least 10 years after the project ends. • [E]: The methylome module shall be containerized and version-controlled (Docker/Singularity) for reproducibility across SPEs and future Federated Computing. • [L]: This aspect still requires additional investigation to be conducted in the iterative phase of T2.1
Information that supports the user story [W]: • [E]: Meeting between Jaime (Genomic expert) and Geneticians. • [E]: Requirements mapped by T2.4. • [E]: Clinical protocol documentation. • [R]: European Parliament & Council of the European Union. (2025, February 11). *Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847*. Official Journal of the European Union, L 63, 5 March 2025.

Table 46. Variant visualization and familial analysis

43. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Variant visualization and familial analysis
As a Protect child data user (mainly Geneticists)
I want to upload variant files (PED) to the Protect-Child platform and visualize them interactively
So that I can confirm variants, assess inheritance patterns, and filter by clinical relevance or phenotype.
Initial acceptance criteria [E, L]: • [E]: Platform allows uploading PED, VCF and BAM files for familial variant visualization. • [E]: Users can upload and manage custom annotation databases. • [E]: Filtering by phenotype and Human Phenotype Ontology (HPO) is supported. • [E]: Differentiates somatic and germline variants, supporting paired normal/tumor files. • [E]: Supports familial/trio analysis via visual family trees (manual, CSV, or newick input). • [E]: Annotates uploaded variants asynchronously (e.g., through ANNOVAR or VEP). • [E]: Integrates IGV or an equivalent web-based viewer for visual confirmation. • [E]: Displays list of available annotation databases and their license status. • [E]: Ensures scalability for large datasets and multiple concurrent visualizations.
Additional notes [L, W]: • [E]: Familial relationships and paired sample encoding must be clarified. • [L]: Visualization effort may vary depending on required format standardization. • [L]: This aspect still requires additional investigation to be conducted in the iterative phase of T2.1
Information that supports the user story [W]: • [E]: Meeting between Jaime and Geneticians.

Table 47. Review analysis results and summary statistics

44. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Review analysis results and summary statistics
As a Protect child data user
I want to view and interpret the output of my selected analysis method, including cohort-based and center-based breakdowns
So that I can extract meaningful insights from the data and evaluate the clinical relevance of my research findings
Initial acceptance criteria [P]: • [P]: The *Analysis Results* tab within the analysis step displays the algorithm used, cohorts compared, and variables analyzed • [P]: Crosstab outputs (e.g., polygenic risk scores) are shown in an interactive table, with totals and percentages • [P]: The *Data Overview* within the analysis step tab provides summary statistics (e.g., survival time, site of risk scores) per cohort and center • [P]: Variables are aggregated and stratified to support interpretation by both clinical and methodological experts • [P]: Visual elements such as bar charts help identify patterns or outliers (e.g., median survival per center) • [P]: A “Back” button allows returning to the analytic configuration for further refinements or comparisons
Additional notes: • [P, W, LR]: In line with ethical and data protection best practices (e.g., TEHDAS2) the display of subject-level demographic or clinical data should only be permitted for populations exceeding a minimum threshold (e.g., $n > 10$) to prevent re-identification risks.

Information that supports the user story [P]: ● [P]: Previous projects such as IDEA4RC.

Table 48. Data analysis [AI driven process]

45. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Data analysis [AI driven process]
As a Protect child data user
I want to use AI to guide my selection of statistical algorithms based on the types of variables in my dataset and my analysis goals,
so that I can efficiently run appropriate statistical tests and interpret results relevant to clinical and genomic research questions.
Initial acceptance criteria [L]: ● [L]: The platform allows users to select data types (e.g., categorical, continuous, binary, genomic, time-series, survival, multivariate). ● [L]: Users can define a desired analytical goal, such as: ○ Correlation between variables ○ Mean comparison between groups ○ Survival association ○ Prediction or classification ○ Multivariate exploration (e.g., clustering) ● [L]: The platform intelligently suggests appropriate statistical analyses based on selected variables and goals, for instance: Parametric and non-parametric tests (e.g., t-test, ANOVA, Mann-Whitney U, Kruskal-Wallis), Correlation metrics (e.g., Pearson, Spearman) or Regression models (e.g., linear, logistic, Cox proportional hazards). ● [L]: The platform includes assumptions and diagnostic checks for suggested tests (e.g., normality, homogeneity of variance). ● [L]: The system supports multi-omics data analysis, allowing integration of gene expression, variant data, and clinical phenotypes. ● [L]: Machine learning analysis pathways are suggested for prediction goals, including classification (e.g., random forest, SVM) and unsupervised learning (e.g., PCA, clustering).
Additional notes [L]: ● [L]: This proposal outlines a user story intended to enhance support for users during the analysis process. ● [L]: This aspect still requires additional investigation to be conducted in the iterative phase of T2.1
Information that supports the user story [W]: ● [W]: Ideally users would like as much support as possible throughout the user journey.

Table 49. Data download

46: Partners involved: SERMAS-LA Paz, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Data download
As a Protect child data user (mainly Geneticists)
I want to extract pseudoanonymized aggregated data from my CoEs' Secure Processing Environments (Data Capsules)
So that I can share data outside PROTECT-CHILD
Initial acceptance criteria [W, E]: ● [W, E]: Aggregated data is available through download in FAIR-compliant formats (e.g., JSON, CSV, OMOP Tables). ● [W]: Data is pseudoanonymized and aggregated before download.

○ In general, the CoE's FHIR Capsule contains pseudonymized data. ● [L, W, E]: Access rights are provided and logged in accordance with the GDPR. ● [L, W]: Download is logged to comply with GDPR requirements. • The platform also allows Data Users analyzing genomic data to download their analyses in the Phenopackets standard (e.g., JSON or Protobuf formats).
Additional notes [W, E]: • [W, E]: Only Data Users from the CoEs can access to the download of aggregated pseudonymized data. • [W]: Agreements and approvals necessary for regulatory compliance (e.g., GDPR mandated data agreements, ethical approvals, etc.) are to be managed outside the scope of PROTECT-CHILD (i.e. without PROTECT-CHILD Data Permit, and according to usual practices). • [L]: To maintain pseudonymization of genomic data, exported Phenopackets should present age as a range, remove any local hospital identifiers, and follow all GDPR-aligned pseudonymization procedures. • [P, W, LR]: In line with ethical and data protection best practices (e.g., TEHDAS2) the display of subject-level demographic or clinical data should only be permitted for populations exceeding a minimum threshold (e.g., $n > 10$) to prevent re-identification risks.
Information that supports the user story [W]: ● [W]: Clinicians would ideally prefer individual anonymized data to further evaluate the history of patients; however, aggregated data is acceptable for now. ● [E]: Requirements mapped by T2.4 and defined by T2.2. ● [E]: Conversation with bioinformatics expert.

Visualization

Table 50. Visualization of data

47. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Visualization of data
As a Protect child data user (mainly Geneticists)
I want the platform to give a graphical visualization of the results from the analysis conducted
so that I can derive insights from it.
Initial acceptance criteria [S, L]: ● [S]: The platform shows graphical results depending on the analysis conducted. ● [S]: The platform provides interactive graphs to support the exploration of statistical and predictive model results. ● [S]: The platform includes charts (e.g., bar charts, pie charts, boxplots) to summarize clinical variables, outcomes, and subgroup distributions. ● [E]: Allows PCA and GED analysis visualization. ● [S]: The platform includes heatmaps to visualize gene expression patterns, variant clustering, or pathway activation. ● [S]: The platform offers genome sequencing viewers or simplified sequencing tracks to help visualize variants in genomic context. ● [S]: The platform supports network visualizations to display gene-gene or protein-protein interactions and pathway relationships. ● [S]: Visualizations are interactive, allow filtering/subsetting of patient cohorts, and display tooltips or annotations for clinical interpretation. ● [S]: All visual outputs are exportable and can be embedded in reports or shared securely with authorized collaborators. ● [S]: Visualizations comply with privacy rules, showing only aggregated or de-identified data where appropriate. ● [E]: The platform also shows the results through tables.

• [L]: The visualization and tables can be exported in multiple formats, like CSV, XSLX, JPG, PNG.
Additional notes: • [L]: This aspect still requires additional investigation to be conducted in the iterative phase of T2.1
Information that supports the user story [FG, S]: • [FG]: *"If you have a specific question, maybe, to obtain a graphical result or even a descriptive result, maybe it's better, but this is just something very personal."* **— Pediatrician** • [S]: How important are the following visualizations? (Graphs) (82%) • [S]: How important are the following visualizations? (Charts) (76%) • [S]: How important are the following visualizations? (Heat map) (72%) • [S]: How important are the following visualizations? (Sequencing) (70%) • [S]: How important are the following visualizations? (Networks) (74%) • [E]: Conversations with bioinformatics expert.

Research finalization

Reporting

Table 51. Generate and finalize a structured report of analysis results

48. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] Generate and finalize a structured report of analysis results
As a Protect child data user
I want to organize and generate a structured, downloadable report of my analysis
So that I can document findings, support clinical decision-making, and share outcomes with collaborators or for publication

Initial acceptance criteria [P, L]:
• [P]: Users can enter content into key sections of the report, including:
  ○ Project Information
  ○ Open Science Information
  ○ Data Information
  ○ Analysis Information
  ○ Results and Findings
• [P]: Report includes a table showing data quality indicators (e.g., completeness, conformance, plausibility) per center and variable.
• [P]: The system compiles this content into a formatted preview document.
• [P]: Users can download, Print, or finalize the report.
• [P]: A confirmation dialog warns that finalization is irreversible.
• [P]: Finalized reports are archived within the workspace with version control.
• [P]: Project dashboard reflects the status (e.g., "Completed") for the Result Report phase.
• [L]: Export options should include multiple formats to ensure flexibility in sharing reports.

Additional notes: • [L]: This aspect still requires additional investigation to be conducted in the iterative phase of T2.1
• **Information that supports the user story [S]:** • [S]: How important is it for you to share the knowledge produced through the platform with Researchers? (*89%)* • [S]: How important is it for you to share the knowledge produced through the platform with clinical experts? (87*%)*

• [S]: How important is it for you to share the knowledge produced through the platform with the Ethical board? (*69%)* • [S]: How important is it for you to share the knowledge produced through the platform with Policy makers? (*67%)* • [S]: How important is it for you to share the knowledge produced through the platform with patients? (*62%)*

Non-functional requirements

Communication & Support

Table 52. Capsule support

50. Partners involved: SERMAS-LA PAZ, UNIPD, UKE, ISMETT, UPM **Expertise:** Data holders (Managers of the platform) Capsule support
As a Protect child data holder
I want to have documentation and online support to install and update the data capsule
So that I can manage the Protect Child platform and update the data sources.
Initial acceptance criteria [L, E]: • [L]: Documentation and scripts for installing and updating the data capsule are made available and maintained by the central technical team.
Additional notes [L]: • [L]: This aspect still requires additional investigation to be conducted in the iterative phase of T2.1
Information that supports the user story [E]: • [E]: Conversations with technical experts and documentation made for other project tasks.

User experience (UX) and user interface (UI) characteristics

Table 53. UX/UI

52. Partners involved: SERMAS, INETUM, ISMETT, UKE, UNIPD. **Expertise:** Data users [Clinical experts such as pediatricians, hepatologists, geneticists, and lab scientists], researchers [e.g., statisticians, bioinformatics] UX/UI
As a Protect child data user (mainly Geneticists)
I want the platform to provide an intuitive and user-friendly interface for querying cohort data
so that I can efficiently and intuitively discover metadata, organize my research project, request data access, conduct a pre-processing and data analysis and finalize my research project
Initial acceptance criteria [L]: • [L]: Clear and structured UI ensures that even non-technical users can navigate and retrieve insights effortlessly. • [W]: Make the colors accessible for different users. • [W]: Include maps and visual whenever possible.
Additional notes: • [L]: This aspect still requires additional investigation to be conducted in the iterative phase of T2.1
Information that supports the user story [W, FG]: • [FG]: *"Friendly is quite important"* **— Pediatrician** • [FG]: *"If you open something too complex, you close it"* **— Lab scientist**

- [W]: Accessibility would help to make it more usable.
- [W]: *"I would like a map of the different centers to see where the data is located"* -**Data manager**

7. References

1. Albuquerque, Marco A., Bruno Grande, Elie J. Ritch, Prasath Pararajalingam, Selin Jessa, Martin Krzywinski, Jasleen K. Grewal, Sohrab P. Shah, Paul C. Boutros, and Ryan D. Morin. 2017. "Enhancing Knowledge Discovery from Cancer Genomics Data with Galaxy." *GigaScience* 6(5):1–13.
2. Alzu'bi, Amal, and Leming Zhou. 2017. "An Integrated Patient Genomic Information Management and Analysis System for Healthcare Professionals" edited by C. GiraudCarrier, J. Facelli, H. Nakajima, M. Cummins, and G. Meixner. *2017 IEEE INTERNATIONAL CONFERENCE ON HEALTHCARE INFORMATICS (ICHI)* 107–13.
3. Alzu'bi, Amal, Leming Zhou, and Valerie Watzlaf. 2014. "Personal Genomic Information Management and Personalized Medicine: Challenges, Current Solutions, and Roles of Him Professionals." *Perspectives in Health Information Management* 11(Spring):1c.
4. Amer-Yahia, Sihem, Georgia Koutrika, Martin Braschler, Diego Calvanese, Davide Lanti, Hendrik Lücke-Tieke, Alessandro Mosca, Tarcisio Mendes De Farias, Dimitris Papadopoulos, Yogendra Patil, Guillem Rull, Ellery Smith, Dimitrios Skoutas, Srividya Subramanian, and Kurt Stockinger. 2022. "INODE: Building an End-to-End Data Exploration System in Practice." *ACM SIGMOD Record* 50(4):23–29. doi: 10.1145/3516431.3516436.
5. Arneson, Douglas, Anindya Bhattacharya, Le Shu, Ville-Petteri Mäkinen, and Xia Yang. 2016. "Mergeomics: A Web Server for Identifying Pathological Pathways, Networks, and Key Regulators via Multidimensional Data Integration." *BMC Genomics* 17(1):722. doi: 10.1186/s12864-016-3057-8.
6. Agrawal, Shruti, Michal M. Placek, Deborah White, Esther Daubney, Manuel Cabeleira, Peter Smielewski, Marek Czosnyka, Adam Young, Suzanna Watson, Anna Maw, and Peter John Hutchinson. 2023. "Studying Trends of Auto-Regulation in Severe Head Injury in Paediatrics (STARSHIP): Protocol to Study Cerebral Autoregulation in a Prospective Multicentre Observational Research Database Study." *BMJ Open* 13(3):e071800. doi:10.1136/bmjopen-2023-071800.
7. Bhuvaneshwar, K., A. Belouali, V. Singh, R. M. Johnson, L. Song, A. Alaoui, M. A. Harris, R. Clarke, L. M. Weiner, Y. Gusev, and S. Madhavan. 2016. "G-DOC Plus. An Integrative Bioinformatics Platform for Precision Medicine." *BMC Bioinformatics* 17(1). doi: 10.1186/s12859-016-1010-0.
8. Bonomi, Luca, Yingxiang Huang, and Lucila Ohno-Machado. 2020. "Privacy Challenges and Research Opportunities for Genomic Data Sharing." *Nature Genetics* 52(7):646–54. doi: 10.1038/s41588-020-0651-0.
9. Calabria, Andrea, Giulio Spinozzi, Fabrizio Benedicenti, Erika Tenderini, and Eugenio Montini. 2015. "adLIMS: A Customized Open Source Software That Allows Bridging Clinical and Basic Molecular Research Studies." *BMC Bioinformatics* 16(S9):S5. doi: 10.1186/1471-2105-16-S9-S5.
10. Caliskan, Yasar, Brian Lee, Adrian M. Whelan, Fadee Abualrub, Krista L. Lentine, and Arksarapuk Jittirat. 2022. "Evaluation of Genetic Kidney Diseases in Living Donor Kidney Transplantation: Towards Precision Genomic Medicine in Donor Risk Assessment." *Current Transplantation Reports* 9(2):127–42. doi: 10.1007/s40472-021-00340-3.
11. Campbell, E. M., A. Boyles, A. Shankar, J. Kim, S. Knyazev, R. Cintron, and W. M. Switzer. 2021. "MicrobeTrace: Retooling Molecular Epidemiology for Rapid Public Health Response." *PLoS Computational Biology* 17(9). doi: 10.1371/journal.pcbi.1009300.
12. Canakoglu, Arif, Pietro Pinoli, Andrea Gulino, Luca Nanni, Marco Masseroli, and Stefano Ceri. 2021. "Federated Sharing and Processing of Genomic Datasets for Tertiary Data Analysis." *Briefings in Bioinformatics* 22(3):bbaa091. doi: 10.1093/bib/bbaa091.
13. Cappelli, E., F. Cumbo, A. Bernasconi, A. Canakoglu, S. Ceri, M. Masseroli, and E. Weitschek. 2020. "OpenGDC: Unifying, Modeling, Integrating Cancer Genomic Data and Clinical Metadata." *Applied Sciences* 10(18). doi: 10.3390/APP10186367.
14. Cerami, Ethan, Jianjiong Gao, Ugur Dogrusoz, Benjamin E. Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J. Byrne, Michael L. Heuer, Erik Larsson, Yevgeniy Antipin, Boris Reva, Arthur P. Goldberg, Chris Sander, and Nikolaus Schultz. 2012. "The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data." *Cancer Discovery* 2(5):401–4. doi: 10.1158/2159-8290.CD-12-0095.
15. Chen, I.-Min A., Victor M. Markowitz, Krishna Palaniappan, Ernest Szeto, Ken Chu, Jinghua Huang, Anna Ratner, Manoj Pillay, Michalis Hadjithomas, Marcel Huntemann, Natalia Mikhailova, Galina Ovchinnikova, Natalia N. Ivanova, and Nikos C. Kyrpides. 2016. "Supporting Community Annotation and User Collaboration in the Integrated Microbial Genomes (IMG) System." *BMC Genomics* 17(1):307. doi: 10.1186/s12864-016-2629-y.
16. Cleland-Huang, Jane, Raffaella Settimi, Xuchang Zou, and Peter Solc. 2006. "The Detection and

Classification of Non-Functional Requirements with Application to Early Aspects.” Pp. 39–48 in *14th IEEE International Requirements Engineering Conference (RE’06)*.

17. Crawford, Katherine M., Cristina Gallego-Fabrega, Christina Kourkoulis, Laura Miyares, Sandro Marini, Jason Flannick, Noel P. Burtt, Marcin Von Grotthuss, Benjamin Alexander, Maria C. Costanzo, Neil H. Vaishnav, Rainer Malik, Jennifer L. Hall, Michael Chong, Jonathan Rosand, and Guido J. Falcone. 2018. “Cerebrovascular Disease Knowledge Portal: An Open-Access Data Resource to Accelerate Genomic Discoveries in Stroke.” *Stroke* 49(2):470–75. doi: 10.1161/STROKEAHA.117.018922.
18. Danahey, Keith, Brittany A. Borden, Brian Furner, Patrick Yukman, Sheena Hussain, Donald Saner, Samuel L. Volchenboum, Mark J. Ratain, and Peter H. O’Donnell. 2017. “Simplifying the Use of Pharmacogenomics in Clinical Practice: Building the Genomic Prescribing System.” *Journal of Biomedical Informatics* 75:110–21. doi: 10.1016/j.jbi.2017.09.012.
19. Das, S., X. Lecours Boucher, C. Rogers, C. Makowski, F. Chouinard-Decorte, K. Oros Klein, N. Beck, P. Rioux, S. T. Brown, Z. Mohaddes, C. Zweber, V. Foing, M. Forest, K. J. O’Donnell, J. Clark, M. J. Meaney, C. M. T. Greenwood, and A. C. Evans. 2018. “Integration of Omics Data and Phenotypic Data within a Unified Extensible Multimodal Framework.” *Frontiers in Neuroinformatics* 12. doi: 10.3389/fninf.2018.00091.
20. Davis-Turak, Jeremy, Sean M. Courtney, E. Starr Hazard, W. Bailey Glen, Willian A. Da Silveira, Timothy Wesselman, Larry P. Harbin, Bethany J. Wolf, Dongjun Chung, and Gary Hardiman. 2017. “Genomics Pipelines and Data Integration: Challenges and Opportunities in the Research Setting.” *Expert Review of Molecular Diagnostics* 17(3):225–37. doi: 10.1080/14737159.2017.1282822.
21. Demchak, B., T. Hull, M. Reich, T. Liefeld, M. Smoot, T. Ideker, and J. P. Mesirov. 2014. “Cytoscape: The Network Visualization Tool for GenomeSpace Workflows.” *F1000Research* 3. doi: 10.12688/f1000research.4492.2.
22. Duan, Wenjing, Yongguang Yang, Jingge Zhao, Tianzhong Yan, and Xiangyong Tian. 2022. “Application of Metagenomic Next-Generation Sequencing in the Diagnosis and Treatment of Recurrent Urinary Tract Infection in Kidney Transplant Recipients.” *Frontiers in Public Health* 10:901549. doi: 10.3389/fpubh.2022.901549.
23. Duong, Benjamin Q., Meghan J. Arwood, J. Kevin Hicks, Amber L. Beitelshees, Francesco Franchi, John T. Houder, Nita A. Limdi, Kelsey J. Cook, Aniwaa Owusu Obeng, Natasha Petry, Sony Tuteja, Amanda R. Elsey, Larisa H. Cavallari, and Kristin Wiisanen. 2020. “Development of Customizable Implementation Guides to Support Clinical Adoption of Pharmacogenomics: Experiences of the Implementing Genomics in Practice (IGNITE) Network.” *Pharmacogenomics and Personalized Medicine* Volume 13:217–26. doi: 10.2147/PGPM.S241599.
24. European Parliament, and European Council. 2016. *General Data Protection Regulation (GDPR)*.
25. European Parliament, and European Council. 2022. *Proposal for a Regulation of the European Parliament and of the Council on the European Health Data Space*.
26. Fahr, Patrick, James Buchanan, and Sarah Wordsworth. 2019. “A Review of the Challenges of Using Biomedical Big Data for Economic Evaluations of Precision Medicine.” *Applied Health Economics and Health Policy* 17(4):443–52. doi: 10.1007/s40258-019-00474-7.
27. Forsdick, Natalie J., Jana Wold, Anton Angelo, François Bissey, Jamie Hart, Mitchell Head, Libby Liggins, Dinindu Senanayake, and Tammy E. Steeves. 2023. “Journeying towards Best Practice Data Management in Biodiversity Genomics.” in *Molecular Ecology Resources*. Vol. n/a.
28. Gilbert, R. M., D. Sumodhee, N. Pontikos, C. Hollyhead, A. Patrick, S. Scarles, S. Van Der Smissen, R. M. Young, N. Nettleton, A. R. Webster, and J. Cammack. 2022. “Collaborative Research and Development of a Novel, Patient-Centered Digital Platform (Myeyesite) for Rare Inherited Retinal Disease Data: Acceptability and Feasibility Study.” *JMIR Formative Research* 6(1). doi: 10.2196/21341.
29. Gill, Ivan S., Emma J. Griffiths, Damion Dooley, Rhiannon Cameron, Sarah Savić Kallesøe, Nithu Sara John, Anoosha Sehar, Gurinder Gosal, David Alexander, Madison Chapel, Matthew A. Croxen, Benjamin Delisle, Rachelle Di Tullio, Daniel Gaston, Ana Duggan, Jennifer L. Guthrie, Mark Horsman, Esha Joshi, Levon Kearny, Natalie Knox, Lynette Lau, Jason J. LeBlanc, Vincent Li, Pierre Lyons, Keith MacKenzie, Andrew G. McArthur, Emily M. Panousis, John Palmer, Natalie Prystajecky, Kerri N. Smith, Jennifer Tanner, Christopher Townend, Andrea Tyler, Gary Van Domselaar, and William W. L. Hsiao. 2023. “The dataHarmonizer: A Tool for Faster Data Harmonization, Validation, Aggregation and Analysis of Pathogen Genomics Contextual Information.” *Microbial Genomics* 9(1):000908. doi: 10.1099/mgen.0.000908.
30. Glinz, Martin. 2007. “On Non-Functional Requirements.” Pp. 21–26 in *15th IEEE International Requirements Engineering Conference (RE 2007)*.
31. Holtgrewe, Manuel, Oliver Stolpe, Mikko Nieminen, Stefan Mundlos, Alexej Knaus, Uwe Kornak, Dominik Seelow, Lara Segebrecht, Malte Spielmann, Björn Fischer-Zirnsak, Felix Boschann, Ute Scholl, Nadja Ehmke, and Dieter Beule. 2020. “VarFish: Comprehensive DNA Variant Analysis for

Diagnostics and Research." *Nucleic Acids Research* 48(W1):W162–69. doi: 10.1093/nar/gkaa241.
32. Hombach, Daniela, Markus Schuelke, Ellen Knierim, Nadja Ehmke, Jana Marie Schwarz, Björn Fischer-Zirnsak, and Dominik Seelow. 2019. "MutationDistiller: User-Driven Identification of Pathogenic DNA Variants." *Nucleic Acids Research* 47(W1):W114–20. doi: 10.1093/nar/gkz330.
33. Ienca, M., Ferretti, A., Hurst, S., Puhan, M., Lovis, C., & Vayena, E. (2018). Considerations for ethics review of big data health research: A scoping review. *PloS one*, 13(10), e0204937.
34. Jetten, Mijke, Marjan Grootveld, Annemie Mordant, Mascha Jansen, Margreet Bloemers, Margriet Miedema, and Celia W. G. Van Gelder. 2021. *Professionalising Data Stewardship in the Netherlands. Competences, Training and Education. Dutch Roadmap towards National Implementation of FAIR Data Stewardship*. Zenodo. doi:10.5281/ZENODO.4486423.
35. Kho, Abel N., Luke V. Rasmussen, John J. Connolly, Peggy L. Peissig, Justin Starren, Hakon Hakonarson, and M. Geoffrey Hayes. 2013. "Practical Challenges in Integrating Genomic Data into the Electronic Health Record." *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 15(10):772–78. doi: 10.1038/gim.2013.131.
36. Kounelis, F., A. Kanterakis, A. Kanavos, M. T. Pandi, Z. Kordou, O. Manusama, G. Vonitsanos, T. Katsila, E. E. Tsermpini, V. M. Lauschke, M. Koromina, P. J. van der Spek, and G. P. Patrinos. 2020. "Documentation of Clinically Relevant Genomic Biomarker Allele Frequencies in the Next-Generation FINDbase Worldwide Database." *Human Mutation* 41(6):1112–22. doi: 10.1002/humu.24018.
37. Kuzmenkov, A. Y., I. V. Trushin, A. G. Vinogradova, A. A. Avramenko, M. V. Sukhorukova, S. Malhotra-Kumar, A. V. Dekhnich, M. V. Edelstein, and R. S. Kozlov. 2021. "AMRmap: An Interactive Web Platform for Analysis of Antimicrobial Resistance Surveillance Data in Russia." *Frontiers in Microbiology* 12. doi: 10.3389/fmicb.2021.620002.
38. Lau, Jessica W., Erik Lehnert, Anurag Sethi, Raunaq Malhotra, Gaurav Kaushik, Zeynep Onder, Nick Groves-Kirkby, Aleksandar Mihajlovic, Jack DiGiovanna, Mladen Srdic, Dragan Bajcic, Jelena Radenkovic, Vladimir Mladenovic, Damir Krstanovic, Vladan Arsenijevic, Djordje Klisic, Milan Mitrovic, Igor Bogicevic, Deniz Kural, and Brandi Davis-Dusenbery. 2017. "The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized. A New Paradigm in Large-Scale Computational Research." *Cancer Research* 77(21):e3–6. doi: 10.1158/0008-5472.CAN-17-0387.
39. Li, Lishi, Yunyun An, Li Ma, Mengqi Yang, Pengxiang Yuan, Xiaojian Liu, Xin Jin, Yu Zhao, Songfa Zhang, Xin Hong, and Kun Sun. 2022. "Msuite2: All-in-One DNA Methylation Data Analysis Toolkit with Enhanced Usability and Performance." *Computational and Structural Biotechnology Journal* 20:1271–76.
40. Li, N. T., N. C. Wu, R. Cao, J. L. Cadavid, S. Latour, X. Lu, Y. Zhu, M. Mijalkovic, R. Roozitalab, N. Landon-Brace, F. Notta, and A. P. McGuigan. 2022. "An Off-the-Shelf Multi-Well Scaffold-Supported Platform for Tumour Organoid-Based Tissues." *Biomaterials* 291. doi: 10.1016/j.biomaterials.2022.121883.
41. Li, Yanni. 2021. *Genetic and Environmental Effects on Liver Diseases*. [Groningen]: University of Groningen.
42. Ma, Clement, Madhumitha Sridharan, Hasan Al-Sayegh, Anran Li, Dongjing Guo, Mark Auclair, Vani Kuragayala, Chandini Bandaru, Dana Milne, Hugh Cruse, Ron Beaudoin, John Orechia, Jonathan Bickel, and Wendy B. London. 2021. "Building a Harmonized Datamart by Integrating Cross-Institutional Systems of Clinical, Outcome, and Genomic Data: The Pediatric Patient Informatics Platform (PPIP)." *JCO Clinical Cancer Informatics* (5):202–15. doi: 10.1200/CCI.20.00083.
43. Martínez-García, M., & Hernández-Lemus, E. (2022). Data integration challenges for machine learning in precision medicine. *Frontiers in medicine*, 8, 784455.
44. McLeod, Clay, Alexander M. Gout, Xin Zhou, Andrew Thrasher, Delaram Rahbarinia, Samuel W. Brady, Michael Macias, Kirby Birch, David Finkelstein, Jobin Sunny, Rahul Mudunuri, Brent A. Orr, Madison Treadway, Bob Davidson, Tracy K. Ard, Arthur Chiao, Andrew Swistak, Stephanie Wiggins, Scott Foy, Jian Wang, Edgar Sioson, Shuoguo Wang, J. Robert Michael, Yu Liu, Xiaotu Ma, Aman Patel, Michael N. Edmonson, Mark R. Wilkinson, Andrew M. Frantz, Ti-Cheng Chang, Liqing Tian, Shaohua Lei, S. M. Ashiqul Islam, Christopher Meyer, Naina Thangaraj, Pamella Tater, Vijay Kandali, Singer Ma, Tuan Nguyen, Omar Serang, Irina McGuire, Nedra Robison, Darrell Gentry, Xing Tang, Lance E. Palmer, Gang Wu, Ed Suh, Leigh Tanner, James McMurry, Matthew Lear, Alberto S. Pappo, Zhaoming Wang, Carmen L. Wilson, Yong Cheng, Soheil Meshinchi, Ludmil B. Alexandrov, Mitchell J. Weiss, Gregory T. Armstrong, Leslie L. Robison, Yutaka Yasui, Kim E. Nichols, David W. Ellison, Chaitanya Bangur, Charles G. Mullighan, Suzanne J. Baker, Michael A. Dyer, Geralyn Miller, Scott Newman, Michael Rusch, Richard Daly, Keith Perry, James R. Downing, and Jinghui Zhang. 2021. "St. Jude Cloud: A Pediatric Cancer Genomic Data-Sharing Ecosystem." *Cancer Discovery* 11(5):1082–99. doi: 10.1158/2159-8290.CD-20-1230.
45. Mohr, Christopher, Andreas Friedrich, David Wojnar, Erhan Kenar, Aydin Can Polatkan, Marius Cosmin

Codrea, Stefan Czemmel, Oliver Kohlbacher, and Sven Nahnsen. 2018. “qPortal: A Platform for Data-Driven Biomedical Research” edited by F. Lisacek. *PLOS ONE* 13(1):e0191603. doi: 10.1371/journal.pone.0191603.

46. Murtagh, Madeleine J., Andrew Turner, Joel T. Minion, Michaela Fay, and Paul R. Burton. 2016. “International Data Sharing in Practice: New Technologies Meet Old Governance.” *Biopreservation and Biobanking* 14(3):231–40. doi: 10.1089/bio.2016.0002.
47. Nanni, Luca, Pietro Pinoli, Arif Canakoglu, and Stefano Ceri. 2019. “PyGMQL: Scalable Data Extraction and Analysis for Heterogeneous Genomic Datasets.” *BMC Bioinformatics* 20(1):560. doi: 10.1186/s12859-019-3159-9.
48. Ochoa, D., A. Hercules, M. Carmona, D. Suveges, A. Gonzalez-Uriarte, C. Malangone, A. Miranda, L. Fumis, D. Carvalho-Silva, M. Spitzer, J. Baker, J. Ferrer, A. Raies, O. Razuvayevskaya, A. Faulconbridge, E. Petsalaki, P. Mutowo, S. MacHlitt-Northen, G. Peat, E. McAuley, C. K. Ong, E. Mountjoy, M. Ghoussaini, A. Pierleoni, E. Papa, M. Pignatelli, G. Koscielny, M. Karim, J. Schwartzentruber, D. G. Hulcoop, I. Dunham, and E. M. McDonagh. 2021. “Open Targets Platform: Supporting Systematic Drug-Target Identification and Prioritisation.” *Nucleic Acids Research* 49(D1):D1302–10. doi: 10.1093/nar/gkaa1027.
49. Ochoa, David, Andrew Hercules, Miguel Carmona, Daniel Suveges, Jarrod Baker, Cinzia Malangone, Irene Lopez, Alfredo Miranda, Carlos Cruz-Castillo, Luca Fumis, Manuel Bernal-Llinares, Kirill Tsukanov, Helena Cornu, Konstantinos Tsirigos, Olesya Razuvayevskaya, Annalisa Buniello, Jeremy Schwartzentruber, Mohd Karim, Bruno Ariano, Ricardo Esteban Martinez Osorio, Javier Ferrer, Xiangyu Ge, Sandra Machlitt-Northen, Asier Gonzalez-Uriarte, Shyamasree Saha, Santosh Tirunagari, Chintan Mehta, Juan María Roldán-Romero, Stuart Horswell, Sarah Young, Maya Ghoussaini, David G. Hulcoop, Ian Dunham, and Ellen M. McDonagh. 2023. “The Next-Generation Open Targets Platform: Reimagined, Redesigned, Rebuilt.” *Nucleic Acids Research* 51(D1):D1353–59. doi: 10.1093/nar/gkac1046.
50. Osmond, Matthew, Taila Hartley, David A. Dyment, Kristin D. Kernohan, Michael Brudno, Orion J. Buske, A. Micheil Innes, Kym M. Boycott, Kym Boycott, Michael Brudno, Francois Bernier, Clara van Karnebeek, David Dyment, Kristin Kernohan, Micheil Innes, Ryan Lamont, Jillian Parboosingh, Deborah Marshall, Christian Marshall, Roberto Mendoza, James Dowling, Robin Hayeems, Bartha Knoppers, Anna Lehman, and Sara Mostafavi. 2022. “Outcome of over 1500 Matches through the Matchmaker Exchange for Rare Disease Gene Discovery: The 2-Year Experience of Care4Rare Canada.” *Genetics in Medicine* 24(1):100–108. doi: 10.1016/j.gim.2021.08.014.
51. Quinn, P., Ellyne, E., & Yao, C. (2024). Will the GDPR restrain health data access bodies under the European Health Data Space (EHDS)? *Computer Law & Security Review*, 54, 105993.
52. Pang, Z., J. Chong, G. Zhou, D. A. De Lima Morais, L. Chang, M. Barrette, C. Gauthier, P. E. Jacques, S. Li, and J. Xia. 2021. “MetaboAnalyst 5.0: Narrowing the Gap between Raw Spectra and Functional Insights.” *Nucleic Acids Research* 49(W1):W388–96. doi: 10.1093/nar/gkab382.
53. Pavia, M. J., A. Chede, Z. Wu, H. Cadillo-Quiroz, and Q. Zhu. 2023. “BinaRena: A Dedicated Interactive Platform for Human-Guided Exploration and Binning of Metagenomes.” *Microbiome* 11(1). doi: 10.1186/s40168-023-01625-8.
54. Pavlenko, E., Strech, D., & Langhof, H. (2020). Implementation of data access and use procedures in clinical data warehouses. A systematic review of literature and publicly available policies. *BMC medical informatics and decision making*, 20(1), 157.
55. Pearce, Thomas M., Marina N. Nikiforova, and Somak Roy. 2019. “Interactive Browser-Based Genomics Data Visualization Tools for Translational and Clinical Laboratory Applications.” *The Journal of Molecular Diagnostics* 21(6):985–93. doi: 10.1016/j.jmoldx.2019.06.005.
56. Post, Andrew R., Nancy Ho, Erik Rasmussen, Ivan Post, Aika Cho, John Hofer, Arthur T. Maness, Timothy Parnell, and David A. Nix. 2023. “Hypermedia-Based Software Architecture Enables Test-Driven Development (TDD).” *JAMIA Open* 6(4).
57. Prasser, F., Spengler, H., Bild, R., Eicher, J., & Kuhn, K. A. (2019). Privacy-enhancing ETL-processes for biomedical data. *International journal of medical informatics*, 126, 72-81.
58. Ramirez, Sergio, Antonio Munoz-Merida, Johan Karlsson, Maximiliano Garcia, Antonio J. Perez-Pulido, M. Gonzalo Claros, and Oswaldo Trelles. 2010. “MOWServ: A Web Client for Integration of Bioinformatic Resources.” *Nucleic Acids Research* 38(2):W671–76.
59. Rao, Shruti, Beth Pitel, Alex H. Wagner, Simina M. Boca, Matthew McCoy, Ian King, Samir Gupta, Ben Ho Park, Jeremy L. Warner, James Chen, Peter K. Rogan, Debyani Chakravarty, Malachi Griffith, Obi L. Griffith, and Subha Madhavan. 2020. “Collaborative, Multidisciplinary Evaluation of Cancer Variants Through Virtual Molecular Tumor Boards Informs Local Clinical Practices.” *JCO Clinical Cancer Informatics* (4):602–13. doi: 10.1200/CCI.19.00169.
60. Raudvere, U., L. Kolberg, I. Kuzmin, T. Arak, P. Adler, H. Peterson, and J. Vilo. 2019. “G:Profiler: A

Web Server for Functional Enrichment Analysis and Conversions of Gene Lists (2019 Update).” *Nucleic Acids Research* 47(W1):W191–98. doi: 10.1093/nar/gkz369.

61. Raveendran, K., N. H. Freese, C. Kintali, S. Tiwari, P. Bole, C. Dias, and A. E. Loraine. 2022. “BioViz Connect: Web Application Linking Cyverse Cloud Resources to Genomic Visualization in the Integrated Genome Browser.” *Frontiers in Bioinformatics* 2. doi: 10.3389/fbinf.2022.764619.
62. Rawashdeh, Adnan, and Bassem Matalkah. 2006. “A New Software Quality Model for Evaluating COTS Components.” *Journal of Computer Science* 2(4):373–81. doi: 10.3844/jcssp.2006.373.381.
63. Reiff, Sarah B., Andrew J. Schroeder, Koray Kırlı, Andrea Cosolo, Clara Bakker, Luisa Mercado, Soohyun Lee, Alexander D. Veit, Alexander K. Balashov, Carl Vitzthum, William Ronchetti, Kent M. Pitman, Jeremy Johnson, Shannon R. Ehmsen, Peter Kerpedjiev, Nezar Abdennur, Maxim Imakaev, Serkan Utku Öztürk, Uğur Çamoğlu, Leonid A. Mirny, Nils Gehlenborg, Burak H. Alver, and Peter J. Park. 2022. “The 4D Nucleome Data Portal as a Resource for Searching and Visualizing Curated Nucleomics Data.” *Nature Communications* 13(1):2365. doi: 10.1038/s41467-022-29697-4.
64. Reska, Daniel, Marcin Czajkowski, Krzysztof Jurczuk, Cezary Boldak, Wojciech Kwedlo, Witold Bauer, Jolanta Koszelew, and Marek Kretowski. 2021. “Integration of Solutions and Services for Multi-Omics Data Analysis towards Personalized Medicine.” *Biocybernetics and Biomedical Engineering* 41(4):1646–63. doi: 10.1016/j.bbe.2021.10.005.
65. Rodchenkov, I., O. Babur, A. Luna, B. A. Aksoy, J. V. Wong, D. Fong, M. Franz, M. C. Siper, M. Cheung, M. Wrana, H. Mistry, L. Mosier, J. Dlin, Q. Wen, C. O’Callaghan, W. Li, G. Elder, P. T. Smith, C. Dallago, E. Cerami, B. Gross, U. Dogrusoz, E. Demir, G. D. Bader, and C. Sander. 2020. “Pathway Commons 2019 Update: Integration, Analysis and Exploration of Pathway Data.” *Nucleic Acids Research* 48(D1):D489–97. doi: 10.1093/nar/gkz946.
66. Sante, Tom, Sarah Vergult, Pieter-Jan Volders, Wigard P. Kloosterman, Geert Trooskens, Katleen De Preter, Annelies Dheedene, Frank Speleman, Tim De Meyer, and Björn Menten. 2014. “ViVar: A Comprehensive Platform for the Analysis and Visualization of Structural Genomic Variation” edited by F. Lisacek. *PLoS ONE* 9(12):e113800. doi: 10.1371/journal.pone.0113800.
67. Sauria, Michael E. G., Jennifer E. Phillips-Cremins, Victor G. Corces, and James Taylor. 2015. “HiFive: A Tool Suite for Easy and Efficient HiC and 5C Data Analysis.” *GENOME BIOLOGY* 16.
68. Sommerville, Ian. 2016. *Software Engineering*. Tenth edition. Boston Columbus Indianapolis New York San Francisco Hoboken Amsterdam Cape Town Dubai London: Pearson.
69. Staunton, Ciara, Mahsa Shabani, Deborah Mascalzoni, Signe Mežinska, and Santa Slokenberga. 2024. “Ethical and Social Reflections on the Proposed European Health Data Space.” *European Journal of Human Genetics* 32(5):498–505. doi:10.1038/s41431-024-01543-9.
70. Strauss, Alexandra T., Carolyn N. Sidoti, Tanjala S. Purnell, Hannah C. Sung, John W. Jackson, Scott Levin, Vedant S. Jain, Daniel Malinsky, Dorry L. Segev, James P. Hamilton, Jacqueline Garonzik-Wang, Stephen H. Gray, Macey L. Levan, Joseph R. Scalea, Andrew M. Cameron, Ahmet Gurakar, and Ayse P. Gurses. 2022. “Multicenter Study of Racial and Ethnic Inequities in Liver Transplantation Evaluation: Understanding Mechanisms and Identifying Solutions.” *Liver Transplantation* 28(12):1841. doi: 10.1002/lt.26532.
71. Staunton, Ciara, Mahsa Shabani, Deborah Mascalzoni, Signe Mežinska, and Santa Slokenberga. 2024. “Ethical and Social Reflections on the Proposed European Health Data Space.” *European Journal of Human Genetics* 32(5):498–505. doi:10.1038/s41431-024-01543-9.
72. Suciu, Radu M., Emir Aydin, and Brian E. Chen. 2015. “GeneDig: A Web Application for Accessing Genomic and Bioinformatics Knowledge.” *BMC Bioinformatics* 16(1):67. doi: 10.1186/s12859-015-0497-0.
73. Takai-Igarashi, Takako, Kengo Kinoshita, Masao Nagasaki, Soichi Ogishima, Naoki Nakamura, Sachiko Nagase, Satoshi Nagaie, Tomo Saito, Fuji Nagami, Naoko Minegishi, Yoichi Suzuki, Kichiya Suzuki, Hiroaki Hashizume, Shinichi Kuriyama, Atsushi Hozawa, Nobuo Yaegashi, Shigeo Kure, Gen Tamiya, Yoshio Kawaguchi, Hiroshi Tanaka, and Masayuki Yamamoto. 2017. “Security Controls in an Integrated Biobank to Protect Privacy in Data Sharing: Rationale and Study Design.” *BMC Medical Informatics and Decision Making* 17(1):100. doi: 10.1186/s12911-017-0494-5.
74. Taraborelli, Dario. 2003. “Feature Binding and Object Perception. Does Object Awareness Require Feature Conjunction?”
75. Ullah, S., F. Ullah, W. Rahman, D. A. Karras, A. Ullah, G. Ahmad, M. Ijaz, and T. Gao. 2022. “The Cancer Research Database (CRDB): Integrated Platform to Gain Statistical Insight into the Correlation between Cancer and COVID-19.” *JMIR Cancer* 8(2). doi: 10.2196/35020.
76. Verbruggen, Steven, and Gerben Menschaert. 2019. “mQC: A Post-Mapping Data Exploration Tool for Ribosome Profiling.” *Computer Methods and Programs in Biomedicine* 181:104806. doi: 10.1016/j.cmpb.2018.10.018.
77. Ware, A. P., K. Satyamoorthy, and B. Paul. 2024. “CmirC Update 2024: A Multi-Omics Database for

Clustered miRNAs." *Functional & Integrative Genomics* 24(4):133. doi: 10.1007/s10142-024-01410-2.
78. Warner, Jeremy L., Ishaan Prasad, Makiah Bennett, Monica Arniella, Alicia Beeghly-Fadiel, Kenneth D. Mandl, and Gil Alterovitz. 2018. "SMART Cancer Navigator: A Framework for Implementing Asco Workshop Recommendations to Enable Precision Cancer Medicine." *JCO Precision Oncology* (2):1–14. doi: 10.1200/PO.17.00292.
79. Wood, A., Denholm, R., Hollings, S., Cooper, J., Ip, S., Walker, V., ... & Sudlow, C. (2021). Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. *bmj*, 373.
80. Wolf, Beat, Pierre Kuonen, Thomas Dandekar, and David Atlan. 2015. "DNAseq Workflow in a Diagnostic Context and an Example of a User Friendly Implementation." *BioMed Research International* 2015(1):403497. doi: 10.1155/2015/403497.
81. Wünsch, C., H. Banck, C. Müller-Tidow, and M. Dugas. 2020. "AMLVaran: A Software Approach to Implement Variant Analysis of Targeted NGS Sequencing Data in an Oncological Care Setting." *BMC Medical Genomics* 13(1). doi: 10.1186/s12920-020-0668-3.
82. Xia, J., M. J. Benner, and R. E. W. Hancock. 2014. "NetworkAnalyst. Integrative Approaches for Protein-Protein Interaction Network Analysis and Visual Exploration." *Nucleic Acids Research* 42(W1):W167–74. doi: 10.1093/nar/gku443.
83. Yousif, Ayman, Nizar Drou, Jillian Rowe, Mohammed Khalfan, and Kristin C. Gunsalus. 2020. "NASQAR: A Web-Based Platform for High-Throughput Sequencing Data Analysis and Visualization." *BMC Bioinformatics* 21(1):267. doi: 10.1186/s12859-020-03577-4.
84. Yukselen, Onur, Osman Turkyilmaz, Ahmet Rasit Ozturk, Manuel Garber, and Alper Kucukural. 2020. "DolphinNext: A Distributed Data Processing Platform for High Throughput Genomics." *BMC Genomics* 21(1):310. doi: 10.1186/s12864-020-6714-x.

Appendixes

Appendix 1. Overview of functional and non-functional requirements mentioned by authors in each selected article in the Literature Review.

Table 54. Functional and non-functional requirements mentioned by authors in each selected article

ID	Main groups of functional requirements			Main group of non-functional requirements			
	General data management	Data processing and analysis	Data visualization and reporting	Communication and support	Platform technical infrastructure	UI/UX characteristics	Security and Compliance
(Demchak et al. 2014)	✓	✓	✓	✓	✓	✓	✓
(Xia, Benner, and Hancock 2014)	✓	✓	✓	✓	-	✓	-
(Sante et al. 2014)	✓	✓	✓	-	✓	✓	-
(Suciu et al. 2015)	✓	✓	✓	-	-	✓	-
(Sauria et al. 2015)	✓	✓	-	✓	✓		-
(Calabria et al. 2015)	✓	✓	✓	-	✓	✓	-
(Wolf et al. 2015)	✓	✓	✓	-	-	✓	✓
(Bhuvaneshwar et al. 2016)	✓	✓	✓	-	✓	✓	✓
(Murtagh et al. 2016)	✓	✓	-	✓	✓	✓	✓
(Arneson et al. 2016)	✓	✓	✓	✓	-	✓	-
(Chen et al. 2016)	✓	✓	✓	✓	-	✓	✓
(Albuquerque et al. 2017)	✓	✓	-	✓	-	✓	-
(Davis-Turak et al. 2017)	✓	✓	✓	-	✓	✓	✓
(Takai-Igarashi et al. 2017)	✓	✓	✓	✓	-	-	✓
(Danahey et al. 2017)	✓	✓	✓	✓	✓	✓	✓
(Lau et al. 2017)	✓	✓	✓	✓	✓	✓	✓
(Crawford et al. 2018)	✓	✓	✓	-	-	-	✓
(Das et al. 2018)	✓	✓	✓	-	-	✓	✓
(Warner et al. 2018)	✓	✓	✓	✓	✓	✓	✓
(Mohr et al. 2018)	✓	✓	✓	✓	✓	✓	✓
(Pearce et al. 2019)	✓	✓	✓	✓	✓	✓	-
(Xia et al. 2014)	✓	✓	✓	✓	-	✓	-
(Raudvere et al. 2019)	✓	✓	✓	✓	✓	✓	-
(Canakoglu et al. 2021)	✓	✓	✓	✓	-	✓	-

(Hombach et al. 2019)	✓	✓	✓	✓	-	✓	-
(Nanni et al. 2019)	✓	✓	✓	✓	✓	✓	✓
(Verbruggen and Menschaert 2019)	✓	✓	✓	✓	-	✓	-
(Wünsch et al. 2020)	✓	✓	✓	✓	-	✓	✓
(Cappelli et al. 2020)	✓	✓	✓	✓	✓	✓	-
(Kuzmenkov et al. 2021)	✓	✓	✓	-	-	✓	✓
(Kounelis et al. 2020)	✓	✓	✓	✓	✓	✓	✓
(Rodchenkov et al. 2020)	✓	✓	✓	-	-	✓	-
(Bonomi, Huang, and Ohno-Machado 2020)	✓	✓	✓	✓	✓	✓	✓
(Yousif et al. 2020)		✓	✓	-	✓	✓	✓
(Yukselen et al. 2020)	✓	✓	✓	✓	-	✓	✓
(Holtgrewe et al. 2020)	✓	✓	✓	✓	-	✓	✓
(Ma et al. 2021)	✓	✓	✓	✓	✓	✓	✓
(Reska et al. 2021)	✓	✓	✓	✓	✓	✓	✓
(Pang et al. 2021)	✓	✓	✓	✓	✓	✓	✓
(Campbell et al. 2021)	✓	✓	✓	✓	✓	✓	✓
(Ochoa et al. 2021)	✓	✓	✓	-	✓	✓	-
(McLeod et al. 2021)	✓	✓	✓	✓	✓	✓	✓
(Ullah et al. 2022)	✓	✓	✓	✓	-	✓	✓
(Raveendran et al. 2022)	✓	✓	✓	-	✓	✓	✓
(Gilbert et al. 2022)	✓	✓	✓	✓	-	✓	✓
(L. Li et al. 2022)	✓	✓	✓	✓	✓	✓	✓
(Osmond et al. 2022)	✓	✓	✓	✓	✓	✓	✓
(Reiff et al. 2022)	✓	✓	✓	✓	✓	✓	-
(Amer-Yahia et al. 2022)	✓	✓	✓	✓	✓	✓	-
(Pavia et al. 2023)	✓	✓	✓	✓	-	✓	-
(Post et al. 2023)	✓	✓	-	-	✓	✓	✓
(Gill et al. 2023)	✓	✓	✓	✓	✓	✓	-

Appendix 2. Overview of requirements identified in the literature [1]

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**
1.1.1. Data Aggregation of Different Sources	2.1.1. Inspection of Data Quality	3.1.1. Networks
1.1.2. Data Upload	2.1.2. Data Normalization	3.1.2. Scatter Plots
		3.1.3. Genome Browser
1.2. Data Standardization	**2.2. Data Analysis**	3.1.4. Heatmaps
	2.2.1. Multiple Types of Analysis [Examples: RNA-Seq Analysis, Single-end (SE) or Paired-end (PE) Sequencing, Gene Set Enrichment Analysis, ChIP-seq]	
1.2.1. Usage of Libraries and APIs		3.1.5. Pie Charts
1.2.2. Usage of Pipelines to Standardize Data	2.2.2. Sequencing Data	3.1.6. Histograms
1.2.3. Usage of Algorithms to Standardize Data	2.2.3. Statistical Analysis	3.1.7. Trend Charts
1.2.4. Usage of Dashboard to Explore and Standardize the Data	2.2.4. Pathway Analysis	3.1.8. Dashboards
1.2.5. Usage of Customized Data Models to Standardize Data	2.2.5. Network Analysis	
1.2.6. DataSHaPER Harmonization Model	2.2.6. Automated Pipelines	**3.2. Generate Reports**
1.2.7. Integrate PATRIC and RCBPR	2.2.7. Customized Pipelines	
1.2.8. Genomics Data Commons (GDC)	2.2.8. Command-Line Options for Advanced Analysis	**3.3. Download Data**
1.2.9. Ensembl		
1.2.10. Variant Call Format		**3.4. Knowledge Dissemination**
1.2.11. Adopting Nomenclature Standards		
1.2.12. EUROCAST		**3.5. Citation Buttons**
1.2.13. Global Ontologies		
1.3. Data Sharing		
1.3.1. Data Sharing Capabilities (Data Distribution, Information Sharing)		
1.3.2. Collaboration Features (Users Can View or Collaborate on Specific Datasets)		
1.3.3. Customizable Access Permissions (e.g., grants access to case annotation, experiment groups, have read only access or full edit rights)		
1.3.4. User Application for Data Access (Data Access Application)		

Figure 6. Summary of Functional Requirements

Non Functional Requirments			
1. Communication and Support	**2. UX/UI Features**	**3. Security and Compliance**	**4. Platform technical infrastructure**
1.1. Feedback from Users	2.1. User-friendly Interface	3.1. Privacy-Protective Measures	4.1. Federated Infrastructure
1.2. Documentation	2.2. Searching for Data	3.2. User Registration	4.2. Scalability and Adaptability
1.3. Notifications	2.3. Highlighting Specific Regions	3.3. User Authentication	4.3. Application Programming Interfaces (APIs)
1.4. Communication Between Patients and Clinicians	2.4. Zooming in on Graphs	3.4. Consent Management	
	2.5. Hover Effects		
	2.6. Dropdown Menus		
	2.7. Drag-and-Drop Functionality		
	2.8. Mobile-friendly Interfaces		
	2.9. Supporting Multiple Languages		

Figure 7. Summary of Non-Functional Requirements

[1] This appendix is composed by two consecutive figures: Figure 1 and Figure 2.

Appendix 3. Key requirements from technical, clinical and legal experts[2]

Table 55. Key requirements for IT and technical experts

Requirement	Mean %	SD %	IQR	Respondents %	Agreement (1)	Importance (2)	Assessment (3)
Importance of privacy-preserving techniques	92%	11%	1	89%	Yes	Yes	Include
Network Topologies: Importance of peer-to-peer topology	76%	22%	1,5	83%	Yes	Yes	Include
Network Topologies: Importance of hierarchical topology	62%	27%	2,5	78%	No	Yes	Further testing
Network Topologies: Importance of centralized topology	71%	31%	2	78%	No	Yes	Further testing
Network Topologies: Importance of hybrid topology	62%	34%	2,5	83%	No	Yes	Further testing
Data Partitioning & Learning Strategies: Commonality of horizontal partitioning	96%	7%	0,25	89%	Yes	Yes	Include
Data Partitioning & Learning Strategies: Commonality of vertical partitioning	68%	29%	3	89%	No	Yes	Further testing
Data Partitioning & Learning Strategies: Commonality of transfer learning	52%	28%	1	89%	Yes	Yes	Include
Data Partitioning & Learning Strategies: Commonality of hybrid learning	51%	32%	4	89%	No	Yes	Further testing
Participation Model: Commonality pre-determined governance	97%	7%	0	89%	Yes	Yes	Include
Participation Model: Commonality Open participation	23%	27%	1,5	89%	Yes	No	Further testing
Participation Model: Commonality hybrid approach	34%	32%	3	89%	No	No	Exclude
Commonality of using HTTP-based API	82%	14%	0,5	89%	Yes	Yes	Include
Commonality of using gRPC	36%	37%	3,5	89%	No	No	Exclude
Commonality of using custom protocols	32%	25%	2,25	89%	No	No	Exclude
Commonality of using Event Bus (e.g., Kafka)	29%	29%	3	89%	No	No	Exclude
Support for Federated Learning algorithms	68%	27%	1	89%	Yes	Yes	Include
Support for Federated Analytics [Interoperability]	68%	28%	1,25	89%	Yes	Yes	Include
Importance of standardized ontologies/frameworks	82%	25%	1,25	89%	Yes	Yes	Include
Importance of reproducibility	88%	16%	1	89%	Yes	Yes	Include

[2] Requirements are provided in three consecutive tables below: Table 1, Table 2 and Table 3.

Importance of non-IID data	75%	24%	1,25	89%	Yes	Yes	Include
Importance of IID data	75%	16%	1	89%	Yes	Yes	Include
Importance of LLM training	51%	29%	2,25	89%	No	Yes	Further testing
Framework used: TensorFlow Federated	55%	35%	3	78%	No	Yes	Further testing
Framework used: PySyft	44%	32%	3	67%	No	No	Exclude
Framework used: Flower FL	59%	27%	0	72%	Yes	Yes	Include
Framework used: Vantage6	74%	31%	2	78%	Yes	Yes	Include
Framework used: PyTorch	55%	35%	3	72%	No	Yes	Further testing
Framework used: NVIDIA FLARE	33%		0	6%	Yes	No	Further testing
Framework used: Custom-built	50%	38%	3,75	56%	No	No	Exclude
Importance of synchronous updates [Metadata]	60%	34%	2,5	89%	No	Yes	Further testing
Importance of asynchronous updates [Metadata]	61%	22%	2	89%	No	Yes	Further testing
Importance of mixed update strategies	53%	24%	0,75	89%	Yes	Yes	Include
Importance of differential privacy	68%	26%	1,5	83%	Yes	Yes	Include
Importance of homomorphic encryption	63%	26%	2	83%	No	Yes	Further testing
Importance of secure multi-party computation [interoperability]	58%	30%	1,5	89%	Yes	Yes	Further testing
Importance of computational efficiency [interoperability]	65%	28%	2	89%	No	Yes	Further testing
Importance of communication efficiency [interoperability]	67%	21%	2	89%	No	Yes	Further testing
Importance of model accuracy [interoperability]	83%	14%	0,25	89%	Yes	Yes	Include
Importance of privacy guarantees [interoperability]	89%	15%	1	89%	Yes	Yes	Include

Note: The mean values are converted from a 1–7 Likert scale to percentages, and the standard deviation is adjusted accordingly to match the transformed scale.
[1]*Agreement* is considered "Yes" when the Interquartile Range (IQR) is less than 2, indicating consensus among respondents.
[2]*Importance* is marked as "Yes" when the average rating exceeds 50%.
[3] *Assessment criteria* is based on the alignment between agreement and important.

Table 56. Key requirements for Clinical Experts and Healthcare providers

Requirement	Mean %	SD %	IQR	Respondents %	Agreement (1)	Importance (2)	Assessment (3)
Importance of acquiring genomic data from multiple sources.	83%	16%	1	56%	Yes	Yes	Include
Importance of uploading your own genomic data into a system.	79%	16%	0	56%	Yes	Yes	Include
Relevance of the PATRIC data model. [interoperability]	44%	19%	0	34%	Yes	No	Further testing

Relevance of the Genomic Data Model. [INTEROPERABILITY]	63%	14%	1,25	38%	Yes	Yes	Include
Relevance of the DataSHaPER model. [INTEROPERABILITY]	48%	17%	0	34%	Yes	No	Further testing
Relevance of the OMOP model. [INTEROPERABILITY]	62%	14%	1	41%	Yes	Yes	Include
Relevance of the FHIR model. [INTEROPERABILITY]	63%	17%	1	41%	Yes	Yes	Include
Relevance of the VCF model. [INTEROPERABILITY]	65%	22%	2	41%	Yes	Yes	Include
Relevance of the PHENOPACKET model. [INTEROPERABILITY]	58%	14%	0,75	19%	Yes	Yes	Include
Importance of VCF file format. [INTEROPERABILITY]	64%	22%	2	41%	Yes	Yes	Include
Importance of FASTQ file format. [INTEROPERABILITY]	62%	18%	2	41%	Yes	Yes	Include
Importance of BAM file format. [INTEROPERABILITY]	64%	21%	2	44%	Yes	Yes	Include
Importance of IDAT file format. [INTEROPERABILITY]	67%	24%	2	16%	Yes	Yes	Include
Importance of efficiency in data sharing decisions.	84%	12%	0,75	56%	Yes	Yes	Include
Importance of security in data sharing decisions.	82%	13%	1,75	56%	Yes	Yes	Include
Importance of project scope in data sharing decisions.	84%	15%	2	56%	No	Yes	Further testing
Importance of inspecting data quality before analysis.	87%	11%	1	56%	Yes	Yes	Include
Importance of automated checks for data completeness.	81%	13%	1	56%	Yes	Yes	Include
Importance of epidemiological analysis.	68%	25%	2	56%	No	Yes	Further testing
Importance of predictive modelling.	81%	25%	1	56%	Yes	Yes	Include
Importance of statistical analysis.	83%	25%	1,75	56%	Yes	Yes	Include
Importance of data visualization.	74%	25%	1	56%	Yes	Yes	Include
Importance of exploratory data analysis.	76%	23%	1	56%	Yes	Yes	Include
Importance of AI modelling.	75%	22%	1	56%	Yes	Yes	Include
Importance of preventive models.	63%	20%	1,75	31%	Yes	Yes	Include
Importance of reproducibility in research.	81%	22%	1,75	56%	Yes	Yes	Include
Importance of command-line tools for analysis.	67%	21%	1,75	56%	Yes	Yes	Include

IMPORTANCE OF GRAPHS.	82%	19%	1	56%	Yes	Yes	Include
IMPORTANCE OF CHARTS.	76%	22%	1	56%	Yes	Yes	Include
IMPORTANCE OF HEAT MAPS.	72%	22%	1,75	56%	Yes	Yes	Include
IMPORTANCE OF SEQUENCING VISUALIZATIONS.	70%	22%	1,75	56%	Yes	Yes	Include
IMPORTANCE OF NETWORK VISUALIZATIONS.	74%	23%	3	56%	No	Yes	Further testing
IMPORTANCE OF BEING ABLE TO DOWNLOAD/EXPORT DATA.	81%	27%	1	56%	Yes	Yes	Include
IMPORTANCE OF SHARING KNOWLEDGE WITH CLINICIANS.	87%	25%	1	56%	Yes	Yes	Include
IMPORTANCE OF SHARING KNOWLEDGE WITH RESEARCHERS.	89%	24%	1	56%	Yes	Yes	Include
IMPORTANCE OF SHARING KNOWLEDGE WITH PATIENTS.	62%	24%	2	56%	Yes	Yes	Include
IMPORTANCE OF SHARING KNOWLEDGE WITH POLICY MAKERS.	67%	27%	2	56%	No	Yes	Further testing
IMPORTANCE OF SHARING KNOWLEDGE WITH ETHICAL BOARDS.	69%	27%	1,75	56%	Yes	Yes	Include
IMPORTANCE OF PROTECTING DATA PRIVACY IN RESEARCH.	77%	28%	1,5	72%	Yes	Yes	Include
IMPORTANCE OF STAYING UPDATED ON GENOMIC DATA SECURITY STANDARDS.	76%	24%	2	72%	No	Yes	Further testing
IMPORTANCE OF A MOBILE-FRIENDLY PLATFORM.	63%	26%	2	72%	No	Yes	Further testing
IMPORTANCE OF MULTILINGUAL PLATFORM SUPPORT.	67%	30%	2	72%	No	Yes	Further testing
IMPORTANCE OF RECEIVING PLATFORM NOTIFICATIONS.	65%	27%	0,5	72%	Yes	Yes	Include
IMPORTANCE OF ACCESS TO FEDERATED COMPUTING INFRASTRUCTURE.	79%	16%	1,5	72%	Yes	Yes	Include
IMPORTANCE OF COMPUTATIONAL EFFICIENCY.	86%	13%	1	72%	Yes	Yes	Include
IMPORTANCE OF COMMUNICATION EFFICIENCY.	83%	12%	1	72%	Yes	Yes	Include
IMPORTANCE OF MODEL ACCURACY.	88%	13%	1	72%	Yes	Yes	Include
IMPORTANCE OF PRIVACY GUARANTEES.	85%	19%	2	72%	Yes	Yes	Include
IMPORTANCE OF HORIZONTAL PARTITIONING.	76%	21%	1,5	72%	Yes	Yes	Include
IMPORTANCE OF VERTICAL PARTITIONING.	70%	27%	2,5	72%	No	Yes	Further testing
IMPORTANCE OF TRANSFER LEARNING.	70%	26%	2	72%	Yes	Yes	Include
IMPORTANCE OF HYBRID PARTITIONING.	74%	26%	2,5	72%	No	Yes	Further testing

IMPORTANCE OF PRE-DETERMINED PARTICIPANT SELECTION.	67%	27%	1	72%	Yes	Yes	Include
IMPORTANCE OF OPEN PARTICIPANT SELECTION.	60%	29%	2	72%	Yes	Yes	Include
IMPORTANCE OF USING A MIXED APPROACH.	65%	24%	2	72%	Yes	Yes	Include
RELEVANCE OF TENSORFLOW.	60%	26%	2	50%	Yes	Yes	Include
RELEVANCE OF PYSYFT.	56%	23%	1	50%	Yes	Yes	Include
RELEVANCE OF FLOWER.	56%	25%	0,5	47%	Yes	Yes	Include
RELEVANCE OF VANTAGE6.	58%	27%	1,25	50%	Yes	Yes	Include
RELEVANCE OF CUSTOM-BUILT FRAMEWORKS.	63%	19%	2	50%	Yes	Yes	Include
RELEVANCE OF PYTORCH.	61%	26%	2	47%	Yes	Yes	Include

Note: The mean values are converted from a 1–7 Likert scale to percentages, and the standard deviation is adjusted accordingly to match the transformed scale.
[1]*Agreement* is considered "Yes" when the Interquartile Range (IQR) is less than 2, indicating consensus among respondents.
[2]*Importance* is marked as "Yes" when the average rating exceeds 50%.
[3] *Assessment criteria* is based on the alignment between agreement and importance.

Table 57. Key requirements for data governance and compliance experts

Requirement	Mean %	SD %	IQR	Respondents %	Agreement (1)	Importance (2)	Assessment (3)
IMPORTANCE OF DEVELOPING STANDARDIZED EU BEST PRACTICES FOR FEDERATED COMPUTING TO ENSURE COMPLIANCE WITH DATA SOVEREIGNTY REGULATIONS.	69%	31%	1,25	100%	Yes	Yes	Include
RELEVANCE OF EU-WIDE STANDARDS FOR FEDERATED COMPUTING TO MINIMIZE PRIVACY RISKS IN CROSS-BORDER DATA TRANSFERS AND SUPPORT INNOVATION.	76%	23%	1,5	88%	Yes	Yes	Include
CRITICALITY OF CONSISTENT EU STANDARDS FOR FEDERATED COMPUTING TO HARMONIZE DATA GOVERNANCE ACROSS JURISDICTIONS	67%	32%	2,5	88%	No	Yes	Further testing
VALUE OF AN EU-WIDE APPROACH TO FEDERATED COMPUTING FOR TRANSPARENCY, AUDITABILITY, AND TRUST IN DECENTRALIZED PROCESSING.	71%	34%	3	88%	No	Yes	Further testing
IMPORTANCE OF EU GUIDELINES FOR FEDERATED COMPUTING IN BALANCING DATA SOVEREIGNTY WITH CROSS-BORDER HEALTHCARE RESEARCH.	67%	25%	2	88%	No	Yes	Further testing
IMPORTANCE OF EU-WIDE BEST PRACTICES FOR FEDERATED COMPUTING TO ENSURE GDPR COMPLIANCE (E.G., DATA MINIMIZATION, PURPOSE LIMITATION).	83%	19%	1,5	88%	Yes	Yes	Include
RELEVANCE OF STANDARDIZED EU GUIDELINES FOR FEDERATED COMPUTING TO PROMOTE PRIVACY-BY-DESIGN AND PRIVACY-BY-DEFAULT.	71%	30%	2	88%	No	Yes	Further testing
CRITICALITY OF ALIGNING PRIVACY-PRESERVING TECHNIQUES WITH EU STANDARDS IN FEDERATED COMPUTING	69%	33%	2,5	88%	No	Yes	Further testing

VALUE OF EU STANDARDS FOR FEDERATED COMPUTING TO MITIGATE RISKS SUCH AS DATA BREACHES.	74%	36%	3	88%	No	Yes	Further testing
IMPORTANCE OF STANDARDIZED EU PROTOCOLS FOR FEDERATED COMPUTING TO SUPPORT REGULATORY APPROVAL AND BUILD TRUST.	69%	33%	2,5	88%	No	Yes	Further testing
IMPORTANCE OF DATA SHARING FOR RESEARCH AND INNOVATION TO IMPROVE HEALTHCARE IN EUROPE.	67%	17%	1	38%	Yes	Yes	Include
IMPORTANCE OF CONDUCTING RISK ASSESSMENTS BEFORE DATA SHARING DECISIONS IN FEDERATED SYSTEMS.	67%	20%	1	63%	Yes	Yes	Include
CRITICALITY OF RETAINING DATA ON-PREMISES FOR REGULATORY COMPLIANCE.	67%	18%	0,75	75%	Yes	Yes	Include
NEED TO ADDRESS REGULATORY RISKS IN FEDERATED COMPUTING: ENSURING DATA PROTECTION AND PRIVACY COMPLIANCE.	88%	16%	1,5	88%	Yes	Yes	Include
NEED TO ADDRESS REGULATORY RISKS IN FEDERATED COMPUTING: POTENTIAL LACK OF TRANSPARENCY OR AUDITABILITY.	62%	28%	2,5	88%	No	Yes	Further testing
NEED TO ADDRESS REGULATORY RISKS IN FEDERATED COMPUTING: CHALLENGES IN ACCOUNTABILITY AND OVERSIGHT.	69%	31%	2,5	88%	No	Yes	Further testing
NEED TO ADDRESS REGULATORY RISKS IN FEDERATED COMPUTING: RISK OF NON-COMPLIANCE WITH LEGAL FRAMEWORKS.	71%	33%	3	88%	No	Yes	Further testing
IMPORTANCE OF ENABLING RAPID IDENTIFICATION OF CROSS-NATIONAL HEALTH ISSUES THROUGH FEDERATED COMPUTING	67%	18%	2	75%	No	Yes	Further testing
IMPORTANCE OF ENABLING EFFICIENT DATA SHARING ON RARE DISEASES VIA FEDERATED COMPUTING SYSTEMS.	72%	17%	1,5	75%	Yes	Yes	Include
IMPORTANCE OF MODEL ACCURACY IN FEDERATED COMPUTING, ESPECIALLY IN HEALTHCARE USE CASES.	78%	23%	2,25	75%	No	Yes	Further testing
IMPORTANCE OF STRONG PRIVACY GUARANTEES IN FEDERATED COMPUTING, SUCH AS ANONYMIZATION AND DATA PROTECTION.	81%	25%	2,5	75%	No	Yes	Further testing

Note: The mean values are converted from a 1–7 Likert scale to percentages, and the standard deviation is adjusted accordingly to match the transformed scale.
[1]*Agreement* is considered "Yes" when the Interquartile Range (IQR) is less than 2, indicating consensus among respondents.
[2]*Importance* is marked as "Yes" when the average rating exceeds 50%.
[3] *Assessment criteria* is based on the alignment between agreement and importance.

Appendix 4. Frequency of codes discussed during first workshop

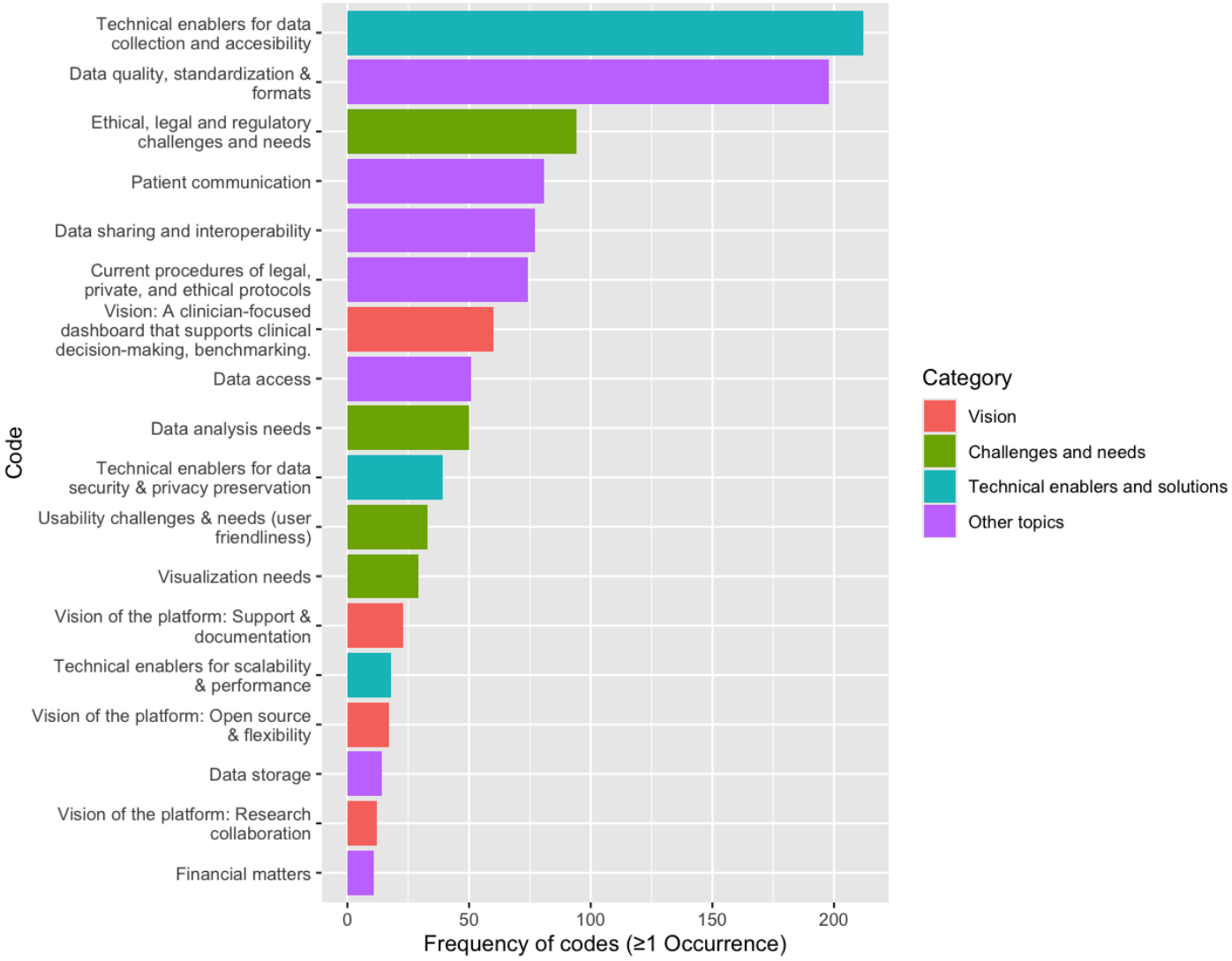

Figure 8. Frequency of discussed topics per category

Appendix 5. Legends and acronyms used throughout the Protect-Child project

L: Logic
LR: Literature Review
P: Previous work (e.g., IDEARC, previous similar European projects)
FG: Focus group
W: Workshop
E: Formal and informal expert insights
S: Survey
R: Regulation (e.g., EHDS)
Stakeholders: A group of experts involved in the development of Protect Child (e.g., clinical experts and healthcare providers, researchers, IT and technical experts, data governance and compliance).
CoE: Center of Expertise
Data: Electronic health records, genomic data, etc.
Cohort: A group of individuals having a statistical factor (such as age or class membership) in common in a demographic study.
Metadata: Structured and descriptive information about data. In protect child this include information such as ownership of the data, data sources, description of variables (e.g., viral, bacterial infections, drug related toxicity, organ failure), data quality (conformance, completeness, plausibility), patient consent, ethical requirements
Data holder: Under the EHDS, a health data holder is a public sector body responsible for managing and granting access to health data while ensuring compliance with ethical and legal standards. In alignment with these principles, the PROTECT-CHILD Project oversees the PROTECT-CHILD-SPE capsules, which function as data holders.
Data capsule: The data holder in protect child managed by the centers of expertise. The data capsules providing reference implementation methods and architectures for standard and interoperable integration of disparate private data-oriented environment and multiple application paradigms, from conventional data queries to the training and validation of AI models.
Data manager: The data managers oversee the collection, curation, validation, and secure storage of high-quality data, serving as a bridge between technical systems and research or clinical teams to ensure reliable and compliant data use.
Data user: As defined by the EHDS data users include researchers, healthcare professionals, innovators, and policymakers who access health data under an authorized permit. Within the PROTECT-CHILD Project, this category specifically includes researchers and clinical experts such as geneticists, pediatricians, and epidemiologists.
Dataset: A structured collection of data organized and stored together for analysis or processing. The data within a dataset is typically related in some way and taken from a single source or intended for a single project. In the case of Protect Child a dataset is a collection of healthcare data, sometimes including genomic data.
Workspace (for research project management): A workspace is a digital environment where data users can organize, manage, and track all aspects of a research project, including data access requests, analyses, and reports, enabling streamlined workflow, reproducibility, and project oversight.
OMOP: The Observational Medical Outcomes Partnership (OMOP) is a common data model for organizing healthcare data from various sources.
FHIR: Fast Healthcare Interoperability Resources (FHIR) is a healthcare data standard with an application programming interface (API) for representing and exchanging electronic health records (EHR)
ETL: Extract, Transform, Load
NLP: Natural language processing
NLU: Natural language understanding

SPE: Secure Processing Environment.
Fractal architecture: A distributed system operated through microservice, which composes smaller, self-similar components, called fractal leaves. In this architecture, the fractal leaves share the same properties, which enables them to be combined in different ways to create larger microservices to perform more complex tasks.
Federated Learning (FL): An advanced machine learning approach that enables collaborative model training across decentralized devices or entities without the need to share raw data. Instead, models are trained locally on each device using their respective data, and only model updates (gradients) are sent to a central server for aggregation.
Federated Analytics (FA): Focuses on statistical analysis of distributed data without sharing sensitive information. Aggregated statistics are computed locally and shared, preserving data privacy.
Federated Computing (FC): A privacy-preserving approach where data stays on local devices, creating information asymmetry between servers and clients.
Differential Privacy (DP): A privacy-preserving technique that protects individual data by adding controlled noise to datasets.
Centralized architecture: A server that aggregates results from clients.
Hierarchical architecture: An approach that introduces an intermediate layer, where devices act as both clients and servers, increasing redundancy and robustness, ensuring insights are still generated even if a cluster fails.
Decentralized (peer-to-peer) architecture: A type of architect where there is no central server, making it more complex to manage and challenging for ML models to converge.
Mesh: A service mesh or simply mesh is an infrastructure layer that enables managed, observable, and secure communication between workload instances.
Microservice: An architectural and organizational approach to software development where software is composed of small independent services that communicate over well-defined APIs.
Protect-Child Orchestrator: This is a secure processing environment (SPE) that presents the metadata of each Centers of Expertise (CoE), allows for the access to the data once the data permit application is granted, presents the data aggregated for the preprocessing and data analysis and allows to export the results of the data analysis.
Protect-Child Ecosystem: It is a combination of components such as the EHD Capsule that supports the secure processing of the data.
FAIR principes: Findability, Accessibility, Interoperability and Reusability
Data quality dimensions: conformance, completeness, plausibility
EHDS: European Health Data Space Regulation
GDPR: General Data Protection Regulation
HDAB: Health Data Access Body

Appendix 6. Overview of stakeholders to be further involved in the process of requirements elicitation and negotiation

Table 58. Overview of stakeholders to be further involved in the process of requirements

Category	*Stakeholders*	*Involvement Phase*	*Notes*
Clinical Experts (Data Users)	• Pediatricians • Immunologists • Nephrologists • Clinical Pharmacologists • Hepatologists • Surgeons • Geneticists	During & After Project	Core data users across clinical domains
Potential Future Stakeholders (Data Users)	• Clinical Researchers (including specialists above) • Epidemiologists • Pharmacologists • Pharma Companies • Biotech Companies • Biomolecular Researchers / Biologists • Hospital Managers • Healthcare Policymakers • Insurance Companies • Patient Associations	After Project	To be evaluated by data providers for acceptance or rejection
Data Providers (During Project & New Centres)	• IT Experts Chairperson & Staff *(capsule setup, automated extraction)* • Data Managers *(data extraction and annotation)* • Data Domain Experts *(in collaboration with clinicians)* • Bioinformaticians *('omics data processing, biomarker extraction)*	During Project & Onboarding	Supply and maintain project data
Data Security & Governance Stakeholders	• IT Security Office Experts • Data Protection Officers • Legal and TTO Representatives • Ethical Experts *(from ethics committees)*	Continuous	Oversee governance, privacy, and compliance
Data Processors (Project Development)	• IT Experts *(data models, capsule population, NLP systems)* • Bioinformaticians *('omics data processing)* • Geneticists *(biological materials & data extraction)* • IT Security Department *(data backup, protection)*	During Project	Responsible for processing, analysis, and technical infrastructure