

Health Data Sharing for Secondary Use: a Critical Overview^{*}

Alessio Famiani¹, Ruggero G. Pensa^{1,*}

¹Università di Torino, Dipartimento di Informatica, Via Pessinetto, 12, 10149 Torino TO)

Abstract

Data has been playing a crucial role in healthcare. Its digitisation and exchange have improved the delivery of healthcare to patients and fostered scientific research. However handling the workflow and governance of such data is not a trivial matter, due to legal and technical barriers. The aim of this paper is to explore existing health data sharing schemes to identify trends and desiderata for a hypothetical system that meets the legal requirements for handling data for scientific research. This discussion highlights the importance of interoperability to address the fragmented nature of data and emphasises the need for data owners' autonomy and empowerment, alongside technical trends that can help achieve these goals.

Keywords

health data, privacy, biomedical research, data spaces

1. Introduction

Data has increasingly become a valuable resource in many sectors, if not all, and has played a crucial role in the healthcare setting. Over the years, the digitisation of health data and its exchange have improved the delivery of healthcare to patients, facilitated the collaboration of various stakeholders within the sector, and also fostered scientific research [1]. Nowadays, patients have the ability to view and download their health records, including laboratory results, exams, visits, and drug prescriptions, using online platforms. Additionally, they can monitor their health status by using wearables or other IoT devices that collect and process various metrics. Doctors and specialists can issue certifications, prescriptions, and documentation regarding patients using these same platforms. In addition, health institutions can also collect and share these records for collaborating with scientists. However, handling the workflow and governance of this type of data is not a trivial matter.

We tackle this issue in PADS4Health (*Privacy-Aware Data Sharing Model for Health Data*), a project funded by the Italian Ministry of University and Research through the National Recovery and Resilience Plan (PNRR) under the Next Generation EU program, whose objective is to identify a scalable and privacy-compliant data management solution that addresses the challenges associated with the re-use of health data for research purposes. Our primary focus is to raise awareness about the importance of data privacy and enhance the workflow of scientific research, thereby improving its workflow and times.

In this position paper, we present the perspective of PADS4Health regarding the privacy-aware secondary use of health data for scientific research. To this purpose, we first point out both technical and legal barriers that can arise during the usage or development of a system that manages health data, particularly when re-using it for research purposes (Section 2). Then, in Section 3, we explore some data sharing schemes proposed in literature, together with some real-life examples of systems and resources for interoperability. Finally, in Section 4, we summarise the relevant works in literature and identify the main trends and desired properties/requirements a system handling health data should have.

ITADATA2025: The 4th Italian Conference on Big Data and Data Science, September 9–11, 2025, Turin, Italy

*Corresponding author.

✉ alessio.famiani@unito.it (A. Famiani); ruggero.pensa@unito.it (R. G. Pensa)

ORCID 0000-0001-5145-3438 (R. G. Pensa)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Technical and legal burdens

There are many burdens, of different nature, that users, developers, and researchers might face during the use or development of systems dealing with such data.

Technical burdens From a technical point of view, the main challenges come from the **lack of interoperability** among systems belonging to different medical sites and organisations, as well as applications related to healthcare and/or wellbeing. **Data is scattered across multiple locations** and services which usually do not communicate with each other, which may use different schemas, standards, or data formats. This makes it difficult for patients to receive healthcare from different medical institutions and get a clear overall picture of their own health [1, 2]. An issue that is worth mentioning is related to the obvious **concerns about data security and privacy** for both in-transit and at-rest data [3]. Breaches involving such sensitive data can have harmful consequences on the people data is about and their lives. This ranges from identity theft, discrimination, reputation or financial damages, to the inability of accessing data on them and so on. This also **negatively influences the level of trust** of the system and the related institution, making people reluctant to share their data [3]. However, when adopting potential countermeasures, the ability to access one's data in emergency situations should be taken into account, a feature also known as **break the glass** [4, 5], and, more generally, some **delegation mechanisms** [6]. For what concerns scientific research, the aforementioned fragmentation can cause **difficulties in retrieving high quality data**, since data can encounter different barriers during its collection. For instance, it can be challenging to **obtain and combine all the data belonging to the same individual** [2]. Different medical sites or data sources may have varying indexing and techniques for identifying users within their systems, which can differ from each other. Caution should also be exercised in the **assessment, and eventually in the mitigation, of possible biases and errors** that can be found within the collected data [7]. A non-trivial aspect is also represented by the ability to **find and query data in a privacy-respectful** and safe manner, also under the assumption of **not having access to it yet**. Core point is **how to perform research on data while respecting people privacy and data security** (i.e., it should not be possible to infer that someone was included or not in the research or perform research on encrypted data).

Legal burdens When it comes to legal barriers, given our involvement in the *PADS4Health* project, we'll focus on the European framework. The European legal context on matters of privacy, share and reuse of health data for secondary purposes is largely influenced by the *General Data Protection Regulation* (GDPR) [8]. GDPR is the most relevant European law on privacy, which defines the rights individuals, known as data subjects, have on their data. It imposes obligations for controllers and processors, and fines for unlawful behaviours. Plus, it also helped raise awareness on privacy issues. The GDPR is a pervasive regulation since it applies to all personal data, protects EU citizens globally, and has brought legal harmonisation across Member States. Thus, its scope of application is broader and more consistent compared to other legal contexts around the world. For instance, in the USA, the landscape is fragmented, with privacy laws regulating specific sectors, resulting in a narrower range of protection. A US privacy law regulating health data sharing privacy and security is represented by the *Health Insurance Portability and Accountability Act* (HIPAA) [9]. It is a federal law which regulates the privacy of the so called *Protected Health Information* (PHI) treated by specified entities. It protects PHI within US borders and not worldwide like GDPR does.

Additionally, over the years, laws and regulations have been enacted to discipline privacy and data exchange in the public sector or for public interests, including scientific research. Some are built upon the GDPR, denoting again the relevance of this regulation. The *Open Data Directive* [10] and *Data Governance Act* (DGA) [11] encourage the re-use of data in the public sector and its secure exchange. Interestingly, DGA defines and regulates the figures of data intermediaries and the concept of data altruism, which refers to data donated by individuals for general interest purposes. A more recent regulation, the *European Health Data Space* (EHDS) [12], aims at establishing a common space for health data for both **primary use**, concerning the delivery of healthcare within EU borders, and **secondary**

use, referring to the use of data for public interest purposes (e.g. scientific research). It also institutes various intermediary figures for data exchange and access, other than interoperability aspects. Art. 110 of the *Italian Privacy Code* [13] norms the use of health data for medical, biomedical and epidemiological research, also in circumstances where data subjects cannot give consent. Referring to the technical issues previously discussed, having data scattered across different locations makes it **harder for data subjects to exercise control over their data and their rights** [14]. In addition, the redundancy coming from the collection of the same details of data subjects by multiple parties **colludes with the principles of data minimisation and once-and-only** [8, 15]. The collection, exchange and re-use of data also has to be compliant with existing regulations, other than respecting and protecting people rights and freedom. Eventually, it should also be taken into account the need of **keeping track of different legal bases** aside consent [8, 13]. Especially in the secondary use case, **research needs to be privacy-preserving and respectful**. All this by taking into account the eventual need to **protect intellectual property** of researches, institutions, and related [16].

3. Data sharing schemes in literature and real life examples

Given the fragmented nature and the vulnerability centralised solutions face due to potential breaches exposing large amounts of data, developers came up with various schemes over the years to effectively and safely exchange health data for research. In the following paragraphs, we will explore various health data sharing systems in literature, examining their **architectures** (their overall structure), **data flow patterns** (how data is shared and handled across/within systems), and **security and privacy considerations** (how sensitive information is protected). At the end of this section, we will also tackle a few interoperability resources and briefly illustrate two real life examples of such systems in action.

Architectures One example is SHRINE [17], an open source project whose goal is to aggregate patient observations scattered across hospitals in a standard manner in order to re-use these information for research activities. Its architecture consists in a P2P network with no central authority, where each medical site handles its own storage, security and verifies its own researchers. Local warehouses are run by i2b2¹ instances. [16] instead supports collaborative research by enabling federated analysis on individual-level data stored on various databases, each one curated and maintained by a participant site. The architecture consists of a central computational node performing the analysis, capable of issuing commands to several nodes where data is stored and queried locally in a parallel fashion. In [18] the authors developed a federated system which enables clinical sites to safely outsource their data for distributed genomic and clinical analysis. Each institution can choose its preferred storage solution: its own, a governmental one or a cloud provider's. This way sites can offload maintenance and availability burdens. These units form a secure, federated and interoperable network. PHT [19] is a system that implements FAIR principles and machine readability is at the core: interpretable workflows, services, data and metadata. PHT is organised in: stations, the participant sites, where data is stored; trains, analytical workflows performed within stations; tracks, intermediaries maintained by trusted parties that enforce rules and connect trains and stations. The authors of [20] rely on *Federated Learning* (FL) in a cross-device setting, generalisable to larger federation units, for performing analysis on distributed individual-level health data perturbed by *Differential Privacy* (DP) techniques. [21] proposed a system backed by permissioned blockchain technology, specifically designed for handling research of Covid-19 electronic medical records. Blockchain access is restricted: nodes, such as hospitals and research institutions, need to register and authenticate before transmissions and query. [22] is a toolkit² made for aiding researchers to perform various privacy-preserving federated analysis on genomic data. It consists of a web server and command line interface (CLI), and enables users to create, set up and run collaborative analysis. A coordinator handles the studies. In [23], the system is made of study participants (SPs), individuals, institutions or data custodians that own the data, and computing parties

¹<https://i2b2transmart.org/>

²<https://sfkit.org>

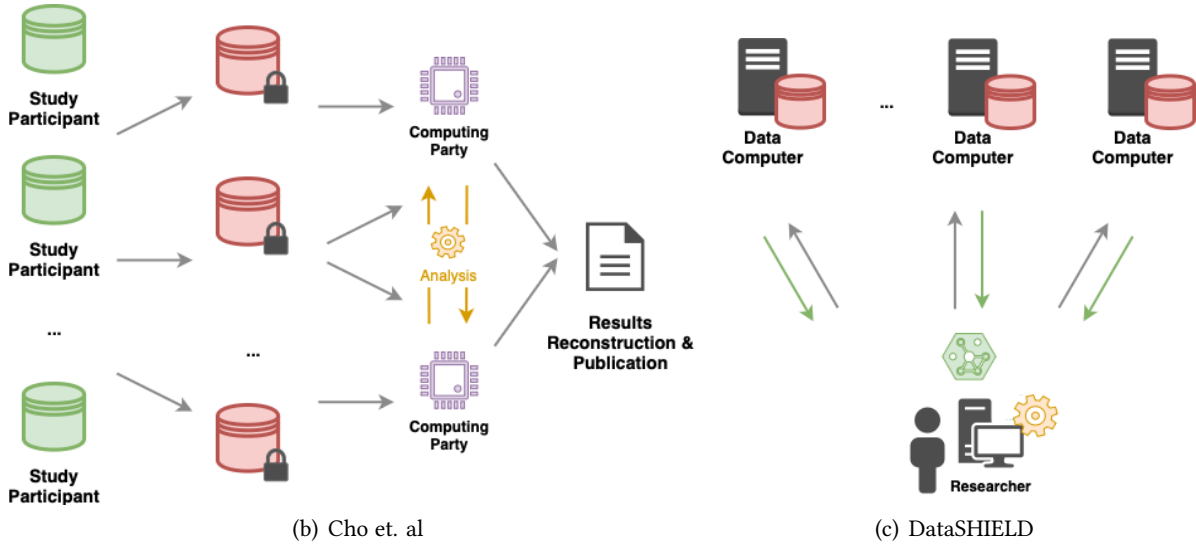
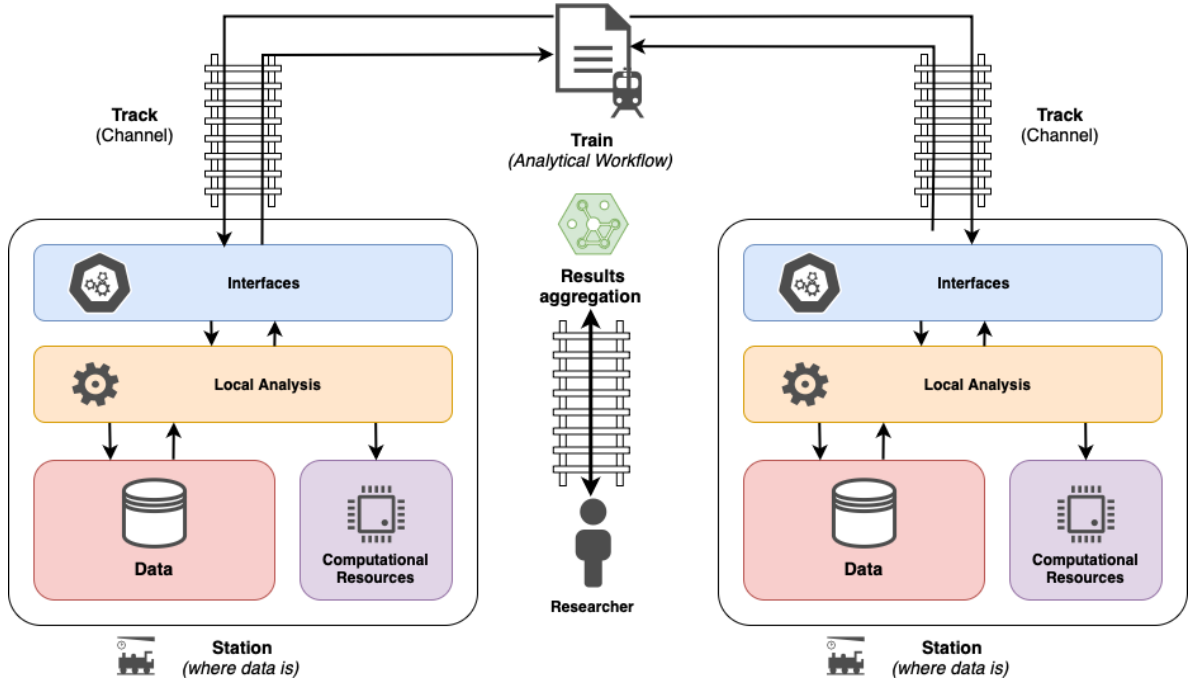


Figure 1: Simplified data flow and resources allocation of different architectures.

(CPs), entities with appropriate computing powers which jointly carry out the analysis. A different scheme instead is used in [24], where data providers and queriers are interconnected. Each site perform local computation and encrypts intermediate results. These results are later aggregated and redistributed for further computation iteratively until convergence is reached.

Data flows In [17] only authorised analysts perform searches for patient populations matching some criteria, specified using terms of an ontology defining a set of standardised medical concepts. Each peer needs to map its own terms to this common ontology. Queries are broadcasted to participant institutions and computed locally. Anonymised and aggregated results are then presented to the investigator progressively as they become available. [16] uses another type of approach (see Figure 1b): at the beginning of the study relevant features are chosen and data harmonised but still partitioned. A researcher can issue commands that specify which kind of operations have to be performed locally.

Local databases are powered by Opal³. Non-identifiable results are then sent back to the investigator and aggregated in such a way that returned outputs resemble non-disclosive study level statistics. This might be an iterative process. In [18] each site transforms data from its private system to match a predefined schema and a set of medical ontologies concepts for interoperability purposes. Local warehouse are run by ib2b instances. Investigators can then access and query encrypted data as if it's coming from a single source, without the need of decrypting it first. In PHT [19] stations host and manage their own data, run their own system, define interfaces for executing queries and provide computational resources for performing analysis within a secure environment (see Figure 1a). Technical choices are encoded into metadata for interoperability and discoverability purposes. Researchers build and maintain trains, objects containing all the information needed to run distributed analysis algorithms. Researchers are entirely decoupled from the computation phase. Tracks orchestrate workflows end-to-end and are responsible for train forwarding, results aggregation, transaction management and rules enforcement. Every object above has a unique ID and is tracked in a registry. Data in [20] is distributed across repositories and never sent outside of them. Sub-models are learned locally and then aggregated. Updates are performed locally. In [21] anonymised records of patients are stored on the chain upon previous consent in order to make them searchable. Results of research can also be stored on the platform for keeping track of progress in the field and protect intellectual property. Researchers issue query and/or upload requests. In [23] each SP securely shares their data with the CPs, which jointly execute an interactive protocol to accomplish the required analysis task (see Figure 1c). Finally, CPs combine their results together to obtain the final statistics and publish them. While in [24] a queries sends a query in clear, data providers perform the requested query locally on clear data and then send to each other encrypted results. Then, these results are aggregated and the process is repeated until convergence. Final results are sent to the querier.

Security and privacy details In [17] only authorised investigators can perform requests for queries, which are then verified at destination institutions. SHRINE uses digital certificates and signatures to secure communication and for identifying participants. Results are anonymised by default. In [16] researchers can issues only safe and approved operations in R and only non-identifiable outputs are returned. [18] uses *Homomorphic Encryption* (HE) and obfuscation techniques to provide privacy and security guarantees. Data remains encrypted end-to-end and authorised researchers are able to decrypt just the results. Every participant is involved in the encryption key generation process, so not even a compromised site can decrypt data alone. Depending on different privileges levels, it may also be impossible for an investigator to trace back responses to the original clinical site. As in other works, in [19] data never leaves its origin and is processed in a secure environment provided by its location. Data sovereignty is emphasized in the system and results are communicated via open protocols mandating authentication and authorisation procedures. Consortium blockchain technology in [21] and access control performed by the nodes of the alliance ensure only authorised users can access and upload data on the chain. Data is uploaded only upon consent of the patient, who directly has to authorise the operation. Integrity and identities check is ensured by the nature of the blockchain. Records are anonymised, and patients' identities are replaced with pseudonyms in the chain. [23] uses MPC protocol and during exchange inputs stay private. No CP can learn anything about the raw data aside from the final published results. Cryptographic methods secure data during its exchange and processing. [24] uses *Multiparty HE* (MHE), so all results, both intermediate and final, are encrypted and privacy is ensured end-to-end. Obfuscation techniques are used on final result to preserve accuracy.

Real life systems Around the world, many countries have adopted systems for managing health data for better assisting patients within their national healthcare system, and also for fostering scientific research from data previously shared by patients.

Some initiatives rely on entities called data enclaves or data custodians [25], trusted parties where individual-level data from multiple sources are safely stored and where privacy-friendly research is

³<https://github.com/obiba/opal>

performed. In Scotland, for example, within the *Scotland National Safe Haven* project [26], secure physical locations under some administrative arrangements safely store data coming from the *National Health Service* (NHS) for research purposes. The environments offer various pre-installed software and high-powered computing resources for performing analysis on pseudonymised individual-level data or aggregates. There is the possibility of granting remote access to investigators but no possibility of downloading data. Release of outputs are subject to human screening and statistical disclosure methods. Data retention policies have also been put in place.

A more advanced and recent system is adopted in Estonia. The e-Health Record [27] is a blockchain-backed [28] healthcare system which retrieves data as necessary from various providers that potentially use different systems and presents it in a standard format. This thanks to a secure distributed information exchange platform: X-road [29]. The system follows the principle of once-only: information is never asked twice. The system is backed by the *Keyless Signature Infrastructure* (KSI) blockchain, a scalable alternative to *Public Key Infrastructure* (PKI) for verifying integrity and authenticity of data and processes in zero-trust applications. Anonymised data is also made available for research purposes, plus the data collected within the system helped build a few databases for such objectives.

Resources for interoperability As seen previously, many systems interface with some biomedical electronic resources available online, while others adhere to some standards of the sector. Example of this are: the *Observational Health Data Sciences and Informatics* (OHDSI, pronounced Odyssey) [30], a project whose goal is bringing out the value of health data through large scale analytics. The main contribution is represented by the development of a common data model and the help in its adoption. Data stored in different observational databases can be converted into this standard for enabling the reliable large-scale analysis. As cited before, the Estonian X-road [29] is another case. This data exchange platform allows secure and standardised communication between different information systems, both private and public. It is capable of transmitting large quantities of data and performs searches simultaneously across several information systems. In [17, 31], an ontology, *The Shrine Core Ontology*, is used for data harmonisation. It is a collection of different concepts available in other resources concerning demographics, diagnoses, medications and laboratory tests. For instance, *ICD-9-CM* [32] is used for categorising diseases, which is the *International Classification of Diseases* standardised by the *World Health Organisation* (WHO). Another attempt worth mentioning is the *National Center for Biomedical Ontology* (NCBO) repository [33], which gives open access to various biomedical ontologies and also includes mappings between terms of different resources.

4. Discussion

As the authors of [25] argue, the approaches usually fall into three main categories: **Distributed data analysis**: where parties locally execute computations on their data, exchange partial results, and aggregate them across locations; **Cryptographic Secure Multi-party Computation Systems**: exchange of data is encrypted and only aggregated statistics can be decrypted; **Data enclaves**: data from multiple sources is merged in a single curated repository hosted by a trusted party. In addition, Blockchain technology could fall into a category of its own.

Despite the diversities of the illustrated works, several trends emerge. Firstly, the majority of the systems acknowledge the **distributed nature of the healthcare domain by employing federated, distributed and/or decentralised architectures** for enabling health data sharing for scientific research while respecting participants' privacy. Some works addressed this aspect by highlighting the **need for interoperability** via common standards and mapping tools. For example, in [17] participants institutions map their data schemes to a unified ontology of medical concepts, providing standardised ways for describing diseases, symptoms and drugs. This ensure data is harmonised across heterogeneous sites. However, it's important to note that converting data formats and mapping schemes can cause bottlenecks in the data flow. Plus, [17] and [18] are **built on widely used frameworks** for clinical research. The approach used in [19] implemented **FAIR principles** [34] in order to make every aspect

(i.e., data and analyses) of the workflows machine-readable, thus interoperable. It also clearly emerges the **respect for participants autonomy and data sovereignty**. For a medical site, it is usually desirable to have flexibility regarding the EHR system to use, data administration and governance, and also which computing environment or hosting solution to use. Data usually never leaves its original location, or at least not in a plain format, and it is queried in a federated manner, giving the perception that different data reside in the same place. In certain architectures, such as [35] **data owners empowerment** is also emphasized. For what concerns studies performed in a privacy-friendly manner, widely used technologies are *Federated Learning* (FL), *Differential Privacy* (DP) for obfuscating results and perturbing individual level data. However, a rising trend is given by **cryptographic techniques** like *Secure Multiparty Computation* (MPC) which enables participants to jointly perform computations without disclosing raw inputs, and *Homomorphic Encryption* (HE) that make able to perform operations on data without the need of decrypting it first. They are costly from a computationally perspective but avoid many legal barriers since encrypted data is anonymised, thus not regulated under the GDPR.

Identified requirements Besides what concerns the primary use of health data, starting from the analysis of the related literature, we have identified the following requirements for a system dealing with health data for research. A patient-centric approach can **give power and ownership back to data subjects**, since they can exercise their data rights intuitively and in a more effective manner. For example, they can revoke consents for treatments of their data where applicable, request and perform updates on their information or pause data processing selectively. Seeing and managing the flow of their own data can enhance individuals' awareness on the topic of privacy and their related rights. Data collection should be in accordance with the principles of **data minimisation and once-only** (gathered data can be shared with others later). This can help decrease management costs and improve privacy and trust. Enforcing **data retention policies** is also another crucial point. As other laws already require, logging and tracing every activity on data and access to it could help in the auditing processes and in being more transparent in general. This could also help in demonstrating the compliance with the GDPR and other laws, and for holding institutions and people accountable. It can also be helpful to keep track of legal bases (the most "popular" one being data subject's consent) and purposes of data processing. This can be beneficial also to data processors and data holders for demonstrating compliance. For what directly impacts scientific research, researchers could perform **searches on data without having yet access to it**, in order to retrieve data having certain properties without leaking information about individuals owning the data. Alternatively, they could have the ability to open "data campaigns" which are **containers populated by crowdsourcing health information** donated by data subjects (informed about the platform following criteria specified by the researchers). To achieve this, **data subjects should have the ability to donate their data** for research or for other public interest objectives upon previous anonymisation (when feasible) or pseudonymisation, automatically performed to protect their privacy and identity. This way, researchers can have access to large sets of complete and up-to-date health-related data upon which they can **study and perform learning while respecting people's privacy and privacy principles** (e.g., not retaining data more than needed). Generally, approaches like this can substantially **improve the workflow of scientific research**, resulting in faster scientific progress and better results. A broader list of requirements, along with some possible technical solutions, can be found in Table 1.

5. Conclusions

In the last few years, huge advancements have been made in the exchange of health data and its re-use for biomedical studies. The selected works covered a diverse spectrum of privacy-preserving architectures, data flows, security and privacy-related practises, thereby highlighting certain trends. The distributed nature of this domain underscores the importance of interoperability via common standards, shared resources or applications, and open protocols. The respect for participants' autonomy in managing their own systems and data, coupled with the need for data owners' empowerment,

Table 1

Technical measures to address the identified issues/requirements.

Issue/Requirement	Technological Fixes
Primary use	
Break-the-glass [4, 5]	Proxy re-encryption, Attribute Based Encryption, Broadcast Encryption, Smart Contracts
Support Healthcare-related Decision Making (<i>e.g. drug-to-drug interactions</i> [36])	Smart Contracts, Integration with electronic medical resources
Secondary use	
Safe Search on data	Symmetric Searchable Encryption, Zero-Knowledge Proof
Privacy-Preserving Analysis	Federated Learning, Homomorphic Encryption, Synthetic Data Generation, Differential Privacy, Statistical Disclosure Control
Bias and errors detection and mitigation [7]	Data quality dashboards, Pattern Recognition/Data Mining, Anomaly Detection
Improving workflow of scientific research [2]	Machine Learning Techniques for pre-processing
Protect intellectual property [16]	Watermarking, Secure Multiparty Computation, Digital Signature, Encryption, Virtualisation and Sandboxing
Impossibility of downloading data [26, 37, 38]	Virtualisation, Sandboxing
Data Governance, awareness and compliance	
Respect institutions autonomy [19, 17, 18]	Federated Learning, Secure Multiparty Computation, Blockchain
Keep data subjects aware and informed [8]	Dashboards, "Privacy Labels", Smart Contracts
Processing Ledgers [8]	Blockchain
Data portability and interoperability [1, 2]	Ontologies
Delegation Mechanisms [6]	Smart Contracts, Proxy re-encryption, Attribute Based Encryption, Broadcast Encryption
Records Linkage and deduplication support [39]	Hashing, Probabilistic Record Linkage
Management of legal bases [13, 8]	Ontologies, Vocabularies, Technology for the Semantic Web, Smart Contracts
Security & Privacy	
Security of at-rest and in-transit data	Symmetric key encryption, Public-key encryption, Identity-Based Encryption, Attribute Based Encryption, Proxy re-encryption
Non-repudiation	Digital Signature, Digital Certificates, Blockchain
Integrity	Hashing, Encryption, Digital signature
General robustness from attacks	Proof of Work (hashing), Encryption
Availability, Disaster Recovery and Fault Tolerance	Cloud, Blockchain, Proof of Work (hashing), Backup solutions
Auditing, activities tracing and logging	Blockchain
Anonymisation and pseudonymisation procedures	Differential Privacy, Symmetric Encryption
Data minimisation and once-only principles [8, 15]	Single Sign-On, Data Masking and Anonymisation
Access control, authorisation and authentication	Smart contracts (blockchain), Proxy re-encryption, Attribute Based Encryption, Attribute Based Access Control, Access Control Lists, Single Sign-On, Smart Contracts

are particularly emphasised. Furthermore, there are also emerging technological trends in privacy-friendly analysis, such as the usage of cryptographic techniques, federated learning, and differential privacy. In conjunction with legal frameworks, these developments have facilitated the identification of specific requirements for a solution that effectively addresses both technical and legal obstacles initially elucidated. Finally, this collaborative effort has resulted in the compilation of a draft of technical solutions tailored to the identified requirements.

Acknowledgments

The work presented in this paper is funded by the European Union – Next Generation EU, Mission 4 Component 2 Investment 1.1 CUP D53D23022370001 (GA n. P2022MSMAW), PRIN 2022 PNRR “PADS4Health”.

Declaration on Generative AI

During the preparation of this work, the author(s) used Apple Intelligence in order to: Grammar and spelling check.

References

- [1] A. Torab-Miandoab, T. Samad-Soltani, A. Jodati, P. Rezaei-Hachesu, Interoperability of heterogeneous health information systems: a systematic literature review, *BMC medical informatics and decision making* 23 (2023) 18.
- [2] M. Lehne, J. Sass, A. Essenwanger, J. Schepers, S. Thun, Why digital medicine depends on interoperability, *NPJ digital medicine* 2 (2019) 79.
- [3] I. Keshta, A. Odeh, Security and privacy of electronic health records: Concerns and challenges, *Egyptian Informatics Journal* 22 (2021) 177–183.
- [4] D. Povey, Optimistic security: a new access control paradigm, in: *Proceedings of the 1999 workshop on New security paradigms*, 1999, pp. 40–45.
- [5] A. D. Brucker, H. Petritsch, Extending access control models with break-glass, in: *Proceedings of the 14th ACM symposium on Access control models and technologies*, 2009, pp. 197–206.
- [6] M. Joshi, K. P. Joshi, T. Finin, Delegated authorization framework for ehr services using attribute-based encryption, *IEEE Transactions on Services Computing* 14 (2019) 1612–1623.
- [7] D. Cirillo, S. Catuara-Solarz, C. Morey, E. Guney, L. Subirats, S. Mellino, A. Gigante, A. Valencia, M. J. Rementeria, A. S. Chadha, et al., Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare, *NPJ digital medicine* 3 (2020) 81.
- [8] European Parliament, Council of the European Union, Regulation (EU) 2016/679 of the European Parliament and of the Council, 2016. URL: <https://data.europa.eu/eli/reg/2016/679/oj>.
- [9] The 104th United States Congress, Health insurance portability and accountability act, 1996. Public Law 104-191.
- [10] European Parliament, Council of the European Union, Directive (eu) 2019/1024 of the european parliament and of the council of 20 june 2019 on open data and the re-use of public sector information (recast), 2019. URL: <http://data.europa.eu/eli/dir/2019/1024/oj>.
- [11] European Parliament, Council of the European Union, Regulation (eu) 2022/868 of the european parliament and of the council of 30 may 2022 on european data governance and amending regulation (eu) 2018/1724 (data governance act) (text with eea relevance), 2022. URL: <http://data.europa.eu/eli/reg/2022/868/oj>.
- [12] European Parliament, Council of the European Union, 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 (text with eea relevance), 2025. URL: <http://data.europa.eu/eli/reg/2025/327/oj>.

- [13] P. of the Italian Republic, Italian personal data protection code - legislative decree no. 196 of 30 june 2003, 2003. URL: https://www.gazzettaufficiale.it/atto/serie_generale/caricaDettaglioAtto/originario?atto.dataPubblicazioneGazzetta=2003-07-29&atto.codiceRedazionale=003G0218.
- [14] A. Ghorbel, M. Ghorbel, M. Jmaiel, Privacy in cloud computing environments: a survey and research challenges, *J. Supercomput.* 73 (2017) 2763–2800.
- [15] European Parliament, Council of the European Union, Communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions eu egovernment action plan 2016-2020 accelerating the digital transformation of government, 2016. URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52016DC0179>.
- [16] A. Gaye, Y. Marcon, J. Isaeva, P. LaFlamme, A. Turner, E. M. Jones, J. Minion, A. W. Boyd, C. J. Newby, M.-L. Nuotio, R. Wilson, O. Butters, B. Murtagh, I. Demir, D. Doiron, L. Giepmans, S. E. Wallace, I. Budin-Ljøsne, C. Oliver Schmidt, P. Boffetta, M. Boniol, M. Bota, K. W. Carter, N. deKlerk, C. Dibben, R. W. Francis, T. Hiekkalinna, K. Hveem, K. Kvaløy, S. Millar, I. J. Perry, A. Peters, C. M. Phillips, F. Popham, G. Raab, E. Reischl, N. Sheehan, M. Waldenberger, M. Perola, E. van den Heuvel, J. Macleod, B. M. Knoppers, R. P. Stolk, I. Fortier, J. R. Harris, B. H. Woffenbuttel, M. J. Murtagh, V. Ferretti, P. R. Burton, Datashield: taking the analysis to the data, not the data to the analysis, *International Journal of Epidemiology* 43 (2014) 1929–1944. URL: <https://doi.org/10.1093/ije/dyu188>. doi:10.1093/ije/dyu188.
- [17] A. J. McMurphy, S. N. Murphy, D. MacFadden, G. Weber, W. W. Simons, J. Orechia, J. Bickel, N. Wattanasin, C. Gilbert, P. Trevvett, S. Churchill, I. S. Kohane, SHRINE: Enabling Nationally Scalable Multi-Site Disease Studies, *PLOS ONE* 8 (2013) 1–11. URL: <https://doi.org/10.1371/journal.pone.0055811>. doi:10.1371/journal.pone.0055811.
- [18] J. L. Raisaro, J. R. Troncoso-Pastoriza, M. Misbach, J. S. Sousa, S. Pradervand, E. Missiaglia, O. Michielin, B. Ford, J. Hubaux, MedCo: Enabling Secure and Privacy-Preserving Exploration of Distributed Clinical and Genomic Data, *IEEE ACM Trans. Comput. Biol. Bioinform.* 16 (2019) 1328–1341. URL: <https://doi.org/10.1109/TCBB.2018.2854776>. doi:10.1109/TCBB.2018.2854776.
- [19] O. Beyan, A. Choudhury, J. van Soest, O. Kohlbacher, L. Zimmermann, H. Stenzhorn, M. R. Karim, M. Dumontier, S. Decker, L. O. B. da Silva Santos, A. Dekker, Distributed analytics on sensitive medical data: The personal health train, *Data Intell.* 2 (2020) 96–107. URL: https://doi.org/10.1162/dint_a_00032. doi:10.1162/DINT\A_00032.
- [20] A. Sadilek, L. Liu, D. Nguyen, M. Kamruzzaman, S. Serghiou, B. Rader, A. Ingerman, S. Mellem, P. Kairouz, E. O. Nsoesie, J. Macfarlane, A. Vullikanti, M. V. Marathe, P. Eastham, J. S. Brownstein, B. A. y Arcas, M. D. Howell, J. Hernandez, Privacy-first health research with federated learning, *NPJ Digit. Medicine* 4 (2021). URL: <https://doi.org/10.1038/s41746-021-00489-2>. doi:10.1038/s41746-021-00489-2.
- [21] K. Yu, L. Tan, X. Shang, J. Huang, G. Srivastava, P. Chatterjee, Efficient and privacy-preserving medical research support platform against COVID-19: A blockchain-based approach, *IEEE Consumer Electron. Mag.* 10 (2021) 111–120. URL: <https://doi.org/10.1109/MCE.2020.3035520>. doi:10.1109/MCE.2020.3035520.
- [22] S. Mendelsohn, D. Froelicher, D. Loginov, D. Bernick, B. Berger, H. Cho, sikit: a web-based toolkit for secure and federated genomic analysis, *Nucleic Acids Res.* 51 (2023) 535–541. URL: <https://doi.org/10.1093/nar/gkad464>. doi:10.1093/NAR/GKAD464.
- [23] H. Cho, D. J. Wu, B. Berger, Secure genome-wide association analysis using multiparty computation, *Nature biotechnology* 36 (2018) 547–551. URL: <https://doi.org/10.1038/nbt.4108>. doi:10.1038/nbt.4108.
- [24] D. Froelicher, J. R. Troncoso-Pastoriza, J. L. Raisaro, M. A. Cuendet, J. S. Sousa, H. Cho, B. Berger, J. Fellay, J.-P. Hubaux, Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption, *Nature communications* 12 (2021) 5910. URL: <https://doi.org/10.1038/s41467-021-25972-y>. doi:10.1038/s41467-021-25972-y.
- [25] F. N. Wirth, T. Meurers, M. Johns, F. Prasser, Privacy-preserving data sharing infrastructures for medical research: systematization and comparison, *BMC Medical Informatics Decis. Mak.* 21 (2021) 242. URL: <https://doi.org/10.1186/s12911-021-01602-x>. doi:10.1186/s12911-021-01602-x.

- [26] P. H. Scotland, National Safe Haven (NSH), <https://publichealthscotland.scot/resources-and-tools/health-intelligence-and-data-management/electronic-data-research-and-innovation-service-edris/national-safe-haven-nsh/>, 2025. [Accessed 28-07-2025].
- [27] E. Business, I. Agency, Estonian e-Health Records — e-estonia.com, <https://e-estonia.com/solutions/e-health-2/e-health-records/>, 2025. [Accessed 24-06-2025].
- [28] K. Mäeots, KSI Blockchain Stack: Zero Trust Applications - DigiExpo — digiexpo.e-estonia.com, <https://digiexpo.e-estonia.com/cyber-security/ksi-blockchain-stack-zero-trust-applications/>, 2025. [Accessed 24-06-2025].
- [29] E. Business, I. Agency, X-Road - e-Estonia — e-estonia.com, <https://e-estonia.com/solutions/interoperability-services-x-road/x-road/>, 2025. [Accessed 24-06-2025].
- [30] G. Hripcsak, J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, M. A. Suchard, R. W. Park, I. C. K. Wong, P. R. Rijnbeek, et al., Observational health data sciences and informatics (ohdsi): opportunities for observational researchers, in: MEDINFO 2015: eHealth-enabled Health, IOS Press, 2015, pp. 574–578.
- [31] Harvard Catalyst, Development - SHRINE - open.catalyst wiki — open.catalyst.harvard.edu, <https://open.catalyst.harvard.edu/wiki/display/SHRINE/Development>, 2025. [Accessed 24-06-2025].
- [32] U.S. Centers for Disease Control and Prevention, ICD - ICD-9-CM - International Classification of Diseases, Ninth Revision, Clinical Modification — archive.cdc.gov, https://archive.cdc.gov/www_cdc_gov/nchs/icd/icd9cm.htm, 2025. [Accessed 24-06-2025].
- [33] NCBO, Ncbo bioportal, <https://bioportal.bioontology.org>, 2025. [Accessed 24-06-2025].
- [34] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, *Scientific data* 3 (2016) 1–9. URL: <https://doi.org/10.1038/sdata.2016.18>. doi:10.1038/sdata.2016.18.
- [35] Y. Zhuang, L. R. Sheets, Y. Chen, Z. Shae, J. J. P. Tsai, C. Shyu, A patient-centric health information exchange framework using blockchain technology, *IEEE J. Biomed. Health Informatics* 24 (2020) 2169–2176. URL: <https://doi.org/10.1109/JBHI.2020.2993072>. doi:10.1109/JBHI.2020.2993072.
- [36] K. Kõnd, A. Lilleväli, E-prescription success in estonia: The journey from paper to pharmacoeconomics., *Eurohealth* 5 (2019) 18–20.
- [37] Research Data Assistance Center, CCW Virtual Research Data Center (VRDC) | ResDAC — resdac.org, <https://resdac.org/cms-virtual-research-data-center-vrdc>, 2025. [Accessed 24-06-2025].
- [38] Research Data Assistance Center, CCW Virtual Research Data Center (VRDC) FAQs | ResDAC — resdac.org, <https://resdac.org/virtual-research-data-center-vrdc-faqs>, 2025. [Accessed 24-06-2025].
- [39] S. Padmanabhan, L. Carty, E. Cameron, R. E. Ghosh, R. Williams, H. Strongman, Approach to record linkage of primary care data from clinical practice research datalink to other health-related patient data: overview and implications, *European journal of epidemiology* 34 (2019) 91–99.