frontiers | Research Topics

Unlocking the potential of health data spaces with the proliferation of new tools, technologies and digital solutions

Edited by
Gokce Banu Laleci Erturkmen, Dagmar Krefting and Oya Beyan

Coordinated by
Adamantios Koumpis

Published in
Frontiers in Medicine

ISSN 1664-8714
ISBN 978-2-8325-6095-2
DOI 10.3389/978-2-8325-6095-2

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Unlocking the potential of health data spaces with the proliferation of new tools, technologies and digital solutions

Topic editors

Gokce Banu Laleci Erturkmen — Software Research and Development Consulting, Türkiye

Dagmar Krefting — University Medical Center Göttingen, Germany

Oya Beyan — University Hospital of Cologne, Germany

Topic coordinator

Adamantios Koumpis — University Hospital of Cologne, Germany

Citation

Erturkmen, G. B. L., Krefting, D., Beyan, O., Koumpis, A., eds. (2025). *Unlocking the potential of health data spaces with the proliferation of new tools, technologies and digital solutions*. Lausanne: Frontiers Media SA.

doi: 10.3389/978-2-8325-6095-2

Table of contents

05 Will it run?—A proof of concept for smoke testing decentralized data analytics experiments
Sascha Welten, Sven Weber, Adrian Holt, Oya Beyan and Stefan Decker

18 Unlocking the potential of big data and AI in medicine: insights from biobanking
Kaya Akyüz, Mónica Cano Abadía, Melanie Goisauf and Michaela Th. Mayrhofer

24 Health data space nodes for privacy-preserving linkage of medical data to support collaborative secondary analyses
Martin Baumgartner, Karl Kreiner, Aaron Lauschensky, Bernhard Jammerbund, Klaus Donsa, Dieter Hayn, Fabian Wiesmüller, Lea Demelius, Robert Modre-Osprian, Sabrina Neururer, Gerald Slamanig, Sarah Prantl, Luca Brunelli, Bernhard Pfeifer, Gerhard Pölzl and Günter Schreier

41 Cybersecurity policy framework requirements for the establishment of highly interoperable and interconnected health data spaces
Christian Luidold and Christoph Jungbauer

53 Artificial intelligence based data curation: enabling a patient-centric European health data space
Isabelle de Zegher, Kerli Norak, Dominik Steiger, Heimo Müller, Dipak Kalra, Bart Scheenstra, Isabella Cina, Stefan Schulz, Kanimozhi Uma, Petros Kalendralis, Eno-Martin Lotman, Martin Benedikt, Michel Dumontier and Remzi Celebi

65 OHDSI-compliance: a set of document templates facilitating the implementation and operation of a software stack for real-world evidence generation
Felix N. Wirth, Hammam Abu Attieh and Fabian Prasser

72 Transforming evidence-based clinical guidelines into implementable clinical decision support services: the CAREPATH study for multimorbidity management
Mert Gencturk, Gokce B. Laleci Erturkmen, A. Emre Akpinar, Omid Pournik, Bilal Ahmad, Theodoros N. Arvanitis, Wolfgang Schmidt-Barzynski, Tim Robbins, Ruben Alcantud Corcoles and Pedro Abizanda

94 Streamlining intersectoral provision of real-world health data: a service platform for improved clinical research and patient care
Katja Hoffmann, Igor Nesterow, Yuan Peng, Elisa Henke, Daniela Barnett, Cigdem Klengel, Mirko Gruhl, Martin Bartos, Frank Nüßler, Richard Gebler, Sophia Grummt, Anne Seim, Franziska Bathelt, Ines Reinecke, Markus Wolfien, Jens Weidner and Martin Sedlmayr

103 Showcasing the Saudi e-referral system experience: the epidemiology and pattern of referrals utilising nationwide secondary data
Nawfal A. Aljerian, Abdullah A. Alharbi, Reem S. AlOmar, Meshary S. Binhotan, Hani A. Alghamdi, Mohammed S. Arafat, Abdulrahman Aldhabib and Mohammed K. Alabdulaali

114 A reference architecture for personal health data spaces using decentralized content-addressable storage networks
Toomas Klementi, Gunnar Piho and Peeter Ross

136 A scalable and transparent data pipeline for AI-enabled health data ecosystems
Tuncay Namli, Ali Anıl Sınacı, Suat Gönül, Cristina Ruiz Herguido, Patricia Garcia-Canadilla, Adriana Modrego Muñoz, Arnau Valls Esteve and Gökçe Banu Laleci Ertürkmen

156 Seeing the primary tumor because of all the trees: Cancer type prediction on low-dimensional data
Julia Gehrmann, Devina Johanna Soenarto, Kevin Hidayat, Maria Beyer, Lars Quakulinski, Samer Alkarkoukly, Scarlett Berressem, Anna Gundert, Michael Butler, Ana Grönke, Simon Lennartz, Thorsten Persigehl, Thomas Zander and Oya Beyan

167 Bridging health registry data acquisition and real-time data analytics
Johannes Schmidt, Sita Arjune, Volker Boehm, Franziska Grundmann, Roman-Ulrich Müller and Philipp Antczak

frontiers | Frontiers in Medicine

TYPE Technology and Code
PUBLISHED 08 January 2024
DOI 10.3389/fmed.2023.1305415

OPEN ACCESS

EDITED BY
Elisio Costa,
University of Porto, Portugal

REVIEWED BY
Ana Corte-Real,
University of Coimbra, Portugal
Mert Gencturk,
Software Research and Development Consulting, Türkiye

*CORRESPONDENCE
Sascha Welten
✉ welten@dbis.rwth-aachen.de

RECEIVED 01 October 2023
ACCEPTED 14 December 2023
PUBLISHED 08 January 2024

CITATION
Welten S, Weber S, Holt A, Beyan O and Decker S (2024) Will it run?—A proof of concept for smoke testing decentralized data analytics experiments.
Front. Med. 10:1305415.
doi: 10.3389/fmed.2023.1305415

Will it run?—A proof of concept for smoke testing decentralized data analytics experiments

Sascha Welten[1]*, Sven Weber[1,2], Adrian Holt[1], Oya Beyan[2,3] and Stefan Decker[2,3]

[1]Chair of Computer Science 5, Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen University, Aachen, Germany, [2]Institute for Biomedical Informatics, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany, [3]Fraunhofer Institute for Applied Information Technology FIT, St. Augustin, Germany

The growing interest in data-driven medicine, in conjunction with the formation of initiatives such as the European Health Data Space (EHDS) has demonstrated the need for methodologies that are capable of facilitating privacy-preserving data analysis. Distributed Analytics (DA) as an enabler for privacy-preserving analysis across multiple data sources has shown its potential to support data-intensive research. However, the application of DA creates new challenges stemming from its distributed nature, such as identifying single points of failure (SPOFs) in DA tasks before their actual execution. Failing to detect such SPOFs can, for example, result in improper termination of the DA code, necessitating additional efforts from multiple stakeholders to resolve the malfunctions. Moreover, these malfunctions disrupt the seamless conduct of DA and entail several crucial consequences, including technical obstacles to resolve the issues, potential delays in research outcomes, and increased costs. In this study, we address this challenge by introducing a concept based on a method called Smoke Testing, an initial and foundational test run to ensure the operability of the analysis code. We review existing DA platforms and systematically extract six specific Smoke Testing criteria for DA applications. With these criteria in mind, we create an interactive environment called Development Environment for AuTomated and Holistic Smoke Testing of Analysis-Runs (DEATHSTAR), which allows researchers to perform Smoke Tests on their DA experiments. We conduct a user-study with 29 participants to assess our environment and additionally apply it to three real use cases. The results of our evaluation validate its effectiveness, revealing that 96.6% of the analyses created and (Smoke) tested by participants using our approach successfully terminated without any errors. Thus, by incorporating Smoke Testing as a fundamental method, our approach helps identify potential malfunctions early in the development process, ensuring smoother data-driven research within the scope of DA. Through its flexibility and adaptability to diverse real use cases, our solution enables more robust and efficient development of DA experiments, which contributes to their reliability.

KEYWORDS

decentralized applications, federated learning, machine learning, software testing, simulation, web services

1 Introduction

Data-driven analyses, such as basic statistics or Machine Learning (ML)-based approaches, have been extensively used for analyzing data in a variety of applications such as medical diagnosis and treatment or financial business intelligence (1–3). Traditionally, data is collected from several sources, stored in a central location, and analyzed by scientists. However, data centralization poses several challenges (4). For example, due to the exponential growth of data, the gathered data volume might not allow central storage, or in some cases, it would be too expensive (5). Besides these technical challenges, regulations such as the General Data Protection Regulation (GDPR) in the European Union[1] prohibit or limit the centralization of personal data due to privacy concerns and its level of sensitivity. This issue is particularly present in domains such as healthcare, where personal data is protected (5). In the context of the European Health Data Space (EHDS)[2], the issue of accessing fragmented and silo-ed data is intended to be resolved through the implementation of Federated Health Data Networks (FHDNs) that consist of decentralized and interconnected nodes, allowing data to be analyzed by participants of the FHDNs (6). In order to enable data analysis across multiple nodes, key technologies for DA [such as Federated Learning (FL)] have been considered as indispensable and proposed as a solution by omitting the need for data centralization (7, 8). Here, the analysis code is executed at the data source(s), and only the (intermediate) analysis results, such as aggregated statistics or, in ML-terms, model parameters, are transmitted between the data providers rather than sharing actual data instances. DA provides solutions for several legal considerations such as patient data ownership or data control (9). This includes compliance with measures such as the GDPR. Furthermore, ensuring transparent and accountable access to this data is crucial to uphold privacy and security standards (9). Since it addresses challenges, such as data privacy, high storage costs, or long transfer times, Distributed Analytics (DA) has recently gained attention and has found applications in various use cases, including skin cancer classification, predictive modeling using radiomics for lung cancer, brain tumor segmentation, and breast cancer detection (5, 10–14).

Before analyses can deliver their full potential, several steps must be taken to build an error-free and robust analysis code. Among other steps, we recognize three essential phases: Development, testing, and execution phase (Figure 1) (15). The development phase involves implementing the code, covering a data pre-processing routine and the analysis script. During the testing phase, there may be two types of testing scenarios: one is testing from a software perspective that ensures the code is executable. The other is analysis validation using test data to assess performance. The execution phase covers the application of the analyses on real data to obtain actual analysis outcomes. At this point, it becomes evident that these standard workflows assign an essential role to the availability of data: Without sufficient data, fast prototyping through, e.g., trial-and-error and software tests, can be only conducted on a limited basis. Moreover, up to now and to the best of our knowledge, how DA code is tested has been left to the developer's responsibility and intuition, showcasing a lack of clearly defined testing criteria and capabilities in the domain of DA. This circumstance entails a specific degree of uncertainty regarding the analysis code during its execution: Will it run? The consequence is that insufficiently tested analysis code is susceptible to single points of failure (SPOFs) during the execution phase, such that another development round is needed to fix the code (Figure 1). Due to the decentralized nature of DA, any kind of errors during the execution require the analysis code to be re-built, re-distributed to the data holders, and re-executed (Figure 1). This re-distribution is time-consuming and potentially involves multiple parties, e.g., in the medical domain, where the analysis has to be verified before interacting with data. Thus, there is a need for adequate testing criteria and capabilities that identify potential malfunctions in the code before its execution.

1.1 Objectives

To establish an initial foundation for testing in DA, we derive requirements for DA code, which should be fulfilled to ensure that the analysis code is operational. We aim to define criteria for a testing approach called *Smoke Testing* to support developers in their development process (16, 17). These criteria constitute the minimum requirements for DA code that must be guaranteed before its execution. We hypothesize that without these requirements the DA code will definitely fail or cause undesired behavior. In summary, we evaluate the following research question:

RQ1 What are suitable Smoke Testing criteria for DA executions?

Secondly, we intend to develop a Smoke Testing suite as a Proof of Concept (PoC), specifically designed to evaluate analysis code according to our defined criteria. Since data is essential to test data-driven analysis code properly, we aim for a Smoke Testing suite capable of generating data instances that can be used for Smoke Testing, making our approach less reliant on prior data-sharing. Regarding this, we hypothesize that a simulation-based Smoke Testing suite reduces the dependence on data providers. One of our core assumptions is that data schema details are shared, while actual sensitive data instances can be kept under seal by the data providers. To reach this goal, we will evaluate the following research questions:

RQ2 What is necessary to enable Smoke Testing on DA code?

RQ2.1 How can privacy-preserving testing of DA algorithms be enabled?

RQ2.2 How can the execution of DA algorithms be (Smoke) tested without a real DA environment?

1 GDPR: www.gdpr-info.eu.

2 https://www.europarl.europa.eu/RegData/etudes/STUD/2022/740054/IPOL_STU(2022)740054_EN.pdf

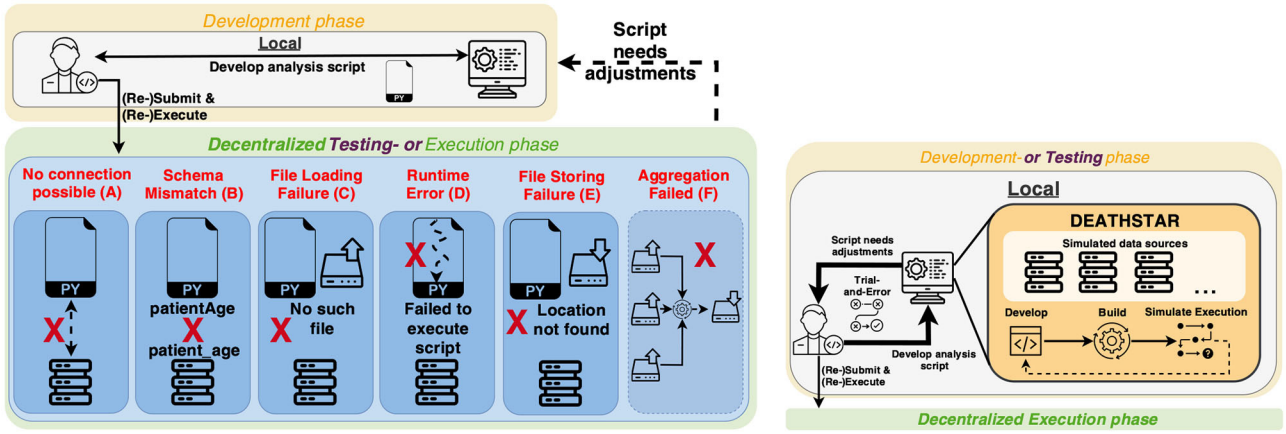

FIGURE 1
Proposed concept for local smoke tests. Up to now, the (smoke) testing phase of the analysis code has been postponed to the actual execution, making it difficult to differentiate between the testing and execution phases **(left)**. As a result, any code errors can only be identified by running the code at each data provider. This dependence on distributed data providers can make creating DA code time-consuming and cumbersome. Instead of smoke testing during the execution phase itself, we propose DEATHSTAR to iteratively trial-and-error the analysis script locally **(right)**.

1.2 Contributions and findings

Aligned with our objectives from the previous section, this work presents the following contributions:

- We propose six criteria for Smoke Testing that we derive from a literature review of DA infrastructure implementations. Those criteria must be met by DA analyses in order to ensure their operability.
- We developed and implemented a Smoke Testing suite, called Development Environment for AuTomated and Holistic Smoke Testing of Analysis-Runs (DEATHSTAR)[3]. DEATHSTAR employs a *testing-through-simulation* approach to identify potential malfunctions in the analysis code by systematically validating our six criteria. This PoC, inspired by Integrated Development Environments (IDEs), allows the prototyping and simulation of DA experiments on synthetic or (real) sample data.
- We conduct a User-Study with 29 participants to evaluate the effectiveness of our criteria and the usability of DEATHSTAR.
- We lastly present a technical evaluation demonstrating the flexibility and adaptability of our approach by successfully repeating and reproducing three real-world use cases.

Overall, we find that almost all DA algorithms (96.6%), developed and (Smoke) tested by participants of our User-Study using our approach, terminated with no errors in a real DA execution. These results suggest that the six criteria we proposed are sufficient for ensuring the operability of the analysis code. Additionally, we achieved a System Usability Scale (SUS) score of 88.3 in our User-Study, which is considered to be "excellent" (18). The outcomes of the second part of our evaluation show that our concept can support DA-driven research under real circumstances and is flexible enough to serve various data types and sources.

2 Method

In the previous section, it became apparent that the essential element of DA approaches is the analysis code. As these analyses are executable software fragments, they can consequently be vulnerable to unexpected failure during the execution, like any software product (19). For example, the algorithm might not be compatible with a specific data source version or contain a logical error that needs to be resolved before the execution (see Figure 1). As the most widespread method to verify software quality, testing can prevent such failures (19). Moreover, the importance of testing is also evident when reviewing so-called Software Development Life Cycles (SDLCs) (20). These SDLC models describe systematic processes on how software should be developed and what steps should be taken in the SDLC (21). Consequently, an SDLC model can control costs, reliability, performance, and functionality of the developed software (21). As a result, various SDLC models have been developed and play a significant role in software engineering (15). It is worth noting that each SDLC model embraces a testing phase, which emphasizes that testing is indispensable in professional software development (15). Specifically for DA, the necessity of testing capabilities has already been formulated in work by Bonawitz et al. who state that an environment for testing and simulation of analysis algorithms is a requirement for DA platforms (22). One specific testing method playing a major role in this work is called *Smoke Testing* (16, 23, 24). This term stems from the industry and includes an initial and fundamental test run to ensure that a program—here: the analysis—is operational, executes successfully, and does not *end up in smoke*. For example, Herbold and Haar successfully applied Smoke Testing to find problems in analytics software libraries and algorithms (16). Specifically, they

3 The code of the PoC is available as Open Source, including screenshots, a screencast, examples, the developed data schema, and explanations at: https://github.com/PADME-PHT/playground.

designed a total of 37 Smoke Tests for classification- and clustering algorithms (16).

The methodology of this paper is inspired by the work of Cannavacciuolo and Mariani (17), who applied Smoke Testing to cloud systems, intending to validate whether a system is operational post-deployment, which helped to determine if more sophisticated tests can be conducted. As part of their work, they propose several Smoke Testing criteria that can be used as a foundation for creating Smoke Testing suites in the scope of cloud systems (17). Since the relevant DA platforms discussed in Section 2.2 are typically not deployed in cloud systems, and our primary emphasis is on (Smoke) testing analysis code rather than an entire infrastructure, these criteria are not applicable to our specific scenario. Nevertheless, they have specified three key characteristics of Smoke Tests, that serve as an inspiration for our work. Those characteristics define the way *how* Smoke Testing criteria should be validated:

- **Shallow:** Smoke Tests should be kept at a higher abstraction level and not overly detailed. This means that only a system's or software's basic functionality and operability should be validated. It is just serving as a prerequisite for more sophisticated testing methods.
- **Fast:** Smoke Tests must be fast in their execution since they are performed before other test runs or, in our scenario, the analysis execution.
- **Automatic:** As an extension to the *fast* characteristic, Smoke Tests should be fully automated to reduce manual intervention.

To realize Smoke Testing suites, so-called playgrounds or prototyping environments may provide a possible solution (22). Here, the term playground refers to services that allow users to interact and *play* with software without prior complex setup or configuration (25). Moreover, these playgrounds enable users to iteratively (i.e., trial-and-error) develop and priorly test their entire implementation or specific modules (25, 26). Because playgrounds have successfully enabled testing approaches in other settings, our work pursues a similar approach (25–28).

We begin the conceptualization of such a Smoke Testing suite by abstracting and formalizing the scenario, focusing on the relevant steps in which the analysis execution might fail based on related works in the DA domain (Section 2.2). Moreover, our approach aims for a user-centric design, so we initially describe the problem statement from a user perspective (Section 2.1). The outcome of this abstraction is a formal model that describes the analysis process of the code, which is distributed within a DA infrastructure. Based on the steps in the process model, we derive our set of Smoke Test criteria that aim to ensure that each step can be executed (Section 2.3). We aim to keep the set of criteria as "shallow" as possible to comply with the defined characteristics of Smoke Testing (see above). Subsequently, we present a PoC implementation that can apply Smoke Tests to analysis code based on our defined criteria (Section 2.4). We aim for a "fast" and "automated" solution consistent with the Smoke Testing characteristics. Lastly, we evaluate the effectiveness of our solution, its usability, and we apply it to three distinct use cases as part of our technical evaluation (Section 3). For the implementation and evaluation, we use the DA platform PADME as infrastructure to execute the analyses (7).

2.1 User-centered problem description

Initially, developers or scientists who intend to conduct a DA experiment need to develop the code for the analysis, which is designed to analyze data provided by decentralized data holders (see Figure 1). The development process usually occurs locally or on a machine the developer can access. It is vital to test the analysis code to ensure its proper operation after the development (or even during it, through a trial-and-error approach). While certain parts and components of the code can be tested on a module-by-module basis, the presented setting has a shortcoming: To conduct a complete test of the code, the developer requires (sample) data to execute the developed algorithms on. However, the availability of sufficient and potentially sensitive data for testing purposes is not guaranteed due to the mentioned data protection and privacy regulations. As a result, researchers are left with two options. In case sample data is available, following an *ad hoc* testing approach might not cover all criteria that are needed to ensure the operability of the code. Secondly, in the worst case, the developer is obliged to submit the analysis script to each data provider and wait for its execution on their data in order to identify potential issues in the code. These circumstances result in an inefficient development process since the developer is reliant on the data providers, and even minor malfunctions (such as `Index-out-of-Bounds`, `Nullpointer`, `TypeCast` exceptions) can cause a new development round. From an abstract perspective on this scenario, the testing phase is closely coupled with the actual execution phase, which causes the mentioned inefficiency (see Figure 1). Usually, the testing phase is designed to support the development phase to allow for fast code updates and trial-and-error development. Therefore, in this work, we aim to separate the testing and execution phases and provide a solution that facilitates Smoke Testing during or after the development phase (see Figure 1, right).

2.2 Abstract workflow

Our initial step involves examining how the analysis code operates on a conceptual and abstract level. In general, two execution policies exist that enable DA: A parallel and a sequential approach (sometimes referred to as FL and Institutional Incremental Learning (IIL), respectively) (13, 29). In IIL, the data holders are arranged in a sequence, and the analysis code is sent from institution to institution until the last institution sends the final (and aggregated) results back. The procedure for FL repeats the following steps: First, the analysis algorithm is simultaneously distributed to all participating data holders. Then, each data holder executes the analysis algorithm on the local data and sends the result of this analysis back to the central component. The central component aggregates all partial results, combining the results of all participants. This aggregated result is either the final or intermediate result for the next so-called communication or federated round. The conduct of a DA experiment generally

TABLE 1 Applicability of the six steps identified in this paper to different DA infrastructures.

References	S1	S2	S3	S4	S5	S6
PHT (IIL) (7, 32, 33)	✓	✓	✓	✓	✓	
DS (FL) (31)	✓	✓	✓	✓	✓	✓
Swarm Learning (P2P) (35)	✓	✓	✓	✓	✓	✓
SMPC (P2P) (36)	✓	✓	✓	✓	✓	(✓)

Steps required by an infrastructure (row) are shown as checkmarks in the respective column. All infrastructures require connecting to a data source (S1), querying data (S2), loading previous results (S3), executing the analysis (S4), and storing results (S5). Some infrastructures require result aggregation (S6).

requires an infrastructure that orchestrates the analysis and transmits the code to the data holder according to one of the foundational execution policies mentioned above. In recent years, several implementations of DA have been proposed. DataSHIELD (DS) is an open-source solution that follows the FL approach and uses the programming language *R*, often used in statistics[4] (30, 31). Another emerging concept is the Personal Health Train (PHT), which follows the sequential paradigm. The PHT uses software containers[5] to distribute the analysis code to each data provider. Some implementations following the PHT concept are Vantage6, PHT-meDIC, and PADME by Welten et al. (7, 32, 33). Besides FL and IIL, additional (hybrid) approaches for DA exist: Swarm Learning (SL) and Secure Multiparty Computation (SMPC), which use Peer-To-Peer (P2P) communication instead of relying on a central component (34, 35). These infrastructures, founded on the dispatching paradigms, such as IIL and FL, serve as the source for our abstraction.

After systematically reviewing these infrastructures, studies conducted with them, and our personal experiences from DA experiments, we have identified six abstract steps (S1–S6) that the analysis code performs during its execution, as shown in Table 1. We transformed our findings into a process diagram for a better overview of the abstract workflow (Figure 2). Despite how the (intermediate) results are finally combined, the infrastructures do not differ in their workflow on the conceptual level. First, the developed code must establish a database connection (S1). Then, the analysis queries the data (S2) and loads the intermediate results (S3) from previous execution rounds. The queried data from Step 2 and the previous results from Step 3 serve as the input for a generic analysis code. During the data analysis (S4), the queried data is used to compute updated analysis results. Once the analysis terminates, the updated results are stored (S5). In the IIL-setting, the results are stored in the analysis payload, which is then transmitted to the next data provider. In contrast, for FL, the results are directly transmitted to a central aggregation component, where the intermediate results of all analysis replicas are aggregated into a single global result (S6). As each approach we examined is round-based, these six steps are repeated in each subsequent round. In the IIL scenario, a new round starts after the analysis has been sent to the next data holder. On the other hand, in the FL scenario, a round begins after the aggregator has combined all results. Hence, the approaches following the paradigm of parallel analysis executions undergo an additional step.

2.3 Criteria definition

Now that we have our abstract workflow model, we define six criteria that must be fulfilled to ensure that the analysis code is operational in every of our derived execution steps. For each requirement, we linked the corresponding step in our workflow.

Requirement A: Proper connection interface. The analysis code should be able to establish a connection to the data source without any issues. This necessitates that the algorithm's configuration is compatible with the data source's connection interface(s). Proper configuration implies that all connection parameters (e.g., file path, hostname, port number, or database type) are correct and available (S1).

Requirement B: Matching schema. The analysis code should be able to send syntactically correct queries to the data store and receive corresponding results in response. Hence, the expected data schema of the analysis code must match the actual data schema of the data source. Note that Requirement A focuses on the technical aspect of connecting to the data source. Requirement B refers to successfully establishing a connection based on data (schema) compatibility (S2).

Requirement C: Load previous (intermediate) results. Loading the (intermediate) results from previous executions into the analysis code is necessary to enable result updates, representing the core functionality of DA. In the first round, we require a successful initialization if necessary (S3).

Requirement D: Analysis execution without errors. If the Requirements **A**, **B**, and **C** hold, the actual DA algorithm should run without encountering any errors. An error-free execution is indicated by, e.g., the `exit code 0` (S4).

Requirement E: Successful result storage. The analysis code should save the analysis results in the appropriate location and format. The term "correct location" refers to emitting the results as either a file or a processable bit string for transmission. This guarantees extractable analysis results, which the researcher can inspect after the execution (S5).

Requirement F: Successful result aggregation. In aggregation-based approaches (e.g., FL), we additionally require that the central aggregation of the intermediate results computed and stored in steps 4&5 terminates without an error (S6).

It is worth noting that we interpret these six requirements as the root causes of SPOFs and as the fundamental factors that must be met for an analysis to terminate properly. As such, these requirements only represent a subset (see "shallow" criterion) of potential additional criteria. To illustrate, it may be necessary to ensure a reliable and low-latency connection between the entities involved in DA to guarantee the proper transmission of the analysis code. However, we argue that such criteria are mainly subject to the responsibility of the DA infrastructure providers rather than the developers of the analysis code. Consequently, we have only considered requirements that developers and the analysis codes can directly influence. Additionally, we do not check for the plausibility of the results. Since DA can cover a wide spectrum of analysis types, we argue that validating the result's plausibility might contradict the "shallow" and the "fast" criteria since possible tests might be

4 Further DS studies are available at: www.datashield.org/about/publications.

5 Open Container Initiative: https://opencontainers.org.

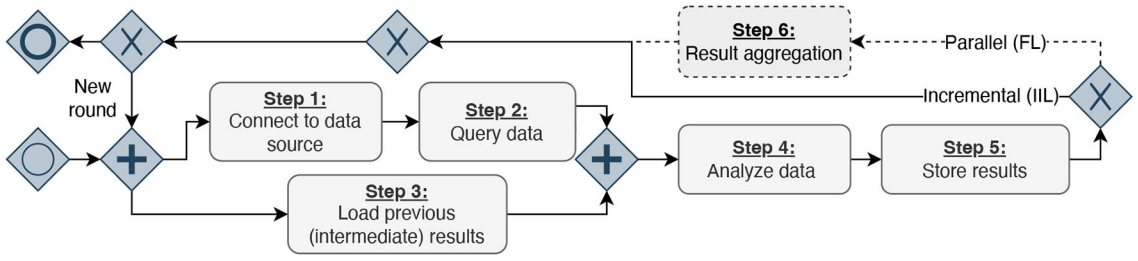

FIGURE 2
Process diagram inspired by the business process model and notation, displaying the identified six steps performed in DA experiments. First, the analysis code needs to connect to a data source and query analysis data (Steps 1 & 2). Simultaneously, the code can load results from previous executions or initial models and weights (Step 3). Afterward, the analysis is executed, and the results are stored (Steps 4 & 5). The results must be aggregated depending on the DA architecture (Step 6). Finally, either a new execution round is triggered, or the execution finishes.

too detailed in our DA setting. For example, Smoke Tests for classification and clustering algorithms have already been proposed by Herbold and Haar (16).

2.4 Implementation of DEATHSTAR

With the foundations established in the previous section, we proceed to our PoC implementation that we refer to as DEATHSTAR. This prototype evaluates the analysis code as per our six criteria. According to the key characteristics of Smoke Testing, DEATHSTAR should offer capabilities for "fast" and "automated" Smoke Testing. To accomplish this, we adopt a *testing-through-simulation* approach, which simulates an entire DA execution with multiple rounds and data sources to detect possible non-compliances with our six criteria. Beyond this aspect of fast test automation, we also focus on a user-centric design that is inspired by IDEs and playgrounds as common tools in software engineering. To provide an overview, we have provided a top-level architectural diagram in Figure 3.

We developed a containerized web application in Node.js, using the client-server paradigm (see Figure 3), which enables the integration into other ecosystems via the provided API (component ❶). Through the use of containerization this application can be run platform independent. Moreover, the provided API can also be used in CI/CD pipelines and other IDEs, enabling developers to integrate the functionalities of DEATHSTAR into broader development processes. The User Interface (UI) includes elements that support developers in writing code and monitoring the simulations via log outputs. Our implementation is accessible under the MIT license via the repository associated with this paper. This repository offers technical descriptions, screenshots, and a video demonstrating the described features. The following sections provide a more detailed description of the architectural design.

2.4.1 Data schema model ❷

We assume that the developer has access to the data schema information and the technical details of the data sources. In this work, we intend to replicate the real data sources of a DA infrastructure for our simulation and fill each replicated data source with synthetic data following the same schema and format. As no real data is involved, we claim that this approach is privacy-preserving and satisfies our aforementioned objectives. It should be noted that we consider the term *real* to be associated with sensitive and non-shareable data. In some instances, such as data donations, the developer may have access to *real* sample data, which can be used for our Smoke Testing scenario. In the latter case, we also demonstrate that our targeted approach can handle *real* sample data beyond the synthetic data we generate. Describing the structure of data sources used for data analysis is challenging because of the sheer amount of data storage technologies, data types, and their combinations. For these reasons, our goal is to find a solution that can enable the initialization of the database, the management and creation of the data structure, and the insertion of synthetic data while allowing extensions to support different data sources and data structures in the future.

A common way of specifying data structures and data formats are Data Schema Models (37, 38). We have decided to use the widely used and well-established Resource Description Framework (RDF) and its serialization Turtle[6] (39). RDF is very flexible regarding extendability, adaptability, and granularity level. By utilizing RDF, we can model the hierarchical fashion of data sources (see Figure 3), starting from the database technology, via the inlying tables to the atomic data types of attributes. Moreover, RDF's graph-based nature enables us to model more complex data structures with interconnections between data entities by additional arcs and nodes added to the graph. Further, we used RDF in conjunction with the Web Ontology Language (OWL) to model and represent data structures, making it a versatile tool that facilitates interoperability and reusability on data-level[7]. An integral part of RDF are IRIs, which uniquely identify the entities described in the RDF model. In our case, this means that data sources or atomic data types are represented by an IRI. Two example IRIs are depicted in the Data Schema Model in Figure 3. IRI A represents the identifier for a specific data source technology, whereas IRI B refers to the atomic data type `double`. For the sake of simplicity, Figure 3 only shows the model for one specific data source, i.e., a data provider. To represent multiple data providers, which might participate in a DA execution, additional Data Schema Models in the same format can be added. The Data Schema

6 Terse RDF Triple Language: www.w3.org/TR/turtle/.

7 Schema and examples: https://github.com/PADME-PHT/playground.

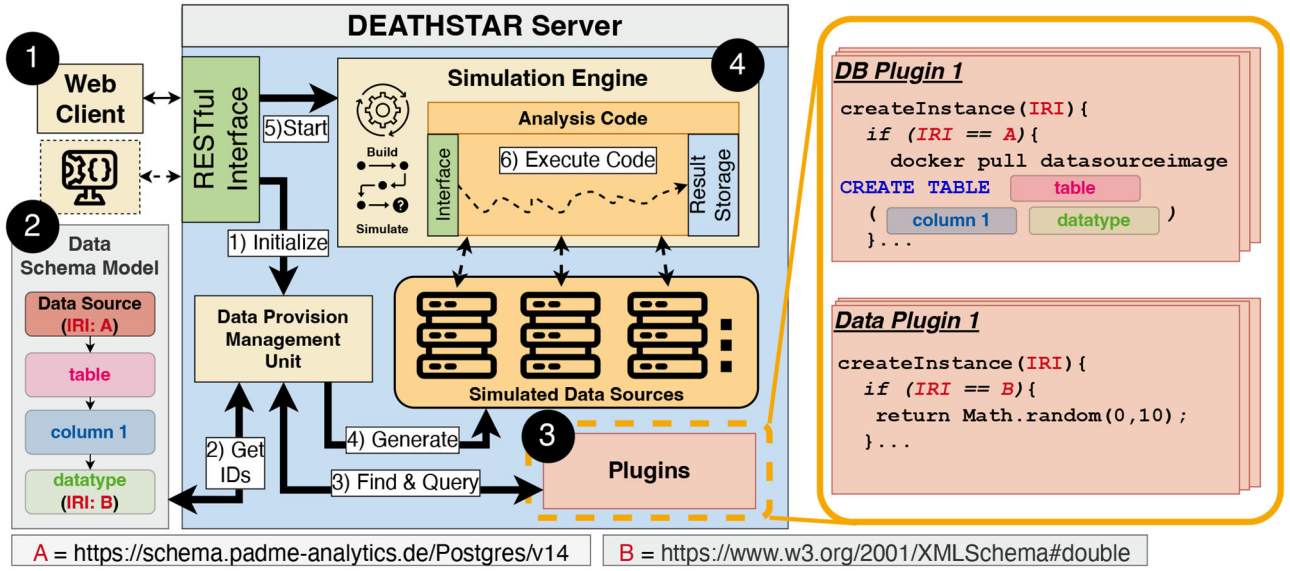

FIGURE 3
Overview of the DEATHSTAR architecture, containing a web application following the client-server paradigm. The server is a monolithic application that implements each sub-component as a library. Moreover, the overview depicts the process of generating simulated data sources. This process leverages Internationalized Resource Identifier (IRI) to find and query plugins that generate synthetic data and the data source instances. After the simulated data sources have been setup, the Simulation Engine 4 simulates a distributed execution of the provided code.

Model is usually specific for one DA use case involving multiple data sources. Therefore, it is mandatory to initially model each data provider manually or with semi-automated means. While our schema as mentioned above only models the structure of the data source, we further need a mechanism to instantiate actual data sources and generate data.

2.4.2 Plugin system 3

We decided to leverage a module-based plugin system with standardized interfaces to handle the instantiation and generation of multiple data sources and synthetic data (see Figure 3). There are two general types of plugins: The first type, called Database Plugins (DB Plugins), manages the data sources (e.g., PostgreSQL) and their underlying structures (e.g., tables and columns). The second type, the Data Plugins, produces new data instances of a specific data type. Both types of plugins are available and provided as `Node.js` modules within the DEATHSTAR server and loaded when the application starts. Therefore, the benefits of using IRIs have become apparent at this point: Each modeled data source and type is linked to exactly one instantiation function of a plugin via an IRI.

Consequently, we can explicitly define how to instantiate a data source or generate a data instance. Developers can leverage the flexible plugin system to establish databases according to the "mix-and-match" principle, allowing them to combine complementary data plugins to populate the database. Our collection of 30 plugins for the most common atomic data types are available open-source[8] for reuse or can be used as templates for the development of new plugins.

To manage the various types of storage technology, we rely on software containers, more specifically Docker containers[9], to create a new instance of a data source through our DB plugins mentioned above. This approach allows us, for example, to instantiate a separate container for each required data source using a single Docker API call. Moreover, most data sources like PostgreSQL, MongoDB, MinIO, or Opal already provide images of various versions for the Docker environment that can be used as a starting point. Further, containers provide standardized connection interfaces, which facilitate the insertion of data instances into the database. We argue that this approach is versatile enough to support highly-customized storage technologies since containers can also be pulled from private repositories. Additionally, developers are also able to use *real* data samples with DEATHSTAR by using a custom plugin that either provides a proxy for the connection to an already existing data source or creates a data source that uses the *real* data samples instead of the generated ones.

2.4.3 Simulation engine 4

The task of the Simulation Engine is to take analysis code and simulate a DA execution on the data sources, which have been introduced in the previous sections. At this point, we face another challenge regarding the analysis code that could range from basic statistics to even complex code for ML model training, including a data-preprocessing pipeline, and can be written in different programming languages. Hence, our solution must be independent of the analysis complexity and the technology stack used. In

8 Plugins: https://github.com/PADME-PHT/playground/tree/main/src/backend/src/lib/data-generator/plugins.

9 Docker: www.docker.com.

order to achieve this goal, we make use of the containerization technology again and containerize the analysis code before the actual simulation. This means that the developer has all the necessary degrees of freedom to develop the analysis code with DEATHSTAR. For example, our concept is compatible with all widely used ML frameworks such as PyTorch[10] or Scikit-Learn[11]. Apart from the analysis code, we only need the image building file (e.g., `Dockerfile`), which gives the instructions for building the container. To simplify this process, we offer `Dockerfile` templates for the most popular programming languages used in data science, such as Python[12] and R[13].

We chose to implement the IIL and FL paradigm in our Simulation Engine, giving us one representative of DA approaches with and without aggregation. Moreover, we argue that the implementation can be extended, if needed. For the simulation of the IIL paradigm, the developer has to provide the mentioned `Dockerfile` and the analysis code. In the FL scenario, we additionally require code for the aggregation component. The Simulation Engine manages the simulation process, which builds the analysis container(s). The simulation proceeds as follows: Upon building the analysis container, the engine injects DB-plugin-provided connection credentials through environment variables into the container. It then launches the analysis container, which executes the analysis code. It should be noted that in FL, these preliminary steps may occur simultaneously for each replica of an analysis container. The analysis itself adheres to the abstract workflow presented in Figure 2. It takes the received credentials and establishes a connection to the simulated data source (S1). The analysis code queries the data (S2), loads previous results if available from the filesystem of the analysis container (S3), processes, and analyzes the queried data (S4). The computed analysis results are saved in the container, which is then stopped by the Simulation Engine. A new container is instantiated from the stopped container, which carries out steps S1–S5 using the previous results and the next data source. This represents the transfer from one data source to the next, enabling us to simulate the IIL paradigm. On the other hand, in the FL case, the engine initiates a container containing the aggregator code, which has to be provided by the developer. This container gets the intermediate results produced by each replicated analysis container from the Simulation Engine, which extracts them from a pre-defined path. The aggregation container then combines the provided intermediate results into a single global result (S6) before a new analysis round begins. It is important to mention that each data source is simulated within its own virtual network. This approach prevents any side effects, like duplicated hostnames between institutions, and ensures the simulation accurately reflects the real execution environment. Moreover, using virtual networks, the Simulation Engine can be adjusted for the FL case to exchange intermediate results trough the network.

10 PyTorch: https://pytorch.org/.

11 Scikit-Learn: https://scikit-learn.org.

12 Python: https://www.python.org/.

13 R: https://www.r-project.org/.

3 Results

In order to evaluate our Smoke Testing approach, we divided our evaluation into two parts to assess different aspects of our concept. First, we invited potential users and conducted a User-Study with an accompanying survey (Section 3.1). Through this User-Study, we investigate the effectiveness of our criteria. Secondly, as part of a technical evaluation, we replicate several real-world use cases to evaluate the fitness of our realization *in operando* (Section 3.2)[14].

3.1 Evaluation of the effectiveness

This part of our evaluation has two goals. Firstly, we want to determine the effectiveness of our defined criteria through DEATHSTAR by conducting an exemplary DA use case (called User-Study, see Figure 4). Besides this, we want to assess the contribution of our concept to the development phase of DA experiments from a user perspective and surveyed the users after their development. It should be noted that the scope of this User-Study is limited to the development of a basic statistical query rather than a complex ML model. This is due to the potential difficulty and complexity of conducting a User-Study for the latter. However, we argue that the six criteria established in this study remain relevant and applicable, regardless of the level of complexity involved in the analysis, or more specifically, in S4 (Figure 2). In either scenario, data must be queried and processed, and the results must be stored.

3.1.1 Setup

We designed an exemplary use case that might occur in a real clinical study[15]. The use case aims to determine the number of patients in two hospitals that are at least (≥) 50 years old. Since we assume that these two hospitals, i.e., data providers, exist in our real ecosystem, we consequently need to re-model these, called *Hospital A* and *Hospital B*, with DEATHSTAR. Both offer a relational PostgreSQL database that provides patient information. The database at *Hospital A* contains data on patients and their treatment history, while *Hospital B* provides data on patients and their insurance information. At this point, it is worth mentioning that we explicitly introduce data heterogeneity and schema mismatches as potential sources of error in DA. The idea behind introducing those differences has been to investigate DEATHSTAR's capabilities to aid users in detecting potential malfunctions in the code. In our case, both relations about the relevant patient information have different names (*patients* on *Hospital A*, *patient_info* on *Hospital B*) and offer varying additional attributes. Participants are expected to identify these differences and adjust their code accordingly to pass the evaluation.

14 Detailed results and resources for replicating the evaluations to are provided at: https://github.com/PADME-PHT/playground/tree/main/evaluation.

15 For a video demonstration of the use case task see https://github.com/PADME-PHT/playground.

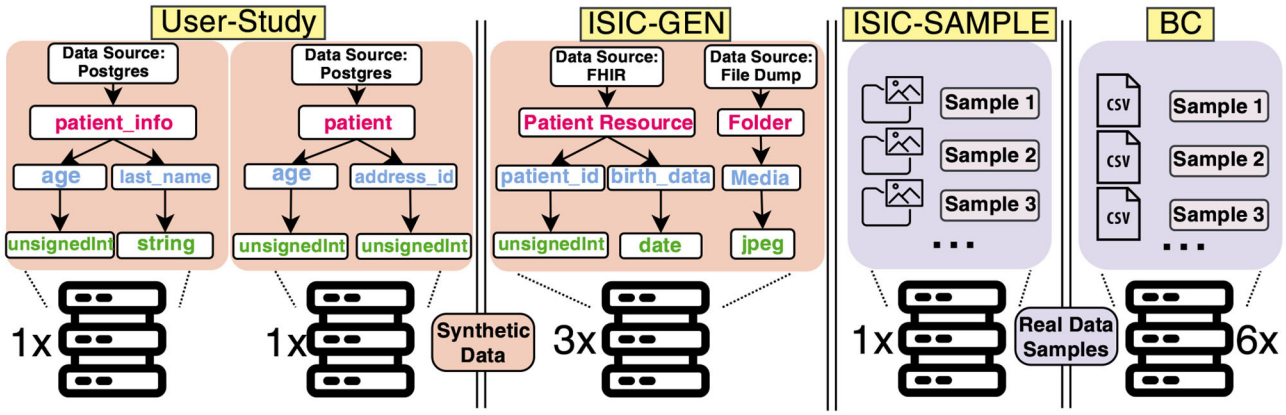

FIGURE 4
Our evaluation involved RDF data structures in the user-study and three application scenarios: ISIC-GEN, ISIC-SAMPLE, and the BC use case. The user-study used two distributed data sources with synthetic data, while ISIC-GEN used three data sources with synthetic data, and ISIC-SAMPLE used one data source with real sample data (13). Lastly, the BC use case leveraged six data sources with real sample data (14).

3.1.2 User task description and survey

All participants were provided with a task description document to implement the DA code for this use case with the programming language Python and the query language SQL[16]. The programming and query language has been selected arbitrarily as our concept leverages programming language-agnostic containers. After a short oral tutorial explaining the interaction with DEATHSTAR, participants were asked to develop the analysis code for the scenario mentioned above. Alongside this main task, users were encouraged to explore the DEATHSTAR's features and functionalities. However, no further guidance or hints have been provided regarding possible issues during the development and the participants are unrestricted in how they fulfill the task. Especially, the intentionally introduced mismatch problem needs to be identified by the participants only with the help of DEATHSTAR. After the development was completed, we asked each participant to submit the code. The submitted code was then distributed and executed within the actual infrastructure. We also aimed to assess the quality of our solution from the users' perspective. Therefore, we conducted a survey upon completion of the use case implementation. The survey consisted of three parts and was conducted via an anonymous online questionnaire. The questionnaire is based on the SUS as a metric to measure the usability of a system (18, 40). The SUS consists of ten questions that are answered on a scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree) (40). From the answers to these questions, a score is calculated that ranges from 0 to 100 and indicates the system's usability, with 100 being the best reachable score (40). The final part of the evaluation consisted of six custom questions regarding the comprehensiveness and usefulness of DEATHSTAR, using the same scale as the SUS.

In total, the evaluation involved 29 participants[17] from diverse backgrounds, such as researchers, developers, and those with experience in DA algorithm development. The evaluation sessions lasted 30–60 min on average, and the study was completed within one month. Of the participants, 11 (37.9%) reported prior experience with DA, while 18 (62.1%) stated having no prior experience. Of the 29 code submissions, 28 were executed successfully (96.6%) in the real ecosystem. All participants found the intentionally introduced schema mismatch at the two data providers and adjusted their code accordingly. However, one submission failed to establish a connection to the database since a connection parameter had been misconfigured (non-compliance with Requirement A). We have received 28 survey submissions—one submission was invalid. Based on these, we calculated the SUS according to Brooke (40). Overall, we reached a SUS score of 88.3, indicating a high level of usability. Moreover, the question, stating *"The playground solves the problem of [Smoke] testing distributed analysis algorithms"*, has an average of 4.11. Tables 2, 3 provide an overview of the user ratings. Additionally, the supplemental material[18] provide the raw data and scripts to calculate the ratings.

3.2 Real-world use cases

In order to showcase the adaptability and flexibility of our approach, we intend to technically evaluate it further by replicating three real-world application scenarios with more complex data structures, schemas, and data types (see Figure 4). We aim to collect performance benchmarks of DEATHSTAR, assessing its suitability for a range of scenarios with varying complexity levels of the analyses involved, usage of data instances, and (simulated) data sources. We further demonstrate the compatibility of our PoC to various underlying hardware options and perform the Smoke Tests using the CPU or the GPU. The selected use cases were previously conducted by Mou et al. and Welten et al. (13, 14). We refer to these cited references for further details about the DA experiments.

ISIC-GEN (Summary: 10 synthetic data instances per source, three data sources, GPU only). The open-source dataset used for

16 PostgreSQL Syntax: https://www.postgresql.org.

17 Raw data and details about the evaluation: https://github.com/PADME-PHT/playground/tree/main/evaluation/user_study.

18 Supplemental material can be found here: https://github.com/PADME-PHT/playground.

TABLE 2 Average (Avg) and standard deviation (SD) per statement of the System Usability Scale (SUS) (n = 28).

Question	Avg	SD
I think that I would like to use the Playground frequently	4.21	±0.79
I found the Playground unnecessarily complex	1.43	±0.50
I thought the Playground was easy to use	4.57	±0.69
I think that I would need the support of a technical person to be able to use the Playground	1.57	±0.84
I found that the various functions in the Playground were well integrated	4.64	±0.56
I thought that there was too much inconsistency in the Playground	1.14	±0.36
I would imagine that most people would learn to use the Playground very quickly	4.46	±0.74
I found the Playground very awkward to use	1.79	±1.10
I felt very confident using the Playground	4.54	±0.58
I needed to learn a lot of things before I could get going with the Playground	1.18	±0.48

Each question could be answered on a scale from 1 (strongly disagree) to 5 (strongly agree).

TABLE 3 Average (Avg) and standard deviation (SD) per question regarding the Playground's comprehensiveness and usefulness (n = 28).

Question	Avg	SD
The Playground offers the relevant tools needed to test distributed analysis algorithms	4.50	±0.75
The schema information provided in the Playground offers all the needed information to develop an analysis task on the described data before its actual execution/deployment	4.54	±0.69
The Playground facilitates access to the schema information, which is usually sealed within the institution	4.82	±0.39
Using the Playground improves the development process—compared to deploying the analysis algorithms without the Playground	4.50	±0.75
The Playground helps with discovering possible problems in the execution, like differences in data schemas between Stations, before the execution	4.64	±0.73
The Playground solves the problem of testing distributed analysis algorithms	4.11	±0.79

Each question could be answered on a scale from 1 (strongly disagree) to 5 (strongly agree).

the skin lesion analysis is sourced from the ISIC[19] and comprises image and patient metadata. Mou et al. distributed this data across three institutions in a real DA setting and conducted an experiment. In our scenario, we aim to re-model the data provision. However, this use case presents a challenge as we need to model two interlinked data sources for each data holder: A Fast Healthcare Interoperability Resource (FHIR)[20] server for patient data and an object storage system for the skin images (as shown in Figure 4). We first developed the plugin for the FHIR server instance, and, secondly, we modeled a basic file dump to store image data. Finally, we need plugins for each modeled data type. We have decided to create plugins that generate random data, including random `strings` or `integers`, datatypes according to the FHIR standard, and even images with no semantics. Our plugins support the FHIR resource types `Patient`, `Media`, and `ImagingStudy` required in this use case, which are randomly filled. The chosen data type for dermoscopic images is `jpeg`, as it matches the format of the original images. For the `jpeg`-plugin, we obtained 70 placeholder images from an external service used for websites[21]. After the plugin is instantiated, these images are stored in the file dump mentioned earlier. Revisiting our main objective, we strive to offer a concept that enables Smoke Testing of algorithms. Therefore, we consider the synthetic data instances as placeholders that can be queried and processed to test the analysis, but it is not intended for producing plausible analysis results.

ISIC-SAMPLE (Summary: 8,444 sample data instances, one data source, GPU only). To demonstrate that DEATHSTAR is capable of managing real) sample data and custom data sources, we replicated the ISIC-GEN use case using actual plausible sample data obtained from the ISIC repository mentioned earlier. To achieve this, we set up an external data source similar to the real setting by Mou et al. in a network accessible from DEATHSTAR's host machine instead of using our provided mechanism for data source replication.

BC (Summary: 539 sample data instances, six data sources, CPU only). We conducted another use case with real data samples about BC characteristics, following a similar approach as in the previous use case. In their work, Welten et al. distributed CSV data across six institutions in a real DA setting and conducted a DA experiment on this BC dataset. We set up external storage for the CSV data, which is accessible to DEATHSTAR.

After re-modeling the required data sources, we need to develop the analysis code with DEATHSTAR. For the ISIC use cases, we developed the same image classification model, which classifies the images into benign and malign. In contrast, for BC, we implement code that trains a logistic regression model to predict BC. We implemented the analyses according to both executions paradigms, i.e., one IIL and two FL versions. Note that, regarding the FL paradigm, we implemented one fully parallelized version (original version) and one version, called FL-INC, which executes at most one analysis simultaneously. In other words, FL-INC performs IIL but updates the analysis results at the end of the round. At this point, we have provided all necessities to perform Smoke Tests on each use case. We choose three, one, and six instances for each respective scenario (as shown in Figure 4) and start the simulation. Once we successfully executed the code in the simulated environment, indicating a successful Smoke Test, we ran the DA algorithms in the PADME platform to evaluate their operability in a real-world setting. We state that all executions were as expected and successful.

4 Discussion

The outcomes of our first evaluation (see Section 3.1) show the effectiveness of our criteria. We observed that almost all executions of the participant's algorithms were successful. Overall, the high number of successful executions shows that our solution can

19 ISIC Challenge: www.isic-archive.com.

20 FHIR standard: https://hl7.org/fhir.

21 LoremFlickr CC): www.loremflickr.com.

indeed provide Smoke Testing capabilities for DA. The outcomes of our survey further reinforce this claim: The participants rated DEATHSTAR positively and acknowledged that it effectively *"solves the problem of [Smoke] Testing DA algorithms"* and *"offers the relevant tools needed to [Smoke] Test"* (Table 3). Beyond the results about the effectiveness, the accompanying user survey demonstrates that our realization was well-received by our study group. This result is also reflected in the SUS score of 88.3 (Table 2), placing our realization clearly above the mean score of 68 (41). Moreover, according to Bangor et al. this score can be described with an adjective rating of "excellent", placing it in the highest out of four quartiles (18). When we investigate the cohorts, including participants with and without prior experience, only a small difference in the SUS score is visible: Participants with a background in DA rated our concept with a score of 86.6 compared to a rating of 89.8 by the unfamiliar users. All participants have been able to *"discover possible problems in the execution, like differences in data schemas, before the execution"* with DEATHSTAR. Additionally, the participants appreciated the ability to employ a trial-and-error approach during development.

In the second and more technical evaluation, we assessed the flexibility of our approach by applying it to real-world use cases. We have been able to use DEATHSTAR for generating data and creating complex, interlinked data sources, indicating that its concept is capable of working with very distinct settings such as structured data, images or textual data. We would like to emphasize that the same code used for ISIC-GEN also worked for ISIC-SAMPLE, indicating that our approach involving synthetic data was able to successfully replicate data sources used in the real-world use case (ISIC-SAMPLE). During our technical evaluation, we additionally measured the duration of each Smoke Test (i.e., simulation). Note that each analysis code has to be containerized before the simulation. As this factor might also count as part of the Smoke Test, we also measured the image-building time (see Table 4). All builds have been executed without pulling the overarching Python image for the analysis container, and the needed dependencies have been downloaded with a connection speed of 900 MBits. In the scope of this technical evaluation, DEATHSTAR has been deployed on a server with 4×3.60 GHz CPU, 128 GB RAM, and a TITAN XP GPU.

Based on these measured times, we can derive three factors that influence the Smoke Tests:

1. **Analysis complexity:** While the Smoke Test of the User-Study case terminates almost immediately, the more complex data analyses ISIC-GEN, ISIC-SAMPLE, and BC need more time since these involve ML model training, whose duration is usually influenced by the number of epochs or the complexity of the to be trained model itself. Additionally, we can identify another effect, which is the number of required dependencies used for the analyses. Due to our design based on containerization, DEATHSTAR builds an image for each analysis. Hence, each dependency has to be included. This results in the BC analysis image needing more time to be built than the ISIC images since the BC image covers more packages. However, note that many packages can be cached once an image has been built. This caching reduces the build times to >2 s.
2. **Dataset size:** Similar to the analysis complexity, the number of used data instances for the Smoke Tests influence its duration. While the analysis code for User-Study and ISIC-GEN only processes 10 instances per provider (fastest), BC processes 539 instances, and the ISIC-SAMPLE analysis queried 8,444 images (slowest).
3. **Number of simulated data sources:** The more providers are involved in the Smoke Test, the longer the duration. This can be explicitly seen in ISIC-GEN and BC, where we involved three and six providers, respectively. Thus, the simulation duration is directly influenced by a factor proportional to the number of data sources.

Regarding the three characteristics of Smoke Testing, we can derive the following connections and conclusions from our evaluation results. By simulating the analyses, DEATHSTAR can identify potential issues and problems in the algorithm's functionality without having to perform an exhaustive and extensive test. This contributes to the "shallow" characteristic, and the high number of error-free executions underpin the effectiveness of our criteria. Regarding the "fast" characteristic, we face a trade-off between the duration of the Smoke Tests and three factors that influence the simulation, as discussed above. At this point, we argue that the Smoke Test can be optimized, for example, by using fewer data sources (e.g., in the case of homogeneous data sources) or fewer data instances. For example, the ISIC-SAMPLE use case also works using a fraction of the 8,444 images, which might reduce the Smoke Test duration significantly (see ISIC-GEN). Furthermore, there is potential for improvement in implementing the FL paradigm. While executing the fully parallelized version (FL) in the BC use case, we encountered a slowdown of the Smoke Test due to the increased loads produced by the parallel execution. An alternative that circumvents the concurrency issues and therefore offers faster Smoke Testing could be FL-INC, which exhibits similar performance to IIL. Finally, regarding the "automated" characteristic, we found that through our simulation-based approach, we enable a fully automated Smoke Test with minimal manual intervention. Each Smoke Testing criterion mentioned above is automatically validated by our Simulation Engine, contributing to a seamless use of DEATHSTAR, partially shown by our survey results.

4.1 Threats to validity

Some limitations have become apparent that can be attributed to our design decisions. While DEATHSTAR fully automates the Smoke Tests, some prior efforts still have to be devoted to collecting the schema information from each data source, which could pose a bottleneck. This especially holds for the creation of plugins and the data re-modeling in case sample data is unavailable for Smoke Testing. Although we included the aspect of reusability in our design decisions ("mix-and-match") and our already developed assets can be used as foundations, the aspect of re-modeling data sources might still be a time-consuming and error-prone factor. Since our main objective has been the definition of Smoke Testing criteria for DA analyses, we mainly focused on the effectiveness of our criteria. Hence, our evaluation does not cover the aspect of data re-modeling, and this question remains open. The second threat is our implementation as such. Our simulation might produce an overhead in the Smoke Testing strategy that

TABLE 4 Each row represents the measured duration for the building times of the images, the time for one single provider, and the time for a complete Smoke Test.

Use case	Build	One Data Source	Smoke Test IIL	Smoke Test FL	Smoke Test FL-INC
User-study	23 s	6 s	12 s	15 s	17 s
ISIC-GEN	1 m 39 s	24 s	1 m 6 s	56 s	1 m 15 s
ISIC-SAMPLE	1 m 39 s	4 m 31 s	–	–	–
BC	6 m 6 s	48 s	4 m 33 s	11 m 53 s	4 m 51 s

Note that the ISIC-SAMPLE use case has only been conducted on one data source.

might validate additional requirements implicitly, which influences the effectiveness of our approach. The defined criteria can be tested through another approach beyond simulation that tests each criterion individually. This threat has also been analogously stated in work by Cannavacciuolo and Mariani (17). We have chosen a *testing-through-simulation* approach to comply with the "automated" characteristic and the iterative manner of DA analyses. Hence, we argue that our approach provides the flexibility to master the sheer amount of data source technologies, schemas, or analysis types. Validating each criterion separately for each DA scenario might impede this flexibility. However, the benchmarking of our concept against other similar approaches remains open.

In summary, in this work, we addressed the issue of lacking Smoke Testing criteria for the validation of DA code. We have pointed out that insufficiently tested analysis code is susceptible to SPOFs, which causes a complicated and time-consuming development process due to the inherently decentralized nature of DA infrastructures and the dependence on the data providers during development. In order to tackle this issue, we propose six criteria that must be guaranteed to ensure the operability of the analysis code, representing a successful Smoke Test (RQ1). Based on these criteria, we developed a PoC, called DEATHSTAR, that locally performs Smoke Testing on DA code following a *testing-through-simulation* approach by simulating an entire DA experiment (RQ2). Since the application of Smoke Testing to data analyses is dependent on the availability of sufficient sample data, we leveraged a flexible and adaptable plugin system, which allows the semi-automated creation of synthetic test data, which can be used for Smoke Testing (RQ2.1 & RQ2.2). Hence, we developed a solution that allows users to develop iteratively (i.e., trial-and-error) and (Smoke) test their analysis code by simulating its execution on re-modeled data sources. We evaluated DEATHSTAR in a two-folded evaluation. First, we conducted a User-Study with 29 participants to evaluate the effectiveness of our criteria. We found that 96.6% of all developed DA analyses that were initially Smoke Tested could be successfully executed in a real DA environment. Furthermore, our accompanying survey resulted in a SUS score of 88.3, giving DEATHSTAR an "excellent" usability rating. Secondly, we applied DEATHSTAR to three real-world use cases in the scope of a technical evaluation. The technical results of our evaluation show that our concept is flexible enough to serve for different use cases and complies with the three characteristics of Smoke Testing: Shallow, Fast, and Automatic. In conclusion, within the scope of our work, the contribution of our PoC fuels research by reducing obstacles in conducting DA studies.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

Author contributions

SWel: Conceptualization, Formal analysis, Methodology, Validation, Writing – original draft, Writing – review & editing. SWeb: Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. AH: Software, Validation, Writing – review & editing. OB: Supervision, Writing – review & editing. SD: Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

We sincerely thank all the study participants for their valuable contributions to this research. Additionally, we extend our gratitude to Laurenz Neumann for his valuable support during the development of the metadata schema.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Negash S, Gray P. Business intelligence. In: *Handbook on Decision Support Systems 2*. Berlin; Heidelberg: Springer (2008). p. 175–93.

2. Dilsizian SE, Siegel EL. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Curr Cardiol Rep*. (2014) 16:1–8. doi: 10.1007/s11886-013-0441-8

3. Jamshidi M, Lalbakhsh A, Talla J, Peroutka Z, Hadjilooei F, Lalbakhsh P, et al. Artificial intelligence and COVID-19: deep learning approaches for diagnosis and treatment. *IEEE Access*. (2020) 8:109581–95. doi: 10.1109/ACCESS.2020.3001973

4. Stavrinides GL, Karatza HD. The impact of data locality on the performance of a SaaS cloud with real-time data-intensive applications. In: *2017 IEEE/ACM 21st International Symposium on Distributed Simulation and Real Time Applications (DS-RT)*. Rome: IEEE (2017). p. 1–8.

5. Hong L, Luo M, Wang R, Lu P, Lu W, Lu L. Big data in health care: applications and challenges. *Data Inf Manag*. (2018) 2:175–97. doi: 10.2478/dim-2018-0014

6. Hallock H, Marshall SE, t'Hoen PAC, Nygård JF, Hoorne B, Fox C, et al. Federated networks for distributed analysis of health data. *Front Public Health*. (2021) 9:712569. doi: 10.3389/fpubh.2021.712569

7. Welten S, Mou Y, Neumann L, Jaberansary M, Ucer YY, Kirsten T, et al. A privacy-preserving distributed analytics platform for health care data. *Methods Inf Med*. (2022) 61:e1–e11. doi: 10.1055/s-0041-1740564

8. Beyan O, Choudhury A, van Soest J, Kohlbacher O, Zimmermann L, Stenzhorn H, et al. Distributed analytics on sensitive medical data: the personal health train. *Data Intelligence*. (2020) 2:96–107. doi: 10.1162/dint_a_00032

9. Corte-Real A, Nunes T, Santos C, Rupino da Cunha P. Blockchain technology and universal health coverage: Health data space in global migration. *J For Legal Med*. (2022) 89:102370. doi: 10.1016/j.jflm.2022.102370

10. Chen J, Qian F, Yan W, Shen B. Translational biomedical informatics in the cloud: present and future. *Biomed Res Int*. (2013) 2013:1–8. doi: 10.1155/2013/658925

11. Sheller MJ, Reina GA, Edwards B, Martin J, Bakas S. Multi-institutional deep learning modeling without sharing patient data: a feasibility study on brain tumor segmentation. *Brainlesion*. (2019) 11383:92–104. doi: 10.1007/978-3-030-11723-8_9

12. Shi Z, Zhovannik I, Traverso A, Dankers FJWM, Deist TM, Kalendralis P, et al. Distributed radiomics as a signature validation study using the personal health train infrastructure. *Sci Data*. (2019) 6:218. doi: 10.1038/s41597-019-0241-0

13. Mou Y, Welten S, Jaberansary M, Ucer Yediel Y, Kirsten T, Decker S, et al. Distributed skin lesion analysis across decentralised data sources. *Public Health Inf*. (2021) 281:352–6. doi: 10.3233/SHTI210179

14. Welten S, Hempel L, Abedi M, Mou Y, Jaberansary M, Neumann L, et al. Multi-institutional breast cancer detection using a secure on-boarding service for distributed analytics. *Appl Sci*. (2022) 12:4336. doi: 10.3390/app12094336

15. Mishra A. A comparative study of different software development life cycle models in different scenarios. *Int J Adv Res Comp Sci Manag Stud*. (2013) 1:64–9.

16. Herbold S, Haar T. Smoke testing for machine learning: simple tests to discover severe bugs. *Emp Softw Eng*. (2022) 27:45. doi: 10.1007/s10664-021-10073-7

17. Cannavacciuolo C, Mariani L. Smoke testing of cloud systems. In: *2022 IEEE Conference on Software Testing, Verification and Validation (ICST)*. Valencia (2022). p. 47–57. doi: 10.1109/ICST53961.2022.00016

18. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. *Int J Hum Comput Interact*. (2008) 24:574–94. doi: 10.1080/10447310802205776

19. Tuteja M, Dubey G. A research study on importance of testing and quality assurance in software development life cycle (SDLC) models. *Int J Soft Comp Eng*. (2012) 2:251–7.

20. Rovce DWW. Managing and Techniques the development of large software systems: concepts. In: *Proceedings of the 9th International Conference on Software Engineering*. Monterey, CA; Washington, DC: IEEE Computer Society Press (1987). p. 11.

21. Davis AM, Bersoff EH, Comer ER. A strategy for comparing alternative software development life cycle models. *IEEE Transact Softw Eng*. (1988) 14:1453–61. doi: 10.1109/32.6190

22. Bonawitz K, Eichner H, Grieskamp W, Huba D, Ingerman A, Ivanov V, et al. Towards federated learning at scale: system design. *Proc Mach Learn Syst*. (2019) 1:374–88. doi: 10.48550/arXiv.1902.01046

23. Hooda I, Chhillar RS. Software test process, testing types and techniques. *Int J Comput Appl*. (2015) 111:10–14.

24. Sneha K, Malle GM. Research on software testing techniques and software automation testing tools. In: *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*. Chennai. (2017). p. 77–81.

25. Khalyeyev D, Hnetynka P, Bures T. A virtual playground for testing smart cyber-physical systems. In: *2018 IEEE International Conference on Software Architecture Companion (ICSA-C)*. Seattle, WA (2018). p. 85–8.

26. Prakash S, Callahan T, Bushagour J, Banbury C, Green AV, Warden P, et al. CFU playground: full-stack open-source framework for tiny machine learning (tinyML) acceleration on FPGAs. *arXiv*. (2022). doi: 10.1109/ISPASS57527.2023.00024

27. Risdianto AC, Ling TC, Tsai PW, Yang CS, Kim J. Leveraging open-source software for federated multisite SDN-cloud Playground. In: *2016 IEEE NetSoft Conference and Workshops (NetSoft)*. Seoul: IEEE (2016). p. 423–7.

28. Kim B, Kashiba Y, Dai S, Shiraishi S. Testing autonomous vehicle software in the virtual prototyping environment. *IEEE Embed Syst Lett*. (2017) 9:5–8. doi: 10.1109/LES.2016.2644619

29. Bonawitz K, Eichner H, Grieskamp W, Huba D, Ingerman A, Ivanov V, et al. Towards federated learning at scale: system design. In: Talwalkar A, Smith V, Zaharia M, editors. *Proceedings of Machine Learning and Systems*. Vol. 1. Stanford, CA (2019). p. 74–88.

30. Ihaka R, Gentleman R. R a language for data analysis and graphics. *J Comput Graph Stat*. (1996) 5:299–314. doi: 10.1080/10618600.1996.10474713

31. Gaye A, Marcon Y, Isaeva J, LaFlamme P, Turner A, Jones EM, et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. *Int J Epidemiol*. (2014) 43:1929–44. doi: 10.1093/ije/dyu188

32. Moncada-Torres A, Martin F, Sieswerda M, Van Soest J, Geleijnse G. VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange. *AMIA Annu Symp Proc*. (2021) 2020:870–7.

33. Herr MdAB, Graf M, Placzek P, König F, Bötte F, Stickel T, et al. Bringing the algorithms to the data-secure distributed medical analytics using the personal health train (PHT-meDIC). *arXiv*. (2022). doi: 10.48550/arXiv.2212.03481

34. Lindell Y, Pinkas B. Secure multiparty computation for privacy-preserving data mining. *J Priv Confld*. (2009) 1:40. doi: 10.29012/jpc.v1i1.566

35. Warnat-Herresthal S, Schultze H, Shastry KL. Swarm learning for decentralized and confidential clinical machine learning. *Nature*. (2021) 594:265–70. doi: 10.1038/s41586-021-03583-3

36. Wirth FN, Kussel T, Müller A, Hamacher K, Prasser F. EasySMPC: a simple but powerful no-code tool for practical secure multiparty computation. *BMC Bioinformat*. (2022) 23:531. doi: 10.1186/s12859-022-05044-8

37. Balaji B, Bhattacharya A, Fierro G, Gao J, Gluck J, Hong D, et al. Brick: Towards a unified metadata schema for buildings. In: *Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments*. Palo Alto, CA: ACM (2016). p. 41–50.

38. Bodenreider O, Cornet R, Vreeman D. Recent Developments in Clinical Terminologies —SNOMED CT, LOINC, and RxNorm. *Yearb Med Inform*. (2018) 27:129–39. doi: 10.1055/s-0038-1667077

39. Ali W, Saleem M, Yao B, Hogan A, Ngomo CAN. A survey of RDF stores & SPARQL engines for querying knowledge graphs. *VLDB J*. (2022) 31:1–26. doi: 10.1007/s00778-021-00711-3

40. Brooke J. SUS - A quick and dirty usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL, editors. *Usability Evaluation in Industry. 1st ed*. Taylor & Francis (1996). p. 189–94.

41. Lewis J, Sauro J. Item benchmarks for the system usability scale. *J Usability Stud*. (2018) 13:158–67.

frontiers | Frontiers in Medicine

TYPE Perspective
PUBLISHED 31 January 2024
DOI 10.3389/fmed.2024.1336588

OPEN ACCESS

EDITED BY
Gokce Banu Laleci Erturkmen,
Software Research and Development Consulting, Türkiye

REVIEWED BY
Bertrand De Meulder,
European Institute for Systems Biology and Medicine (EISBM), France

*CORRESPONDENCE
Kaya Akyüz
✉ kaya.akyuez@bbmri-eric.eu

RECEIVED 11 November 2023
ACCEPTED 19 January 2024
PUBLISHED 31 January 2024

CITATION
Akyüz K, Cano Abadía M, Goisauf M and Mayrhofer MT (2024) Unlocking the potential of big data and AI in medicine: insights from biobanking.
Front. Med. 11:1336588.
doi: 10.3389/fmed.2024.1336588

Unlocking the potential of big data and AI in medicine: insights from biobanking

Kaya Akyüz*, Mónica Cano Abadía, Melanie Goisauf and Michaela Th. Mayrhofer

Department of ELSI Services and Research, BBMRI-ERIC, Graz, Austria

Big data and artificial intelligence are key elements in the medical field as they are expected to improve accuracy and efficiency in diagnosis and treatment, particularly in identifying biomedically relevant patterns, facilitating progress towards individually tailored preventative and therapeutic interventions. These applications belong to current research practice that is data-intensive. While the combination of imaging, pathological, genomic, and clinical data is needed to train algorithms to realize the full potential of these technologies, biobanks often serve as crucial infrastructures for data-sharing and data flows. In this paper, we argue that the 'data turn' in the life sciences has increasingly re-structured major infrastructures, which often were created for biological samples and associated data, as predominantly data infrastructures. These have evolved and diversified over time in terms of tackling relevant issues such as harmonization and standardization, but also consent practices and risk assessment. In line with the datafication, an increased use of AI-based technologies marks the current developments at the forefront of the big data research in life science and medicine that engender new issues and concerns along with opportunities. At a time when secure health data environments, such as European Health Data Space, are in the making, we argue that such meta-infrastructures can benefit both from the experience and evolution of biobanking, but also the current state of affairs in AI in medicine, regarding good governance, the social aspects and practices, as well as critical thinking about data practices, which can contribute to trustworthiness of such meta-infrastructures.

KEYWORDS

biobanks, artificial intelligence, big data, European Health Data Space, infrastructures

1 Introduction

Life sciences knowledge production is increasingly structured by big data approaches, internationalization of research and closer coupling between research and applications, where biobanks comprise a major form of infrastructure in the current research ecosystems. For decades, biobanks have efficiently ensured access to biological samples and associated health data, which is being produced, collected and used in various ways, such as for medical research and public health databases as the two broad categories of population-based and clinical biobanks reflect (1). The historical development of the biobanks and their diversification over time contrast starkly with the current efforts for standardization, harmonization, integration, globalization and most significantly datafication. They have evolved from mere repositories to trusted infrastructures in sharing biomaterials and data (2), highlighting their crucial role in

data-intensive research. These efforts for facilitating the movement of data materialized into platforms, infrastructures and guiding principles to enable the exchange of data that is compliant with ethical, legal and societal considerations.

With artificial intelligence (AI), renewed discussions are taking place due to the idiosyncrasies of AI, the speed and consequences of the implementation of such technologies in biobanking and other domains (3, 4). Over the last decade, the development of national and transnational biobank networks or infrastructures have made such infrastructures instrumental to international research consortia (5–7). In addition, meta data infrastructures called health data spaces are developed that have the potential to significantly transform the life sciences, medicine and healthcare. Back in December 2020, the European Commission published the roadmap for the European Health Data Space (EHDS) initiative inviting public responses and presenting a first draft in May 2022. Currently discussed in the European Council and the European Parliament, the ambitious goal remains to complete the legislative process by the end of 2023 but no later than within the current Commission's mandate to ensure the implementation by 2025 (8). The EHDS will undoubtably transform the health sector in Europe. It remains to be seen in which form it will be realized, especially as expectations are high across various stakeholder groups, such as patient advocacy groups, researchers from academia and industry as well as policy makers (9). At the same time, infrastructures such as biobanks have a wealth of experience regarding the collection and use of health data for research purposes in an ethically and legally compliant way (10). The perspective we present here builds on the observation that many biobanks are already going through a transformation in becoming bio(data)banks and are entangled in trials of various data practices that can inform both the debates around AI's use in life sciences and health research and emerging meta infrastructures considering developments, such as EU's upcoming Artificial Intelligence Act. Although there has been a provisional agreement as of December 9th, 2023, among negotiators from EU's Parliament and Council, the legal text will be implemented when the two institutions provide their approval and, if so, with its risk-based categorization and the accompanying requirements, the AI Act may have an impact on many aspects of AI's use in health research and applications, such as on data governance, explainability, requirements, practicing human-in-the-loop among others with potential effect also on the EHDS (11). In light of these recent developments, we argue that it is timely to look back at the practice of biobanking, especially the so-called data turn, and the current momentum in biobanking and medicine regarding AI and its implementation into research and technology, for insights on health data spaces and their development.

2 Data turn in life sciences: biobanks as data infrastructures

Biomedical research has become increasingly data-intensive and undergone a process of datafication (12). Central to this datafication are biobanks. As infrastructures, they can be characterized as vital entities in organizing practices, as embedded in other structures, social arrangements and technologies (13). In this capacity, biobanks support medical innovation, such as personalized medicine and genomic research, with scholars noting the molecularization and computerization sustaining both (14, 15).

The molecularization and data turn in the focus of biobank research in the last two decades deserves more attention. For instance, infrastructures have been created that gather genetic data from commercial and clinical sources, enabling population-based genetics research to be conducted (16). The outcome of such research, especially in genomics, raises hopes with a better understanding of the genetic bases of health conditions such as coronary artery disease, ideally based on diverse populations (17). However, the genomic data and infrastructures raise also concerns, especially regarding phenomena, such as sexual orientation, which received renewed attention in the search for a genetic basis (18) and also harbor emerging risks that are radically different than the previous ones due to intensive datafication, for instance, risks of genomic identifiability (19).

The existence of efforts towards standardization and interoperability in biobanking as reflected in the acronyms SPREC (20), BRISQ (21), MIABIS (22, 23) and others show the centrality of these notions for the data turn, but also harmonization regarding samples, technical infrastructures and practices. The relevant research contributes to developments such as specific algorithms for post-analytical use, which may bridge the differences between distinct types of blood samples originally stored for different uses (24, 25). Such developments are especially salient considering that biobanks are not independent of the broader infrastructures of medicine and healthcare. From disease categorization to defining and standardizing biomarkers at a time wearable devices, sensors and emerging forms of data are increasingly being embedded into entire ecosystems often in the digital (26), the existing samples and data with different conditions of collection, annotation, consent status and storage, as well as variations across institutions are still part of the picture. Biobanks are expanding with both typical samples and data (e.g., blood, BMI) and further kinds (e.g., epigenetic, microbiome, etc.) being integrated and standardized, expanding the data in both dimensions of volume and diversity.

In attempts towards datafication, practices around samples such as in pathology are also being transformed, exemplified by "digital pathology" where whole slide images that are once created may decrease the need to store samples or increase the findability by turning images into data collected (27). Scholars observe along a trend of consolidation emergence of virtual biobanks brings together resources from multiple biobanks (28, 29), though such cataloging examples also include efforts of broader research infrastructures, such as BBMRI-ERIC (30). Similarly, in the genomics world, efforts to standardize and make genomic data accessible such as summary statistics of genome-wide association studies is picking up pace (31, 32) as well as the development of trusted research environments despite critique (33) with specific tools, such as DataSHIELD (34).

3 AI in medicine and new beginnings for biobanking

Large amounts of data are needed to advance biomedical knowledge generation as well as big data analytics and new data-driven technologies in AI. While the history of AI in medicine goes back half a century with the initiation of computational tools and technical infrastructures as well as events devoted to the topic (35), it has gained pronounced attention and applicability in recent years in line with its intensive use in other domains. Medical AI is seen as a

promising innovation for uses such as screening, diagnosis, risk assessment, clinical decision-making, management planning, and precision medicine, with available tools ranging from chatbots to clinical decision support (36). The hope is that AI systems will reduce human bias and improve performance, as has been demonstrated in certain areas such as radiology (37), by improving accuracy in medical image analysis and easing the workload in screening (38), or for AI-driven polygenic risk scores (PRS) which may enable greater accuracy, performance and prediction (39). AI can also bring improvements when it comes to clinical measurements (40), interpretation of tests (41), decision making for intensive care unit admission (42), or embryo implantation (43), among others. However, it is important to note that AI is not a one-size-fits-all solution, and its benefits may not be realized in every application.

The development and implementation of medical AI involves numerous key challenges. First, AI is data hungry. Large amounts of data are needed to train AI and access to these data is challenging for technical, legal, and practical reasons, along with emerging issues regarding computational power and infrastructures and alternatives such as federated learning, which bring their own challenges and opportunities (44). One salient challenge in this respect relates to the tradeoff between data access and data privacy, the resolution of which necessitates bottom-up, democratic and engaging processes (3) in consideration of commitment for findable, accessible, interoperable and reusable data as often referred to with the acronym FAIR (45) and further FAIR principles (e.g., https://www.go-fair.org/fair-principles/). Second, despite the immense potential benefits, the risks revolve around perpetuation or even amplification of societal inequality and injustices due to potentially biased datasets as well as certain data practices (46). Third, practitioners require practical recommendations for applying AI (47). Furthermore, patients' preference for human agents or human supervision, possible strain between patients and treating physicians, especially in relation to privacy, data security and potential vulnerabilities related to AI tools need attention as do the implementation of guidelines and frameworks to ensure bioethical principles [e.g., (48)] are upheld and monitored (49). These call for engagement of multiple stakeholders in the resolution of ethical and legal issues, sharing similarities with biobanking, though at a different scale.

Biobanks, as key entities for providing access to large amounts of high-quality data, are central to the development of new data-based technologies such as AI. Similar to AI in medicine, the early developments in the use of AI in biobanking often focus on biobank participants' health conditions as reviewed elsewhere (50). These include developments such as, identifying and categorizing Alzheimer's disease patients (51), calculating risks scores for conditions such as age-related macular degeneration (52) or cardiovascular diseases (53), aiding in classification of disease subtypes (54) as well as providing predictions at individual level for COVID-19 (55, 56) or potential conditions due to therapeutic agents such as aromatase inhibitor-related arthralgia (57). However, biobanks are not merely support structures for healthcare or repositories for medical data. Biobanks have the potential to handle the data turn as they pursue data-driven practices in a standardized, industrialized manner (58). As research infrastructures, biobanks, may benefit from AI in the collection of biological samples and data, such as analysis of the scholarly literature for development of criteria for sampling, analysis, interpretation, data extraction, even engagements with biobank participants, from consent process to research process; however, AI can also contribute to purely *managerial tasks* including storage space optimization or *upstream research processes*, such as suggesting samples and data for research proposals based on content and methods, as well as *downstream research evaluation*, assessing the "value" of samples and data based on the scholarly literature (59). AI's potential impact on biobanking may also include possible increases in the use of biobank samples and data, thus contributing to sustainability and speed of research as well as aiding biobanks in identification and recruitment of participants, training, annotation of samples and data, increasing interoperability, visibility, and access (60).

AI is central to the idea of "biobanks for the future" (61) though challenges in implementation of AI in biobanking range from difficulties aligning standards not only across data in the long run, but also samples, workflows, ethics management, legal and governance-related aspects, from transparency to informed consent (28) as well as justice, both epistemically and ethically (14). There are efforts such as workshops or collections of best practices to increase the "readiness" of these infrastructures for AI (60) with calls, checklists, tools and frameworks for ethical use of AI in medicine/biobanking (47, 62). New and alternative forms of governance are needed for a new form of biobanking that revolves around big data considering the increasing widening of the scope of data from social media to devices capturing bodily function, resulting in streams of data over *time* and analytical capacity over *space* (63). Biobanks' positioning at the in practice often gray intersection of healthcare and research can inform the discussions on health data spaces, in light of the recent developments.

4 Discussion

The ways in which risks are approached in biobanking and the normative arguments regarding how they should, such as future-proofing the governance of biobanks (64) and adaptive risk governance (65), suggest biobanking may be helpful in identifying key questions medical AI and health data spaces are facing from informed consent, representation in datasets, to risks associated with data protection and responsibility. While acknowledging the digital divide and its consequences, the increased ability of participants to follow and engage with biobanking and healthcare infrastructures are leading to reconfigurations of "traditional boundaries between the public domain (healthcare systems, medical research, and clinical practice) and the private one (patients and citizens)" which necessitate new approaches to fostering trust (63). Health data spaces bring such observations to a new level.

Trust and trustworthiness have become keywords that are often attached to how AI should be, with limited discussion of what this entails. Despite the burgeoning literature on ethics of AI in medicine, three areas relevant for trust are problematic (46): limited analytical accuracy and conceptual slippages, inadequate analysis of the contexts in which medical AI tools are embedded, and scarcity of interdisciplinary approaches. Considering trust central to societal functioning as "a fundamental principle for interpersonal interactions" (66), it cannot be considered unidirectional. Rather, it needs to be understood as a complex, situated, context-dependent, and relational concept that involves several trustor/trustee relationships, such as trust in persons (e.g., scientists who trust each other, patients who trust scientists and clinicians), technology, and institutions (67, 68). Trust or

more precisely trusting relationships are fragile and require continuous work, which means that they need to be actively established and sustained. In this sense, we see three main considerations from biobanking – a domain that should be built on trust – that can contribute to better medical AI and health data spaces.

Regulations may provide guidance, but good governance is an active process that comprises more than following regulations. Efforts towards regulating and guiding AI have been abundant with 'AI Ethics' becoming a buzzword (69, 70) along with the legal frameworks such as the proposed Artificial Intelligence Act of the EU (11). Considering international standards, overseeing organizations, national legislations, as well as practices, from engaging participants to consents, biobanks have accumulated over decades experiences related to intensified transnational data sharing, international collaborations, including public-private partnerships, access to and reuse of data, and efforts to harmonize data, ethical/legal standards and societal aspects. Hence, biobanking incorporates knowledge of the "ethics work" that is an integral part of data flows (71) and necessitates thinking critically about potential issues that go beyond individual institutions, such as identifiability risks in a datafied world both in regards to genomic (19) and medical imaging data (72). Thus, necessary good governance involves more than procedure-following.

Infrastructures are not merely technical, i.e., *buildings, data repositories, but also social – involving practices*. A recent study (73) with biobank professionals and experts indicates that expectations towards biobanks in view of data processing are going beyond their status as repositories. They see biobanks in a more active role when it comes to providing information and communicating and engaging with biobanks participants and point to the need to improve consent procedures and the role of biobanks in sharing samples and data with industry partners and different countries. Considering that participants are the origin of the data, as key stakeholders they should be involved in the development and governance, just as staff in biobanks should be included (74). Decades of biobanking show that *the concerns of citizens cannot be ignored*. In the case of AI in health, these not only relate to the general concerns regarding AI. On the contrary, as suggested by the PRS and AI, ethical, legal and societal issues necessitate a layered understanding due to increasing complexity bringing new relevance to concepts such as explainability and interpretability, both for the users and the broader society (39). Considering the drivers of AI in medicine, such as identification and management of potential patients that can be "high-risk" but also "high-cost" (75), the developments may not benefit individuals who may otherwise develop conditions that are harder to treat or identify and manage emerging outbreaks in real-time, and such AI tools may cause further burdens on the individuals. These necessitate societal debates and empowering citizens, including involving potential non-users, as part of bringing infrastructures to life (76).

Not only are data not always perfect due to inherent finite categorization of potentially infinite diversity, but their capacity to represent should always be continuously problematized. Against the biobanking professionals' concerns, the tendency to see biobanks as data repositories and medicine as increasingly digital (27, 63) can result in a false sense of security in the imaginary of increasing data interoperability and connectedness at the peril of ignoring what D'Ignazio and Klein (77) rightly note the existence of "problems that cannot be represented—or addressed—by data alone" (p. 10). Risks accompany the opportunities in a datafied world. The existence of data should not automatically lead to testing of any potential association and scholars have been trying to identify ways of coping with such issues of reproducibility, e.g., for PRS (78, 79). In this regard, the "curse of dimensionality" in biobanking due to multitude of secondary data even in cases of low sample sizes, can also be seen as an opportunity to think outside of the box to overcome issues even in smaller sample size situations (80). Furthermore, AI may also exacerbate the existing big data issues that are yet to be resolved. While the uses may relate to privacy with unintended access to data from patient implants, sensors and other devices that collect and transfer multiple forms of data, they may also lead to spurious correlations and false positives, tacit assumptions regarding individual behavior based on limited data, sampling issues due to replacement of traditional ways of data collection as well as resulting in injustices due to resource mismanagement and allocation, especially in case of public health issues (81). With health data spaces, these issues will likely need more attention.

Projectified ways of health infrastructuring often restrict the outcome in many ways, through visions and expectations for whom and which purposes the infrastructure is to be developed even in cases where the aim is to involve stakeholders in co-creation processes (76). In this paper we have shown the wealth of knowledge generated through the use of AI in medicine and the evolution of biobanking. We argue, when taken into account, these can positively impact the future European Health Data Space, but also similar establishments, giving power to the citizen, strengthening governance, breaking down potential silos and contributing to trustworthiness of such meta-infrastructures.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

KA: Conceptualization, Writing – original draft, Writing – review & editing. MA: Conceptualization, Writing – original draft, Writing – review & editing. MG: Conceptualization, Writing – original draft, Writing – review & editing. MM: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This publication was funded by BBMRI-ERIC in the context of the activities of BBMRI-ERIC's ELSI Services and Research Unit.

Acknowledgments

Where authors are identified as personnel of the Biobanking and BioMolecular resources Research Infrastructure (BBMRI-ERIC), the authors alone are responsible for the views

expressed in this article and they do not necessarily represent the decisions, policy, or views of BBMRI-ERIC.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Parodi B. Biobanks: a definition In: D Mascalzoni, editor. *Ethics, law and governance of biobanking: National, European and international approaches*. Dordrecht: Springer (2015).

2. De Souza YG, Greenspan JS. Biobanking past, present and future: responsibilities and benefits. *AIDS*. (2013) 27:303–12. doi: 10.1097/QAD.0b013e32835c1244

3. Bak M, Madai VI, Fritzsche M-C, Mayrhofer MT, McLennan S. You Can't have AI both ways: balancing health data privacy and access fairly. *Front Genet*. (2022) 13:929453. doi: 10.3389/fgene.2022.929453

4. Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: mapping the debate. *Big Data Soc*. (2016) 3:205395171667967. doi: 10.1177/2053951716679679

5. Saunders G, Baudis M, Becker R, Beltran S, Béroud C, Birney E, et al. Leveraging European infrastructures to access 1 million human genomes by 2022. *Nat Rev Genet*. (2019) 20:693–701. doi: 10.1038/s41576-019-0156-9

6. Mate S, Kampf M, Rödle W, Kraus S, Proynova R, Silander K, et al. Pan-European data harmonization for biobanks in ADOPT BBMRI-ERIC. *Appl Clin Inform*. (2019) 10:679–92. doi: 10.1055/s-0039-1695793

7. Mayrhofer MT, Prainsack B. Being a member of the club: the transnational (self-) governance of networks of biobanks. *IJRAM*. (2009) 12:64–81. doi: 10.1504/IJRAM.2009.024130

8. European Commission (2023). Directorate-general for health and food safety. European Health Data Space: European Commission. Available at: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en

9. Marelli L, Stevens M, Sharon T, Van Hoyweghen I, Boeckhout M, Colussi I, et al. The European health data space: too big to succeed? *Health Policy*. (2023) 135:104861. doi: 10.1016/j.healthpol.2023.104861

10. BBMRI-ERIC (2023). Public statement: Recommendations for the realisation of the EHDS (European health data spaces) for biobanking from the viewpoint of patient advocates and patient representatives: BBMRI-ERIC. Available at: https://www.bbmri-eric.eu/wp-content/uploads/BBMRI-ERIC-EHDS-Statement-310823.pdf

11. Council of the European Union. Artificial intelligence act: Council and parliament strike a deal on the first rules for AI in the world 2023. Available at: https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/

12. Ruckenstein M, Schüll ND. The Datafication of health. *Annu Rev Anthropol*. (2017) 46:261–78. doi: 10.1146/annurev-anthro-102116-041244

13. Star SL, Ruhleder K. Steps toward an ecology of infrastructure: design and access for large information spaces. *Inf Syst Res*. (1996) 7:111–34. doi: 10.1287/isre.7.1.111

14. Brault N, Aucouturier E. Ethical horizons of biobank-based artificial intelligence in biomedical research In: A Saxena and N Brault, editors. *Artificial intelligence and computational dynamics for biomedical research*. Berlin: De Gruyter (2022).

15. Lemoine M. Neither from words, nor from visions: understanding p-medicine from innovative treatments. *Lato Sensu*. (2018) 4:12–23. doi: 10.20416/lsrsps.v4i2.793

16. Hoeyer K, Bauer S, Pickersgill M. Datafication and accountability in public health: introduction to a special issue. *Soc Stud Sci*. (2019) 49:459–75. doi: 10.1177/0306312719860202

17. Tcheandjieu C, Zhu X, Hilliard AT, Clarke SL, Napolioni V, Ma S, et al. Large-scale genome-wide association study of coronary artery disease in genetically diverse populations. *Nat Med*. (2022) 28:1679–92. doi: 10.1038/s41591-022-01891-3

18. Goisauf M, Akyüz K, Martin GM. Moving back to the future of big data-driven research: reflecting on the social in genomics. *Humanit. Soc. Sci.* (2020) 7:55. doi: 10.1057/s41599-020-00544-5

19. Akyüz K, Goisauf M, Chassang G, Kozera Ł, Mežinska S, Tzortzatou-Nanopoulou O, et al. Post-identifiability in changing sociotechnological genomic data environments. *BioSocieties*. (2023):1–28. doi: 10.1057/s41292-023-00299-7 [Online ahead of print].

20. Lehmann S, Guadagni F, Moore H, Ashton G, Barnes M, Benson E, et al. Standard Preanalytical coding for biospecimens: review and implementation of the sample PREanalytical code (SPREC). *Biopreserv Biobank*. (2012) 10:366–74. doi: 10.1089/bio.2012.0012

21. Moore HM, Kelly A, Jewell SD, McShane LM, Clark DP, Greenspan R, et al. Biospecimen reporting for improved study quality. *Biopreserv Biobank*. (2011) 9:57–70. doi: 10.1089/bio.2010.0036

22. Eklund N, Andrianarisoa NH, van Enckevort E, Anton G, Debucquoy A, Müller H, et al. Extending the minimum information about BIobank data sharing terminology to describe samples, sample donors, and events. *Biopreserv Biobank*. (2020) 18:155–64. doi: 10.1089/bio.2019.0129

23. Merino-Martinez R, Norlin L, van Enckevort D, Anton G, Schuffenhauer S, Silander K, et al. Toward global biobank integration by implementation of the minimum information about BIobank data sharing (MIABIS 2.0 Core). *Biopreserv Biobank*. (2016) 14:298–306. doi: 10.1089/bio.2015.0070

24. Zhuang Y-J, Mangwiro Y, Wake M, Saffery R, Greaves RF. Multi-omics analysis from archival neonatal dried blood spots: limitations and opportunities. *Clin Chem Lab Med*. (2022) 60:1318–41. doi: 10.1515/cclm-2022-0311

25. Kaushal A, Zhang H, Karmaus WJJ, Ray M, Torres MA, Smith AK, et al. Comparison of different cell type correction methods for genome-scale epigenetics studies. *BMC Bioinformatics*. (2017) 18:216. doi: 10.1186/s12859-017-1611-2

26. Califf RM. Biomarker definitions and their applications. *Exp Biol Med*. (2018) 243:213–21. doi: 10.1177/1535370217750088

27. Bonizzi G, Zattoni L, Fusco N. Biobanking in the digital pathology era. *Oncol Res*. (2021) 29:229–33. doi: 10.32604/or.2022.024892

28. Kozlakidis Z. Biobanks and biobank-based artificial intelligence (AI) implementation through an international Lens In: A Holzinger, R Goebel, M Mengel and H Müller, editors. *Artificial intelligence and machine learning for digital pathology: State-of-the-art and future challenges*. Cham: Springer International Publishing (2020)

29. Vande Loock K, Van der Stock E, Debucquoy A, Emmerechts K, Van Damme N, Marbaix E. The Belgian virtual Tumorbank: a tool for translational Cancer research. *Front Med*. (2019) 6:6. doi: 10.3389/fmed.2019.00120

30. Holub P, Swertz M, Reihs R, van Enckevort D, Müller H, Litton J-E. BBMRI-ERIC directory: 515 biobanks with over 60 million biological samples. *Biopreserv Biobank*. (2016) 14:559–62. doi: 10.1089/bio.2016.0088

31. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. (2018) 47:D1005–12. doi: 10.1093/nar/gky1120

32. Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, et al. The NHGRI-EBI GWAS catalog: knowledgebase and deposition resource. *Nucleic Acids Res*. (2023) 51:D977–85. doi: 10.1093/nar/gkac1010

33. Mackenzie G, Richard M, Paige F, Mark S. Trust and the Goldacre review: why trusted research environments are not about trust. *J Med Ethics*. (2022) 49:670–3. doi: 10.1136/jme-2022-108435

34. Marcon Y, Bishop T, Avraam D, Escriba-Montagut X, Ryser-Welch P, Wheater S, et al. Orchestrating privacy-protected big data analyses of data from different resources with R and DataSHIELD. *PLoS Comput Biol*. (2021) 17:e1008880. doi: 10.1371/journal.pcbi.1008880

35. Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. *Gastrointest Endosc*. (2020) 92:807–12. doi: 10.1016/j.gie.2020.06.040

36. Chen M, Decary M. Artificial intelligence in healthcare: an essential guide for health leaders. *Healthc Manage Forum*. (2019) 33:10–8. doi: 10.1177/0840470419873123

37. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. *Nature*. (2020) 577:89–94. doi: 10.1038/s41586-019-1799-6

38. Mudgal KS, Das N. The ethical adoption of artificial intelligence in radiology. *BJR|Open*. (2019) 2:20190020. doi: 10.1259/bjro.20190020

39. Fritzsche M-C, Akyüz K, Cano Abadía M, McLennan S, Marttinen P, Mayrhofer MT, et al. Ethical layering in AI-driven polygenic risk scores—new complexities, new challenges. *Front Genet*. (2023) 14:1098439. doi: 10.3389/fgene.2023.1098439

40. Niel O, Bastard P, Boussard C, Hogan J, Kwon T, Deschênes G. Artificial intelligence outperforms experienced nephrologists to assess dry weight in pediatric patients on chronic hemodialysis. *Pediatr Nephrol*. (2018) 33:1799–803. doi: 10.1007/s00467-018-4015-2

41. Topalovic M, Das N, Burgel P-R, Daenen M, Derom E, Haenebalcke C, et al. Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests. *Eur Respir J*. (2019) 53:1801660. doi: 10.1183/13993003.01660-2018

42. Carrano FM, Wang B, Sherman SE, Makarov DV, Berman RS, Newman E, et al. Artificial intelligence outperforms clinical judgment in triage for postoperative ICU care: prospective preliminary results. *J Am Coll Surg*. (2019) 229:S141–2. doi: 10.1016/j.jamcollsurg.2019.08.312

43. Hariton E, Dimitriadis I, Kanakasabapathy MK, Thirumalaraju P, Gupta R, Pooniwala R, et al. A deep learning framework outperforms embryologists in selecting day 5 euploid blastocysts with the highest implantation potential. *Fertil Steril*. (2019) 112:e77–8. doi: 10.1016/j.fertnstert.2019.07.324

44. Rahman A, Hossain MS, Muhammad G, Kundu D, Debnath T, Rahman M, et al. Federated learning-based AI approaches in smart healthcare: concepts, taxonomies, challenges and open issues. *Clust Comput*. (2023) 26:2271–311. doi: 10.1007/s10586-022-03658-4

45. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*. (2016) 3:160018. doi: 10.1038/sdata.2016.18

46. Goisauf M, Cano AM. Ethics of AI in radiology: a review of ethical and societal implications. *Front Big Data*. (2022) 5:850383. doi: 10.3389/fdata.2022.850383

47. Muller H, Mayrhofer M, Veen EV, Holzinger A. The ten commandments of ethical medical AI. *Computer*. (2021) 54:119–23. doi: 10.1109/MC.2021.3074263

48. Beauchamp TL, Childress JF. Principles of biomedical ethics: marking its fortieth anniversary. *AJOB*. (2019) 19:9–12. doi: 10.1080/15265161.2019.1665402

49. Prakash S, Balaji JN, Joshi A, Surapaneni KM. Ethical conundrums in the application of artificial intelligence (AI) in healthcare — a scoping review of reviews. *J Pers Med*. (2022) 12:1914. doi: 10.3390/jpm12111914

50. Battineni G, Hossain MA, Chintalapudi N, Amenta F. A survey on the role of artificial intelligence in biobanking studies: a systematic review. *Diagnostics*. (2022) 12:1179. doi: 10.3390/diagnostics12051179

51. Tian J, Smith G, Guo H, Liu B, Pan Z, Wang Z, et al. Modular machine learning for Alzheimer's disease classification from retinal vasculature. *Sci Rep*. (2021) 11:238. doi: 10.1038/s41598-020-80312-2

52. Yan Q, Jiang Y, Huang H, Swaroop A, Chew EY, Weeks DE, et al. Genome-wide association studies-based machine learning for prediction of age-related macular degeneration risk. Translational vision. *Sci Technol*. (2021) 10:29. doi: 10.1167/tvst.10.2.29

53. Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK biobank participants. *PLoS One*. (2019) 14:e0213653. doi: 10.1371/journal.pone.0213653

54. Schulz M-A, Chapman-Rounds M, Verma M, Bzdok D, Georgatzis K. Inferring disease subtypes from clusters in explanation space. *Sci Rep*. (2020) 10:12900. doi: 10.1038/s41598-020-68858-7

55. Dabbah MA, Reed AB, Booth ATC, Yassaee A, Despotovic A, Klasmer B, et al. Machine learning approach to dynamic risk modeling of mortality in COVID-19: a UK biobank study. *Sci Rep*. (2021) 11:16936. doi: 10.1038/s41598-021-95136-x

56. Jimenez-Solem E, Petersen TS, Hansen C, Hansen C, Lioma C, Igel C, et al. Developing and validating COVID-19 adverse outcome risk prediction models from a bi-national European cohort of 5594 patients. *Sci Rep*. (2021) 11:3246. doi: 10.1038/s41598-021-81844-x

57. Reinbolt RE, Sonis S, Timmers CD, Fernández-Martínez JL, Cernea A, de Andrés-Galiana EJ, et al. Genomic risk prediction of aromatase inhibitor-related arthralgia in patients with breast cancer using a novel machine-learning algorithm. *Cancer Med*. (2018) 7:240–53. doi: 10.1002/cam4.1256

58. Mayrhofer MT. About the new significance and the contingent meaning of biological material and data in biobanks. *Hist Phil Life Sci*. (2013) 35:449–67.

59. Lee J. Artificial intelligence in the future biobanking: current issues in the biobank and future possibilities of artificial intelligence. *Biomed J Sci Tech Res*. (2018) 7:5937–9. doi: 10.26717/BJSTR.2018.07.001511

60. Grossman GH, Henderson MK. Readiness for artificial intelligence in biobanking. *Biopreserv Biobank*. (2023) 21:119–20. doi: 10.1089/bio.2023.29121.editorial

61. Garcia DL. ISBER President's message: ISBER's 20th anniversary—celebrating the journey. *Biopreserv Biobank*. (2019) 17:375–6. doi: 10.1089/bio.2019.29056.dlg

62. Kargl M, Plass M, Müller H. A literature review on ethics for AI in biomedical research and biobanking. *Yearb Med Inform*. (2022) 31:152–60. doi: 10.1055/s-0042-1742516

63. Tozzo P, Delicati A, Marcante B, Caenazzo L. Digital biobanking and big data as a new research tool: a position paper. *Healthcare*. (2023) 11:1825. doi: 10.3390/healthcare11131825

64. Gille F, Vayena E, Blasimme A. Future-proofing biobanks' governance. *Eur J Hum Genet*. (2020) 28:989–96. doi: 10.1038/s41431-020-0646-4

65. Akyüz K, Chassang G, Goisauf M, Kozera Ł, Mezinska S, Tzortzatou O, et al. Biobanking and risk assessment: a comprehensive typology of risks for an adaptive risk governance. *Life Sci Soc Policy*. (2021) 17:1–28. doi: 10.1186/s40504-021-00117-7

66. Ryan M, Stahl BC. Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *J Inf Commun Ethics Soc*. (2020) 19:61–86. doi: 10.1108/JICES-12-2019-0138

67. Bijker EM, Sauerwein RW, Bijker WE. Controlled human malaria infection trials: how tandems of trust and control construct scientific knowledge. *Soc Stud Sci*. (2016) 46:56–86. doi: 10.1177/0306312715619784

68. Wyatt S, Harris A, Adams S, Kelly SE. Illness online: self-reported data and questions of Trust in Medical and Social Research. *Theory Cult Soc*. (2013) 30:131–50. doi: 10.1177/0263276413485900

69. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Mind Mach*. (2018) 28:689–707. doi: 10.1007/s11023-018-9482-5

70. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell*. (2019) 1:389–99. doi: 10.1038/s42256-019-0088-2

71. Hoeyer K, Tupasela A, Rasmussen MB. Ethics policies and ethics work in cross-national genetic research and data sharing. *Sci Technol Hum Values*. (2017) 42:381–404. doi: 10.1177/0162243916674321

72. Lotan E, Tschider C, Sodickson DK, Caplan AL, Bruno M, Zhang B, et al. Medical imaging and privacy in the era of artificial intelligence: myth, fallacy, and the future. *J Am Coll Radiol*. (2020) 17:1159–62. doi: 10.1016/j.jacr.2020.04.007

73. Goisauf M, Martin G, Bentzen HB, Budin-Ljøsne I, Ursin L, Durnová A, et al. Data in question: a survey of European biobank professionals on ethical, legal and societal challenges of biobank research. *PLoS One*. (2019) 14:e0221496. doi: 10.1371/journal.pone.0221496

74. Akyüz K, Goisauf M, Martin GM, Mayrhofer MT, Antoniou S, CharalambidouG, etal. Risk mapping for better governance in biobanking: The case of biobank.Cy. In press.

75. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Aff*. (2014) 33:1123–31. doi: 10.1377/hlthaff.2014.0041

76. Felt U, Öchsner S, Rae R, Osipova E. Doing co-creation: power and critique in the development of a European health data infrastructure. *J Responsible Innov*. (2023) 10:2235931. doi: 10.1080/23299460.2023.2235931

77. D'Ignazio C, Klein LF. *Data Feminism*. Cambridge: MIT Press (2020).

78. Lambert SA, Gil L, Jupp S, Ritchie SC, Xu Y, Buniello A, et al. The polygenic score catalog as an open database for reproducibility and systematic evaluation. *Nat Genet*. (2021) 53:420–5. doi: 10.1038/s41588-021-00783-5

79. Wand H, Lambert SA, Tamburro C, Iacocca MA, O'Sullivan JW, Sillari C, et al. Improving reporting standards for polygenic scores in risk prediction studies. *Nature*. (2021) 591:211–9. doi: 10.1038/s41586-021-03243-6

80. Narita A, Ueki M, Tamiya G. Artificial intelligence powered statistical genetics in biobanks. *J Hum Genet*. (2021) 66:61–5. doi: 10.1038/s10038-020-0822-y

81. Strang KD, Sun Z. Hidden big data analytics issues in the healthcare industry. *Health Informatics J*. (2019) 26:981–98. doi: 10.1177/1460458219854603

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 10 April 2024
DOI 10.3389/fmed.2024.1301660

OPEN ACCESS

EDITED BY
Gokce Banu Laleci Erturkmen,
Software Research and Development Consulting, Türkiye

REVIEWED BY
Sascha Welten,
RWTH Aachen University, Germany
Stefano Abbate,
University of Naples Federico II, Italy

*CORRESPONDENCE
Martin Baumgartner
✉ martin.baumgartner@ait.ac.at

RECEIVED 25 September 2023
ACCEPTED 21 March 2024
PUBLISHED 10 April 2024

CITATION
Baumgartner M, Kreiner K, Lauschensky A, Jammerbund B, Donsa K, Hayn D, Wiesmüller F, Demelius L, Modre-Osprian R, Neururer S, Slamanig G, Prantl S, Brunelli L, Pfeifer B, Pölzl G and Schreier G (2024) Health data space nodes for privacy-preserving linkage of medical data to support collaborative secondary analyses. *Front. Med.* 11:1301660. doi: 10.3389/fmed.2024.1301660

Health data space nodes for privacy-preserving linkage of medical data to support collaborative secondary analyses

Martin Baumgartner[1,2]*, Karl Kreiner[1], Aaron Lauschensky[1], Bernhard Jammerbund[1], Klaus Donsa[1], Dieter Hayn[1,3], Fabian Wiesmüller[1,2,3], Lea Demelius[4,5], Robert Modre-Osprian[6], Sabrina Neururer[7,8], Gerald Slamanig[9], Sarah Prantl[9], Luca Brunelli[10], Bernhard Pfeifer[8,11], Gerhard Pölzl[10] and Günter Schreier[1,2]

[1]Center for Health and Bioresources, AIT Austrian Institute of Technology, Vienna, Austria, [2]Institute of Neural Engineering, Graz University of Technology, Graz, Austria, [3]Ludwig Boltzmann Institute for Digital Health and Prevention, Salzburg, Austria, [4]Institute of Interactive Systems and Data Science, Graz University of Technology, Graz, Austria, [5]Know-Center GmbH, Graz, Austria, [6]telbiomed Medizintechnik und IT Service GmbH, Graz, Austria, [7]Department of Clinical Epidemiology, Tyrolean Federal Institute for Integrated Care, Tirol Kliniken GmbH, Innsbruck, Austria, [8]Division for Digital Health and Telemedicine, UMIT TIROL—Private University for Health Sciences and Technology, Hall in Tyrol, Austria, [9]Tirol Kliniken GmbH, Innsbruck, Austria, [10]Department of Internal Medicine III, Cardiology and Angiology, Medical University of Innsbruck, Innsbruck, Austria, [11]Tyrolean Federal Institute for Integrated Care, Tirol Kliniken GmbH, Innsbruck, Austria

Introduction: The potential for secondary use of health data to improve healthcare is currently not fully exploited. Health data is largely kept in isolated data silos and key infrastructure to aggregate these silos into standardized bodies of knowledge is underdeveloped. We describe the development, implementation, and evaluation of a federated infrastructure to facilitate versatile secondary use of health data based on Health Data Space nodes.

Materials and methods: Our proposed nodes are self-contained units that digest data through an extract-transform-load framework that pseudonymizes and links data with privacy-preserving record linkage and harmonizes into a common data model (OMOP CDM). To support collaborative analyses a multi-level feature store is also implemented. A feasibility experiment was conducted to test the infrastructures potential for machine learning operations and deployment of other apps (e.g., visualization). Nodes can be operated in a network at different levels of sharing according to the level of trust within the network.

Results: In a proof-of-concept study, a privacy-preserving registry for heart failure patients has been implemented as a real-world showcase for Health Data Space nodes at the highest trust level, linking multiple data sources including (a) electronical medical records from hospitals, (b) patient data from a telemonitoring system, and (c) data from Austria's national register of deaths. The registry is deployed at the tirol kliniken, a hospital carrier in the Austrian state of Tyrol, and currently includes 5,004 patients, with over 2.9 million measurements, over 574,000 observations, more than 63,000 clinical free text notes, and in total over 5.2 million data points. Data curation and harmonization processes are executed semi-automatically at each individual node according to data sharing policies to ensure data sovereignty, scalability, and privacy. As a

feasibility test, a natural language processing model for classification of clinical notes was deployed and tested.

Discussion: The presented Health Data Space node infrastructure has proven to be practicable in a real-world implementation in a live and productive registry for heart failure. The present work was inspired by the European Health Data Space initiative and its spirit to interconnect health data silos for versatile secondary use of health data.

KEYWORDS

data-driven healthcare, privacy-preservation, record linkage, advanced analytics, interoperability, machine learning, artificial intelligence, European Health Data Space

1 Introduction

Real-world data (RWD) is typically gathered over a patient's lifetime for the purpose of patient care (*primary use*). However, beyond its original use, RWD can be used for other analyses (*secondary use*) to generate additional real-world evidence (1). Among other aspects, secondary use proved to be valuable for cost-effectiveness analysis (2), data exploration (3), clinical outcomes research (4, 5), data validation (6) and data aggregation (7). However, medical data is sensitive by nature. Strict legal frameworks around highly sensitive data impose challenging demands on data holders (e.g., healthcare organizations). On top of that, as opposed to RWD, collecting data in clinical trials is eminently expensive and the resulting data is therefore highly valuable to those who hold it. Both privacy and security considerations as well as the associated costs of health data make data holders exceedingly reluctant to share any data with a health ecosystem. Sharing data also has implications regarding data sovereignty (i.e., who owns and controls data). This is further complicated by the fact, that many countries have not yet fully defined ownership of medical data in their legal frameworks (8). Consequently, health data of different sources is often kept in isolated data silos, and its value for further secondary analyses remains underutilized (9, 10). Connecting silos can accomplish both *vertical linkage* (i.e., more data for one patient) as well as *horizontal linkage* (i.e., more patients for specific data) and thus provide more holistic views on patients and diseases increasing the data's value for research even further.

An example of secondary use of health data was an analysis of data from *HerzMobil Tyrol* (HMT), which is a telehealth-supported disease management program for heart failure patients in Tyrol, Austria for which patients are recruited after an episode of acute heart failure and receive optimized disease management care by a network of health professionals (11). In HMT, patients are given measurement equipment (e.g., a bodyweight scale, blood pressure cuff), which is connected to an app, through which patients can record daily physiological (e.g., bodyweight), fitness (e.g., steps per day) and self-reported (e.g., wellbeing) data. In Tyrol, over 1,000 patients have been monitored by this telehealth system and the data is highly valuable for secondary analyses. To investigate the clinical effectiveness of the program, electronic medical records (EMR), and clinical outcome data from HMT patients and a control group were compiled for a secondary use analysis (5). For this analysis, data from three different sources were required: (1) telehealth data from the HMT system itself, (2) EMRs from the patients' hospitals' information system and (3) information about time and cause of death from Austria's national register of deaths. This resulted in an aggregated dataset containing more than 80 variables and while reduced mortality for patients in the telehealth program compared to conventional care has been found (5), several challenges were encountered:

- Data linkage—The analysis required linkage of data from three different data sources including hospital information systems (HIS), the *HerzMobil* telehealth system, and the Austrian register of deaths. Data linkage had to be done manually, as the data sources did not share a unique alpha-numeric identifier. Additionally, although the laboratory information systems were part of the same hospitals, they also used their specific identifiers.
- Privacy preservation—To achieve privacy preservation, personally identifiable information (PII) had to be manually removed from the datasets.
- Unstructured data—RWD data used in the analysis contained both structured and unstructured data. The latter imposed additional challenges for the de-identification of text for secondary use.
- Interoperability—While data sources provided coded data (e.g., ICD-10 codes) for various data elements, they did not adhere to one harmonized coding vocabulary or a common data model for the resulting dataset making the individual data sources not interoperable.
- Collaboration—Different data sources and different data types (e.g., unstructured data) required a team of researchers compiling the aggregated datasets using various analysis pipelines, which made intensive communication and exchange of intermediate results necessary.
- Traceability—With more than 80 variables involved in the analysis, tracing all involved algorithms and processing steps used to derive a specific variable proved to be difficult.
- Extensibility—Necessity for both *vertical linkage* of more data sources from out-patient domains as well as *horizontal linkage* of data for comparison with identical *HerzMobil* systems in the states of Styria and Carinthia to improve the data analysis was identified for future studies.
- Automation—To increase repeatability, having the possibility to easily rerun analyses on a regular basis is required. This was not possible with the aforementioned manual labor required.

Comparing the experience from this retrospective view on the challenges encountered during the HMT effectiveness analysis with published literature, a general trend of similar, regularly occurring problems can be observed. Privacy, interoperability, data governance, organizational coordination, data quality and funding considerations are frequently being mentioned as the most pressing issues (12–15). A more detailed view on these challenges is given in the following list.

1.1 Privacy, security and data linkage

Health information is highly sensitive data and therefore access is regulated through data protection and security frameworks. To mitigate the data's sensitivity and to keep with the spirit of the EU's General Data Protection Regulation's (GDPR) (16) principle of data minimization, any identifying elements not required for analysis (e.g., names, specific date of birth) should be removed from the dataset in advance. However, removing this information complicates record linkage, which is necessary to associate data with the correct individuals across different contexts and to avoid duplication of subjects. Furthermore, medical free texts (e.g., clinical messages, nursing documentation) typically include references to personal information (e.g., names, addresses) that also infringe on patient privacy and increase the risk of re-identification.

1.2 Standardization and interoperability

While interoperability might not be of utmost importance when working with isolated data silos, it becomes a core necessity when connecting data from multiple silos. Source systems store data in different data formats (i.e., data models) and use different vocabularies and thus datasets are frequently not interoperable originally. This requires time-consuming, manual effort to map different elements from the sources into a common dataset (i.e., a feature matrix) and to translate values into a mutual standard vocabulary.

1.3 Data quality and availability

As health data is often entered or administered manually, source data needs to be verified to avoid erroneous data. Furthermore, related to the aforementioned interoperability aspects, some elements of data are ambiguously encoded or worded. Also, in some instances, not all source data is available in digital form or complete at all times. These factors require regular contact with data holders for clarification. Lastly, sometimes additional context is necessary for analysis (e.g., labels for supervised machine learning), which is also time-consuming and is known to be associated with it its own unique challenges (17).

1.4 Stakeholder management and data sovereignty

Data linkage requires collaboration of multidisciplinary teams of clinicians, nurses, administrators, and engineers. These groups have different interests (e.g., data sovereignty, workload management) that need to be aligned. Dedicating medical and engineering staff to set up, provide and maintain infrastructure to link and harmonize data is generally associated with costs (18). Since health infrastructure projects are often non-profit oriented and executed with public funds, a certain political and institutional will is often required, Also, as data can originate from different sources, datasets can be subject to different data sovereignty spheres and legislation.

1.5 Collaboration during data analysis

In complex real-world scenarios, multiple data engineers, data analysts and machine learning engineers are working on the same data. This requires extensive communication and coordination to avoid redundant work on data processing, feature engineering and model development processes. Many experiments require the same standard data and feature engineering algorithms, which are at risk of being duplicated by multiple team members, which ultimately results in less efficient collaborative analysis. To improve collaboration, the concept of feature stores has gained popularity recently (19). The idea is to collect feature extraction algorithms over multiple experiments to nurture a growing repository of re-usable features, which can be made accessible for all team members to speed up machine learning analyses. Additionally, machine learning operations (MLOps) to aid in model deployment requires suitable infrastructure. Kim discussed the software engineering difficulties concerning MLOps, such as complex software stacks and distributed data (20). Due to the intricacies of MLOps for health data, Khattak et al. introduced the term "*Machine Learning Healthcare Operations*" (MLHOps) (21).

Tayefi et al. (14) concluded that key infrastructure technology to facilitate secondary use of health data addressing these challenges is required but still underdeveloped. A typical approach to implementing such infrastructure is the introduction of an enterprise data warehouse or integrated data repository (IDR). Gagalova et al. (22) have described architectural principles of IDRs in the clinical domain distinguishing centralized approaches (General architecture), biobank-driven architectures and federated approaches. They also identified the need for a common data model (CDM) to represent data. Solutions following these approaches are described in literature. For example, *DataSHIELD* is a federated platform by an international consortium of researchers that facilitates distributed analysis to avoid data exchange entirely with a client–server infrastructure for data analysis (23). The *Personal Health Train* (PHT) is another federated infrastructure solution to reuse medical data for secondary use (24). The *PHT* aims to establish FAIR data stations that can be governed by data holders and accessed by analysts whereas trains travel from station to station carrying algorithms that are executed in the FAIR data stations. Secure multiparty computation (25–27) and more recently, blockchain-based concepts (28–32) have also gained popularity to increase data security in privacy-preserving trustless systems. Although keeping data distributed across multiple sources is privacy-minded, performance of machine learning models still suffers in federated learning settings compared to conventional centralized learning (33–35). Therefore, another architectural approach is to accumulate data in a centralized point (i.e., a clinical data warehouse) with secure and privacy-oriented infrastructure. Wirth et al. (36) and Jin et al. (37) both provide a comprehensive overview and analysis of a selection of privacy-minded data sharing networks in their works. CDMs are important for data warehouses to serve as a common

denominator when multiple heterogenous data sources are to be linked and standard vocabularies ensure interpretability of data values. A specific successful example of medical data sharing is the open-source software platform *informatics for integrating biology and the bedside* (i2b2) developed by Harvard Medical School (38) to drive clinical research. The partnership between *i2b2* and *tranSMART* (39), an open-source data warehouse developed by a consortium of private pharmacological companies resulted in the *i2b2 tranSMART* foundation (40). Further literature examples include *GIFT-Cloud* (sharing medical image data) (41), the *Shariant* platform (sharing clinical genetic data-testing data) (42) and *IMPROVE-PD* (sharing peritoneal dialysis data) (43). However, it has been outlined clearly that many currently existing solutions are limited to one specific use case (44). Gruendner et al. made use of best-practice principles and established the *KETOS* platform, which is a containerized (Docker) solution with standard vocabularies (SNOMED & LOINC) and the Observational Medical Outcomes Partnership common data model (OMOP CDM) for a more general development environment (44).

While these solutions work well for their intended purposes, they do not completely fulfill our requirements. Blockchain-based distributed systems are proven effective in multiple studies (28–32), however suffer from the slow pace at which this technology is adopted in the health sector, which ultimately makes them impractical currently. While *DataSHIELD* is an excellent example of a framework that enables federated analyses, it is not intended to also support machine learning (e.g., federated learning). The *PHT* is based on data trains containerized with Docker to be sent to data stations where code is executed. In our experience, system administrators of healthcare organizations are hesitant about this form of code execution on their environments even though there are containerized, mostly because they lack control over the code and thus data sovereignty becomes a concern. Furthermore, although the *PHT* could support a form of federated learning, studies have shown, that performance of ML models trained by federated learning can trail behind centrally trained models (33–35). Therefore, for optimal AI applications, data is required to be aggregated in a central point to train models to their full potential, for which key infrastructure is required. While the *KETOS* platform aims to fulfill exactly that, in *KETOS*, privacy and security by limiting data storage to remain within a hospital information system. Therefore, linkage to other data sources is restricted, which is a key requirement for our system.

In this study we propose a federated node-based system architecture called Health Data Space (HDS) nodes. These nodes aim at facilitating linkage (*horizontal* and *vertical*) between multiple, decentralized data sources. The architecture supports privacy-preserving record linkage (PPRL) and additional de-identification algorithms. For interoperability, we outline how we harmonized heterogenous data into the OMOP CDM, which is suitable since our data is mostly observational health data. We further propose how a multi-level feature store can be realized to support collaborative data analytics. We also present preliminary experiments to assess the nodes' feasibility of supporting MLOps in future developments. We hope to utilize this solution to facilitate time-efficient analyses to answer clinical research questions (e.g., efficiency, health economics) quicker and allow data linkage to scale with related systems (e.g., *HerzMobil Styria* and *HerzMobil Carinthia*).

As a proof-of-concept, we describe a real-world application of a heart failure registry established in Austria with HDS nodes with three different data sources. We further discuss the organizational considerations of developing such multidisciplinary infrastructure. In particular, the following contributions are to be highlighted.

1.5.1 Pseudonymization concept and free text de-identification

To adhere to strict legal frameworks like GDPR, respect patient privacy and minimize risk of exposure, the HDS nodes use a PPRL system to avoid storing quasi-identifiers. In this spirit, an additional de-identification algorithm is in place to remove identifying references from free text data, while aiming to retain context by applying basic entity recognition logic.

1.5.2 Multi-level feature store based on the OMOP CDM

A feature store based on the OMOP CDM is used to avoid repeated feature engineering and improve experiment repeatability. The feature store allows features on multiple levels (e.g., on patient level like age and sex, but also on daily observational level like blood pressure). These features can then be linked into a feature matrix and accessed for later ML experiments.

1.5.3 Case study of sharing secondary data in a heart failure registry

The HDS nodes are used in a real-world case study for a registry for chronic heart failure patients, in which health data from three different sites are linked.

2 Materials and methods

We introduce the concept of HDS nodes as fundamental building blocks of health data spaces. The HDS node components are illustrated in Figure 1. Python and the Django Web framework for server components (45) were chosen due to the large popularity of Python in data analysis. During development, only modules and libraries were selected that allowed HDS nodes to be infrastructure-agnostic, meaning they are compatible with deployment on different cloud environments (e.g., Microsoft Azure or Amazon AWS), but can also be deployed on-premises. They also support a variety of relational databases (e.g., MySQL, PostgreSQL). For evaluation, PostgreSQL was used as primary database technology.

Data can be submitted from a source to an HDS node by data holders via a public application programming interface (API) which forwards the data to the HDS node's Extract-Transform-Load (ETL) framework. The ETL framework consists of a collection of individual ETL classes, that act as converters and first pseudonymize and then transform incoming data into the OMOP CDM. ETL classes are implemented as plain Python classes. No visual editors are used, but instead all steps in the workflow are expressed as code, that digests new data submitted to a node. Data submission can either be automated in regular intervals (e.g., via cron jobs) or manually executed on demand. The data engineering pipeline as seen in Figure 1 starts with pseudonymization (1), which is followed by harmonization (2) after which data is saved in a data store based on the OMOP CDM. We chose the OMOP CDM because (a) it is increasingly adopted in clinical research for observational health data, (b) it provides a large variety of standard terminologies, and (c) it is based

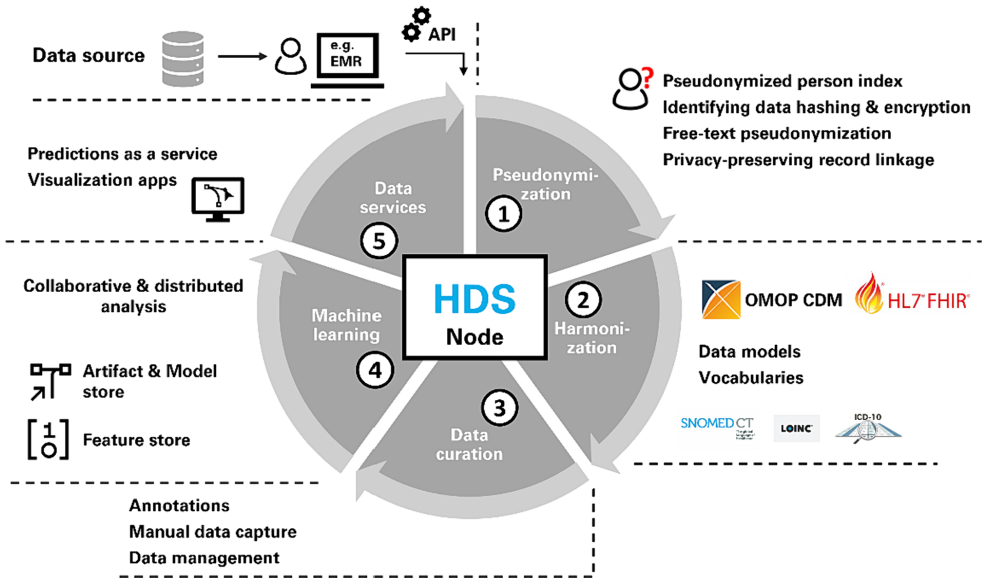

FIGURE 1
Each health data space (HDS) node consists of the same capabilities for pseudonymization, harmonization, data curation, machine learning support and data services.

TABLE 1 Entity types of the D4Health Heart Failure Registry that are pseudonymized into master records.

Master entity record	Identity traits
Patients	First name, last name, date of birth, social security number (if available)
Healthcare professionals	First name, last name, date of birth
Clinical sites	Clinical site's name (e.g., a center, department)

on a comparatively flat data model. Data curation services (3) allow for (a) manual data entry through an HL7 FHIR-based electronic data capture system (EDC) and (b) manual annotation and labeling of data. To facilitate ML, a feature store (19) and a model store are implemented for collaborative analysis (4). Finally, data services (5) support the creation of data and visualization apps as well as providing predictions as web services to other applications (e.g., used for primary use of health data). The last part (data services and model deployment) is mainly focus of future work and largely out of scope of this study as further work to mature this aspect is still needed. A data node may use all these components or only a subset of the functionalities. The individual parts of the data engineering pipeline are described in detail in the subsequent chapters.

2.1 Pseudonymization

For the pseudonymization component, we expanded the work of the European Patient Identity Services (EUPID) (45), introducing a hash-based pseudonymized person index for patients and healthcare professionals. We further identified clinical sites as additional entities that require pseudonymization. All entity types that are pseudonymized in the HDS nodes are listed in Table 1.

Every record (e.g., patient, clinician) has specific identity traits that uniquely identify them. For pseudonymization, they are transformed into record-level hashes by concatenating the string values of all traits to one large string and applying a hash function to the result. A variety of record-level hash algorithms are already provided by EUPID (including HMAC512, Argon2, Bloom filters) and could be used in the HDS nodes. However, to enable similarity matching, we use locality-sensitive cryptographic long-term key (CLK) Bloom filter (BF) hashes (46). To ensure scaling performance in large networks, we applied MinHash (47) in combination with the Bloom filters. With this blocking strategy, hashes are only compared to the most similar ones instead of all available hashes. This drastically reduces the amount of redundant Bloom filter comparisons, which can get computationally expensive once large quantities of records are available. Identity traits are hashed into a 459-bit BF vector and then associated with a randomly generated alpha-numeric pseudonym. As an additional layer of security, HDS nodes operate two independent databases: One to store the actual health data from the data sources without personal data (i.e., the data store) and a separate one to store pseudonymized identity traits (i.e., the pseudonymized person index). The link between data and identity traits is achieved via the alpha-numeric pseudonym, which is available in both databases. As an additional layer of privacy, all records (e.g., patients) are given context-specific pseudonyms (i.e., one pseudonym per node). For example, a patient will have pseudonym *P1* in one node, pseudonym *P2* in another and if both data sources for this patient are linked in a central node, will be assigned pseudonym *P3*. While this connection is traceable in the person index, it will not be visible for data scientists only working with the health-related data. To increase security, BFs are encrypted at rest in the database using AES256 encryption. The encryption key and the HMAC keys required for BF generation are stored outside the databases. For record linkage, the Jaccard distance is applied to all possible pairs of BFs in the person index to identify potential duplicates. Depending on a threshold decision, full matches

and partial matches (e.g., typographic errors) are identified and logged. While full matches are automatically consolidated, partial matches are flagged to be resolved at the data source by human administrators to ensure correctness.

Pseudonymization is also applied on free text data (e.g., clinical notes) with an advancement of a previously developed algorithm (48), which relies on name dictionaries (public and internal), common precursors for names and regular expressions to remove personal references such as names, phone numbers, locations, addresses, email addresses and websites. Public name dictionaries were scraped from Wikipedia articles of category *Person* and the publicly available search tool for physicians in Tyrol. The internal dictionary is comprised of all names within the available data sources. A basic rule-based entity recognition is applied to retain context after removing potentially valuable information by de-identification. The entities of *healthcare professional*, *patient*, *person*, *location*, *phone number*, *e-mail address*, *address*, *ZIP code* and *website* are recognized, and corresponding pseudonyms are assigned, which are consistent throughout the entire text corpus.

2.2 Harmonization

For each data type or dataset that is to be digested into the data store, individual harmonizing ETL classes must be developed manually in advance. In essence, these harmonizer classes read the data they are designed for and map data points to suitable OMOP CDM fields. For further interoperability, the ETL classes also map values of data to standardized vocabularies of the ICD-10, SNOMED-CT, LOINC and ATC terminologies. Any data that is processed like this by an ETL class is tracked to enable version control for the data store. These ETL classes can either be integrated into the HDS node to populate data automatically into the data store if the corresponding data is regularly updated, or resort to outside ETL processes if data is simply imported once without expected regular updates. Mapping all incoming data into the OMOP CDM with standard vocabularies created a scalable data store that can be extended should any new data sources be connected to the HDS node.

2.3 Data curation

The ETL process framework is mainly intended for importing and harmonizing of RWD from primary data sources (i.e., the data's origin). On many occasions, additional data is collected that does not originate from the primary care system, such as quality-of-life data [e.g., MacNew questionnaires (49)]. For this reason, we implemented a basic electronic data capture (EDC) system. As each HDS node provides a FHIR repository, we used FHIR Questionnaires to define EDC forms and FHIR CarePlans to express typical workflows. Entered forms and their completion statuses are stored as FHIR Questionnaire responses. The EDC component is tied to the pseudonymization component, so that patients can be registered manually and linked to existing patients from primary care data sources with the record linkage algorithm. For enhanced privacy, subjects in the EDC system receive their own pseudonym which is automatically linked to the pseudonym used in the OMOP database. Both the FHIR Questionnaires as well as the FHIR Questionnaire responses are transformed via ETL classes into the OMOP CDM. We defined functions that transform them into the OMOP entities *VisitOccurrence* (the action of completing a form), *SurveyConduct* (details on the questionnaire itself) and *Observation* (the actual questions) and store them in the OMOP database.

Our experience with HMT has shown that some critically valuable data is only available in unstructured form. For instance, in the telemonitoring setting of HMT, physicians and nurses make extensive use of free text notes to capture additional insights into patients' condition and treatment. Similarly, in the patients' EMR, discharge letters contain free text diagnoses and discharge medication prescriptions. Based on previous work (50), we integrated (a) a tool to create annotation corpora from OMOP data, and (b) a multi-annotator tool for manually annotating text data on both the sentence and the full-text levels. Annotated corpora can be accessed through APIs like data from the data store for further analyses (e.g., training classification algorithms).

2.4 Collaborative analyses

Typically, data analysis and ML tasks are complex, iterative processes with multiple steps involving an interdisciplinary group of experts (20). Depending on their specific role, experience, and training, team members might prefer different tools (e.g., Python, MATLAB). To support the usage of said tools, the HDS nodes provide a dedicated API to extract pseudonymized data via SQL queries from the nodes' data store. We developed functions for Python, R and MATLAB to (a) access an HDS node's data store via the API, and (b) transform the received data into native data formats, including *Pandas DataFrames* (Python), *data.frames* (R), and *tables* (MATLAB). This platform-agnostic way of accessing data allows data scientists to rely on their preferred tool chain they are familiar with to develop algorithms and models. When given access to an HDS node through the permission management system, data scientists can browse the data available (see Figure 2 for an example) and simple descriptive statistics (e.g., distribution of sex and age) are provided via a dashboard. An SQL editor allows data scientists to understand the database scheme and test SQL queries before executing them in their processes. SQL queries are tracked for audits and can be saved for repeated executions. Data scientists are also given access to a collection of already developed feature extraction algorithms, called feature store. Figure 3 illustrates how a typical workflow involving feature generation, model development and model deployment involving a data engineer and a data scientist could be executed.

2.4.1 Feature store

Once the required data is extracted via API queries, data analysis often requires the calculation or engineering of features (i.e., derived values from raw data). These represent information-dense data points to be used for machine learning modeling. Since medical datasets are relatively sparse, typically multiple people work on the same data. However, on occasion, different analyses by different data scientists can require the same features. For example, with the available blood pressure data of systolic and diastolic values, it will often be required to calculate the pulse pressure. The nodes' feature store allows data scientists to upload the algorithms' code they have developed into a

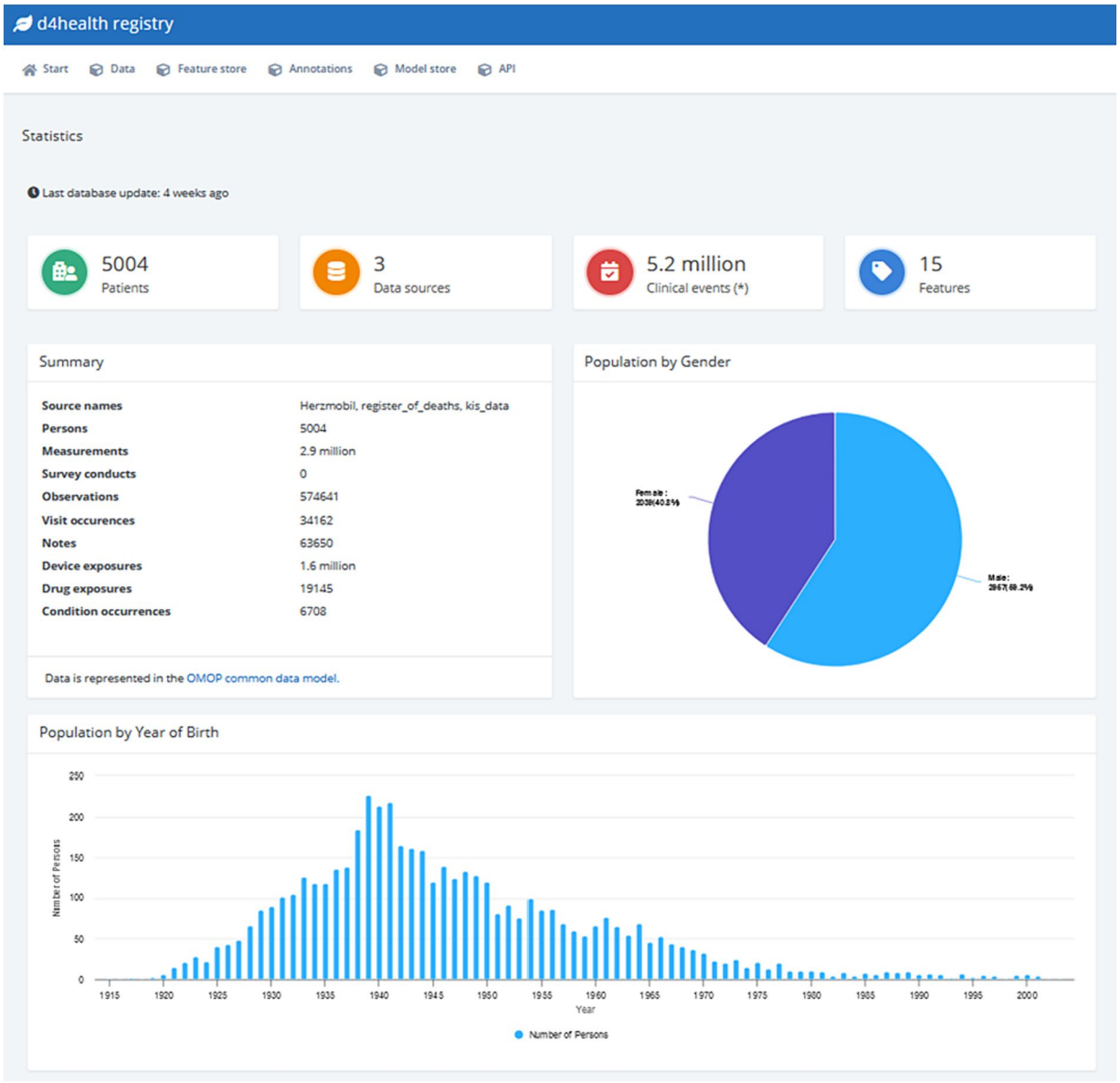

FIGURE 2
Screenshot of the default dashboard of the D4Health Heart Failure Registry.

so-called feature store. The feature store's main purposes are first, to reduce the risk of repeated developed of the same feature engineering algorithms and second, to provide future analysts with a large number of useful features already developed by other team members, that grows over time. This should facilitate collaborative and efficient data analysis. At the time of writing, the feature store supports feature development in Python. Feature engineering algorithms are documented (e.g., author, date, description), versioned and deployed within HDS nodes. Any feature generators uploaded into a node are quarantined initially and only deployed after an audit by an administrator for any malicious code.

Features can be calculated on different levels (e.g., daily level like blood pressure, patient level like height). The OMOP CDM already supports features related to the patient-level including source code for feature generation based on the *CohortDefinition* entity and its associated attributes (*AttributeDefintion*). For the feature store, we extended this functionality to support features on other levels and to support further meta data (e.g., author, timestamp, source code, technology, description). Each entity in the OMOP data model (e.g., *Person*, *Observation* or *Measurement*) has a counterpart in the feature store so that features can be calculated on *Person-level* (e.g., number of re-hospitalizations in the last 3 years), on *Observation-level* (e.g., daily medication adherence) or on *Measurement-level* (e.g., blood pressure). The feature store communicates with the data store, and is notified of all data updates, so that features are re-calculated whenever new data arrives, or existing data is updated. Features are stored in a compact JSON data structure to accommodate use cases with high numbers of variables.

Data scientists can explore available features through a web-based interface. The interface gives a superficial description and overview of

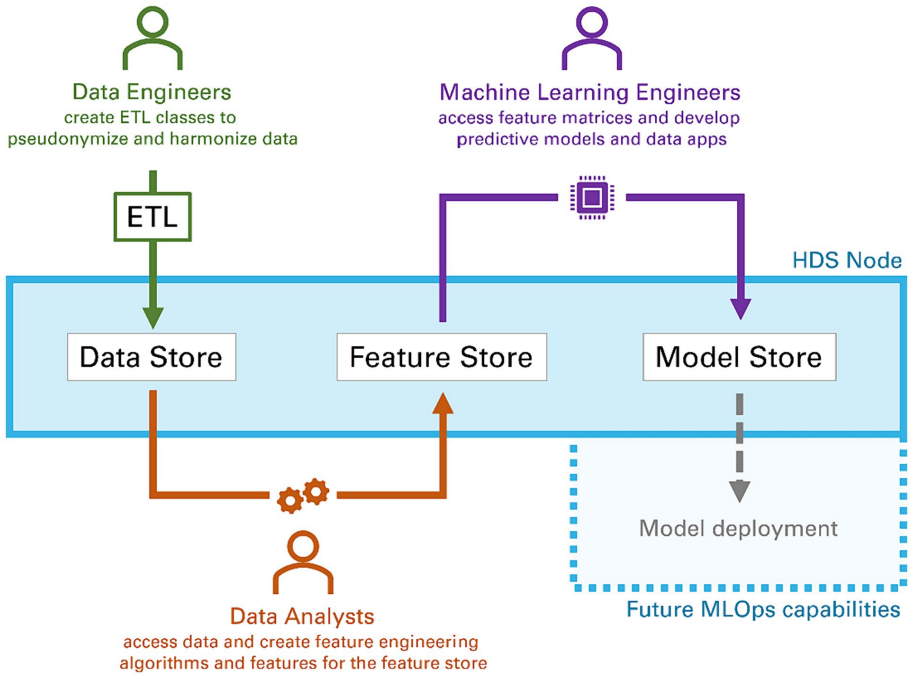

FIGURE 3
Typical workflow of different data scientists (data engineers, data analysts and machine learning engineers) collaborating within the components of a Health Data Space node to process raw data, extract features and develop models. The final step model deployment is subject to future work.

each feature (e.g., availability and distribution of values) to give analysts quick insight whether a feature might be useful for their analyses. Features can be accessed through a dedicated API similar to that of the data store. When features of a given level (e.g., *Person-level*) are accessed, all features on this level are aggregated into one feature matrix.

2.4.2 Model store

Analogous to the feature store, HDS nodes also support a model store, which is a collection of models developed outside of the HDS node's infrastructure. At the time of writing, the model store can digest any model developed with Python's scikit-learn (51) module via manual upload to the model store by the use of model serialization through the built-in pickle module. These model and their required artifacts are accessible via API. In future, this model store should serve as the basic framework for supporting MLOPs. Models and other data apps (e.g., visualization apps, dashboards) are planned to be deployed in this store to provide specific functionalities as services (e.g., prediction as a service).

2.5 HDS nodes in a network

HDS nodes are self-contained units that are linked to one data source (e.g., an EMR or a subsystem) and are pseudonymizing, harmonizing and providing data in an analysis-friendly way. Aggregating health data in one place, thus populating a node with data from multiple sources is particularly difficult if data sources are in different institutions or even countries. We have therefore designed the nodes in a way that collaborators can share artifacts and data according to defined data policies and trust in the system, thus forming a health data space enabling versatile data governance schemes. Healthcare organizations are thus enabled to meet the requirements of local data sovereignty legislation by controlling exactly what data is shared with whom. We have defined 4 layers of sharing elements depending on the level of trust between the nodes (see Table 2). Sharing of elements is done through a dedicated REST API with the HDS nodes' ETL process framework. For instance, on level 4, an HDS node might share specific raw patient data points with another HDS node. In this case, the corresponding ETL process can be activated to allow sharing as long as valid endpoint and credentials for the other HDS node are provided. While the ETL process itself is still executed locally at the source's node (transformation into an OMOP observation), its results are relayed to the other HDS node where they are stored. For levels 2, 3 and 4, it is essential that patients existing in both HDS nodes are correctly associated and linked. Therefore, both HDS nodes must agree on (a) a common set of identity traits and (b) a certain hashing strategy, including related secrets (e.g., a secret key in case of Bloom filters).

2.6 Evaluation in a real-world application

The HDS nodes and various configurations can be helpful in different use cases. We explored the feasibility of the HDS node solution in a real-world scenario in the Austrian federal state of Tyrol, connecting data sourced from three origins (one healthcare organization, one telehealth system and Austria's national register of deaths) into a registry for heart failure patients. To evaluate the architecture's readiness to deploy ML models in the future, a simple use case of a natural language processing (NLP) experiment was tested. For this, free text messages exchanged between healthcare

TABLE 2 Information sharing options depending on level of trust.

Trust level	Requirements	Sharing	Possible use cases
1: Artifacts	-	A data node can share artifacts (e.g., feature extraction algorithms or trained models) with other data nodes.	Sharing algorithms for a federated analysis task. Data itself stays in the HDS.
2: Feature information	PPRL strategy needs to be aligned	A data node may share information (e.g., which features are available) and extraction algorithms of generated features	Increasing findability of data of interest for participating network partners, which then can specifically requested or consent can be requested.
2: Features only	PPRL strategy needs to be aligned	A data node may share generated features with other data nodes	Aggregating selected data from the same patient population in a single place without revealing the raw data (e.g., a node might extract data from clinical notes and only provide extracted data without revealing the clinical notes themselves)
4: All data	PPRL strategy needs to be aligned	OMOP CDM data can be shared	Aggregating data from the same patient population in a single place.

professionals and patients from HMT were de-identified. This de-identification was based on an improved algorithm of a previously developed pseudonymization algorithm (48), which removes meta data (i.e., author, corresponding patient) and identifying references from the corpus (e.g., names, addresses) from the texts. This algorithm was evaluated on a stratified subsample of 200 messages. Subsequently, messages were annotated by human experts and an ML classification model based on Latent Dirichlet Allocation (52) was trained. The model was deployed on the network and the result was presented via a web service based on the open-source visualization library Dash (53).

3 Results

Three main results are presented in the following chapters: (1) the four levels at which data can be shared depending on the level of trust of the participating partners in an HDS node network, (2) a real-world case study of an implemented network at the highest trust level and (3) preliminary results from a MLOps feasibility study with an NLP use case.

3.1 Levels of trust in an HDS node network

To comply with different expectations and agreements of trust between participating partners, we designed HDS nodes in a way that they enable four levels of possible data sharing (summarized in Table 2):

1. Trust level 4: All data and artifacts (e.g., feature engineering algorithms, models) of a node is shared with all other nodes, including raw data from the data store and all available features. In this setting, data is typically aggregated in a central HDS data node. This use case would be helpful for scenarios, where data from the same patient population is to be aggregated in a single place for centralized machine learning.
2. Trust level 3: While trust level 4 is feasible in a setup where all nodes belong to the same data holder, in a cross-institutional network data holders might hesitate to share their transformed OMOP database. As a result, nodes can form a trust level 3 network. At this level, each node performs pseudonymization on the pre-defined elements of the patient record but keeps the data in the OMOP database locally. In contrast to a trust level 4 network, each node computes its own feature matrix (e.g., on patient-level, on daily-level) and then only shares the results along with the code used to compute the features with the heart failure registry node. This prevents sharing of any raw data from the data store. For example, any features generated from clinical messages can be exchanged for analysis without actually sharing the texts themselves.
3. Trust level 2: If the calculated features from the feature store should also not be shared, a trust level 2 network can be used. At this level, the connected HDS node only provides other nodes with the information, which features it has for a given patient, similar to the FAIR principles. To achieve this, trust level 2 connected nodes participate in PPRL, meaning consistent patient identifiers exist throughout the network. A trust level 2 network can be used to make data more findable for participating partners of the network. If specific data is found, which is required for analysis, partners can contact the corresponding data holders and patients to inquire about consent to access the data.
4. Trust level 1: At the lowest level of trust, no data is exchanged. The connected nodes only inform others that it exists and provides meta data about the contents (i.e., what kind of data is available). For this, also no PPRL across nodes is required. The only shared contents are any produced artifacts. A trust level 1 network could be used as infrastructure for federated analyses by sharing feature engineering algorithms.

3.2 HDS nodes in a data-sharing network for a heart failure registry (trust level 4)

The HDS node solution was evaluated in close partnership with tirol kliniken (Tirol Kliniken GmbH). Data from three different sites

TABLE 3 Data sources connected within the D4Health Heart Failure Registry.

Data site	Description	Type of data	No. of ETL processes
tirol kliniken hospital information system	Electronic medical record (EMR) data	Demographic data (age, gender), height, date of admission, discharge and possible readmission, laboratory values from the laboratory information system, diagnoses (ICD-10 coded), NYHA class	8
HerzMobil telehealth data	Daily physiological values measured by patients themselves using medical devices, transmitted to smartphone via Bluetooth and symptoms	Blood pressure, heart rate, bodyweight, medication information (prescription and self-reported intake adherence) and self-reported wellbeing score ("good," "medium," and "bad"), clinical notes by physicians and nurses	18
National Austrian Register of Deaths	Export of register of deaths records	Date of death	1

TABLE 4 Performance of individual ETL converter classes with at least 1,000 data points transformed.

ETL class	Data points per second
Visitation	2092.24
Device exposure	1444.89
Drug exposure	1276.43
Observation	1268.78
Measurement	1172.98
Note	1151.05
Condition occurrence	420.14
Observation period	229.87
Person	57.96

was extracted to an HDS node, respectively (see Table 3): EMR data from the tirol kliniken's hospital information system, HMT telehealth data, and an export of Austria's national register of deaths. Data transfer specifications were defined with cardiologists to select which EMR data elements are required.

For these three sites individual HDS nodes were installed, which were linked to a "D4Health Heart Failure Registry," represented by a fourth HDS node, forming a trust level 4 network (see Figure 4). While the three HDS nodes related to the sources could contain unstructured, identifying data (e.g., discharge letters), only selected, de-identified data was shared with the D4Health Heart Failure Registry HDS node according to the data transfer specifications. In this specific application, the central data node was deployed within the institutional borders of tirol kliniken.

Each node performs pseudonymization of its own identifiers (first name, last name and date of birth of patient, optional social security number where available) by computing a Bloom filter of the corresponding identifier and sharing it with the central D4Health Heart failure registry node. Here, feature ETL classes have been deployed to calculate features.

The HDS node network constituting a trust level 4 network was deployed, operating in a routine care environment and at the time of writing, is continuously linking the data from the three data sources to the registry in a privacy-preserving manner. Record linkage also consolidated duplicated patients. The PPRL found in the *HerzMobil* telehealth node 9 full matches and 19 partial matches, resulting in a duplication rate of 0.70%. The partial matches were subsequently assessed by human experts and found to be all false positives. The hospital information system as well as the Austrian register of deaths nodes had no duplicates since they already used a unique identifier in their respective systems.

At the time of writing, the D4Health Heart Failure Registry HDS node contains data from 5,004 patients, over 2.9 million measurements, over 570,000 observations and more than 63,000 clinical free text notes. In total, over 5.2 million clinical events (i.e., individual data points) are accessible. Figure 2 shows a screenshot of the default dashboard of the D4Health Heart Failure Registry HDS node, which displays basic descriptive statistics to provide an overview of the included data, which can be adapted, according to specific use cases and preferences. To assess performance and scalability, the execution time of individual ETL converters has been recorded. The ETL classes that have transformed most frequent data types were *measurement* (1,173 data points/s), *observation* (1,269 data points/s), *device exposure* (1,445 data points/s), *observation period* (230 data points/s) and *note* (1,151 data points/s). A full list of performance of ETL classes with at least 1,000 data points is presented in Table 4. With increasing amounts of patients, registration slows down significantly as the PPRL framework requires increasingly more comparisons since new patients have to be compared to all registered patients. In our experiments, the application of MinHash (47), increased the speed of registration from 2 per second to 40 per second.

As some of the patients also had coronary heart disease (CHD), another node was established to collect quality of life information from them via FHIR Questionnaires. 60 CHD patients were included in a preliminary node, which is not connected to the registry at the time of writing. Patients completed the MacNew quality of life questionnaire at the start of the telemonitoring phase and once again at the end of the phase to track improvements in the quality of life during the program. These FHIR Questionnaire responses are mapped

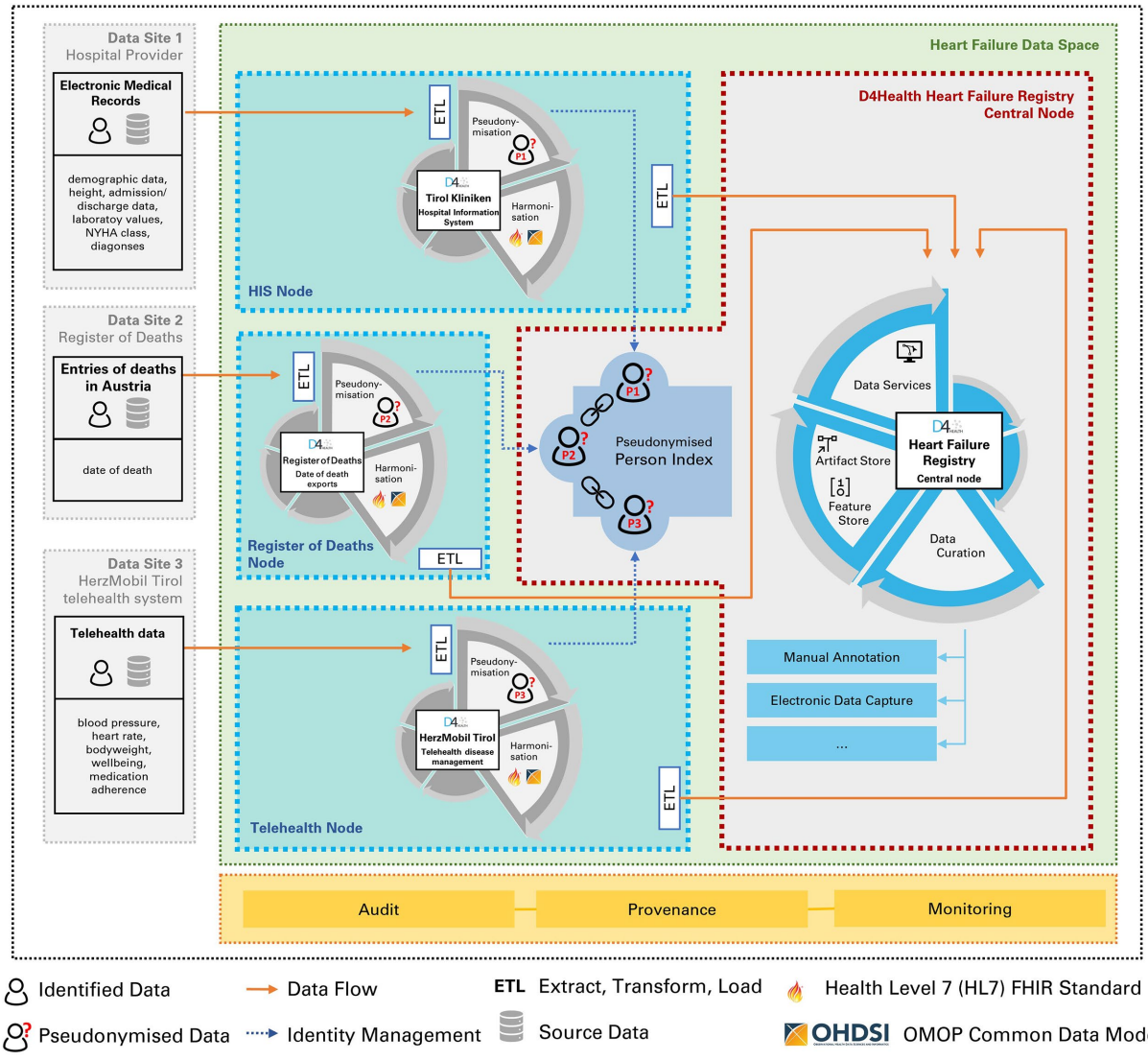

FIGURE 4
Three Health Data Space nodes (tirol kliniken, HerzMobil Tirol, Register of Deaths) are linked to a fourth, central node, in which the registry is located. Identity management and record linkage is done via the pseudonymized person index.

into the OMOP CDM and are planned to be linked into the D4Health Heart Failure Registry in the future.

3.3 Feasibility experiment of deployment of a natural language processing model

To evaluate the capability of our approach to deploy ML models, a basic NLP use case was successfully executed. The pseudonymization algorithm achieved high performance (accuracy: 93.99%, sensitivity: 0.94, specificity: 0.93). Subsequently, the messages were labeled by 9 expert observers using the HDS node's annotation tool (50). Finally, the labeled data was extracted via API to a Python development environment, in which the LDA model was trained. The artifacts produced by the model were successfully deployed within the infrastructure and were reachable via API queries from outside with corresponding permissions. A specific visualization tool could successfully be deployed for exploring and quality-controlling the model (see Figure 5).

4 Discussion

We presented the Health Data Space nodes as flexible system architecture units, which we evaluated in a real-world application called the D4Health Heart Failure Registry. The results obtained from this case study confirm the infrastructure's utility. The processes of linking, harmonizing and analyzing data have proven to be functional. Feature engineering and modeling have been explored experimentally and have shown promising, preliminary results in a proof-of-concept natural language processing use case. Extending the functionality of MLHOps (especially model deployment) to industry-level readiness is subject of future research and development.

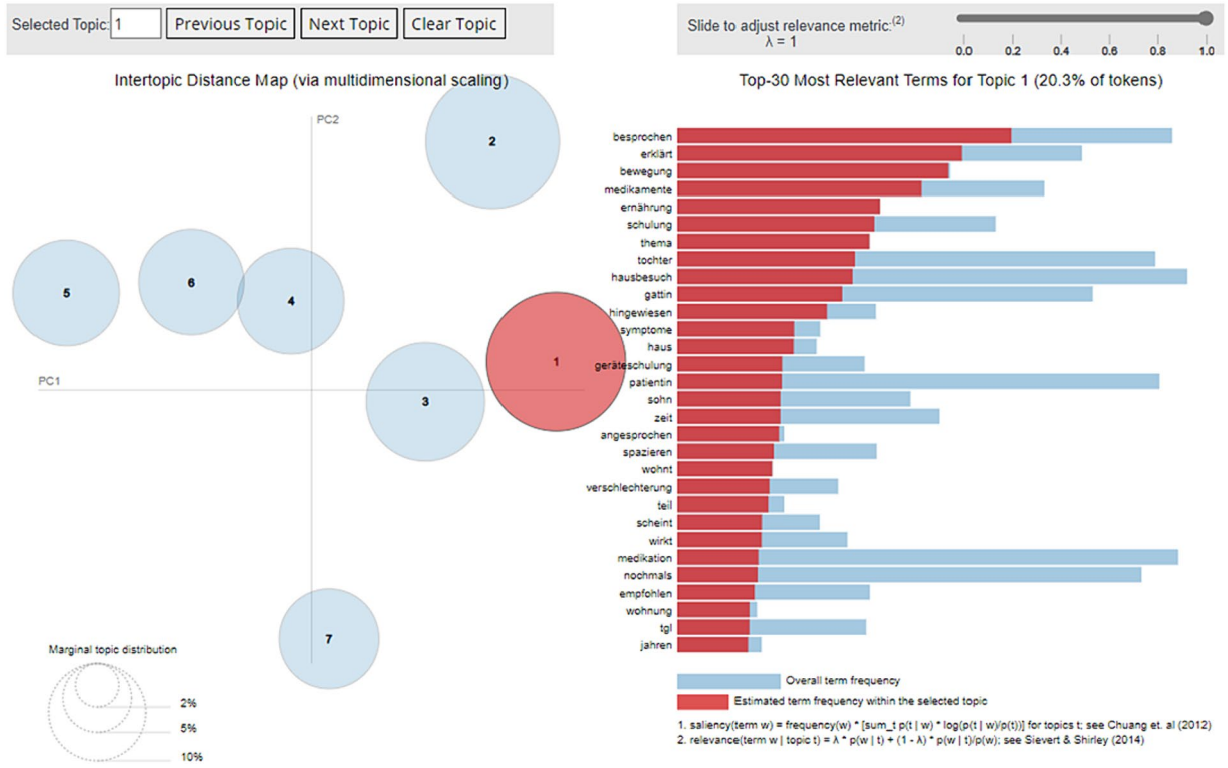

FIGURE 5
Screenshot of the results of the Latent Dirichlet Allocation model developed from all free-text clinical notes of the telehealth service and deployed to the infrastructure's central. Circles on the left illustrate the 7 identified topics in an inter-topic distance map via multidimensional scaling. Bars on the right show the top-30 most relevant terms (in German) for the selected topic 1 "council/training" and their frequency within the related topic (red) and overall, within the corpus (blue).

Although other approaches that address parts of our requirements exist, we hope to contribute new approaches to the complex challenges of sharing and linking medical data with a strong focus on privacy-preservation. *DataSHIELD* enables federated analyses but is not intended to aggregate data into a common feature matrix for centralized machine learning. The *PHT* does that but concerns of containerized code execution with Docker containers make it ultimately nonviable for our application. While other privacy-preserving frameworks were applied to medical data (e.g., *KETOS* platform) and have used Bloom filters [e.g., (54)], linkage assisted with Bloom filters across multiple sources of medical data has not been demonstrated yet. As an additional layer of privacy, we proposed node-specific pseudonyms to avoid using the same pseudonym in multiple contexts, which risk exposing patients by linkage attacks. Privacy-preservation was further focused on by including automated free text de-identification as part of the framework. This is noteworthy, as the rise of large language models (e.g., ChatGPT) has renewed interest in medical free text recently. The application of MinHash (47) with the Bloom filters ensured scalability of the PPRL strategy. To assist in organizational coordination of privacy-oriented data sharing, we introduced four levels of trust within a data sharing network (see Table 2) to provide guidelines for real-world applications. Another novel contribution of our presented architecture is the implementation of a multi-level feature store with the increasingly popular OMOP CDM, which also has not been described in literature. Although, the OMOP CDM supported features on a patient-level with the tables *AttributeDefintion* and *CohortDefinition*, we extended this capability to also represent features that change on a daily basis (e.g., blood pressure).

To summarize our efforts, we combined established techniques (e.g., PPRL, ETL frameworks) with novel ideas (e.g., multi-level OMOP feature store, trust levels, context-specific pseudonyms) to create a starting point for the development of a "full suite" for collaborative analyses of medical data that assists in the entire data science process from start to finish. The HDS nodes have tools for data collection (e.g., FHIR Questionnaires), data cleaning (e.g., de-identification, data annotation), data exploration (e.g., dashboards) and feature engineering (e.g., feature store) and we are aiming to complete the process by implementing sufficient model deployment (e.g., model store) in the future.

To demonstrate the real-world feasibility of this architecture, an ensemble of HDS nodes was applied in a data sharing network for a real-world heart failure registry. Establishing such an infrastructure requires close collaboration between multiple partners, whose interests must be balanced. This concerns not only data governance considerations, but also varying requirements of (a) functionality, (b) processing tools and (c) jurisdiction.

a. To address different functionality requirements, HDS nodes are designed in a modular, flexible and scalable way. This not only refers to including data sources currently, but also to apps and

services like predictive models and visualization in the future. Linkage to other, similar infrastructures and data sharing with other HDS nodes is supported to different degrees depending on the level of cooperation and trust within a network.

b. To enable data scientists to work within their own familiar environments, development of analyses tools is decoupled from the infrastructure. Relying on overly generalized tools can be problematic and enabling data scientists to work with their domain-specific tools is preferable.

c. Individual identity holders are able to fully create and control their credentials. Each jurisdiction operating an HDS node is able control the inputs, processing steps and outputs of the node. Data sovereignty is also part the EU's European Strategy for data (55).

Apart from organizational challenges to coordinate stakeholder interests, we also addressed interoperability on four levels: (1) Syntactic interoperability: ETL processes automatically import and transform source data into the registry. Export functions for JSON, CSV and Microsoft Excel are provided for external use. (2) Semantic interoperability: Data is harmonized using the OMOP CDM. Standard vocabularies are used for further interoperability (SNOMED, LOINC, ICD-10, ATC). (3) Pragmatic interoperability: Linking data also means linking institutions, partners, and pre-existing networks. Data sharing was realized with specifically designed data sharing policies for transparent collaboration process, over which the source data's managers still have control. (4) Legal interoperability: To comply with legal frameworks like GDPR and ethical considerations, the architecture is based on a pseudonymization and privacy-preserving record linkage infrastructure. HDS nodes can be connected on different trust levels (see Table 2).

Mapping different data structures and models into the OMOP CDM and encoding into the SNOMED vocabulary proved to be a major challenge. For example, the telehealth system included information about prescribed medication usually in brand names as available in Austria. However, SNOMED as an international vocabulary did not necessarily provide these exact names and thus a correct mapping was not always possible. As a workaround, medication was encoded according to their active ingredients (i.e., the chemical compounds). Furthermore, for physiological values from the telehealth system (e.g., blood pressure) multiple SNOMED concepts were available. For example, SNOMED provides multiple blood pressure concepts depending on the body position during measurement (e.g., lying, sitting, standing). However, in the telehealth setting, patients measure data without supervision and thus this information is not available. As a compromise, generic concepts were selected at the cost of minor imprecision. Also, telehealth visitations (e.g., by nurses) were simply not available in the OMOP CDM and thus were difficult to represented within this specific CDM.

Limitations

The infrastructure is subject to limitations that need to be discussed. Firstly, at the time of writing, the infrastructure's focus is on observational health data. Other data modalities like time-series, images or genomic data are currently out of scope. Meta data about the D4Health Heart Failure Registry are not made publicly available so far, e.g., via a FAIR Data Point (FDP) as suggested by the FAIR principles (56). Provision of the metadata in an FDP would further improve the visibility and re-usability of the data in the future and enable collaboration with other frameworks (e.g., PHT).

In the presented case study, only one of the sites was a healthcare organization, limiting the scope of the currently demonstrated capabilities. Further, both the EMR data, which is directly from tirol kliniken's HIS, and the data from the HMT telehealth system, which is operated by a subsidiary of tirol kliniken (the *Tyrolean Federal Institute for Integrated Care*) are domain of tirol kliniken. The central node was operated in tirol kliniken's institutional infrastructure to avoid raising concerns over data sovereignty. Linking multiple healthcare organizations complicates the task considerably and increases the necessary technical, organizational and legal effort since data is leaving institutional borders. While the presented HDS nodes are designed to also realize such complex settings from a technical point of view, a real-world implementation remains to be demonstrated and is subject of future studies.

To address the issue of data governance and sovereignty, we have segmented access into four levels according to the trust between sharing partners. As requests by data holders can be extremely specific and legislative framework highly intricate, this simplification might not be appropriate for all use cases. A more granular permission and sharing framework would be required to address this fully.

Further, although access to the HDS nodes is possible via APIs from various data science tools, such as Python, R or MATLAB, feature and model deployment is currently only supported for Python. In specific settings, we have already explored model deployment via the Predictive Modeling Markup Language (PMML) between Python and MATLAB, however, this is not yet deployed in the productive HDS node infrastructure. Furthermore, at the time of writing, the model store only supports models developed with scikit-learn (51).

Lastly, although the free-text de-identification performed satisfactorily well (see chapter 3.3) for clinical messages to protect privacy, it is fine-tuned for this application with specific name dictionaries and regular expressions following local rules (e.g., Austrian phone numbers, Austrian postal codes) and therefore will not translate well into other applications.

4.1 Outlook

The NLP proof-of-concept use case served as first steps of implementing satisfactory MLHOps support in the HDS nodes. Implementing support for additional commonly used ML and industry-leading frameworks (e.g., TensorFlow/Keras, PyTorch) is subject of future development. Once reliable functions for model deployed are implemented, various other use cases present themselves. Two major groups of data services could be useful, which could be developed outside an HDS node (e.g., a local computer) and uploaded to a node:

1 Model interfaces to provide predictions as a service to healthcare professionals and data scientists. Examples include predicting of major cardiac events, risk stratification of the patient population and outcome prognoses. Another

interesting, yet highly specific use case for HMT, would be predicting, which patient would benefit from extending the standard 3 months telehealth disease management program to allocate resources more efficiently. However, further research is necessary to explore the potential of data-driven applications used in the treatment of heart failure patients.

2 Interactive data exploration apps like visualizing dashboards. We provided a basic example with the LDA model implemented with the open-source library Dash (53). Other examples include visual representation of medication adherence or measurement deviations.

Additionally, updates of the HDS nodes based on recent health data can currently either be triggered manually or based on routines in regular intervals. Therefore, any predictions for individual patients would currently face a certain time delay, until all data needed is present in the respective HDS node. Functionalities that trigger data transfers upon updates in the source's database could be explored further in future development, which would enable real-time predictions.

In the future, we will be investigating the expansion of HDS nodes to support privacy-preserving AI with multiple nodes, focusing on federated analysis, secure multiparty computation, exchange of synthetic data and other promising approaches in addition to PPRL. Federated learning is very appealing in medicine and HDS nodes are especially well-suited for it since they provide uniform distributable nodes with standardized data. Developing models locally, without even centralizing data, has the potential to further increase privacy, security, and trust in the system. An additional advantage might be that it serves as incentive for potential partners to join the network and gain access to well-performing models. Furthermore, partners that only contribute small amounts of data could benefit from the knowledge extractable from larger datasets.

We identify considerable potential for the D4Health Heart Failure Registry specifically in adding additional data sources. Further we aim to test the HDS nodes in an actual cross-institutional data sharing setting in future research. This includes first and foremost other *HerzMobil* systems (e.g., in Styria and Carinthia) for horizontal linkage. Furthermore, vertical linkage by including cardiac implantable electronic devices is especially attractive since they are highly relevant for heart failure patients. Besides medical data, health economics information could provide insight into patients' history of procedures and thus to help assessing cost-effectiveness of interventions. To align with the paradigm of patient empowerment and self-governance of medical data, enabling patients to voluntarily include their own data certainly holds potential. Large quantities of health-relevant data are collected with wearable sensors and consumer devices routinely now by many people including physical activity, number of steps, sleep quality and even physiological data like oxygen saturation or single-lead electrocardiograms that can be recorded by smart watches.

Secondary use of health data might be regulated differently in individual countries or governance regions further complicating the issue of data sovereignty. Especially the transatlantic relationship has been strained by the overturning of both the *International Safe Harbor Privacy Principles* in 2015 (57) and the *EU-US Privacy Shield* in 2020 (58) agreements due to concerns of the Court of Justice of the European Union. However, the European Commission has recognized the potential of secondary use and aims to facilitate a common data space inside the European Union. The Commission has published several documents as part of its Data Strategy to work toward a European Health Data Space (EHDS). These concerted efforts are aiming for better utilization of data in both primary and secondary use and more convenience for patients in accessing health services abroad (59). The present work was inspired by this initiative and is intended to contribute to the evolution of the EHDS. Currently, data exchange and linkage policies can already be adapted to support various levels of record linkage across different jurisdictions. With this flexibility, HDS nodes could be linked to the EHDS and service interfaces to existing data space connector solutions such as the Eclipse Dataspace Connector (60) or the International Data Spaces Connector (61), as illustrated in Figure 6. Future work should also consider further development of the HDS node to adhere to specifications coming from initiatives like the EHDS and Gaia-X (62) and also keep different legislative frameworks in mind. Collaboration with similar frameworks like the *Personal Health Train* (24) could also prove fruitful for increasing data availability in the future. Furthermore, the capabilities of Blockchain technology to ensure data immutability could also be topic of future search as it would further increase trust in the system.

Architecture sustainability is always a concern in research projects like this because wide adoption of digital health solutions into regular healthcare settings is notoriously slow. Furthermore, the project-based funding and non-commercial setting of such systems make them inherently at risk of being not fully supported long-term. Many definitions for sustainability in the context of software exist (63). According to Venters et al. (64), sustainability describes a system's *extensibility*, *interoperability*, *maintainability*, *portability*, *reusability*, *scalability*, and *usability*. We outlined that our infrastructure is portable and reusable by relying on common

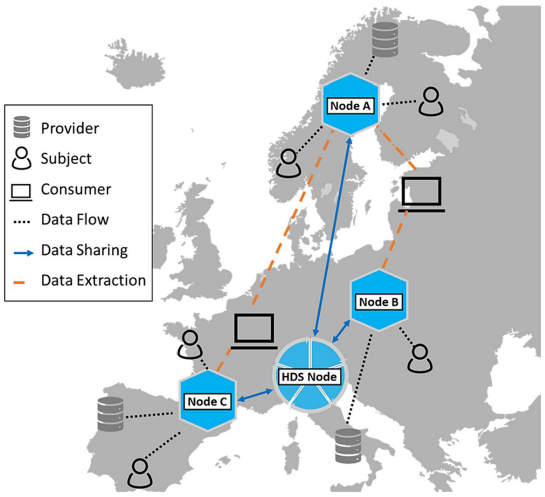

FIGURE 6
An HDS node can act as a national node in the context of a European data market and research infrastructure, such as currently being developed within the European Health Data Space.

and platform-agnostic frameworks (e.g., Python). Further, we also described how we ensured interoperability by utilizing standard vocabulary (e.g., ICD-10, SNOMED, LOINC) and a suitable and commonly used common data model (OMOP CDM). With our PPRL methods, we also focused on the scalability and extensibility of the system by enabling vertical and horizontal linkage across different data sources. The greatest limitation toward scalability and extensibility remaining is the organizational coordination and data sovereignty concerns. To address this, future work could also focus on education and informing stakeholders about the benefits of such technology. Reference projects like the hereby described platform could aid this process. We addressed maintainability by aiming to minimize dependencies on third-party modules and relying on well-maintained open-source modules whenever possible. Since maintainability of our own core components is still a concern, we are also exploring options to potentially open-source parts of our code as well. This would open our developments to interested communities and improve maintainability by possibly increasing the amount of people interested in and working on the software. Usability is currently the least addressed aspect of sustainability in the HDS nodes. Although basic feedback from users (e.g., healthcare professionals, data scientists) has been implemented on occasion, systematic usability tests with stakeholders remain subject of future research. We recognize usability as a core requirement to aid the transition of stakeholders toward digital health solutions and have therefore included thorough usability testing in our development roadmap.

6 Conclusion

We have developed Health Data Space nodes to facilitate the secondary use of health data, which also support privacy-preserving record linkage across data sources to increase data availability. The HDS nodes provide sufficient flexibility to set up application specific infrastructures. With this concept, we realized and presented a pilot case study, including not only development but also deployment of a smart health ecosystem in a real-world infrastructure to establish the D4Health Heart Failure Registry for a routine care setting in Tyrol. With this infrastructure, data can be linked in a privacy-preserving way and be harmonized for interoperability. Preliminary functionality for collaborative feature engineering and model deployment have been tested in simple use cases. In conclusion, we consider these results as the foundation for future developments. Due to the modular architecture, the application of HDS nodes is not restricted to heart failure, but can be applied in various other scenarios.

We believe that such smart health ecosystems which support data management and MLOps and connect data from different health data spaces are the key to successful, efficient and sustainable secondary use of health data. Adhering to privacy standards is not only necessary from a with legal compliance perspective but also helps to improve overall acceptance and is, therefore, considered a must. With the presented case study, we hope to prove the feasibility of such systems and hope to inspire similar pioneering solutions for the upcoming work of building the European Health Data Space.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: The clinical notes used for the NLP use case are considered medical data and therefore are not to be made public. Requests to access these datasets should be directed to martin.baumgartner@ait.ac.at.

Ethics statement

The studies involving humans were approved by Ethics Commission of the Medical University of Innsbruck. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

MB: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. KK: Conceptualization, Data curation, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – review & editing. AL: Conceptualization, Data curation, Investigation, Methodology, Software, Validation, Writing – review & editing. BJ: Conceptualization, Data curation, Investigation, Methodology, Software, Validation, Writing – review & editing. KD: Conceptualization, Investigation, Methodology, Resources, Supervision, Validation, Writing – review & editing. DH: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing. FW: Investigation, Methodology, Software, Writing – review & editing. LD: Investigation, Methodology, Software, Writing – review & editing. RM-O: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing. SN: Conceptualization, Funding acquisition, Project administration, Resources, Validation, Writing – review & editing. GSl: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing. SP: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing. LB: Conceptualization, Investigation, Methodology, Supervision, Validation, Writing – review & editing. BP: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing. GP: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing. GSc: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the D4Health Tirol project which was funded by the state government of the Land Tirol.

Conflict of interest

RM-O and GSc hold shares of telbiomed Medizintechnik und IT Service GmbH, where RM-O is also employed as CEO. LD was employed by Know-Center GmbH. SN, GSl, SP, and BP were employed by Tirol Kliniken GmbH.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-world evidence—what is it and what can it tell us? *N Engl J Med*. (2016) 375:2293–7. doi: 10.1056/NEJMsb1609216

2. Beresniak A, Schmidt A, Proeve J, Bolanos E, Patel N, Ammour N, et al. Cost-benefit assessment of using electronic health records data for clinical research versus current practices: contribution of the electronic health Records for Clinical Research (EHR4CR) European project. *Contemp Clin Trials*. (2016) 46:85–91. doi: 10.1016/j.cct.2015.11.011

3. Bruland P, McGilchrist M, Zapletal E, Acosta D, Proeve J, Askin S, et al. Common data elements for secondary use of electronic health record data for clinical trial execution and serious adverse event reporting. *BMC Med Res Methodol*. (2016) 16:159. doi: 10.1186/s12874-016-0259-3

4. Myers L, Stevens J. Using EHR to conduct outcome and health services research In: *Secondary Analysis of Electronic Health Records*. ed. Celi L., Charlton P., Ghassemi M. M., Johnson A., Komorowski M., Marshall D., Cham (CH): Springer (2016). 61–70.

5. Poelzl G, Egelseer-Bruendl T, Pfeifer B, Modre-Osprian R, Welte S, Fetz B, et al. Feasibility and effectiveness of a multidimensional post-discharge disease management programme for heart failure patients in clinical practice: the HerzMobil Tirol programme. *Clin Res Cardiol*. (2022) 111:294–307. doi: 10.1007/s00392-021-01912-0

6. Hernandez-Boussard T, Monda KL, Crespo BC, Riskin D. Real world evidence in cardiovascular medicine: ensuring data validity in electronic health record-based studies. *J Am Med Inform Assoc*. (2019) 26:1189–94. doi: 10.1093/jamia/ocz119

7. Kannan V, Fish JS, Mutz JM, Carrington AR, Lai K, Davis LS, et al. Rapid development of specialty population registries and quality measures from electronic health record data*. An Agile Framework. *Methods Inf Med*. (2017) 56:e74–83. doi: 10.3414/ME16-02-0031

8. Mirchev M, Mircheva I, Kerekovska A. The academic viewpoint on patient data ownership in the context of big data: scoping review. *J Med Internet Res*. (2020) 22:e22214. doi: 10.2196/22214

9. Miguel Cruz A, Marshall S, Daum C, Perez H, Hirdes J, Liu L. Data silos undermine efforts to characterize, predict, and mitigate dementia-related missing person incidents. *Healthc Manag forum*. (2022) 35:333–8. doi: 10.1177/08404704221106156

10. Alves J, Meneses R. Silos mentality in healthcare services. In: *11th Annual Conference of the EuroMed Academy of Business*. (2018)

11. Ammenwerth E, Modre-Osprian R, Fetz B, Gstrein S, Krestan S, Dörler J, et al. HerzMobil, an integrated and collaborative Telemonitoring-based disease management program for patients with heart failure: a feasibility study paving the way to routine care. *JMIR Cardio*. (2018) 2:e11. doi: 10.2196/cardio.9936

12. Wu H, LaRue EM. Linking the health data system in the U.S.: challenges to the benefits. *Int J Nurs Sci*. (2017) 4:410–7. doi: 10.1016/j.ijnss.2017.09.006

13. Langner I, Riedel O, Czwikla J, Heinze F, Rothgang H, Zeeb H, et al. Linkage of routine data to other data sources in Germany: a practical example illustrating challenges and solutions. *Das Gesundheitswes*. (2020) 82:S117–21. doi: 10.1055/a-0999-5509

14. Tayefi M, Ngo P, Chomutare T, Dalianis H, Salvi E, Budrionis A, et al. Challenges and opportunities beyond structured data in analysis of electronic health records. *WIREs Comput Stat*. (2021) 13:e1549. doi: 10.1002/wics.1549

15. Rizi SAM, Roudsari A. Development of a public health reporting data warehouse: lessons learned. *Stud Health Technol Inform*. (2013) 192:861–5. doi: 10.3233/978-1-61499-289-9-861

16. Regulation (EU) 2016/679 of the European Parliament and of The Council. (2016) Available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj

17. Rädsch T, Reinke A, Weru V, Tizabi MD, Schreck N, Kavur AE, et al. Labelling instructions matter in biomedical image analysis. *Nat Mach Intell*. (2023) 5:273–83. doi: 10.1038/s42256-023-00625-5

18. Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, et al. Hidden technical debt in machine learning systems. *Adv Neural Inf Proces Syst*. (2015) 28:2503–2511

19. Patel J. The democratization of machine learning features. In: *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*. (2020). p. 136–141.

20. Kim M. Software engineering for data analytics. *IEEE Softw*. (2020) 37:36–42. doi: 10.1109/MS.2020.2985775

21. Khattak FK, Subasri V, Krishnan A, Dolatabadi E, Pandya D, Seyyed-Kalantari L. MLHOps: machine learning for healthcare operations. arXiv [Preprint]. arXiv230502474 (2023)

22. Gagalova KK, Leon Elizalde MA, Portales-Casamar E, Görges M. What you need to know before implementing a clinical research data warehouse: comparative review of integrated data repositories in health care institutions. *JMIR Form Res*. (2020) 4:e17687. doi: 10.2196/17687

23. Gaye A, Marcon Y, Isaeva J, LaFlamme P, Turner A, Jones EM, et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. *Int J Epidemiol*. (2014) 43:1929–44. doi: 10.1093/ije/dyu188

24. Beyan O, Choudhury A, Van Soest J, Kohlbacher O, Zimmermann L, Stenzhorn H, et al. Distributed analytics on sensitive medical data: the personal health train. *Data Intell*. (2020) 2:96–107. doi: 10.1162/dint_a_00032

25. Marwan M, Kartit A, Ouahmane H. Applying secure multi-party computation to improve collaboration in healthcare cloud. In: *2016 Third International Conference on Systems of Collaboration (SysCo)*. (2016). p. 1–6.

26. Alghamdi W, Salama R, Sirija M, Abbas AR, Dilnoza K. Secure multi-party computation for collaborative data analysis. In: *E3S Web of Conferences*. EDP Sciences (2023). p. 4034

27. Tso R, Alelaiwi A, Mizanur Rahman SM, Wu M-E, Hossain MS. Privacy-preserving data communication through secure multi-party computation in healthcare sensor cloud. *J Signal Process Syst*. (2017) 89:51–9. doi: 10.1007/s11265-016-1198-2

28. Fan K, Wang S, Ren Y, Li H, Yang Y. MedBlock: efficient and secure medical data sharing via Blockchain. *J Med Syst*. (2018) 42:136. doi: 10.1007/s10916-018-0993-7

29. Xia Q, Sifah EB, Asamoah KO, Gao J, Du X, Guizani M. MeDShare: trust-less medical data sharing among cloud service providers via Blockchain. *IEEE Access*. (2017) 5:14757–67. doi: 10.1109/ACCESS.2017.2730843

30. Azaria A, Ekblaw A, Vieira T, Lippman A. MedRec: using Blockchain for medical data access and permission management. In: *2016 2nd International Conference on Open and Big Data (OBD)*. (2016). p. 25–30.

31. Cerchione R, Centobelli P, Riccio E, Abbate S, Oropallo E. Blockchain's coming to hospital to digitalize healthcare services: designing a distributed electronic health record ecosystem. *Technovation*. (2023) 120:102480. doi: 10.1016/j.technovation.2022.102480

32. Abbate S, Centobelli P, Cerchione R, Oropallo E, Riccio E. Blockchain Technology for Embracing Healthcare 4.0. *IEEE Trans Eng Manag*. (2023) 70:2998–3009. doi: 10.1109/TEM.2022.3212007

33. Baumgartner M, Veeranki SPK, Hayn D, Schreier G. Introduction and comparison of novel decentral learning schemes with multiple data pools for privacy-preserving ECG classification. *J Healthc Informatics Res*. (2023) 7:291–312. doi: 10.1007/s41666-023-00142-5

34. Haggenmüller S, Schmitt M, Krieghoff-Henning E, Hekler A, Maron RC, Wies C, et al. Federated learning for decentralized artificial intelligence in melanoma diagnostics. *JAMA Dermatol*. (2024) 160:303. doi: 10.1001/jamadermatol.2023.5550

35. Tedeschini BC, Savazzi S, Stoklasa R, Barbieri L, Stathopoulos I, Nicoli M, et al. Decentralized federated learning for healthcare networks: a case study on tumor segmentation. *IEEE Access*. (2022) 10:8693–708. doi: 10.1109/ACCESS.2022.3141913

36. Wirth FN, Meurers T, Johns M, Prasser F. Privacy-preserving data sharing infrastructures for medical research: systematization and comparison. *BMC Med Inform Decis Mak*. (2021) 21:242. doi: 10.1186/s12911-021-01602-x

37. Jin H, Luo Y, Li P, Mathew J. A review of secure and privacy-preserving medical data sharing. *IEEE Access*. (2019) 7:61656–69. doi: 10.1109/ACCESS.2019.2916503

38. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc*. (2010) 17:124–30. doi: 10.1136/jamia.2009.000893

39. Szalma S, Koka V, Khasanova T, Perakslis ED. Effective knowledge management in translational medicine. *J Transl Med*. (2010) 8:68. doi: 10.1186/1479-5876-8-68

40. i2b2 tranSMART Foundation. Available at: https://i2b2transmart.org/ (n.d.). (Accessed 12 January 2023).

41. Doel T, Shakir DI, Pratt R, Aertsen M, Moggridge J, Bellon E, et al. GIFT-cloud: a data sharing and collaboration platform for medical imaging research. *Comput Methods Prog Biomed*. (2017) 139:181–90. doi: 10.1016/j.cmpb.2016.11.004

42. Tudini E, Andrews J, Lawrence DM, King-Smith SL, Baker N, Baxter L, et al. Shariant platform: enabling evidence sharing across Australian clinical genetic-testing laboratories to support variant interpretation. *Am J Hum Genet*. (2022) 109:1960–73. doi: 10.1016/j.ajhg.2022.10.006

43. Damgov I, Bartosova M, Marinovic I, Istanbuly O, Kieser M, Lambie M, et al. IMPROVE-PD finder: a web-based platform to search and share peritoneal Dialysis biobank, registry and clinical trial metadata. *Kidney Int Rep*. (2023) 8:912–5. doi: 10.1016/j.ekir.2023.01.003

44. Gruendner J, Schwachhofer T, Sippl P, Wolf N, Erpenbeck M, Gulden C, et al. KETOS: clinical decision support and machine learning as a service–a training and deployment platform based on Docker, OMOP-CDM, and FHIR web services. *PLoS One*. (2019) 14:e0223010. doi: 10.1371/journal.pone.0223010

45. Nitzlnader M, Schreier G. Patient identity management for secondary use of biomedical research data in a distributed computing environment. *Stud Health Technol Inform*. (2014) 198:211–8. doi: 10.3233/978-1-61499-397-1-211

46. Schnell R, Bachteler T, Reiher J. A novel error-tolerant anonymous linking code. *SSRN Electr J*. (2011):3549247. doi: 10.2139/ssrn.3549247

47. Broder AZ. Identifying and filtering near-duplicate documents. In: *Annual symposium on combinatorial pattern matching*. Springer (2000). p. 1–10.

48. Baumgartner M, Schreier G, Hayn D, Kreiner K, Haider L, Wiesmüller F, et al. Impact analysis of De-identification in clinical notes classification. *Stud Health Technol Inform*. (2022) 293:189–96. doi: 10.3233/SHTI220368

49. Dempster M, Donnelly M, O'Loughlin C. The validity of the MacNew quality of life in heart disease questionnaire. *Health Qual Life Outcomes*. (2004) 2:6. doi: 10.1186/1477-7525-2-6

50. Kreiner K, Hayn D, Schreier G. Twister: A Tool for Reducing Screening Time in Systematic Literature Reviews. *Stud Health Technol Inform*. (2018) 255:5–9. doi: 10.3233/978-1-61499-921-8-5

51. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. (2011) 12:2825–30. doi: 10.48550/arXiv.1201.0490

52. Hoffman M, Bach F, Blei D. Online learning for latent dirichlet allocation. *Adv Neural Inf Proces Syst*. (2010) 23:856–864

53. Hossain S, Calloway C, Lippa D, Niederhut D, Shupe D. Visualization of bioinformatics data with dash bio. In: *Proceedings of the 18th Python in Science Conference*. SciPy, Austin, Texas, pp. 126–133 (2019).

54. Randall SM, Ferrante AM, Boyd JH, Bauer JK, Semmens JB. Privacy-preserving record linkage on large real world datasets. *J Biomed Inform*. (2014) 50:205–12. doi: 10.1016/j.jbi.2013.12.003

55. European Commission. A European Strategy for data. (n.d.). Available at: https://digital-strategy.ec.europa.eu/en/policies/strategy-data

56. Wilkinson MD, Dumontier M, IjJ A, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data*. (2016) 3:1–9. doi: 10.1038/sdata.2016.18

57. Court of Justice of the European Union. The court of justice declares that the Commision's US safe harbour decision is invalid. *Maximillian Schrems v Data Prot Comm* (2015) Judgment in Case C-362/14. Available at: http://curia.europa.eu/jcms/upload/docs/application/pdf/2015-10/cp150117en.pdf

58. Court of Justice of the European Union. The court of justice invalidates decision 2016/1250 on the adequacy of the protection provided by the EU-US data protection shield. *Data Prot Comm v Faceb Irel Maximillian Schrems* (2020) 46:n.p. Available at: https://curia.europa.eu/jcms/upload/docs/application/pdf/2020-07/cp200091en.pdf

59. European Commission. European health data space. *eHealth Digit Heal care* (2023) Available at: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en (Accessed 22 May 2023).

60. Eclipse Foundation. Eclipse Dataspace Components. (n.d.). Available at: https://projects.eclipse.org/projects/technology.edc (Accessed 24 August 2023).

61. Nast M, Rother B, Golatowski F, Timmermann D, Leveling J, Olms C. Work-in-Progress: towards an international data spaces connector for the internet of things. In: *2020 16th IEEE International Conference on Factory Communication Systems (WFCS)*. (2020). p. 1–4.

62. Braud A, Fromentoux G, Radier B, Le GO. The road to European digital sovereignty with Gaia-X and IDSA. *IEEE Netw*. (2021) 35:4–5. doi: 10.1109/MNET.2021.9387709

63. Venters CC, Jay C, Lau LMS, Griffiths MK, Holmes V, Ward RR, et al. Software sustainability: the modern tower of babel. In: *CEUR Workshop Proceedings* CEUR (2014). p. 7–12

64. Venters C, Lau LMS, Griffiths M, Holmes V, Ward R, Jay C, et al. The blind men and the elephant: towards an empirical evaluation framework for software sustainability. *J Open Res Softw*. (2014) 2:1–6. doi: 10.5334/jors.ao

frontiers | Frontiers in Medicine

TYPE Policy and Practice Reviews
PUBLISHED 09 May 2024
DOI 10.3389/fmed.2024.1379852

OPEN ACCESS

EDITED BY
Oya Beyan,
University Hospital of Cologne, Germany

REVIEWED BY
Christos Ilioudis,
International Hellenic University, Greece
Nikolaos Polatidis,
University of Brighton, United Kingdom

*CORRESPONDENCE
Christian Luidold
✉ christian.luidold@univie.ac.at

RECEIVED 31 January 2024
ACCEPTED 23 April 2024
PUBLISHED 09 May 2024

CITATION
Luidold C and Jungbauer C (2024) Cybersecurity policy framework requirements for the establishment of highly interoperable and interconnected health data spaces. *Front. Med.* 11:1379852. doi: 10.3389/fmed.2024.1379852

Cybersecurity policy framework requirements for the establishment of highly interoperable and interconnected health data spaces

Christian Luidold* and Christoph Jungbauer

Faculty of Computer Science, Multimedia Information Systems, University of Vienna, Vienna, Austria

This paper examines cybersecurity policy framework requirements for establishing highly interoperable and interconnected health data spaces, with a focus on the European Health Data Space (EHDS) and its corresponding joint action Toward European Health Data Space (TEHDAS). It explores the challenges of ensuring data security within an increasingly digital and collaborative healthcare environment, emphasizing the need for robust policy management to protect sensitive health information across diverse healthcare systems and supply chains. Through an analysis of use cases and held expert workshops, the study identifies key requirements for enhancing cybersecurity measures, fostering cross-border data exchange, and ensuring compliance with regulatory standards. It illustrates the practical implications of cybersecurity policies in a real-world scenario, demonstrating how they can be applied to enhance data security and policy effectiveness.

KEYWORDS

cybersecurity in healthcare, health data interoperability, risk management in health organizations, health data privacy, digital health ecosystems

1 Introduction

In this paper, we analyze cybersecurity policy framework requirements for highly interoperable and interconnected health data spaces, with a focus on the European Health Data Space (EHDS) (EDHS)[1] project "Toward European Health Data Space" (TEHDAS).[2] We explore the significant challenges of securing data within an increasingly digital and collaborative healthcare environment. Our research leverages expert workshops and multiple use cases in a healthcare setting from the SPHINX project (1) to identify key requirements for enhancing cybersecurity measures, supporting cross-border data exchange, and ensuring compliance with regulatory standards. Each contribution is designed to offer actionable insights for policymakers and stakeholders in the healthcare sector.

1 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52022PC0197

2 https://tehdas.eu/

1.1 Cybersecurity policy management at different levels

Effective policy management at all levels includes the development, implementation, monitoring, and enforcement of policies and best practices. This is extended by periodic assessments in order to ensure their relevance and validity supported by collaboration and communication between affected stakeholders. Below we briefly describe cybersecurity policy management at an organizational, interorganizational, and ecosystem level.

Organizational level: The main focus regarding policy management at an organizational level lies in the development and implementation of policies as guidelines pertaining to the organization's cybersecurity processes and practices in order to ensure compliance. Subjects included involve but are not limited to access control and incident response encompassing constant monitoring and enforcement of those policies.

Ecosystem level: Regarding policy management at an ecosystem level, the goal lies in the process of coordination of policies and practices between interconnected organizations in order to address shared risks. Main subjects include but are not limited to risks pertaining to the supply chain and third-party risk management, and involve a given degree of collaboration concerning the development and implementation of policies and practices.

Global level: Policy management at the global level focuses on the process of coordination of policies and practices between interdependent organizations within a broader scale encompassing entire sectors, e.g., critical infrastructures. Main subjects include but are not limited to information sharing and CTI in order to address cybersecurity risks. On the global level the collaboration involves parties from industry and government regarding the development and implementation of cybersecurity policies.

1.2 Aim and context of the research

The goal of our research work was to develop a data driven and risk-aware cybersecurity policy management framework for public organizations with an emphasis on health. The framework takes a systemic-holistic view on policy management, and is driven by organizational and user requirements, building on the integration of proven decision and organizational learning models with artificial intelligence concepts. Previous experience (especially during piloting and evaluation of the CS-AWARE project[3]) has shown that current approaches to policy management are not adequately addressing the dynamic nature of the cybersecurity environment (2) and requires further research in enhancing cybersecurity awareness, as well as in increasing the potential of interoperability of organizations beyond mere data exchange (3).

The dynamic nature of cybersecurity is already challenging on an operational level. It becomes increasingly unmanageable at the policy level, especially for public sector organizations and institutions that handle personal and sensitive data as it is the case of health service providers, hospitals, clinical research and care centers, etc. The need to quickly and dynamically adapt cybersecurity management policies (e.g., relating to risk management and business continuity, incident management) to keep up with the continuously changing threat and attack landscape requires a new and more dynamic approach to policy definition and constant re-evaluation against the requirements defined by the cybersecurity realities, as is reported by threat intelligence provided by, e.g., NIS competent authorities/CSIRTs or threat intelligence communities.

The proposed policy management framework will cover:

1. Support for policy requirement assessment and definition, based on the individual socio-technical requirements of organizations. This will be based on the socio-technical soft systems analysis conducted during the CS-AWARE-NEXT project.[4]
2. A dynamic and data driven continuous re-assessment of policy requirements using AI to dynamically reassess cybersecurity policies through continuous data-driven analysis. By integrating Argyris' double loop learning model (4), it allows for adaptive policy execution and adjustments based on evolving threats, with the inner loop focusing on execution and the outer loop on policy modification itself.
3. A decision support and management model that aids organizations in efficiently implementing and dynamically adjusting policies during cybersecurity incidents. It integrates the OODA Loop–Observe, Orient, Decide, Act–a model (5) suited for rapid and informed decision-making in dynamic environments. It informs adjustments and decision-making by monitoring threat intelligence and internal systems, analyzing risks, and ensuring the explainability of actions through contextualization.

The goal is to evaluate the potential for tighter integration of the dynamic operational cybersecurity management capabilities that CS-AWARE already provides with the organizational component that is defined by the policies. The piloting evaluation of the CS-AWARE project has shown that there is great potential in streamlining those two aspects, which requires a more dynamic approach to policy management.

This paper focuses on the requirements analysis regarding the development of risk-aware cybersecurity policy management. It builds upon the results of the conducted end-user workshops with pilot partners from Larissa in Greece. This section continues with the motivation and relevance pertaining to risk-aware cybersecurity policy management and a brief description of the classification of its usage at different levels. Section 2 we present a state-of-the-art analysis focusing on current trends and advances from a legal standpoint comprising of current standards and guidelines, followed by initiatives from affecting effective policy management. Additionally, key scientific work pertaining to planned design decisions is presented in more detail. Section 3 examines current challenges of implementing effective policy management from the points of transparency, information sharing, and responsibility and accountability. Section 4 presents the main key requirement specification from the results of the end-user workshops, followed by the conclusions in Section 5.

3 https://cordis.europa.eu/project/id/740723

4 https://cordis.europa.eu/project/id/101069543

2 Current trends and advances in risk-aware cybersecurity policy management

Current trends regarding cybersecurity policy management are heavily influenced by the effects of legal frameworks including regulations, standards, directives, and laws. Given a shift toward increasingly placing responsibilities on the individual organizations, particularly the senior management, an increased presence of cybersecurity measures can be noted. A general trend is the usage of machine learning and artificial intelligence to implement and support new and existing cybersecurity measures. Following is a non-exhaustive description of current trends and advances:

- **Proactive risk management** is aimed at preventing cyber attacks before they occur instead of merely responding to them and their fallout as they arise. The main approach lies in implementing effective policies and risk management strategies.
- The **focus on risk assessment** forms a crucial component for effective cybersecurity policy management by supporting organizations to identify potential vulnerabilities and facilitating decision making concerning the prioritization of asset security.
- The **automation of policy management** streamlines the process of creating, implementing, and enforcing policies. By adding processes to include monitoring and maintenance during execution, freed resources can effectively be used to focus on more complex challenges.
- The **integration of threat intelligence** alleviates the efforts of organizations to stay ahead of emerging threats and facilitates timely responses to security incidents. Threat intelligence comprise collections of data and results of corresponding analyses about security incidents and vulnerabilities from various sources shared on various levels by entities including government agencies, CSIRTs, organizations, and communities.
- **Collaboration and information sharing** constitutes a driving factor regarding enhanced cybersecurity resilience and timely response to security incidents. It includes various organizations, governmental bodies, CSIRTs and communities working together involving sharing best practices, collaborating on issues, and coordinating actions pertained to security incidents.

Incorporating the above trends and advances into cybersecurity policy management can help to enhance the resilience of organizations by supporting the protection of assets and increase the preparedness against emerging cybersecurity threats.

2.1 Standards and guidelines

The requirements for CS-AWARE-NEXT are in part heavily influenced by current standards and guidelines with the most prominent being the GDPR, NIS2, ISO27001, and the NIST CSF. The following describes the fundamental aspects of each instance relevant to this project:

Starting with the GDPR (titled "General Data Protection Regulation"),[5] the focus in the context of this document lies in the handling and processing of data by controllers and the associated rights of data subjects.

- **Right to data portability (Art. 20 GDPR)** states that the data subject shall have the right to receive data concerning themselves provided to a controller and transmit the data to another controller in a machine-readable way. The processing has to be carried out in an automated way.
- **Representatives of controllers or processors not established in the Union (Art. 27 GDPR)** states that the controllers or processors need to designate in writing a representative in the European Union, more precisely in a member state, where the data subjects, whose personal data are processed. The obligation of having a designated representative does not apply to public authorities or bodies.
- **Processing under the authority of the controller or processor (Art. 29 GDPR)** states that the processor and any person acting under the authority of the controller or of the processor having access to personal data shall not process those data except on instructions from the controller, unless required to do so by Union or Member State law.
- **Security of processing (Art 32 GDPR)**, specifically Art 32(2) states that in assessing the appropriate level of security account shall be taken in particular of the risks that are presented by processing, in particular from accidental or unlawful destruction, loss, alteration, unauthorized disclosure of, or access to personal data transmitted, stored or otherwise processed.

The NIS2 Directive[6] (titled "Directive on measures for a high common level of cybersecurity across the Union") is an EU-wide legislation on cybersecurity focusing on active risk management. It expands the scope of the original NIS Directive from 8 to 16 sectors and removes the threshold for applicability of the directive regarding the size of an organization of its corresponding sector. Furthermore, it requires improved risk management approaches, more stringent reporting obligations, harmonized sanctions, and enhanced cooperation with authorities and CSIRTs.

The ISO/IEC 27001:2022 (titled "Information security, cybersecurity and privacy protection - Information security management systems - Requirements") is a European standard pertaining to IT security and management systems. It specifies the requirements for establishing, implementing, maintaining and continually improving an information security management system within an organizational context. These requirements are generic and intended to be applicable to all organizations, regardless of type, size, or nature. The corresponding ISO/IEC 27002:2022 (titled "Information security, cybersecurity and privacy protection - Information security controls") standard[7]

5 https://eur-lex.europa.eu/eli/reg/2016/679

6 http://data.europa.eu/eli/dir/2022/2555/oj/eng

7 https://www.iso.org/standard/82875.html

provides a reference set of generic IT security controls including implementation guidance, which can be used for the development of organization-specific information security management guidelines, as well as implementing information security controls according to best practices. These standards are relevant to risk-aware cybersecurity management as they provide a comprehensive framework for managing cybersecurity risks.

The Cybersecurity Framework by the National Institute of Standards and Technology (NIST CSF)[8] provides initial guidelines for improving cybersecurity risk management in critical infrastructures, pointing out its relevance to risk-aware cybersecurity management. It includes five framework functions as its core structure: Identify, Protect, Detect, Respond, and Recover. With the update to CSF2.0,[9] scheduled for Winter 2024, the framework is set to provide a more extensive guidance regarding implementation, including more specific information about definitions, applications, and interoperability. Additionally, a new theme to be included is the consideration of cybersecurity risks in supply chains in the CSF.

2.2 Initiatives from industry, government, and professional organizations

Current initiatives regarding risk-aware cybersecurity policy management can be found from industry, government, and professional organizations. Their goal lies in increasing the resilience against cyberattacks and raising awareness concerning the presently changing threat landscape, as well as best practices, standards and regulations. Initiatives focus on various aspects of risk-aware cybersecurity policy management, including risk assessment, policy development and enforcement, collaboration, information sharing, and decision-making processes.

Prominent examples of initiatives from industry include the NIST Cybersecurity Framework for helping organizations to better understand and improve their management of cybersecurity risks. While it was originally developed for critical infrastructures, many countries across the globe have adopted and adapted the Cybersecurity Framework with some considering its use as mandatory for both private and public sector.

Regarding initiatives from governments, the European Union's GDPR and NIS2 (the latter specifically targeting critical infrastructures) have caused a significant impact on managing cybersecurity risks. They include provisions for data protection and cybersecurity, as well as requiring organizations to implement technical and organizational measures to ensure an appropriate standard concerning cybersecurity.

Initiatives from professional organizations commonly include education material and certification programs in the domain of cybersecurity. Organizations like the Cloud Security Alliance (CSA)[10] additionally provide publications and documents on latest research conducted in the field of cloud security, as well as providing networking opportunities for members.

The latest pivotal initiative in the context of this paper is the EHDS by the EU. The proposition aims for granting natural persons a higher degree of control over their electronic health data. By ensuring a common legal framework across the EU it would enhance the quality of healthcare-related services, as well as creating a single market with agglomerated healthcare data made available in a preprocessed format for researchers, innovators, and policy-makers. This shall be achieved through establishing strong cybersecurity measures focused on the aspect of data exchange within a highly interoperable environment. In conjunction with TEHDAS the new proposition also focuses on an increased stakeholder engagement encompassing different roles and expertise, as well as to support the process of collaboratively developing and implementing effective policies including cybersecurity policies.

2.3 Scientific works

Main considerations pertaining to design decisions and their implementation of a risk-aware cybersecurity policy management framework within this WP are taken from established methods. Taking the standards and guidelines, as well as the current initiatives from the previous subsections into account, the key scientific works is composed of the following research:

The core concept of effective policy management including the performance monitoring of individual policies is defined by the double-loop learning model developed by C. Argyris, which can be applied to a variety of contexts, including education, personal growth and development, and organizational change. It describes a learning process in which individuals and organizations critically examine and question the underlying assumptions and values regarding their actions and decisions. The results of an effective implementation can lead to more efficient and lasting learning and growth effects. In the context of this research two main types of learning were identified: single-loop learning and double-loop learning. Single-loop learning occurs when an individual or organization works to correct or improve actions or outcomes without questioning the underlying assumptions and values regarding their behavior. It is focused on solving problems instead of exploring the causes of those problems. Double-loop learning involves a deeper level of analysis and questioning by requiring individuals and organizations to critically examine their assumptions and values pertaining to their actions and decisions and questioning their validity and appropriateness. Beyond just involving actions regarding correction or improvement, its view also involves questioning and potentially changing the fundamental values (4).

Research conducted by J. Boyd explores mental patterns or concepts of meaning pertaining to individuals to shape and be shaped by a changing environment. The identified basic goal of everyone lies in improving the capacity for independent action. Any level of cooperation or competition exists to satisfy this aim. If a desired level of independence cannot be achieved, compromises are taken, and constraints are developed in order to

8 https://www.nist.gov/cyberframework

9 https://www.nist.gov/cyberframework/updating-nist-cybersecurity-framework-journey-csf-20

10 https://cloudsecurityalliance.org/

collectively pool skills and talents to overcome or remove obstacles. If overcoming or removing still proves to be impossible the group might alienate and lose members for whom these hindrances are deemed important. In order to strengthen alliances pursuing their goals, effective decisions have to be taken and resulting actions are to be monitored. This creates a need for decision models developed for constantly changing environments. Before new models can be implemented, existing models or concepts, which might inhibit the new one need to be separated from the rest of its associated domain and unstructured by a mental concept coined as "destructive deduction". The subsequent restructuring and creation of new models or concepts by piecing together individual bits to conform to given needs was coined "creative induction". The relation and application of these mental concepts are employed to formulate decision models for individuals and groups to determine and monitor actions to address incidents in changing environments and therefore improve their capacity for independent actions (6).

The OODA (Observation-Orientation-Decision-Action) loop introduced by Boyd resulted from the effort to describe the nature of adversarial engagements. OODA time cycle or loop suggests that success in war depends on the ability to out-pace and out-think the opponent, or put differently, on the ability to go through the OODA cycle more rapidly than the opponent. In cybersecurity the process allows stakeholders to learn from previous experiences, feeding lessons learned into the loop activities to achieve better performance contains four steps. Each group of stakeholders must make observations and process those observations through the orientation process, then use orientation in the decision process, then turn the decisions into actions, which in turn change the world being observed. The focus of the OODA loop is not about making faster decisions, but rather about manipulating the environment to "inhibit an adversaries capacity to adapt to such an environment (suppress or distort observations)". The environment is seen as a means of disorientation to disrupt the adversary's decision-making. Rather than operating in isolation, decision and execution cycles take place simultaneously, but not in synchronization, for both sides. The conflict in the minds of the adversaries compromises the cognitive dimension of the information environment. Adding the cognitive dimension to cyberspace changes the analysis of cyberspace operations from a search for vulnerabilities in hardware and software into an engagement including information operations. "Situational awareness" is a term from psychology which describes both a field of study and the coupling of actors to their operating environment. Situational awareness is knowing what's going on around you (7).

3 Special challenges of risk-aware cybersecurity policy management in interdependent health organizations

An important aspect of the EHDS is risk management, as the proposal was specifically designed to take the NIS Directive into account to include measures to mitigate identified risks. Risk management typically focuses on credit risk, market risk, and operations risk. Technology risk constitutes a subset of operations risk, and cybersecurity risk subsequently is a part of technology risk. Given the fact that cybersecurity risk would generally be found on the lower end of the risk hierarchy it is often absent from centralized risk management processes. Despite focusing on technological risks stemming from software, the predominant driving factor for risks in operation is human error. Software engineers more commonly tend to exercise their authority to bypass software restrictions and therefore inhibit developed security measures.

Cybersecurity constitutes a crucial challenge for the health sector since it influences the security, privacy, and quality of the provision of healthcare services, especially in interconnected systems and services, as aimed by the EHDS. Nonetheless, handling cybersecurity risks in interdependent healthcare organizations presents several challenges, which arise from the intricacy, heterogeneity, interconnectivity, dynamics, and resource limitations of the sector. Therefore, a comprehensive and collaborative approach is essential for developing a risk-aware policy management framework, enabling healthcare organizations the identification, assessment, prioritization, and mitigation of cybersecurity risks while considering security and usability requirements. It is crucial to involve all stakeholders and align the framework with industry standards and best practices. Furthermore, the cybersecurity framework ought to possess adaptability and flexibility to effectively manage the dynamic and evolving cyber threats faced by the healthcare industry, while catering to the sector's increasing needs and expectations.

3.1 Key challenges in health organizations

Information security risk assessment focuses on the potential damage to data subjects regarding the confidentiality, integrity, and availability of data. The integration of new security measures is generally decided upon calculating the expected loss through the sustained damage taken and comparing it to the cost of implementation. Problems arise by nature of not knowing the actual performance of those security measures, making the quantification of costs an issue.

Risk assessment as a management tool should be distinguished between risk management and security management. Risk management encompasses strategies involved in decision-making and the subsequent monitoring of the outcomes. Security management encompasses programs, processes, etc. used according to the decisions made from the risk management. Risk management therefore constitutes the integral part for cybersecurity policies and cybersecurity policy management (8).

The most prominent issues pertaining risk management focus on the organizational responsibility to assess risks, individual responsibilities or segregation of duties and the role of the government regarding the assurance of effective risk management practices. Specifically, the shift regarding the placement of responsibility on senior management governed the last years, predominantly through the GDPR, as well as NIS and the upcoming NIS2. This shift was taken into account in defining the proposition of the EDHS in the context of including a broader spectrum of stakeholders, especially regarding policy development and project management.

Managing cybersecurity policies in interdependent health organizations can present unique challenges due to the complex relationships and dependencies that exist between these organizations. Listed below is an overview of special challenges determined during the end-user workshops which can arise in this context:

Varying levels of cybersecurity maturity: Interdependent health organizations may have different levels of cybersecurity maturity and understanding, which can make it difficult to coordinate policies and practices effectively. The difference between small local companies and large organizations might be very large, which can make it challenging to establish a common set of policies and standards.

Limited resources: Small local health organizations may have limited resources to allocate to cybersecurity policy management, which can make it challenging to implement and enforce policies effectively. This can be particularly challenging for smaller healthcare institutions that may not have dedicated cybersecurity staff or budgets.

Complex interdependencies: Different regional organizations may have complex interdependencies that can make it challenging to coordinate policies and practices. For example, a regional healthcare system may rely on multiple local clinics and hospitals to provide patient care, which can make it challenging to establish common cybersecurity policies and practices across the entire system.

Regulatory and compliance requirements: Health organizations may be subject to different regulatory and compliance requirements, which can make it challenging to establish a common set of cybersecurity policies and practices. For example, hospitals are subject to different data protection regulations than organizations in the food industry, which can make it challenging to establish common policies related to data protection.

Communication and coordination challenges: Interdependent health organizations may face communication and coordination challenges when trying to establish common cybersecurity policies and practices. This can be particularly challenging when organizations have different priorities or when there is limited communication and collaboration between stakeholders.

Overall, managing cybersecurity policies in interdependent local and regional organizations requires a collaborative and coordinated approach that takes into account the unique challenges and dependencies that exist between these organizations. This may involve establishing common policies and standards, sharing information and resources, and investing in cybersecurity training and education for staff.

3.2 A comprehensive scenario for secure digital healthcare

The European Health Data Space (EHDS) initiative, implemented by the European Commission, aims to facilitate secure and ethical utilization of health data throughout the EU. The EHDS is designed to improve the quality and efficiency of healthcare services, while promoting research and innovation in the health sector. However, the implementation of the EHDS poses challenges for interdependent healthcare organizations in terms of risk-conscious cybersecurity policy management. In order to demonstrate the importance of cybersecurity management a comprehensive scenario in a healthcare setting was created combining 4 use cases from the Horizon 2020 project SPHINX (1). The scenario combines the following use cases:

1. UC13: Exploiting Remote Patient Monitoring Services,
2. UC24: Theft of Patient Data using the Telemedicine System,
3. UC17: Accessing Health Data from a Fitness Tracker, and
4. UC20: Compromised Workstation Allows the Scanning of Hospital Network.

The complex scenario depicts a combination of exploitation of remote patient monitoring services and vulnerabilities in telemedicine systems leading to unauthorized access of health data, including data from fitness trackers. In conjunction with compromised workstations the scenario evolves into a multi-faceted cyber threat illustrating the dynamics of cybersecurity in healthcare, with a particular focus on emerging technologies and remote healthcare delivery. The unified scenario balances patient monitoring and data management together with cybersecurity measures to represent a necessary standard for integrating technology and security to enhance patient care and privacy.

The following subsections give an overview of the individual use cases followed by an analysis of included issues and proposed relevant cybersecurity policies.

3.2.1 UC13: exploiting remote patient monitoring services

Using a remote patient monitoring service, a patient uses a mobile App to read vital signs captured by medical devices and upload the unencrypted data via a home Wi-Fi router. By cracking the weak password and forcing communications to non-transport layer security (TLS) mode a hacker was able to modify health-related information sent to the server. This resulted in the attacker compromising the trust and data integrity of the provided medical services, creating false alarms and causing emergency actions from the personnel monitoring the patient. An analysis of the relevant policies is depicted in Table 1.

3.2.2 UC24: theft of patient data using the telemedicine system

By exploiting a Web Real Time Communication (WebRTC) bug in a hospitals telemedicine service, an attacker was able to stealthily connect to an active media session between a patient and their doctor using a Man-in-the-Middle (MitM) attack. With this the hacker was not only able to access the audio and video stream of the session but could also access and compromise the patient's Electronic Medical Record (EMR) data. The attacker also introduced a crypto-ransomware into the hospital's network, threatening to destroy patient data. This resulted in the loss of availability of healthcare databases, impacting or preventing IT-based healthcare services for up to 2 months and compromising the trust of patients into the healthcare organization due to violating

TABLE 1 Analysis of policies in UC13.

Policy area	Current state	Recommended policy	Policy management action	Expected outcome
Encryption standards	Patient vital signs data not sent encrypted	Mandatory use of encryption for all data transmissions	Regular security audits to ensure encryption implementation	Enhanced security of patient data transmission
Network access control	Home WiFi router protected by a weak password	Strong password policy for home WiFi router	Implement password strength and complexity checks	Prevention of unauthorized network access
Device authentication	Mobile app connects to the Internet via home WiFi router	Mobile app must authenticate the remote patient monitoring platform before uploading data	Firmware update to enforce platform authentication	Reduction in the risk of man-in-the-middle attacks
Data integrity	Lack of verification of data received by the remote patient monitoring platform	Implementation of data integrity checks	Continuous monitoring for data anomalies	Assurance of accurate patient vital signs data

the confidentiality, integrity, and availability of the patient's data. An analysis of the relevant policies is depicted in Table 2.

3.2.3 UC17: accessing health data from a fitness tracker

An orthopedic center recommends the usage of GNSS-enabled fitness trackers for improving the quality of patient diagnosis by connecting to the centre's WiFi and server. A hacktivist replicates the centre's WiFi SSID and subsequently launches a man-in-the-middle attack, intercepting and manipulating patient data transmitted to the server, as the used encryption was based on a known symmetric algorithm utilizing plain HTTP without TLS. The tampered data registering on the centre's real network server raises alarms among the medical staff, therefore binding additional resources. This attack resulted in the violation of confidentiality and integrity of patient data impacting the centre's quality of services and subsequently the patient's private life, which consequently eroded the centre's credibility. An analysis of the relevant policies is depicted in Table 3.

3.2.4 UC20: compomised workstation allows the scanning of hospital network

By opening an attachment of an email containing a trojan, an employee causes the compromise of a hospital workstation by a hacker, who establishes a backdoor to launch a network scanner. This allows the hacker to gather detailed information about the hospital's IT assets, as well as information about operating systems, browsers, and network protocols in order to exploit vulnerabilities and strengthen the attacker's presence. This access can subsequently be used to impact IT-dependent healthcare services or compromise the confidentiality, integrity, and availability of patient data. An analysis of the relevant policies is depicted in Table 4.

3.3 Cybersecurity policy management and transparency

One of the key challenges in cybersecurity policy management is balancing the need for transparency with the need to protect sensitive information. Reluctance to disclose details about cybersecurity policies and practices for fear of revealing exploitable vulnerabilities is common, which caused a lack of standardized reporting for cybersecurity policy management until legal frameworks took effect. Despite these recent changes, a significant number of organizations struggle to understand and implement guidelines for reporting. As the threat landscape is constantly changing, keeping cybersecurity policies and best practices up-to-date can be challenging.

Many organizations are also subject to regulatory requirements related to cybersecurity, which can create challenges in managing policies and practices. A lack of awareness among stakeholders about the importance of cybersecurity policy management and the risks associated with cyber-attacks can further create barriers to enhance organizational resilience. Addressing these challenges through awareness trainings, dedicated resources, and enforced policies has a significant impact on an organization's cybersecurity resilience and facilitates compliance with legal regulations (8, 9).

3.4 Sharing cybersecurity policy management approaches in interdependent organizations

Sharing cybersecurity policy management approaches as a form of collaboration between interdependent organizations facilitates understanding of risks and risk management, including the identification of areas of concern, aiming at establishing a common baseline regarding policies and practices. One of the key challenges to achieve this objective lies in the heterogeneity of organizations. Differences in organizational structures mean differences in risk strategies and tolerances, which inhibit the development of shared policies and practices.

Another aspect is defined through used infrastructure and technology. Organizations relying on cloud services will have corresponding policies which differ from those organizations utilizing on-premise infrastructure. Paired with different priorities pertaining to individual sectors (e.g., water supply vs. healthcare) establishing a common focus can be difficult.

TABLE 2 Analysis of policies in UC24.

Policy area	Current state	Recommended policy	Policy management action	Expected outcome
Encryption standards	WebRTC bug leaking the customer's IP address	Mandatory use of WebRTC security features	Regular security audits to ensure WebRTC security	Enhanced privacy of patient communication
Network access control	Compromised signaling server	Restricted access to signaling server	Implement network monitoring and access logs	Prevention of unauthorized network access
Device aAuthentication	Lack of verification of peer connection	Implementation of peer identity verification	Firmware update to enforce peer identity verification	Reduction in the risk of man-in-the-middle attacks
Data integrity	Lack of verification of data sent to EMR	Implementation of data integrity checks	Continuous monitoring for data anomalies	Assurance of accurate patient EMR data

TABLE 3 Analysis of policies in UC17.

Policy area	Current state	Recommended policy	Policy management action	Expected outcome
Encryption standards	Use of known symmetric encryption without TLS	Mandatory use of TLS for all communications	Regular security audits to ensure TLS implementation	Enhanced security of patient data transmission
Network access control	Unrestricted WiFi access	Restricted WiFi access with authentication	Implement network monitoring and access logs	Prevention of unauthorized network access
Device authentication	Fitness trackers connecting to any network SSID	Devices must authenticate the network before connecting	Firmware update to enforce network authentication	Reduction in the risk of man-in-the-middle attacks
Data integrity	Lack of verification of data sent to server	Implementation of data integrity checks	Continuous monitoring for data anomalies	Assurance of accurate patient health data
Patient privacy	Potential for patient data and location access	Strict access controls for sensitive data	Training staff on privacy policies and procedures	Protection of patient's private information

An additional challenge lies in regulations and legal constraints. A lack of trust constitutes the inhibiting factor with regard to sharing cybersecurity policy management approaches, predominantly when it comes to sharing sensitive information. Organizations competing in the same industry might further exhibit reluctance in sharing approaches presenting additional barriers for collaboration.

Addressing these challenges through established guidelines for sharing information and dedicated communication channels facilitates the alignment of policies and practices. Furthermore, trust can be built through regular communication and collaboration activities supporting decision making and enhancing cybersecurity resilience of participating organizations (9, 10)

3.5 Responsibility and accountability for cybersecurity policy management

Cybersecurity policy management encompasses a significant amount regarding challenges related to responsibility and accountability as it constitutes a shared responsibility involving multiple stakeholders across an organization. One of the challenges is the lack of clear ownership for cybersecurity policies, which complicates holding individuals or groups accountable for breaches or failures. Another challenge is the existence of blame culture involving individuals or groups being blamed for cybersecurity incidents rather than focusing on addressing the root causes of the incident resulting in the creation of a hostile environment discouraging collaboration and information sharing, further inhibiting efforts to enhance cybersecurity resilience. Furthermore, effective cybersecurity policy management can be resource-intensive, requiring significant investments in technology, training, and personnel. Limited resources combined with issues pertaining to ownership impede the allocation of responsibility and accountability.

Due to the evolving cybersecurity threat landscape effective cybersecurity policy management requires monitoring and maintenance of policies and practices including aspects regarding responsibility and accountability. This is often triggered by changes in compliance and regulatory requirements (e.g., NIS2) affecting cybersecurity policies and practices, possibly creating additional responsibilities and accountabilities pertaining to policy management. Addressing these challenges through establishing clear ownership including a culture of collaboration and information sharing, as well as allocating resources to cybersecurity and actively maintaining cybersecurity policies creates an important baseline for strengthening an organizations cybersecurity posture. Legal compliances and regulations provide goals for implementing clear processes for reporting and investigating cybersecurity incidents further inhibiting the effects of blame culture and facilitating the establishment of a resilient cybersecurity culture (9, 11).

TABLE 4 Analysis of policies in UC20.

Policy area	Current state	Recommended policy	Policy management action	Expected outcome
Email security	Employee opening an email containing a Trojan	Implementation of email filtering and scanning	Regular security training and awareness for employees	Prevention of malware infection via email
Asset management	Lack of information about the hospital's IT assets	Implementation of asset inventory and classification	Continuous monitoring and updating of asset information	Improved visibility and control of IT assets
Data protection	Potential for patient data access, modification, or disclosure by the hacker	Implementation of data encryption, backup, and recovery	Continuous monitoring and reporting of data breaches	Assurance of patient data confidentiality, integrity, and availability

4 Use case

A Use Case based from the CS-AWARE-NEXT project is used to prove the applicability in a real life scenario. The Case handles the response to a stolen Laptop with VPN Access as shown in Figure 1. This use case illustrates the practical implementation and challenges of the cybersecurity strategies and frameworks discussed in Section 3.2. By exploring a real-world scenario, we highlight the need for adaptable and robust cybersecurity measures to effectively address emerging threats, and demonstrate the direct application of risk-conscious cybersecurity policy management in a dynamic healthcare environment.

Scenario In the wake of increased remote work due to the COVID-19 pandemic, a laptop belonging to an employee has been reported stolen. This device has established VPN credentials, providing potential unauthorized access to the organization's secure network.

Actors User (employee from whom the laptop was stolen), IT Security Team, Data Protection Officer (DPO), Network Services Team, Police, Vendor (laptop provider)

Preconditions The employee has been working remotely due to pandemic restrictions and has been using a VPN to access the company's network. The laptop is equipped with the company's standard security features, including VPN access.

Trigger The theft of the laptop is reported by the user to the IT Security Team.

Narrative

- Upon receiving the report of the stolen laptop, the IT Security Team initiates an interview with the user to gather comprehensive information about the incident and the potential data at risk. The team works swiftly to clear the VPN credentials associated with the stolen device to prevent any unauthorized access to the network.
- Simultaneously, the Data Protection Officer is informed of the breach, and instructions are taken to comply with data protection laws and regulations. The DPO initiates the process of legal and notification obligations, including communication with law enforcement.
- Simultaneously, before it is known which data could be accessed through the device, the local authorities, banks, the CSIRTs and internal affairs need to be informed. The DPO is contacted regarding legal guidelines, as well as the manager and the national application/internet provider.
- The Network Services Team jumps into action, conducting an immediate audit of all associated network, email, web, and local services credentials linked to the user's account, as well as personal data stored on the device. They lock down access and initiate a change of all passwords and security protocols as a precautionary measure. The device in question gets completely disabled.
- While the technical teams address the network and system vulnerabilities, the user is advised to change their credentials for personal services that may have been saved or accessed through the stolen laptop, to prevent further personal risks.
- With security measures in place, monitoring is heightened to track any suspicious activity across the system services associated with the user's account. The period of activity from the last known legitimate login to the current time is reviewed to assess any unauthorized actions taken.
- In conjunction with the internal monitoring, the vendor from whom the laptop was sourced is notified, and assistance is requested in tracking the device, if possible, through any built-in location services or tracking technologies that may have been part of the laptop's security features.

Outcome The immediate and coordinated response effectively mitigates the risks associated with the stolen laptop. The company's actions prevent unauthorized access, protecting sensitive data and maintaining compliance with cybersecurity policies. The user is made aware of the steps taken and is educated on the importance of securing personal and professional data. All parties remain vigilant, ready to respond to any subsequent activities related to the incident.

Postconditions The IT Security Team, along with the DPO, reviews the incident to update and refine the organization's security protocols and training, with the aim of preventing similar breaches in the future. Additionally, a follow-up with law enforcement and the vendor is maintained to track the progress on the recovery of the stolen property.

5 Key requirements specification for cybersecurity policy management

A successful cybersecurity policy management framework includes a range of vital components, including risk assessment, policy development and enforcement, collaboration and information sharing, and effective decision-making processes. Furthermore, it requires the involvement of internal and external

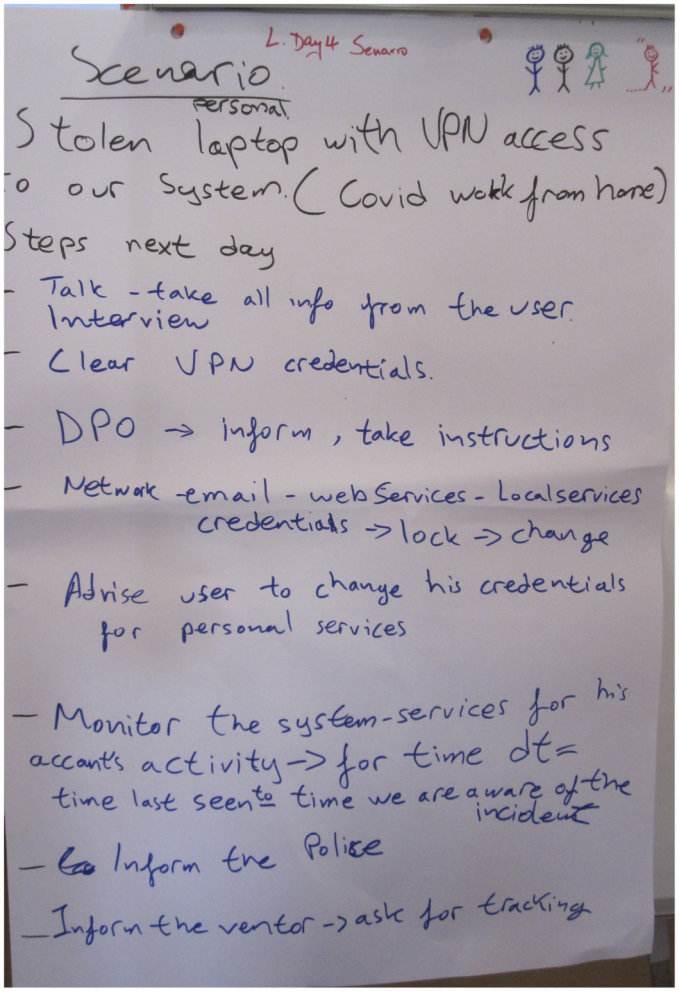

FIGURE 1
Original scenario from the workshop: stolen IT property.

stakeholders, with the latter encompassing government agencies, as well as other organizations.

Adding to the general components, individual requirements from organizations need to be taken into account to ensure a limited degree of restrictions and facilitate the adoption of a cybersecurity policy management framework. The main requirement categories obtained from the end-user workshops were examined and subsets of requirements were defined.

5.1 Basic knowledge and understanding of formalized policies

"Facilitate the understanding of documented formal policies and their advantages." This meta-requirement focuses on raising awareness of stakeholders through training. The involved approach (12) is based on the Erasmus+ project COLTRANE. The sub-requirements derived from this meta-requirement are listed below:

- **Raise awareness of current policies:** Improve dissemination of policies from pure publishing to awareness, understanding and enforcement.
- **Promotion of collaboration and awareness raising:** Build on the COLTRANE approach for promoting collaborative policy management and awareness raising.
- **Simulation and training:** Use a virtual platform to simulate the handling of attack situations. Provide hands-on experience of collaboration- and awareness-driven policy management.
- **Organizational prerequisites for acting on the ecosystem level:** in order to handle policies at the ecosystem level organizations need to provide the necessary basis. The steps toward it have to be identified.

5.2 Formalization of best practices

"The ability to create documentation of best practices & guidelines in the organization to retain expertise and prevent loss of knowledge." This meta-requirement focuses on the collaborative approach involving employees, organizations, communities, and government agencies in order to enhance an organization's resilience. In order to implement effective formalization of best

practices, the cybersecurity policy management component works in concert with the collaboration component in WP2. An overview is listed below:

- Definition of state of the art practices: Facilitate the creation and maintenance of practices depending on current situations.
- Effective applicability and adaptability: Ensure practices are case-type based to provide a best fit for specific environments.

5.3 Shared policy repository

"Enable information sharing through a shared repository." This meta-requirement regarding the provision and usage of a shared knowledge base regarding CTI, reports, as well as information pertaining to legal compliance. In order to realize the requirements of a shared policy repository and its subsequent usage a cybersecurity policy management component demands the support of AI-based quality data assessment and correlation. An overview is listed below:

- Harmonization with governing bodies: Ensure effective collaboration with governing bodies through establishing a common standard for information sharing.
- Provision of information: Make related documents from communities and governing bodies available for improving legal and technical readiness.
- Filter information according to needs: Enable means of distinction between policies according to metadata.
- Highlight current threats and vulnerabilities: Point out trending topics within the organization, community, and governing bodies. Analyze shared enriched CTI.

5.4 Implementation of best practices into workflows

"Enable the adaption of policies and best practices to the needs of the organization and their subsequent adoption into the organizational context." This meta-requirement focuses on the adoption of policies into automatic workflows regarding disaster recovery and business continuity plans, therefore enhancing resilience and supporting legal compliance. In order to effectively implement best practices into organizational workflows the cybersecurity policy management component needs to encompass functionality pertaining to business continuity and disaster recovery. An overview is listed below:

- Provision of core essentials: Ensure the basic needs of an organization are met for legal compliance with governing bodies.
- Policy management life cycle: Provide an environment for creating, managing, enforcing and maintaining policies.
- Adaption of disaster recovery and business continuity plans: Facilitate the integration of policies and best practices into disaster recovery and business continuity procedures. Enable continuous monitoring and adaption of related workflows.
- Provide a basis for decision making: Building on decision making and reflective learning models in support of policy enforcement and maintenance.

5.5 Ensure effective visualization

"Create a visualization supporting the implemented functionalities in an intuitive way." In order to stimulate an active engagement with the cybersecurity policy management component, the user interface and user experience need to appeal to the end user's preferences.

6 Conclusion

The need to adapt cybersecurity management policies quickly and dynamically (e.g., relating to risk management and business continuity, incident management) to keep up with the continuously changing threat and attack landscape requires a new and more dynamic approach to policy definition and constant re-evaluation against the requirements defined by the cybersecurity realities, as is reported by threat intelligence provided by, e.g., NIS competent authorities/CSIRTs or threat intelligence communities.

One of the aspects of collaboration within a shared ecosystem lies in the development of common policies and standards in order to diminish the complexity regarding the management of cybersecurity risks and ensuring actions taken are streamlined according to the same security protocols. Additionally, the implementation of common policies and standards helps to build trust between interdependent organizations and their customers, further increasing the relevance of effective risk-aware cybersecurity policy management.

The main focus regarding policy management at an organizational level lies in the development and implementation of policies as guidelines pertaining to the organization's cybersecurity processes and practices in order to ensure compliance. Main subjects include but are not limited to risks regarding the supply chain and third-party risk management, and involve a given degree of collaboration concerning the development and implementation of policies and practices. Managing cybersecurity policies in interdependent local and regional organizations can present unique challenges due to the complex relationships and dependencies that exist between these organizations. Local and regional organizations may be subject to different regulatory and compliance requirements, which can make it challenging to establish a common set of cybersecurity policies and practices.

Sharing cybersecurity policy management approaches in interdependent organizations has to keep in mind the differences in organizational structure, which can make it challenging to align cybersecurity policies and practices across different organizations. Sharing these sensitive cybersecurity policy management approaches requires a high degree of trust between organizations, which can be difficult to establish and maintain. Ultimately, a shared approach to cybersecurity policy management can help to improve the overall security posture of interdependent organizations and reduce the risk of cyber attacks.

Ongoing research focuses on support for compliance with regulatory bodies and authorities, as well as autonomous adaption to organizational events based on log data. This approach focuses on the use of the double-loop learning model to change minor policy details automatically, or provide decision making support for more substantial changes. Additional focus lies in addressing the development of security and privacy related policies for IoT devices in healthcare. Frameworks targeting compliance with security standards before deployment serve an increased demand in light of legislative plans for fostering data exchange, collaboration, as well as supply chain security.

Author contributions

CL: Writing—original draft, Writing—review & editing. CJ: Writing—original draft, Writing—review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101069543.

Acknowledgments

The authors would like to thank our partners from 5th Regional Health Authority of Thessaly and Sterea for participating in workshops and additional discussions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Manso M. *Use Cases Definition and Pilot Overview Document v3*. (2021). Available online at: https://zenodo.org/records/5052727 (accessed December 7, 2023).

2. Schaberreiter T, Wieser C, Koumpis A, Luidold C, Andriessen J, Cappiello C, et al. Addressing critical issues and challenges for dynamic cybersecurity management in organisations and local/regional networks: the CS-AWARE-NEXT project. In: *2023 Fifth International Conference on Transdisciplinary AI (TransAI)*. (2023). p. 232–236. Available online at: https://ieeexplore.ieee.org/abstract/document/10387599 (accessed January 9, 2024).

3. Luidold C, Schaberreiter T, Wieser C, Koumpis A, Cappiello C, Citro T, et al. Increasing cybersecurity awareness and collaboration in organisations and local / regional networks: the CS-AWARE-NEXT project. In: *Sustainable, Secure, and Smart Collaboration (S3C) Workshop 2023*. (2023). Available online at: http://eprints.cs.univie.ac.at/7835/ (accessed December 19, 2023).

4. Argyris C. Double-loop learning and implementable validity. In: Tsoukas H, Mylonopoulos N, editors. *Organizations as Knowledge Systems: Knowledge, Learning and Dynamic Capabilities*. London: Palgrave Macmillan UK (2004). p. 29–45.

5. Lenders V, Tanner A, Blarer A. Gaining an edge in cyberspace with advanced situational awareness. *IEEE Secur. Privacy*. (2015) 13:65–74. doi: 10.1109/MSP.2015.30

6. Boyd J. *Destruction and Creation*. (1976). Available online at: https://www.semanticscholar.org/paper/Destruction-and-Creation-Boyd/483359fa9420efcddde5a17da597f462c2a788c2 (accessed December 15, 2023).

7. Zager R, Zager J. OODA loops in cyberspace: a new cyber-defense model. *Small Wars J*. (2017). Available online at: https://smallwarsjournal.com/jrnl/art/ooda-loops-cyberspace-new-cyberdefense-model

8. Bayuk JL. Cyber security policy catalog. In: *Cyber Security Policy Guidebook*. Hoboken: John Wiley & Sons, Ltd. (2012). p. 93-210. Available online at: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118241530.ch6 (accessed December 19, 2023).

9. Shaver KG. *The Attribution of Blame: Causality, Responsibility, and Blameworthiness*. Berlin: Springer Science & Business Media. (2012).

10. Pala A, Zhuang J. Information sharing in cybersecurity: a review. *Deci Analy*. (2019) 16:157–237. doi: 10.1287/deca.2018.0387

11. Pollini A, Callari TC, Tedeschi A, Ruscio D, Save L, Chiarugi F, et al. Leveraging human factors in cybersecurity: an integrated methodological approach. *Cogn Technol Work*. (2022) 24:371–90. doi: 10.1007/s10111-021-00683-y

12. Langner G, Andriessen J, Quirchmayr G, Furnell S, Scarano V, Tokola TJ. Poster: the need for a collaborative approach to cyber security education. In: *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. Vienna (2021). p. 719–721. doi: 10.1109/EuroSP51992.2021.00058

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 15 May 2024
DOI 10.3389/fmed.2024.1365501

OPEN ACCESS

EDITED BY
Oya Beyan,
University Hospital of Cologne, Germany

REVIEWED BY
Adamantios Koumpis,
University Hospital of Cologne, Germany
Konstantinos Kalafatakis,
Queen Mary University of London,
United Kingdom

*CORRESPONDENCE
Isabelle de Zegher
✉ isabelle@dezegher.com

RECEIVED 04 January 2024
ACCEPTED 18 March 2024
PUBLISHED 15 May 2024

CITATION
de Zegher I, Norak K, Steiger D, Müller H, Kalra D, Scheenstra B, Cina I, Schulz S, Uma K, Kalendralis P, Lotman E-M, Benedikt M, Dumontier M and Celebi R (2024) Artificial intelligence based data curation: enabling a patient-centric European health data space. *Front. Med.* 11:1365501. doi: 10.3389/fmed.2024.1365501

Artificial intelligence based data curation: enabling a patient-centric European health data space

Isabelle de Zegher[1]*, Kerli Norak[2,3], Dominik Steiger[4], Heimo Müller[5], Dipak Kalra[6], Bart Scheenstra[7], Isabella Cina[8], Stefan Schulz[9,10], Kanimozhi Uma[11], Petros Kalendralis[12], Eno-Martin Lotman[2], Martin Benedikt[13], Michel Dumontier[14] and Remzi Celebi[14]

[1]B!loba, Tervuren, Belgium, [2]North Estonia Medical Centre, Tallinn, Estonia, [3]Department of Health Technologies, Tallinn University of Technology, Tallinn, Estonia, [4]MIDATA Genossenschaft, Zürich, Switzerland, [5]Diagnostics and Research Institute of Pathology, Medical University Graz, Graz, Austria, [6]The European Institute for Innovation Through Health Data, Ghent, Belgium, [7]Department of Cardiothoracic Surgery, Cardiovascular Research Institute Maastricht, Maastricht University Medical Centre, Maastricht, Netherlands, [8]European Heart Network, Bruxelles, Belgium, [9]Averbis GmbH, Freiburg, Germany, [10]Institute for Medical Informatics, Statistics and Documentation, Medical University Graz, Graz, Austria, [11]Faculty of Engineering Science, Department of Computer Science (HCI), Leuven, Belgium, [12]Department of Radiation Oncology (Maastro), GROW School for Oncology and Reproduction, Maastricht University Medical Centre, Maastricht, Netherlands, [13]Department of Internal Medicine, Division of Cardiology, Medical University of Graz, Graz, Austria, [14]Department of Advanced Computing Sciences, Institute of Data Science, Maastricht University, Maastricht, Netherlands

The emerging European Health Data Space (EHDS) Regulation opens new prospects for large-scale sharing and re-use of health data. Yet, the proposed regulation suffers from two important limitations: it is designed to benefit the whole population with limited consideration for individuals, and the generation of secondary datasets from heterogeneous, unlinked patient data will remain burdensome. AIDAVA, a Horizon Europe project that started in September 2022, proposes to address both shortcomings by providing patients with an AI-based virtual assistant that maximises automation in the integration and transformation of their health data into an interoperable, longitudinal health record. This personal record can then be used to inform patient-related decisions at the point of care, whether this is the usual point of care or a possible cross-border point of care. The personal record can also be used to generate population datasets for research and policymaking. The proposed solution will enable a much-needed paradigm shift in health data management, implementing a 'curate once at patient level, use many times' approach, primarily for the benefit of patients and their care providers, but also for more efficient generation of high-quality secondary datasets. After 15 months, the project shows promising preliminary results in achieving automation in the integration and transformation of heterogeneous data of each individual patient, once the content of the data sources managed by the data holders has been formally described. Additionally, the conceptualization phase of the project identified a set of recommendations for the development of a patient-centric EHDS, significantly facilitating the generation of data for secondary use.

KEYWORDS

AI-based data curation, personal health knowledge graph, ontology, catalogue of data sources, EHDS, data intermediary, patient-centricity

1 Introduction

The European Health Data Space (EHDS) draft Regulation published in May 2022 (1) is a ground-breaking initiative which aims to unlock the full potential of health data by facilitating their secure exchange and reuse across the European Union. While the EHDS opens unprecedented opportunities for the management and exploitation of health data, the proposed implementation suffers from two important limitations.

Firstly, the EHDS is designed to benefit the whole population with limited consideration for individuals: it regulates how to manage data for analysis and decision-making across the population, while its usefulness for individual patients in day-to-day care is limited. The main benefit for individual patients, will be the availability of six categories of personal health data—including patient summary, laboratory results, prescribing and dispensing information, imaging reports and discharge summaries—in an interoperable and standardised digital format; this will enable smooth exchange of critical personal health information between healthcare providers across Europe and beyond, primarily for unplanned care needs. Patients will also be able to access their data through National Contact Points for Digital Health (NCPDH); these public health organisations have no direct contact with patients and therefore have little opportunity to establish a relationship of trust at an individual level. While the EHDS will bring benefits to patients, there is a missed opportunity for individuals to actively participate in managing, completing, and improving the quality of their own medical records, which are made of disparate data sources with inconsistencies, gaps and limited interoperability and reuse.

Secondly, the generation of secondary datasets in EHDS will continue to require recurrent curation of potentially identical patient data and provide sub-optimal datasets. Health Data Access Bodies (HDABs), which are also public health organisations, will be granted permission—with opt out possibility for the patients—to process patient data for secondary use by authorities and researchers. As source patient data will remain heterogeneous, there is a risk that the HDABs will process the same data several times for different purposes. Furthermore, as patient data cannot be linked[1] without subjects' consent or in crisis situations, the resulting population datasets can only provide partial views of patients, with sub-optimal data quality.

AIDAVA (2)—a 4-year Horizon Europe project launched in September 2022 with 14 partners, under grant agreement 101057062—proposes a new paradigm in health data management by giving patients greater control and agency (3) over their personal health data through an intelligent virtual assistant (VA). The AIDAVA solution will first help patients to integrate their data collected by hospitals, general practitioners, patient-reported outcome management systems (4), and medical devices. It will then use multiple curation technologies to semi-automatically transform this data into a formal, interoperable representation based on knowledge graph technology (5), called the Personal Health Knowledge Graph (PHKG) (6). Each PHKG is constrained by the AIDAVA reference ontology to ensure interoperability and maximise reuse; the reference ontology (7) will built on ontology frameworks from standards in use in the European Electronic Health Record Exchange Format (EEHRxF) (8)—including HL7 FHIR, SNOMED, LOINC standards—and in clinical research, such as CDISC and OMOP.

During the curation and publishing processes, the VA will request feedback from the individual when full automation cannot be achieved; for complex questions, the VA will request the contribution of an expert data curator. To increase the understanding of the question and the quality of the response, the VA will provide contextual information using metadata regarding the data sources and their transformations and considering the level of health and digital literacy of the patient.

AIDAVA has the potential to implement the 'curate once at patient level, use many times' principle for the benefit of the patients and their care providers. From the interoperable personal longitudinal health record derived from multiple heterogeneous data sources, AIDAVA will be able to generate, on request, the six priority personal health data in EEHRxF format, as well as data extracts complying with national specifications and future versions of EEHRxF. In addition, the availability of multiple, interoperable PHKGs accelerates—with permit or dynamic patient consent—the smooth generation of secondary use datasets, with superior quality because data are linked at the individual level within each PHKG.

This paper first presents the perceived limitations of the EHDS regulation and introduces the potential of data intermediation services described in the Data Governance Act (9) to manage personal health data. It then describes the ongoing research topics developed within the AIDAVA project. Finally, it proposes preliminary recommendations for an innovative digital health infrastructure that promotes seamless data integration, interoperability, and data quality for individual health data, thereby improving patient care, research capabilities and the efficiency of the healthcare system. The authors suggest integrating these preliminary recommendations into the implementing acts currently being drawn up for the deployment of the EHDS.

2 Materials and methods

2.1 Review of EHDS

2.1.1 EHDS is authority and population centric rather than patient-centric

At the heart of health data management are the data holders who collect personal data, including clinical data, social determinants of health and clinical research data[2]. The GDPR data portability right (10) enables individuals to move, copy, or transfer their personal data across data holders; the emerging Data Act (11) will further regulate the portability of data from Internet of Things and medical devices data holders in particular.

1 Privacy Preserving Record Linkage obfuscating or encrypting Personal Identification Information supports record linkage; however, by its nature, it masks personal identifiable information.

2 Inclusion of clinical research data has been requested by the European Parliament in their comments from November 2023.

The EHDS proposes the creation of four different types of organisations within Member States and two at European level, across health care delivery and research (Figure 1).

Organisations on the health care delivery side include: (i) *National Contact Points for Digital Health (NCPDH)* which act as gateway for European citizens to access their data, pooled from data holders, (ii) a *Member State Digital Health Authority* which is responsible for enforcing the lawful use of data in health care delivery, certifying and supervising NCDPHs and cooperating with other Digital Health Authorities and the Commission, and (iii) *MyHealth@EU* which supports the infrastructure for cross-border management of health care delivery data.

Patients have the right to request data holders to transfer their data to a NCDPH, and to access their data from this NCDPH; patients can also request a free copy of their data, in the state they are at the NCDPH. Finally, patients will benefit from six priority categories of identifiable data in a standardised digital format they can share with healthcare providers throughout Europe to ensure safer unplanned care when travelling.

Organisations on the research and policymaking side include: (i) *Health Data Access Bodies (HDAB)* which are responsible for processing health data for secondary use on the basis of the conditions specified in the regulation, (ii) a *Coordinating Health Data Access Body* which enables the cross-border secondary use of electronic health data under the responsibility of each Member State, in cooperation with other coordinating bodies and the Commission and (iii) *HealthData@EU* which supports the infrastructure for cross-border use of research and policymaking data.

Data users, defined as any natural or legal person who have lawful access to personal or non-personal electronic health data for secondary use, may submit a data access application to a HDAB for any purpose identified in the regulation. Patients who wish to understand how their personal health data are used, can access a public website where the HDABs register the permits they have been granted.

Except for data holders and data users, all organisations mentioned above are public organisations or research infrastructure established as a European Research Infrastructure Consortium, funded per Member State and/or the European Commission. Private organisations are not mentioned, while they can bring a wealth of expertise and know-how in processing health data and can stimulate a true data economy benefiting the patients. This is particularly the case for emerging data intermediaries, regulated by the Data Governance Act; they could provide data intermediation services to patients enabling them to exercise their GDPR right to correct errors, and to curate and improve the quality of their health data before it is sent to the NCDPH. Article 13.2. mentions that Clinical Patient Management System may become authorised participants to MyHealth@EU; there is however no further details.

In addition, the EHDS tends to create a barrier between health care delivery, and health research and policymaking. More specifically, Section 2 seems to consider that primary use of data is synonymous with health care delivery (including home care, primary care, secondary care, and tertiary care), while Section 4 considers that secondary use is synonymous with health research & policymaking, where population datasets are generated from data extracted from individuals' clinical data and other, personal and non-personal, data.

This is confusing against the concept of primary use of data, i.e., data collected for a specific purpose, and secondary use of data, i.e., reuse of existing data for a different purpose. As displayed in Figure 2,

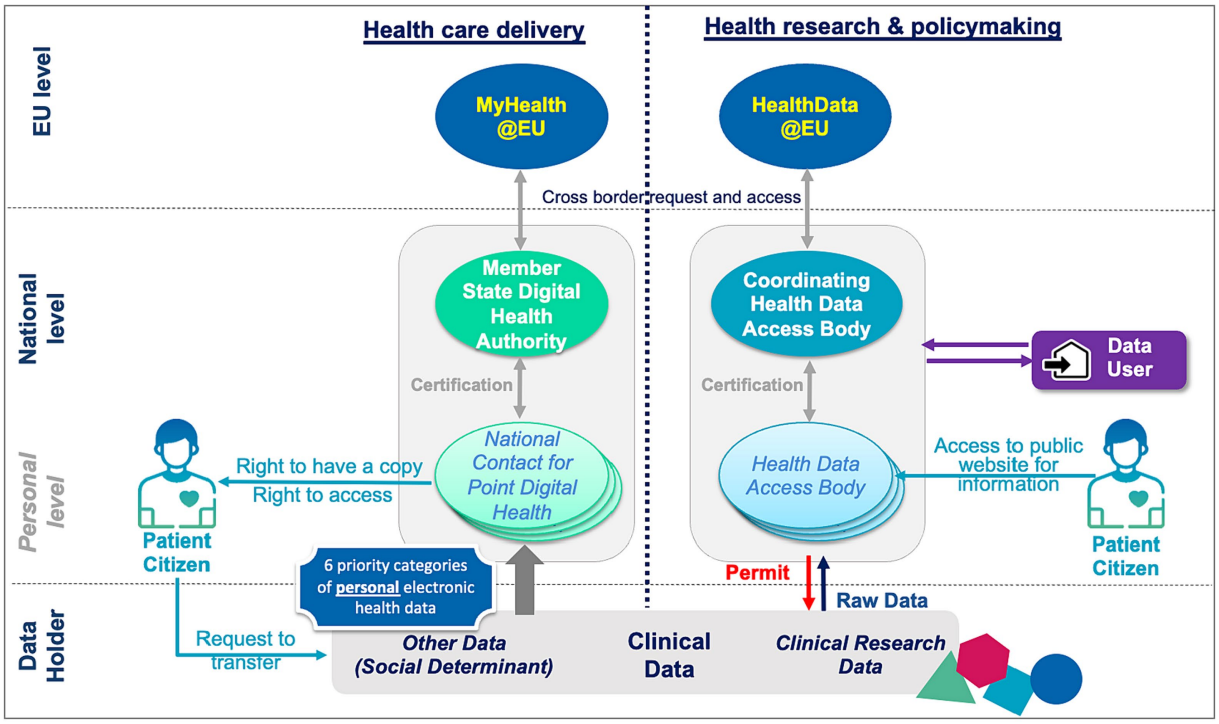

FIGURE 1
Organisations and main information flows within proposed EHDS regulation.

data collection is most often taking place in health care delivery, but it is also happening in research (e.g., interventional clinical trials, adverse events, and clinical registries), and policymaking (e.g., public health surveys). For a true patient-centric EHDS, all personal data related to a patient should be first integrated into their personal longitudinal health record, from which different types of data can be derived for health care delivery as well as research and policymaking.

2.1.2 EHDS does not solve the burden of recurrent curation

Health data are heterogeneous because many legacy systems are still up and running—and may remain so for a long time—with large portions of unstructured variables and narrative text. Additionally, multimodal data (medical imaging, genomic data, EHRs, wearable…) require different types of representation and technologies. Lastly, data standards for clinical care and clinical research have different requirements: for instance, HL7 FHIR is structured vertically, gathering all data for a single patient encounter, while CDISC SDTM and ADaM in clinical research organise the same parameter horizontally, for multiple patients.

For the foreseeable future, health data curation will continue to be necessary. As displayed in Figure 3, the current 'population-based' model relies on expert data stewards extracting pseudonymised data from data sources for a specific purpose and transforming this data into the format required for the analysis. As the GDPR regulation does not permit linkage of personal data without a legal basis or personal consent, and as consent of each relevant individual is difficult to obtain with the existing infrastructure, health data are most often not linked, and the curated data provides only a partial view of the patients. In addition, the number of subjects is often different from one data source to another. Finally, as each secondary use may require slightly different datasets, one individual's data may be curated several times, resulting in massive and unnecessary duplication of effort.

Secondary use in the EHDS follows this model; indeed, the raw data available within the data holders is neither standardised nor linked. Furthermore, while the EHDS regulation introduces basic requirements for quality of the source data, there is no provision for data quality labelling in secondary datasets.

Another concern is that EHDS may become a contributor to additional Greenhouse Gas Emissions (GHGE). After painstakingly generating secondary datasets, HDABs will not be inclined to delete them even though the likelihood of reuse is low; in addition, they might be forced to keep these datasets for liability purposes. Data centres accounted for more than 2.5% of GHGE in 2022, and are

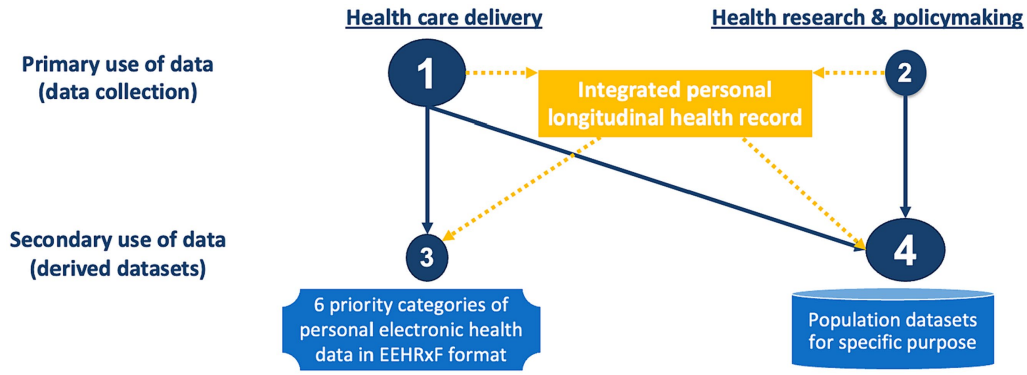

FIGURE 2
The continuum of health data across delivery and research. Orange indicates suggested data and data flows. Missing in the regulation.

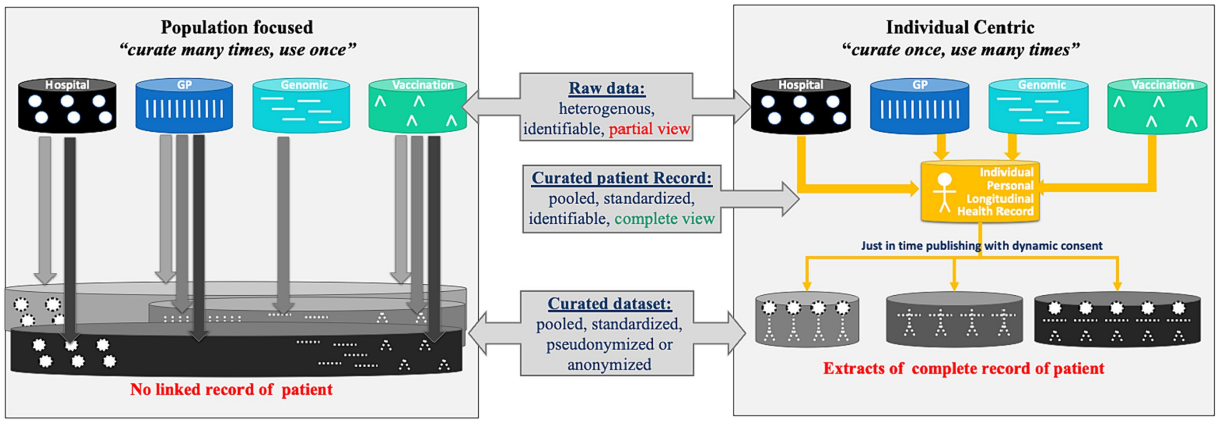

FIGURE 3
Population focused data curation, vs. individual centric data curation.

targeted to rise to 14% by 2040; 30% of the world's data volume is generated in the health sector (12) and is expected to rise to 36% by 2025. More than 90% of the data stored in data centres are not used more than once (13).

In a patient-centric EHDS, it is possible to shift the paradigm towards 'individual centric curation'. The patient, their delegate and/or an agreed expert data curator, curates all their health data, linked across data sources, with the help of an intelligent virtual assistant (VA); the VA orchestrates multiple tools to maximise automation in data curation and quality checks, and involves the patient only when clarifications are required. The result is a personal longitudinal health record, which could be used by attending physicians in the interest of the patient, and by the patient for shared decision-making, second opinion seeking or cross-border care. In addition, if these longitudinal patient records are interoperable, they can be used to generate just-in-time secondary datasets with a quality label derived from the patient records they are extracted from. These datasets could also include metadata—including the programme used to generate them—supporting re-generation of the dataset if needed.

The automated generation of an interoperable, reusable, high-quality, personal longitudinal health record, with and by the patient, is the main objective of the AIDAVA project presented in this paper.

2.2 Data intermediaries and data governance act

The Data Governance Act introduced in November 2019 is in force from September 2023, with a transition period of 2 years. It establishes the foundation for data intermediation services, through public and private data intermediary organisations, for public and business data. It also regulates data altruism, i.e., data voluntarily made available to data altruism organisations for the common good, to reduce the cost of collecting consent and facilitate data portability throughout Europe. The Data Governance Act applies to all sectors, including health.

Although data intermediation services were not initially intended to regulate the sharing of personal data, they can naturally be extended to personal data intermediaries, with a set of structured services as described in the MyData Operators Framework (14), following different business models (15). The draft EHDS regulation only mentions data altruism, which benefits authorities but brings limited value to citizens and patients. As advocated by the AIDAVA project, to be patient-centric, EHDS should include personal health data intermediation services, through dedicated and certified organisations called Health Data Intermediaries. These organisations can serve as trusted partners for patients to control the integration, curation and quality of their data, and to manage their preferences for sharing their data before it is reused in care delivery, research and policymaking.

This approach is the cornerstone of a patient-centric EHDS. If it is easy—hopefully seamless—for patients to manage and curate their data while benefiting from an integrated harmonised health record, they will be more likely to engage in managing their health data and in sharing them for the benefit of the population, and ultimately in managing their personal health.

2.3 Introduction to the AIDAVA project

The main objective of the AIDAVA Horizon Europe project is to deliver and test a prototype intelligent virtual assistant (VA) that will assist patients in curating their heterogeneous, multimodal, personal health data into an interoperable Personal Health Knowledge Graph (PHKG). Individuals' PHKGs can then be transformed into multiple formats for reuse and sharing (16) (Figure 4).

The VA is intended to be used by the patient, or their delegate, and a specialised data curator assisting the patient. The solution aims to maximise automation of the curation process by orchestrating the execution of complementary AI-based curation tools according to the data interoperability issue found in a data source. When automated curation is not achievable, the VA initiates a dialogue with the patient, based on their preferences and skill levels, and provides explanations

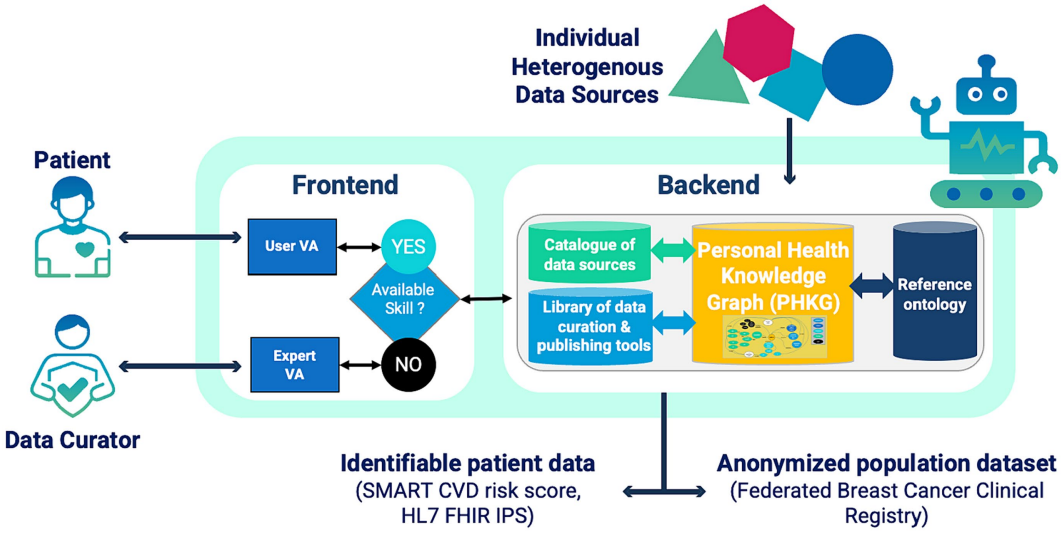

FIGURE 4
Overview of the AIDAVA Virtual Assistant.

of the question at hand. Questions which cannot be answered by the patient are addressed to the supporting data curator.

To demonstrate the 'curate once, use many times' principle, the AIDAVA VA will generate two types of results from the patient's PHKG: (i) fully identifiable data extracted from a single patient's PHKG, in the form of the patient's cardiovascular risk score and International Patient Summary (IPS) in HL7 FHIR format[3], and (ii) anonymized population datasets extracted from multiple PHKGs to form an interoperable, site-specific breast cancer clinical registry that can be federated with other sites.

The project builds on four pillars described in the next section: (i) a structured and repeatable curation process enabling automation by orchestrating the execution of multiple data curation and quality enhancement tools, (ii) a reference ontology as a universal data sharing standard (17), supporting European standards and ensuring interoperability of the resulting PHKGs, (iii) a machine-human interaction module generating personalised explanations of the problem to be solved, and (iv) patient engagement through a trusted health data intermediary.

There will be two generations of the AIDAVA VA prototype. Generation I will include the prototype framework consisting of a Chatbot-like platform as the front-end, and orchestration of a library of data curation & publishing tools at the backend. These tools will preferably be off-the-shelf and open source. Generation II will build on the previous generation: the front end will be extended with an explainability module to increase usability for users less experienced in curating data and in medical content; several curation and publishing tools will be updated with tools developed in the project, including multi-lingual AI based Natural Language Processing (NLP) solutions. Each generation of the prototype will be tested in three clinical sites (Universiteit Maastricht in the Netherlands, Sihtasutus Pohja-Eesti Regionaalhaigla in Estonia, Medizinische Universität Graz in Austria) with the support of two data intermediary organisations (MIDATA Genossenschaft for Estonia and Austria, Digi.me Limited for The Netherlands). As AIDAVA is a prototype, it is not subject to the Medical Device Regulation (18). However, as the prototype will be tested with site patients, the evaluation will follow a strictly defined process, documented in a research protocol which must be approved by the local ethical committees.

To ensure a true patient-centred approach, the project is supported by eight patient 'consultants' from the European Patient Centre Coalition for Breast Cancer and the European Heart Network for Cardiovascular Diseases. These patient consultants are actively involved at regular, well-defined times for a total of 42 person days per patient throughout the project. They ensure the project stays focused on what is important for patients.

3 Results (interim)

The AIDAVA project has been active for 15 months during which the consortium detailed the use cases, the requirements, and the solution architecture, and initiated the development. In parallel, the consortium developed the study research protocol needed for evaluation, as well as the data sharing agreement with data transfer technical specifications for each contributing site. This section describes the interim results.

3.1 Automated curation

The first objective of AIDAVA is to automate as far as possible the curation process, transforming heterogeneous health data into a single, harmonised Personal Health Knowledge Graph (PHKG). The curation process involves resolving interoperability issues across these heterogeneous data. Although interoperability has been widely described in different frameworks (19, 20) and publications (21), automation requires a holistic solution based on a precise classification.

We analysed in more detail the issues that hinder data interoperability, differentiating between issues within individual data sources (single-source data interoperability) and issues when integrating data from multiple sources (cross-sources data interoperability). We identified 11 data interoperability issues based on the analysis of the data sources selected in the project, and literature reviews. The single-source issues comprise digitalisation of paper documents, extraction of structured data from free text, format alignment, transformation of semi-structured and structured data, reference data management, terminology alignment, medical coding, and imaging readability. The cross-sources issues include entity deduplication, semantic inconsistencies, and semantic incompleteness.

For each data interoperability issue, we defined a workflow maximising automation in the transformation of the data into a semantically sound knowledge graph. Each workflow uses one or more curation tools supporting resolution of the issues; candidate tools that could be reused or improved were identified. As new, improved tools [e.g., NLP tools based on Large Language Models (22), data wrangling (23) and AI medical coding (24…)] are emerging, they will be replacing older tools. We also specified within the workflows the need for human intervention to resolve issues; approaches to obtain answers from patients, or their supporting curator, are further described in Section 3.3. Finally, we defined a high-level orchestration workflow to deal with multiple data interoperability issues within one data source.

For semantic inconsistencies and semantic incompleteness, the workflow includes data quality rules with triggers for human intervention in case of errors. Data quality rules represent common sense knowledge (e.g., the discharge time in an hospital must happen after the admission time), physio-pathological knowledge (e.g., a breast tumour must include a laterality) and clinical care pathway information (e.g., diabetes type 1 requires an insulin related treatment). Data quality rules also provide a labelling mechanism to assess the reliability of the curated PHKG. For example, curation through a validated and deterministic tool would score higher than curation through an emerging AI tool. Similarly, human input from staff with high health literacy would have a higher score than input provided by a patient with a limited health literacy. A data quality checker is being implemented, together with a governance process to include new rules or remove existing ones. Governance is particularly important, as knowledge encompassed in data quality rules applies to the whole medicine and requires knowledge elicitation, out of scope of the project.

To maximise automation, we needed a preliminary step called '*data source onboarding*', in which metadata on each data source is

3 It is expected that this will be included in the EEHRxF format.

defined and stored in a dedicated catalogue. This catalogue of data sources includes FAIR metadata enriched with (i) information on the structure and content of the data, such as data type, value restriction and value set, (ii) provenance information related to creation, modification and validation of the source information (25, 26), and (iii) semantic mapping with concepts defined in the reference ontology. Metadata on data sources is collected once, in each data holder organisation; it is used each time the system ingests and curates the data of a specific patient. The AIDAVA catalogue of data sources is being developed on top of DCAT-3 (27); it will be extended with the Data Source Description Vocabulary (28) supporting semantic annotation and the RDF Mapping Language (29) for mapping.

3.2 Personal health knowledge graph: interoperability and reuse through reference ontology

A Personal Health Knowledge Graph (PHKG) is a dynamic, semantic representation, which can harmonise and link multimodal, heterogeneous data during the data curation process. Such a PHKG is ideally positioned to capture the semantics of a data source, independently of its structure; it can also support data integration, data quality enrichment and correction, based on the context. Although a PHKG is personal and contextual, it will be interoperable due to being an instance of the reference ontology. As such, the PHKG constitutes a high-quality, FAIR, longitudinal health record, growing continuously as new data is being ingested. During the data publishing process, the data contained in the PHKG can be made available for multiple purposes in the appropriate format.

Achievement of interoperability is constrained by the availability of a commonly agreed and used reference ontology. AIDAVA identified strategic and content requirements for such an ontology. The strategic requirements include (i) support the European Electronic Health Record Exchange Format (EEHRxF), (ii) maximise potential for reuse of the PHKGs across a large range of use cases, beyond the ones identified in the project, (iii) ensure alignment with standards in place to minimise the need of mapping from and to these different standards, while maximising reusability of the PHKG during and after the project, (iv) support maintainability and extensibility during the project as well as beyond the project, and (v) enable implementation and update of constraints supporting data quality.

In terms of content, the ontology will include (i) standards such as SNOMED CT, LOINC, HL7 FHIR General-Purpose Data Types, and HL7 FHIR resource related to the International Patient Summary (30), (ii) concepts that support mapping and transformation with entities and relationships included in the data sources[4], (iii) predefined mapping supporting transformation to HL7 FHIR IPS and other data exchange messages required by EHDS, and (iv) data quality checks implemented through SHACL rules (31).

We are currently assessing how to use the Swiss Personal Health Network framework (SPHN (32)) as the basic schema of the AIDAVA reference ontology; preliminary results demonstrate that an ontological foundational layer will be needed to support extension of the SPHN schema.

3.3 Human-in-the-loop and the value of explainability

AIDAVA emphasises the importance of making the use of AI solutions transparent, and inherently human-inclusive, with interface components adapted to different types of users. Following user-centred design, the project identified eight user personas across different user groups. Personas are fictional characters who represent the similarities of target user groups and play a pivotal role in ensuring that human-AI interaction is tailored to individual needs, promoting more meaningful engagement. To turn the fictional persona into a tangible, realistic character, and to make it easier for system designers to empathise with the user represented by a persona, the latter is visualised in a one-page layout, called a 'persona canvas', which includes narrative text about the persona's interests, preferences, behaviour patterns and attitudes. Within AIDAVA, personas also serve as the foundation of the explainability and feedback with patients, based on their level of digital and health literacy assessed when setting up the user account and stored in their user profile (33).

Most people are not prepared for unmediated interactions with a digital solution that aims to curate their personal health data. To increase acceptance and democratise personal data curation, AIDAVA aims to maximise automation to minimise user intervention. When automation is not possible, and humans must be brought in the loop, AIDAVA will first decide if the question must be raised to the patient or to the supporting data curator, based on the health and digital literacy levels of the patient. In a second step, the system will raise the questions and generate context-based explanations using the type of issue identified in the workflow, the expected human intervention to solve the issue, the level of digital and health literacy of the target user, and the context of the issue to be solved. Context encompasses all aspects of the data's origin, including information on its creation in the data sources—stored as FAIR metadata available in the catalogue of data sources referred above—and the transformation steps that took place during the curation process. Generation of narrative explanations will be based on canned text translated in each user language for Generation I of the AIDAVA prototype. For Generation II, we are exploring the use of multi-lingual Large Language Models.

3.4 Health data intermediaries

AIDAVA is proposing to provide data intermediation services to patients through organisations called Health Data Intermediaries (HDI), introduced in Section 2.2. These emerging organisations, regulated by the Data Governance Act, are expected to provide three services. First, they should operate as a 'personal health data hub' integrating multimodal data, sourced directly from the patient (wearables, lifestyle data, etc.) or from healthcare providers who treated the patient (34). Second, HDIs should enable dynamic management of consent for data sharing, via a digital app. For

4 If concepts are not available, when onboarding data sources into the system, they are added to the reference ontology following the defined governance process.

example, the patient could specify for which purpose their data (i) can always be shared, without their consent (e.g., public health purpose), (ii) can be shared with their consent when the purpose is clarified (e.g., clinical trials) and (iii) should never be shared (e.g., marketing and commercial research). Finally, HDIs should support the improvement and labelling of the quality of the patients' health data; this would increase the value of reuse of this data, firstly for the patient and their treating physician and secondly for research and policymaking.

NCDPHs described above could become health data intermediaries powered by AIDAVA-like virtual assistants. Alternatively, HDIs could serve as smaller data intermediation organisations that assist patients in integrating and curating their data, before it is transferred at NCDPH level.

Within AIDAVA, we are working with two emerging data intermediaries—MIDATA, and DIGI.me—already active in health with the first two functions (personal health data hub and consent management). We are assessing the opportunity of adding the quality enhancement and labelling tools.

4 Discussion

Although we are still in the conceptualisation phase and have yet to evaluate the prototype in real-life situations with patients, some preliminary conclusions can already be drawn and lead to recommendations for a patient-centred implementation of the EHDS; these recommendations will have to be confirmed as the project develops.

4.1 AIDAVA-like solutions are needed for the benefit of the patients and their treating physicians

In the EHDS, as in many research projects, the focus is on improving the production and quality of targeted secondary datasets for research and policymaking, following the 'population curation' approach described in Figure 3. We argue that a paradigm shift is needed towards 'individual curation', improving the management of patient data at the point of care, and supporting smoother extraction of secondary datasets from these high-quality, interoperable patient records. This is particularly relevant for patients with complex conditions, as their data accumulates across multiple stakeholders and episodes of care over time.

A patient-centric approach makes it possible to prioritise patients' interests and needs for day-to-day care by providing a complete medical record that is easily accessible by attending physicians, thereby reducing their daily workload, which in turn decreases the risk of burnout (35). Regarding the use of secondary data, it has been shown that patients are generally in favour of sharing their health data for the common good (36) provided there is transparency, accountability and no data privacy risk. Therefore, the secondary use of patient data for public health purposes could be the default, with the possibility for patients to opt out.

Today, it is extremely difficult for patients to manage and integrate data across different systems, and thus provide a holistic view of their health status. It is equally difficult for them to share information with their treating healthcare providers. Additionally, there is currently no easy way to opt out of sharing their data whenever used for lawfully agreed public health purposes.

AIDAVA-like solutions, in which all data sources have been onboarded as described previously, would enable the patients to control all their health data, to download them from various data sources, curate them into their PHKG and provide consent for sharing. Through AIDAVA-like solutions, patients, or their delegate, would ensure that their data is integrated and of the highest quality, facilitating medical decision-making. In addition, the availability of interoperable PHKG would facilitate the creation of high-quality datasets for research and policy development.

Recommendation 1. EHDS, as a patient-centric solution seeking to bring benefits to European citizens, should first consider the benefits to each individual patient; and more specifically seek digital solutions that enable every European citizen to maintain an interoperable, high-quality ***personal longitudinal health record***, usable at the point of care and allowing the smooth generation of secondary datasets for lawful public health purposes.

4.2 The major problem in data interoperability and reuse of health data is the lack of documentation on data source

The classical concern about accessing personal identifiable data is local data privacy and protection constraints as well as Ethical Committees' approvals. This is a time-consuming process, though generally well described, clear and manageable. We realised, however, that access to detailed descriptions of health data available within an organisation was unexpectedly difficult; this includes data schema—technical description of each data element collected within the different subsystems of the organisation—data lineage and data quality labels. Without such documentation, automation as proposed in AIDAVA is not possible and the 'curate many times, use once' model will remain the standard, burdensome practise.

Other European projects were faced with the same issue; see for instance 'Deliverable 2.1. Overview of data sources and plan to access available data sources' in Precise4Q (37). Documentation of an extract of the patient data in standardised format—related to the six priority categories of personal data to be exchanged per EHDS—starts to be available in several European countries [e.g., in (38, 39)]. This is not enough; all data sources must be documented. To our knowledge, the only country where detailed description of all collected health data is available is Finland (40) as this is mandated by law since 2013.

Secondary datasets also suffer from the same lack of documentation of data elements, which hampers their reuse. Article 37 (i) of EHDS requires each member state to maintain a catalogue of national datasets with details of the source, scope, main characteristics of the population included in the dataset and conditions of access and use. There are no requirements however to provide a detailed description of the data elements included in the catalogue. The EHDS2 pilot project highlighted the

importance[5] of including such information in national catalogues to facilitate interoperability and reuse (I and R in the FAIR principles) of the datasets generated across Member States.

In AIDAVA, we worked for several months with the clinical evaluation sites, to identify and collect the schema of data elements collected at the point of care, supporting automation and explainability in case of human intervention. This information will be stored in the AIDAVA catalogue of data sources, based on existing standards as described in Section 3.1.

Recommendation 2. In alignment with Article 23.3 (a) and (b) of the EHDS regulation, implement catalogues of data sources with detailed description of each data element collected by relevant data holders.

- Develop a standard describing the content of a catalogue of data sources; this standard should build on existing standards such as DCAT and Data Source Description Vocabulary.
- Provide an appropriate infrastructure to support the implementation and maintenance of these catalogues in each relevant data holder and make them accessible—in a controlled way—to produce secondary datasets.

4.3 Automation potential in data curation should be further explored

The data interoperability issues described in Section 3.1. are well known. The innovation in AIDAVA lies in automating a holistic treatment of all these interoperability issues by means of complementary workflows. One data source may present several data interoperability issues, requiring several workflows. Each workflow may include one or more curation tools as well as requests for human intervention when an issue cannot be solved by the machine. Automation in AIDAVA consists of orchestrating the appropriate workflow for each data source and across data sources, to generate a harmonised PHKG from heterogeneous, multimodal data.

With the emergence of powerful new AI tools, such a Large Language Models (LLM) (22), Neuro-Symbolic AI (41), Generalist Medical AI (42) and Medical Imaging, we can expect more and better tools to be available to support the curation of multimodal data.

Recommendation 3. Formally describe all potential health data interoperability issues that can occur in health data and define a related data curation workflow with description of needed curation tools and human intervention.

Recommendation 4. Maintain a library of data curation tools that can solve the different health data interoperability issues. The library should include an assessment of the tools as well as a formal description of the API, supporting integration.

4.4 Data exchange standards are needed but not sufficient: we need a data sharing standard

Source data in health will remain heterogeneous for the foreseeable future. Different formats are in use and/or will soon be mandated: (i) WHO international classification such as ICD required for billing and epidemiological reporting; (ii) the European electronic health record exchange format (EEHRxF) to be mandated by EHDS to support exchange of personal health data based on HL7 FHIR, SNOMED and LOINC already in place in several European countries; (iii) CDISC supporting data collection in the context of drug related regulatory approval; (iv) OMOP typically used as a target format for secondary datasets in clinical research; and (v) many other—often proprietary—formats exist in research and policymaking databases.

Currently, data sources are mapped directly to the required target output, representing $n*m$ mappings, where n is the number of source formats and m is the number of target formats. This represents a major burden across health and hampers patient care and research. We therefore argue that data exchange standards are needed but not sufficient.

Another possibility is to agree on a data sharing standard, enabling information to be transformed to and from any standard and supporting multiple, but yet unknown, data exchanges; this approach would decrease the number of mappings to $n+m$. This is the objective of the patient Personal Health Knowledge Graph (PHKG) constrained by the concepts defined in the AIDAVA reference ontology, described in Section 3.2. Although the maintenance of such an ontology is beyond the scope of this project, our aim is to demonstrate the value of an interoperable PHKG for multiple types of exchanges and secondary data use, and to identify guidelines to support the development and maintenance of a global reference ontology encompassing all data exchange standards.

Recommendation 5. Develop and maintain an EU-wide (or broader) ontology as the basis for interoperable PHKGs, which supports transformation to main data exchange standards in use (at least EEHRxF and those in use in clinical research such as CDISC, OMOP…)

- Confirm the requirements.
- Review existing/past initiatives (e.g., SNOMED ontological framework, SALUS…) and emerging initiatives (e.g., Precise4Q, EUCAIM Hyper ontology, SPHN…) and develop the European wide foundation layer of the ontology.
- Define and implement a governance process.

4.5 Data sharing requires an assessment of the quality of data

Reusing poor quality data has limited value. When developing the requirements for the AIDAVA curation virtual assistant, data

5 Presentation during the HealthData@EU Pilot—Forum on October 19th.

Recommendation 6. Expanding on Article 23.3 (c) and Article 56, and existing data quality frameworks (43) develop and deploy a quality label framework for each state of data: (i) data sources, (ii) curated data and (iii) published data, with appropriate parameters related to the transformation.

users repeatedly asked the same question: how reliable the data are. The answer differs depending on the state of the data: (i) for data sources, a quality label can be established based on the quality level provided by the data holder—if available—including the credentials of the persons who created and validated the data; (ii) for the curated data (i.e., the PHKG), the quality label will be linked to the quality from the source, the level of quality and certification of the curation tools used during transformation, the level of health and literacy of the humans who provided answers when there were semantic gaps, and the number of data quality checks that could not be resolved; (iii) for published data, the quality label will be linked to the level of the curated data, the compliance with the target format, the completeness of the content, the absence of bias as well as the quality, reliability and certification of the imputation algorithm, if applicable.

Article 23.3 (c) of the EHDS mandates to include a data quality statement, such as the completeness and accuracy of electronic health data. Section 5 on health data quality describes the requirements for the quality and utility label for secondary datasets; these requirements map with the question raised by the AIDAVA data users for curated and published data with two major differences: (i) the EHDS requirements include access constraints not addressed in AIDAVA; (ii) the EHDS merges the concept of curated and published data as it only addresses population datasets. In a patient-centric EHDS, one must distinguish the curated PHKG at patient level, and the published output which can be at patient level (e.g., IPS) or at population level (e.g., clinical registry).

4.6 Health data intermediaries, supported by community curators, are needed

The Data Governance Act regulates the setup and functioning of data intermediation services organisations, or what AIDAVA calls 'health data intermediaries' (HDI) when they manage health data on behalf of the patient. To our knowledge the most advanced business model of HDI has been developed in the Netherlands through 'Persoonlijke gezondheidsomgeving' or Personal Health Environment (44). Such models and organisations, close to the patients, must be further defined and deployed, in alignment with the EHDS regulation, to develop and maintain trust with patients.

To support the patient and their treating physicians, HDIs must equip their customer patients with the appropriate tools to exercise control, agency, and guardianship. This includes a Digital Wallet (45) supporting identity management and linking, dynamic consent management, and data transfer. An AIDAVA-like tool, supported by a catalogue of data sources, will increase the value of data intermediation services by improving the quality of the source data and its value for secondary use, making it a key player in the growing telehealth market, and fostering a genuine health data culture throughout society.

The assumption in AIDAVA is that the automation process will be seamless with maximum automation and minimum of human intervention. When human input is required, it is expected that the patient will be the first person requested to support. The percentage of citizens that will be willing and able to contribute is directly linked to the complexity of the task and will be assessed as part of the prototype evaluation. If we assume that between 5 and 15% of the population will be able to contribute, this means that we need additional support from 'community curators', i.e., persons in the community with a minimum or health and digital literacy that would be specifically trained as expert curators and would offer their services to patients through an HDI. Community curators could be a member of the family that would curate the data of the whole family—parents, siblings and children—for free, or could be a third party who should be rewarded for the work done.

It could be argued that this could increase the gap between patients of high and low socio-economic status. While this risk is always present, different approaches should be explored to fund the community curator and data intermediaries (46). There could be a lump sum per patient and per type of diagnosis from national health funding programmes, as high-quality data should reduce the total cost of illness and the cost of research and policy development. There could also be funding from pharmaceutical companies directly to the patient and their community curator, as the availability of interoperable PHKGs could dramatically decrease the cost of trials — as data would be more readily available, just on time — and reduce the decline in the return on investment for research and development (47).

Recommendation 7. Define and support deployment development of different models of Health Data Intermediaries to ensure patients can be in control of their data, exercise agency and secure guardianship through an actor close to the patient and chosen by him/her. This includes new organisation models or integration of supporting digital solutions, including digital wallet for the patient as well as maintenance of a catalogue of data sources and data curation services to maintain each individual PHKG within the patient digital wallet.

Recommendation 8. Define and pilot the role of community curators, aligned with the Skills data space (48).

5 Conclusion

We argue that a patient-centric EHDS will serve foremost each individual patient, but also the population as a whole and other health stakeholders such as healthcare providers and health researchers and policymakers. This mandates the development

and maintenance of a high-quality, personal longitudinal health record for each patient, resulting from the curation of their data scattered across multiple systems and organisations. This longitudinal record should be formalised in a Personal Health Knowledge Graph (PHKG) which should be interoperable because it is constrained by a reference ontology; the PHKG should also include a data quality label, derived from the quality of the sources data and the transformations that took place during the curation process.

The AIDAVA project implements a combination of AI-based automation and a 'human-in-the-loop' approach, harnessing advanced technologies, human expertise, skill sets, and contextual knowledge to help patients—or their delegates—manage their own data and develop their interoperable, high-quality PHKG. In doing so, patients benefit personally and contribute to the just-in-time production of disposable secondary datasets that promotes research and policymaking. AIDAVA therefore proposes a model that places the patient at the centre of a greener interconnected ecosystem of primary and secondary data use, increasing value for all and for the planet.

Several obstacles need to be overcome to achieve the AIDAVA vision. The first is access to personal health data, not because of data privacy issues, but because of the lack of detailed documentation—including format, data typing, and value restriction—on source data. The definition and enforcement of a catalogue of primary data source should be introduced in EHDS and implemented as a priority. Another important component for the sustainability of AIDAVA-like solutions is the availability of a governed reference ontology, laying the foundations for a global data sharing standard. Additionally, sustainable models for health data intermediaries and supporting community curators need to be defined.

The preliminary results of the AIDAVA project demonstrate that the implementation of a patient-centred EHDS is achievable and beneficial. It requires that the recommendations outlined in this paper are included in the implementing acts being drawn up as part of the EHDS deployment.

Data availability statement

The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding authors.

Author contributions

IZ: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Writing – original draft. KN: Investigation, Writing – review & editing, Methodology. DS: Conceptualization, Funding acquisition, Methodology, Writing – review & editing. HM: Conceptualization, Funding acquisition, Methodology, Writing – review & editing. DK: Conceptualization, Funding acquisition, Methodology, Writing – review & editing. BS: Investigation, Writing – review & editing. IC: Writing – review & editing. SS: Conceptualization, Funding acquisition, Methodology, Writing – review & editing. KU: Writing – review & editing. PK: Investigation, Writing – review & editing. E-ML: Writing – review & editing, Conceptualization, Funding acquisition, Investigation. MB: Writing – review & editing. MD: Conceptualization, Funding acquisition, Investigation, Methodology, Writing – review & editing, Supervision. RC: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the European Union's Horizon Europe research and innovation programme under grant agreements No. 101057062 (AIDAVA). The work of the Swiss Partner (MIDATA) received funding by the Swiss State Secretariat for Education, Research and Innovation (SBFI), subvention contract 22.00093, REF-1131-52104. Parts of the work have received funding from the Austrian Science Fund (FWF), Funder Grant Number: P-32554 explainable Artificial Intelligence.

Acknowledgments

The authors are grateful to all the members of the AIDAVA consortium who diligently support the project with their expertise. A special thanks goes to the eight patient consultants from the ECPC and EHN patient organisations; they keep providing us with insights and a sense of purpose of what is important to the patients that is critically important for the project. Finally, we would like to thank the MyData Global Health working group who provided key insights on an individual centric EHDS.

Conflict of interest

IZ was employed by B!loba. SS was employed by Averbis GmbH.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. EUR-Lex (2022). EUR-Lex-52022PC0197-EN-EUR-Lex. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52022PC0197

2. AIDAVA (2023) Project website. Available at: www.aidava.eu (Accessed November 2023).

3. Mydex (2023) Control, Agency and Guardianship. Available at: https://mydex.org/resources/papers/Control_agency_guardianship/

4. Hughes S, Aiyegbusi OL, Lasserson D, Collis P, Glasby J, Calvert M. Patient-reported outcome measurement: a bridge between health and social care? *J R Soc Med*. (2021) 114:381–8. doi: 10.1177/01410768211014048

5. Peng C, Xia F, Naseriparsa M, Osborne F. Knowledge graphs: opportunities and challenges. *Artif Intell Rev*. (2023) 56:13071–102. doi: 10.1007/s10462-023-10465-9

6. Rastogi N, Zaki MJ (2020). Personal health knowledge graphs for patients. arXiv [Preprint]. Available at: http://arxiv.org/abs/2004.00071

7. Stefan S. (2018). The role of foundational ontologies for preventing bad ontology design. Available at: https://ceur-ws.org/Vol-2205/paper22_bog1.pdf

8. Shaping Europe's digital future (2019). Recommendation on a European electronic health record exchange format. February 2019. https://digital-strategy.ec.europa.eu/en/library/recommendation-european-electronic-health-record-exchange-format

9. EUR-Lex (2022). 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). EUR-Lex-32022R0868-EN-EUR-Lex. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R0868

10. EUR-Lex (2016). 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). EUR-Lex-32016R0679-EN-EUR-Lex. Available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj

11. EUR-Lex (2022). Proposal for a regulation of the European Parliament and of the council on harmonised rules on fair access to and use of data (data act). EUR-Lex-52022PC0068-EN-EUR-Lex. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2022%3A68%3AFIN

12. RBC Capital Markets (2023). The healthcare data explosion. Available at: https://www.rbccm.com/en/gib/healthcare/episode/the_healthcare_data_explosion (Accessed November 2023).

13. Gerry McGovern (2016). Worldwide waste. Available at: https://gerrymcgovern.com/world-wide-waste/

14. MyData Global (2022). Understanding-MyData-Operators-2022. Available at: https://www.mydata.org/wp-content/uploads/2022/07/Understanding-MyData-Operators-2022-1.pdf

15. Micheli M, Farrell E, Carballa SB, Posada SM, Signorelli S, Vespe M (2023). Mapping the landscape of data intermediaries. Available at: https://publications.jrc.ec.europa.eu/repository/handle/JRC133988

16. Abad-Navarro F, Martínez-Costa C. A knowledge graph-based data harmonization framework for secondary data reuse. *Comput Methods Prog Biomed*. (2024) 243:107918. doi: 10.1016/j.cmpb.2023.107918

17. Booth DKnowMED, Inc. (2013). RDF as a universal healthcare exchange language. Available at: http://dbooth.org/2013/rdf-as-univ/rdf-as-univ.pdf

18. EUR-Lex (2017). Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices. 2017. EUR-Lex-32017R0745-EN-EUR-Lex. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32017R0745

19. European Interoperability Framework—Implementation Strategy (2017). Communication from the commission to the European Parliament, the council, the European economic and social committee and the committee of the regions. Available at: https://eur-lex.europa.eu/resource.html?uri=cellar:2c2f2554-0faf-11e7-8a35-01aa75ed71a1.0017.02/DOC_1&format=PDF

20. HIMSS (2023). Interoperability in healthcare. Available at: https://www.himss.org/resources/interoperability-healthcare

21. de Mello BH, Rigo SJ, da Costa CA, da Rosa RR, Donida B, Bez MR, et al. Semantic interoperability in health records standards: a systematic literature review. *Heal Technol*. (2022) 12:255–72. doi: 10.1007/s12553-022-00639-w

22. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. *Commun Med*. (2023) 3:141. doi: 10.1038/s43856-023-00370-1

23. Heer J, Hellerstein JM, Kandel S. *Encyclopedia of Big Data Technologies*. Cham: Springer. (2018).

24. Dong H, Falis M, Whiteley W, Alex B, Matterson J, Ji S, et al. Automated clinical coding: what, why, and where we are? *NPJ Digit Med*. (2022) 5:1–8. doi: 10.1038/s41746-022-00705-7

25. Wittner R, Mascia C, Gallo M, Frexia F, Müller H, Plass M, et al. Lightweight distributed provenance model for complex real-world environments. *Sci Data*. (2022) 9:503. doi: 10.1038/s41597-022-01537-6

26. Plass M, Wittner R, Holub P, Frexia F, Mascia C, Gallo M, et al. Provenance of specimen and data—a prerequisite for AI development in computational pathology. *New Biotechnol*. (2023) 78:22–8. doi: 10.1016/j.nbt.2023.09.006

27. Albertoni R. (2020). Data catalog vocabulary (DCAT)—version 2. Available at: https://www.w3.org/TR/vocab-dcat-2/

28. Ehrlinger L., Schrott J., Wöß W. "DSD: the data source description vocabulary" *Database and Expert Systems Applications—DEXA 2023 Workshops. DEXA 2023. Communications in Computer and Information Science*. 1872. In: G. Kotsis, A. M. Tjoa, I. Khalil, B. Moser, A. Mashkoor, J. Sametinger, et al., editors. Cham: Springer (2023).

29. RDF Mapping Language (RML) (2022). In: A Dimou, M V Sande, editors. Available at: https://rml.io/specs/rml/

30. Profiles defined as part of the IPS Guide. (2024). International patient summary implementation guide v1.1.0. Available at: https://build.fhir.org/ig/HL7/fhir-ips/profiles.html#list-of-profiles

31. SHACL (2017). Advanced features. Available at: https://www.w3.org/TR/shacl-af/

32. SPHN (2023). SPHN—Swiss personalized health network. SPHN Dataset & RDF Schema 2023 release. Available at: https://sphn.ch/2023/03/20/sphn-dataset-rdf-schema-2023-release/

33. Holzinger A, Kargl M, Kipperer B, Regitnig P, Plass M, Muller H. Personas for artificial intelligence (AI) an open source toolbox. *IEEE Access*. (2022) 10:26. doi: 10.1007/978-3-031-39689-2_1

34. Raab R, Küderle A, Zakreuskaya A, Stern AD, Klucken J, Kaissis G, et al. Federated electronic health records for the European health data space. *Lancet Digit Health*. (2023) 5:e840–7. doi: 10.1016/S2589-7500(23)00156-5

35. Hswen Y, Voelker R. Electronic health records failed to make clinicians' lives easier-will AI technology succeed? *JAMA*. (2023) 330:1509–11. doi: 10.1001/jama.2023.19138

36. Ghafur S, Van Dael J, Leis M, Darzi A, Sheikh A. Public perceptions on data sharing: key insights from the UK and the USA. *Lancet Digit Health*. (2020) 2:e444–6. doi: 10.1016/S2589-7500(20)30161-8

37. The PRECISE4Q (2023). Factsheet. Available at: https://precise4q.eu/resources/ (Accessed November 2023).

38. TEHIK (2021). Estonia. Environment for creating data forwarding standards. Available at: https://www.tehik.ee/en/environment-for-creating-data-forwarding-standards

39. KMEHR Data Standards (2023). Portail des services de l'eSanté. Belgium. Available at: https://www.ehealth.fgov.be/standards/kmehr/en (Accessed November 2023).

40. Data resources catalogue (2023). Ladataan, Finland. Available at: https://aineistokatalogi.fi/catalog

41. Sheth A, Roy K, Gaur M (2023). Neurosymbolic AI—why, what and how. arXiv [Preprint]. Available at: http://arxiv.org/abs/2305.00813

42. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ. Foundation models for generalist medical artificial intelligence. *Nature*. (2023) 616:259–65. doi: 10.1038/s41586-023-05881-4

43. Idea4RC (2023). Deliverable 2.5. Metadata Taxonomy. Available at: https://www.idea4rc.eu/wp-content/uploads/2023/11/IDEA4RC_D2.5_Metada_taxonomy_V1_FINAL.pdf

44. Brands MR, Gouw SC, Driessens MHE. Personal health records: a promising tool? *Ned Tijdschr Geneeskd*. (2023) 16:167.

45. Shaping Europe's digital future (2023). EU digital identity wallet pilot implementation. Available at: https://digital-strategy.ec.europa.eu/en/policies/eudi-wallet-implementation

46. Bobev T, Dessers VK, Ducuing C, Fierens M, Palumbo A, Peeters B, et al. (2023). White paper on the definition of data intermediation services. Available at: https://papers.ssrn.com/abstract=4589987

47. Deloitte Switzerland (2023). Deloitte pharma study: drop-off in returns on R&D investments – sharp decline in peak sales per asset. Available at: https://www2.deloitte.com/ch/en/pages/press-releases/articles/deloitte-pharma-study-drop-off-in-returns-on-r-and-d-investments-sharp-decline-in-peak-sales-per-asset.html

48. DS4Skills (2022) Data Space For Skills. Available at: https://www.skillsdataspace.eu/https://www.skillsdataspace.eu/

frontiers | Frontiers in Medicine

TYPE Methods
PUBLISHED 16 May 2024
DOI 10.3389/fmed.2024.1378866

OPEN ACCESS

EDITED BY
Gokce Banu Laleci Erturkmen,
Software Research and Development Consulting, Türkiye

REVIEWED BY
Pantelis Natsiavas,
Institute of Applied Biosciences, Greece
Martin Hofmann-Apitius,
Fraunhofer Institute for Algorithms and Scientific Computing (FHG), Germany

*CORRESPONDENCE
Fabian Prasser
✉ fabian.prasser@charite.de

RECEIVED 30 January 2024
ACCEPTED 02 May 2024
PUBLISHED 16 May 2024

CITATION
Wirth FN, Abu Attieh H and Prasser F (2024) OHDSI-compliance: a set of document templates facilitating the implementation and operation of a software stack for real-world evidence generation.
Front. Med. 11:1378866.
doi: 10.3389/fmed.2024.1378866

OHDSI-compliance: a set of document templates facilitating the implementation and operation of a software stack for real-world evidence generation

Felix N. Wirth, Hammam Abu Attieh and Fabian Prasser*

Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Center of Health Data Science, Berlin, Germany

Introduction: The open-source software offered by the Observational Health Data Science and Informatics (OHDSI) collective, including the OMOP-CDM, serves as a major backbone for many real-world evidence networks and distributed health data analytics platforms. While container technology has significantly simplified deployments from a technical perspective, regulatory compliance can remain a major hurdle for the setup and operation of such platforms. In this paper, we present OHDSI-Compliance, a comprehensive set of document templates designed to streamline the data protection and information security-related documentation and coordination efforts required to establish OHDSI installations.

Methods: To decide on a set of relevant document templates, we first analyzed the legal requirements and associated guidelines with a focus on the General Data Protection Regulation (GDPR). Moreover, we analyzed the software architecture of a typical OHDSI stack and related its components to the different general types of concepts and documentation identified. Then, we created those documents for a prototypical OHDSI installation, based on the so-called Broadsea package, following relevant guidelines from Germany. Finally, we generalized the documents by introducing placeholders and options at places where individual institution-specific content will be needed.

Results: We present four documents: (1) a record of processing activities, (2) an information security concept, (3) an authorization concept, as well as (4) an operational concept covering the technical details of maintaining the stack. The documents are publicly available under a permissive license.

Discussion: To the best of our knowledge, there are no other publicly available sets of documents designed to simplify the compliance process for OHDSI deployments. While our documents provide a comprehensive starting point, local specifics need to be added, and, due to the heterogeneity of legal requirements in different countries, further adoptions might be necessary.

KEYWORDS

health data analytics, real-world evidence, observational health data science, regulatory compliance, data protection

1 Introduction

1.1 Background

Collecting and analyzing data from real-world healthcare settings at a broad scale can provide new insights into patient outcomes, treatment efficacy, and healthcare practices (1). This usually necessitates bringing together data from several healthcare institutions, which requires the implementation of or mapping to data standards, as well as approaches for ethical and data protection compliant access (2). One common solution for the latter challenge is federation, where the analysis is brought to the data instead of bringing the data to the analysis (3). This is, for example, implemented by SHRINE (4), DataSHIELD (5) and the Observational Health Data Sciences and Informatics (OHDSI) (6) initiative. OHDSI is an international, multidisciplinary community of researchers and healthcare professionals to enable data standardization, analysis, and insight discovery from large-scale health datasets, launched in 2013. The community distributes a set of open-source software tools to represent and analyze data in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), which makes extensive use of terminologies and ontologies, such as Logical Observation Identifiers Names and Codes (LOINC) or Systematized Nomenclature of Medicine (SNOMED) Clinical Terms (CT) (7). While the term OMOP describes the now discontinued collaboration that originally developed the CDM, the term OMOP-CDM refers to the further developed version that forms the current technical cornerstone of OHDSI. The EHDEN project has funded the deployment of the OMOP-CDM and the OHDSI software stack across Europe (8). Moreover, the OMOP-CDM will also play an important role in the upcoming European Health Data Space (EHDS; see Section "Discussion"). The EHDS is planned as a large-scale ecosystem facilitating better exchange and access to different types of health data throughout the European Union (EU). EHDS pillar I focuses on primary healthcare data use, i.e., data sharing for healthcare delivery. EHDS pillar II focuses on secondary use of health data, e.g., analysis for research, policy-making or drug safety (9).

Setting up an OHDSI node can involve significant efforts, in particular for the required mapping to standards. However, technical and data integration challenges are not the only obstacles faced when connecting to data sharing networks [for one example for the various technical challenges see (10)]. Legal and regulatory compliance is another important issue (11, 12). National and international data protection laws as well as ethical guidelines must be considered. Important examples include the US Health Insurance Portability and Accountability Act (HIPAA) (13) and the European Union (EU) General Data Protection Regulation (GDPR) (14). To fulfill central requirements, concepts need to be developed and documented for ensuring the confidentiality of the processed healthcare data. An important example is the so-called Record of Processing Activities (ROPA), which needs to be created according to the GDPR, but also according to laws in the United Kingdom (15, 16), Australia (17) or Thailand (18). Amongst other aspects, a ROPA typically describes the processed categories of data and details information flows as well as the technical and organizational security measures implemented, although slight variations might exist between the requirements in different countries. Moreover, information security plays an important role, with relevant standards also requiring documentation of the measures taken (19). Important examples include the International Standards Organization (ISO) Standard 27001 (20), (2) the US National Institute of Standards and Technology (NIST) Cybersecurity Framework (21) or (3) the Health Information Trust Alliance Common Security Framework (HITRUST CSF) (22).

1.2 Objective

It is well known that conceptualizing and documenting the secure operation of data processing platforms can be challenging (23, 24). Research has shown that even reading and comprehending such documents can be difficult (25–27). As a result, different guidelines and templates have been developed (see Section *Comparison with prior work*). However, those are usually generic in nature and not directly applicable to the establishment of an OHDSI node. The objective of the work described in this paper, was to conceptualize an approach specifically for common OHDSI deployments. Moreover, we developed document templates that can be customized to local requirements. We focus on documents for a general OHDSI setup. Depending on the nature of projects that use this infrastructure as well as local requirements, additional documents might be needed for the individual studies performed.

2 Methods

2.1 Overview of the OHDSI tools

The main tools provided by OHDSI are focused on (1) establishing a common data model with clearly defined structure and semantics, as well as (2) assisting medical researchers and data scientists in extracting knowledge from this data. The OMOP-CDM is the central pillar of OHDSI, providing a standardized database schema and a set of terminologies with which heterogeneous data from different sources can be integrated to provide comparability across studies and institutions (28). As a result, OHDSI forms a global network allowing for large-scale distributed studies to be performed. A common database management system for instances of the OMOP-CDM is *PostgreSQL* (29). In addition, the following tools are provided for data mapping:

- *WhiteRabbit* is a tool to scan and describe source data.
- *Rabbit in a Hat* supports structural mapping between source data and the OMOP-CDM.
- *USAGI* has been designed to support semantic standardization and terminology mapping.
- *Athena* is as a publicly available web service providing access to the vocabulary used by the OMOP-CDM.

We note that OHDSI does not provide a standard tool for extracting, transforming and loading (ETL) data, but focuses on tools for specifying the transformations and mappings needed. A common way of deploying a standard OHDSI stack is the container-based *Broadsea* distribution (30). An overview of a typical set of components in Broadsea is provided in Figure 1.

As can be seen, a common installation contains the following additional infrastructure components:

- A *PostgreSQL* database for storing configuration options and study designs.

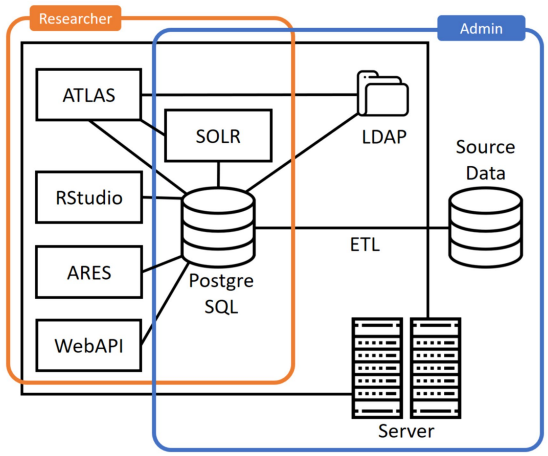

FIGURE 1
Common architecture of an OHDSI implementation.

TABLE 1 Overview of the document templates.

Document title	Short description
Record of processing activities	Description of the data processing activities and protection measures.
Information security concept	Description of information security measures.
Concept of operations	Description of processes and responsibilities when operating the installation.
Authorization concept	Description of groups of user roles and their permissions as well as a description of the process for requesting access to the database.

- *Apache SOLR* for searching through the vocabulary.
- *OpenLDAP* for authentication and authorization.

Based on this basic infrastructure and the CDM, the Broadsea distribution offers further applications for accessing and analyzing the data:

- *WebAPI* is a RESTful service layer for accessing and analyzing data stored in the OMOP-CDM.
- *ATLAS* is a web-based tool for conducting scientific analyses.
- *ARES* is a system facilitating data exploration, characterization, and quality assessments.
- *RStudio* for analyzing data using the statistical programming language R. Broadsea comes with a range of R-packages, such as Shiny for developing interactive web applications and HADES for analyzing data from the OMOP-CDM.

In summary, researchers can work with data stored in the OMOP-CDM through ATLAS and specific R packages. ATLAS provides graphical access to a variety of OHDSI tools and functions, trading usability off against the flexibility of the analyses that can be performed. In addition, analyses can be performed in R using a set of provided packages and APIs, providing more flexibility in working with the data but requiring programming and data science skills.

2.2 Development process

We first identified a set of documents usually required to deploy and operate research systems at German university hospitals. As a basis, these include (1) a description of the processing activities and the technical and organizational measures taken in regards to data protection, (2) an analysis of information security risks and security-related measures taken, (3) a description of processes and responsibilities for maintaining and operating the system. We note that these documents need to be updated regularly following a continuous improvement process.

Next, we related those documents to the systems and processes covered by the common architecture described in the previous section. Data protection aspects were described with a specific focus on systems holding or processing individual-level health data, reflecting requirements by Article 30 GDPR on the content of the description of processing activities. Information security as well as operation of the stack was covered for the complete installation, oriented towards the information security basic protection methodology provided by the German government. Moreover, another document was developed to describe and implement governance processes for use of the data available in the CDM. Finally, we transformed the documents into customizable templates and uploaded them into a version-controlled repository.

3 Results

3.1 Overview

Table 1 provides an overview of the different document templates developed and provided through a GitHub repository (31).

3.2 Record of processing activities

A general description of the software architecture, data flows and processing activities as well as protection measures taken forms the basis of most compliance framework for medical research systems. Thus, as a first component, we developed a template for a Record of Processing Activities (ROPA) for OHDSI installations. As outlined above, ROPAs or related documents are required in most jurisdictions. In this work, we base the content on the requirements outlined in Article 30 of the GDPR and provide information about the personal data processed, the purposes of the processing, retention periods and further relevant details. In the event of legal or data protection audits, the document can be used as a basis to demonstrate compliance and it can also serve as a communication measure for coordinating OHDSI-related activities with an institution's Data Protection Officer.

3.3 Information security concept

While data protection and the ROPA template emphasizes the handling of personal data in a way that respects the rights and

expectations of the data subjects, information security focuses on protecting data from unauthorized access and further threats more relevant to the organization itself than to the data subjects. The well-known ISO/IEC 27000 standard emphasizes confidentiality, integrity, and availability, but also adds further aspects, such as authenticity, accountability, non-repudiation, and reliability (32).

To cover these aspects, we provide a template for describing information security-related properties of OHDSI installations. The template is pragmatic and designed to complement existing information security guidelines at the institution operating the installation. It contains a risk analysis of basic processes carried out with OHDSI installations, such as data transformation, loading, and usage, and systematically describes relevant information security measures. As an example, we use modules from the "Basic Protection" methodology of the Federal Office for Information Security in Germany. While there are some differences to the ISO 27000 set of standards, the "Basic Protection" methodology provides a solid foundation of security controls for achieving ISO 27001 compliance. An organization that already applies ISO 27000 can, for example, benefit from our documents through the included risk assessments and lists of relevant security controls that can inform local information security management processes. The document can also support coordination with an institutions Chief Information Security Officer (CISO).

3.4 Concept of operations

In addition to a sound and secure setup of an OHDSI node, also the operation of the platform needs to be conceptualized and described. Relevant processes also include the continuous improvement process for data protection and information security-related aspects already described above. In addition, the installed components and their configurations need to be kept up to date, user accounts need to be managed and backups need to be performed. The template for an operational concept includes suggestions for those processes, tailored towards the OHDSI components.

3.5 Authorization concept

How access requests by researchers to the OHDSI tools are handled and what governance rules are implemented is an important aspect of compliance. Consequently, we also developed a template for a guideline on how this is implemented. The template describing the access request process describes the duties of administrative personnel responsible for overseeing user access and processes for regular review and removal of outdated permissions. Additionally, it describes the steps researchers must follow to obtain access for conducting studies, including obtaining necessary approvals. In addition to researchers accessing the OHDSI tools, there are further types of personnel involved that need to access the installation for operational purposes. As this is a critical aspect, the proposed template describes all relevant roles, their responsibilities, and access permissions. The template outlines processes for nominating administrators, setting up user access and revoking them upon project completion or staff changes. Moreover, password guidelines and rules for timeouts of sessions are included.

Figure 2 illustrates how the developed document templates cover different components and aspects of a common OHDSI installation. As can be seen, the ROPA focuses on the general setup that processes personal data, while the information security concept and related templates cover all components. Access management focuses specifically on humans involved in the maintenance and use of an installation.

3.6 Customization and document management

We have developed the templates as Markdown files and provide them in the form of a Git repository. Markdown is a lightweight markup language, designed to be easy to write and read, with the ability to present the document content in many different forms. For example, the documents provided can be compiled into PDF files using open-source tools, such as Pandoc. If visual editing is needed, tools like Pandoc can also be used to convert the markdown files into formats suited for word processors, such as the Open Document Format. We recommend to use the templates in their Markdown version, however, as this naturally enables keeping track of changes in versioned repositories, such as Git.

4 Discussion

4.1 Principal results

We presented a set of templates for setting up and maintaining OHDSI installations in compliance with data protection and information security requirements, also covering data governance aspects. The document templates are public available under a permissive license. The templates are meant to provide a starting point and need to be filled out accordingly and potentially extended or modified to comply with local policies or legal requirements. We have successfully executed this process at Charité – Universitätsmedizin Berlin.

4.2 Comparison with related work

Several institutions or research groups have suggested compliance-oriented document templates for data processing in general or for medical research contexts. Examples include data protection guidelines, see (33) for an example, and templates for institutional review board protocols, see (34) for an example, and information security aspects, see (35) as an example. Quite a lot of the documents are tailored towards specific jurisdictions and published in languages other than English [e.g., (33, 36)]. Our work is different in that it focuses on a typical deployment of a common medical research platform and that its content has been, in large parts, abstracted away from country-specific requirements. Previous work has also focused on compliance for deployments of specific research systems (see the work by Wallace et al. (37) and by Budin-Ljøsne et al. (38) for an example on the DataSHIELD software). To the best of our knowledge, our work is the first to target OHDSI deployments. Governance models have also been studied in the literature. For example, Holmes et al. have

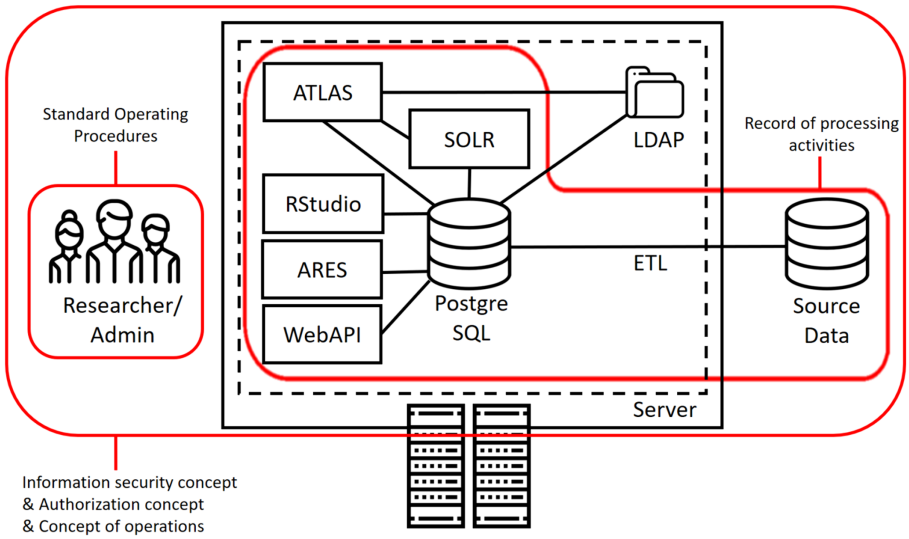

FIGURE 2
Role of the different documents in a common OHDSI deployment.

presented an overview on governance models for federated research (39). The authors propose a framework with which governance models can be assessed and compared considering different aspects. Pavlenko et al. have focused on data governance for health data warehouses (40).

On a more general level, ethical and legal challenges in data-driven biomedical research have also been studied extensively. For instance, Wang et al. discussed several privacy-enhancing technologies and argue that accountability and informed consent are among the most relevant ethical challenges (41). Arellano et al. conduct a review on privacy regulations, patient perspectives as well as consent practices and their interaction with technology (42). They cover questions, such as under which circumstances consent can be considered ethical. Lamas et al. have argued that ethical and legal frameworks are often not fitting well to common scenarios in the secondary use of health data and the development of health data warehouses (43).

Kalkman et al. have studied the sharing practices for compliance-related documentation (44). The authors found that documents like the ones presented in this work is not common.

The OHDSI software stack addressed in the work described in this paper, is expected to play an important role in the upcoming EHDS and is promoted by a range of institutions. For example, the DARWIN initiative - an infrastructure built by the European Medicines Agency (EMA) to enable the secondary use of real-world data - is based on the OMOP-CDM and can be considered one of the first functional parts of the EHDS (45). The Joint Action Towards the European Health Data Space (TEHDAS) is another project with significant contributions to the shaping of the EHDS. Recently, also Health Level Seven (HL7) International and OHDSI have started a collaboration to work on a joint common data model for sharing information for healthcare and research (46).

4.3 Limitations and future work

One limitation of our work is that it has been designed with European and German requirements in mind, although we aimed at generalizing and abstracting away specifics. We note, however, that there are many similarities between relevant laws and regulations in different parts of the world (*cf.* similarities between the California Consumer Privacy Act or the EU-US Data Privacy Framework and the GDPR). We stress again that our templates must hence be regarded as a starting point and might need adaptions. In future work, we hope to be able to extend and adjust our templates based on feedback from their application in different contexts and jurisdictions.

Another limitation of our work is that we currently did not explicitly include a document template for a Data Protection Impact Assessment (DPIA). Under the GDPR a DPIA is necessary for processing activities resulting in a high risk for the privacy of the data subjects. If an institution decides that this is needed for an OHDSI installation, tools, such as the one presented in (47), can be used and information from the documents provided through our work can be reused.

One interested area for future work is to more thoroughly study the compliance of data sharing processes within the OHDSI network. For example, it is not trivial to decide when aggregated statistics can be considered to be anonymous data. The OHDSI collective could be supported by a guideline providing legal and technical assessments of commonly used methods.

5 Summary and conclusion

In this paper, we introduced a set of document templates designed to facilitate the implementation and operation of an OHDSI software stack for generating real-world evidence in compliance with data protection and information security requirements. These templates, tailored for typical OHDSI deployments, include crucial documents, such as a Record of Processing Activities, an Information Security Concept, and an Operational Concept. Our work addresses a significant gap by providing a framework adaptable to different institutional and legal requirements, thereby simplifying compliance processes for OHDSI

deployments. Despite being primarily oriented towards European and German regulations, our templates can serve as an adaptable starting point for organizations worldwide. Future efforts will focus on refining these templates based on feedback received and extending their scope to further compliance aspects.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: The templates created can be found in the associated GitHub repository: https://github.com/BIH-MI/ohdsi-compliance.

Author contributions

FW: Conceptualization, Resources, Writing – original draft, Writing – review & editing. HA: Resources, Writing – original draft, Writing – review & editing. FP: Conceptualization, Resources, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work has been partly funded by the German Federal Ministry of Education and Research under grant number 01ZZ2316B (PrivateAIM).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-world evidence - what is it and what can it tell us? *N Engl J Med*. (2016) 375:2293–7. doi: 10.1056/NEJMsb1609216

2. Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, et al. Electronic health records: new opportunities for clinical research. *J Intern Med*. (2013) 274:547–60. doi: 10.1111/joim.12119

3. Wirth FN, Meurers T, Johns M, Prasser F. Privacy-preserving data sharing infrastructures for medical research: systematization and comparison. *BMC Med Inform Decis Mak*. (2021) 21:242–55. doi: 10.1186/s12911-021-01602-x

4. McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, et al. SHRINE: enabling nationally scalable multi-site disease studies. *PLoS One*. (2013) 8:55811. doi: 10.1371/journal.pone.0055811

5. Gaye A, Marcon Y, Isaeva J, LaFlamme P, Turner A, Jones EM, et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. *Int J Epidemiol*. (2014) 43:1929–44. doi: 10.1093/ije/dyu188

6. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform*. (2015) 216:574–8. doi: 0.3233/978-1-61499-564-7-574

7. Reich C, Ostropolets A, Ryan P, Rijnbeek P, Schuemie M, Davydov A, et al. OHDSI standardized vocabularies—a large-scale centralized reference ontology for international data harmonization. *J Am Med Inform Assoc*. (2024) 31:583–90. doi: 10.1093/jamia/ocad247

8. Voss EA, Blacketer C, van Sandijk S, Moinat M, Kallfelz M, van Speybroeck M, et al. European Health Data & Evidence Network-learnings from building out a standardized international health data network. *J Am Med Inform Assoc JAMIA*. (2023) 31:209–19. doi: 10.1093/jamia/ocad214

9. Shabani M. Will the European health data space change data sharing rules? *Science*. (2022) 375:1357–9. doi: 10.1126/science.abn4874

10. Welten S, Weber S, Holt A, Beyan O, Decker S. Will it run?-a proof of concept for smoke testing decentralized data analytics experiments. *Front Med*. (2023) 10:1305415. doi: 10.3389/fmed.2023.1305415

11. Vis DJ, Lewin J, Liao RG, Mao M, Andre F, Ward RL, et al. Towards a global cancer knowledge network: dissecting the current international cancer genomic sequencing landscape. *Ann Oncol*. (2017) 28:1145–51. doi: 10.1093/annonc/mdx037

12. Khalil R, Macdonald JC, Gustafson A, Aljuburi L, Bisordi F, Beakes-Read G. Walking the talk in digital transformation of regulatory review. *Front Med*. (2023) 10:1233142. doi: 10.3389/fmed.2023.1233142

13. Act Accountability. Health insurance portability and accountability act of 1996. *Public Law*. (1996) 104:191.

14. Regulation Protection. 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. *Off J Eur Union OJ*. (2016) 59:294.

15. UK GDPR. (2020). Available at: https://www.legislation.gov.uk/eur/2016/679/contents

16. Data Protection Act (2018) Available at: https://www.legislation.gov.uk/ukpga/2018/12/contents

17. Privacy Act. (1988). Available at: https://www.legislation.gov.au/Details/C2022C00135

18. OneTrust DataGuidance. Comparing privacy laws: GDPR v. Thai Personal Data Protection Act [Internet]. (2024). Available at: https://www.dataguidance.com/sites/default/files/gdpr_v_thailand_updated.pdf

19. Azmi R, Tibben W, Win KT. Review of cybersecurity frameworks: context and shared concepts. *J Cyber Policy*. (2018) 3:258–83. doi: 10.1080/23738871.2018.1520271

20. ISO/IEC. *Information technology - security techniques - information security management systems - requirements (ISO/IEC 27001:2022)*. Geneva: International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC) (2022).

21. Barrett MP. *Framework for improving critical infrastructure cybersecurity*. Gaithersburg: National Institute of Standards and Technology (2018).

22. HITRUST Alliance. *HITRUST common security framework (CSF version 9.0) [Internet]*. Frisco: HITRUST Alliance (2021).

23. Dierks C, Kircher P, Husemann C, Kleinschmidt J, Haase M. *Data privacy in european medical research: A contemporary legal opinion*. Berlin: MWV Medizinisch Wissenschaftliche Verlagsgesellschaft (2021).

24. International Association of Privacy Professional. Measuring privacy operations [Internet] (2019). Available at: https://iapp.org/media/pdf/resource_center/measuring_privacy_operations_2019.pdf

25. Becher SI, Benoliel U. Law in books and law in action: the readability of privacy policies and the GDPR In: K Mathis and A Tor, editors. *Consumer law and economics*. Berlin: Springer (2021). 179–204. doi: 10.1007/978-3-030-49028-7_9

26. McDonald AM, Cranor LF. The cost of reading privacy policies. *J Law Policy Inf Soc*. (2008) 4:543–68.

27. Benoliel U, Becher S. The duty to read the unreadable. *Boston Coll Law Rev*. (2019) 60:2255–96. doi: 10.2139/ssrn.3313837

28. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc JAMIA*. (2012) 19:54–60. doi: 10.1136/amiajnl-2011-000376

29. Obe RO, Hsu LS. *PostgreSQL: Up and running: A practical guide to the advanced open source database. 3rd* ed. Beijing, Boston, Farnham, Sebastopol, Tokyo: O'Reilly (2018). 294 p.

30. OHDSI. Software Tools (2024). Available at: https://www.ohdsi.org/software-tools/

31. Wirth FN, Abu Attieh H, Prasser F. OHDSI compliance. (2024) Available at: https://github.com/BIH-MI/ohdsi-compliance

32. ISO/IEC. *Information technology - security techniques - information security management systems - overview and vocabulary (ISO/IEC 27000:2020)*. Geneva: International Organization for Standardization (ISO) and international Electrotechnical commission (IEC) (2020).

33. Pommerening K, Drepper J, Helbing K, Ganslandt T. *Leitfaden zum Datenschutz in medizinischen Forschungsprojekten: Generische Lösungen der TMF 2.0*. Berlin: MWV Medizinisch Wissenschaftliche Verlagsgesellschaft (2014).

34. National Institutes of Health. Protocol template for secondary research with biospecimens, data and/or medical records only [Internet]. Available at: https://ohsrp.nih.gov/confluence/download/attachments/67273200/Secondary%20Research%20Protocol%20Template.docx?api=v2

35. Center for Internet Security. NIST Cybersecurity Framework Policy Template Guide [Internet]. Available at: https://www.cisecurity.org/-/jssmedia/Project/cisecurity/cisecurity/data/media/files/uploads/2021/11/NIST-Cybersecurity-Framework-Policy-Template-Guide-v2111Online.pdf

36. Frielitz F, Storm N, Hiort O, Katalinic A, von Sengbusch S. Die Erstellung eines Datenschutzkonzeptes: eine Anleitung für telemedizinische Versorgungsprojekte. *Bundesgesundheitsblatt*. (2019) 62:479:485. doi: 10.1007/s00103-019-02918-w

37. Wallace SE, Gaye A, Shoush O, Burton PR. Protecting personal data in epidemiological research: DataSHIELD and UK law. *Public Health Genomics*. (2014) 17:149–57. doi: 10.1159/000360255

38. Budin-Ljøsne I, Burton P, Isaeva J, Gaye A, Turner A, Murtagh MJ, et al. DataSHIELD: an ethically robust solution to multiple-site individual-level data analysis. *Public Health Genomics*. (2015) 18:87–96. doi: 10.1159/000368959

39. Holmes JH, Elliott TE, Brown JS, Raebel MA, Davidson A, Nelson AF, et al. Clinical research data warehouse governance for distributed research networks in the USA: a systematic review of the literature. *J Am Med Inform Assoc JAMIA*. (2014) 21:730–6. doi: 10.1136/amiajnl-2013-002370

40. Pavlenko E, Strech D, Langhof H. Implementation of data access and use procedures in clinical data warehouses. A systematic review of literature and publicly available policies. *BMC Med Inform Decis Mak*. (2020) 20:157. doi: 10.1186/s12911-020-01177-z

41. Wang S, Bonomi L, Dai W, Chen F, Cheung C, Bloss CS, et al. Big data privacy in biomedical research. *IEEE Trans Big Data*. (2016) 6:296–308. doi: 10.1109/TBDATA.2016.2608848

42. Arellano AM, Dai W, Wang S, Jiang X, Ohno-Machado L. Privacy policy and technology in biomedical data science. *Annu Rev Biomed Data Sci*. (2018) 1:115–29. doi: 10.1146/annurev-biodatasci-080917-013416

43. Lamas E, Barh A, Brown D, Jaulent MC. Ethical, legal and social issues related to the health data-warehouses: re-using health data in the research and public health research. *Stud Health Technol Inform*. (2015) 210:719–23. doi: 10.3233/978-1-61499-512-8-719

44. Kalkman S, Mostert M, Udo-Beauvisage N, van Delden JJ, van Thiel GJ. Responsible data sharing in a big data-driven translational research platform: lessons learned. *BMC Med Inform Decis Mak*. (2019) 19:283. doi: 10.1186/s12911-019-1001-y

45. Arlett P, Kjær J, Broich K, Cooke E. Real-world evidence in EU medicines regulation: enabling use and establishing value. *Clin Pharmacol Ther*. (2022) 111:21–3. doi: 10.1002/cpt.2479

46. OHDSI. HL7 International and OHDSI announce collaboration to provide single common data model for sharing information in clinical care and observational research (2024). Available at: https://www.ohdsi.org/ohdsi-hl7-collaboration/

47. CNIL The open source PIA software helps to carry out data protection impact assessment (2023). Available at: https://www.cnil.fr/en/open-source-pia-software-helps-carry-out-data-protection-impact-assessment

frontiers | Frontiers in Medicine

TYPE Technology and Code
PUBLISHED 27 May 2024
DOI 10.3389/fmed.2024.1386689

OPEN ACCESS

EDITED BY
Christine Gispen-de Wied,
Gispen4RegulatoryScience, Netherlands

REVIEWED BY
Markus Wolfien,
Technical University Dresden, Germany
Manisha Mantri,
Center for Development of Advanced Computing (C-DAC), India

*CORRESPONDENCE
Mert Gencturk
✉ mert@srdc.com.tr

†These authors have contributed equally to this work

RECEIVED 15 February 2024
ACCEPTED 13 May 2024
PUBLISHED 27 May 2024

CITATION
Gencturk M, Laleci Erturkmen GB, Akpinar AE, Pournik O, Ahmad B, Arvanitis TN, Schmidt-Barzynski W, Robbins T, Alcantud Corcoles R and Abizanda P (2024) Transforming evidence-based clinical guidelines into implementable clinical decision support services: the CAREPATH study for multimorbidity management. *Front. Med.* 11:1386689. doi: 10.3389/fmed.2024.1386689

Transforming evidence-based clinical guidelines into implementable clinical decision support services: the CAREPATH study for multimorbidity management

Mert Gencturk[1*†], Gokce B. Laleci Erturkmen[1†], A. Emre Akpinar[1,2], Omid Pournik[3], Bilal Ahmad[3], Theodoros N. Arvanitis[3,4], Wolfgang Schmidt-Barzynski[5], Tim Robbins[4], Ruben Alcantud Corcoles[6,7] and Pedro Abizanda[6,7,8]

[1]SRDC Software Research & Development and Consultancy Corporation, Ankara, Türkiye, [2]Department of Computer Engineering, Middle East Technical University, Ankara, Türkiye, [3]Department of Electronic, Electrical and Systems Engineering, School of Engineering, University of Birmingham, Birmingham, United Kingdom, [4]Digital & Data Driven Research Unit, University Hospitals Coventry & Warwickshire NHS Trust, Coventry, United Kingdom, [5]University Hospital OWL, Bielefeld, Germany, [6]Geriatrics Department, Complejo Hospitalario Universitario de Albacete, Albacete, Spain, [7]CIBER de Fragilidad y Envejecimiento Saludable (CIBERFES), Instituto de Salud Carlos III, Madrid, Spain, [8]Facultad de Medicina de Albacete, Universidad de Castilla-La Mancha, Albacete, Spain

Introduction: The CAREPATH Project aims to develop a patient-centered integrated care platform tailored to older adults with multimorbidity, including mild cognitive impairment (MCI) or mild dementia. Our goal is to empower multidisciplinary care teams to craft personalized holistic care plans while adhering to evidence-based guidelines. This necessitates the creation of clear specifications for clinical decision support (CDS) services, consolidating guidance from multiple evidence-based clinical guidelines. Thus, a co-creation approach involving both clinical and technical experts is essential.

Methods: This paper outlines a robust methodology for generating implementable specifications for CDS services to automate clinical guidelines. We have established a co-creation framework to facilitate collaborative exploration of clinical guidelines between clinical experts and software engineers. We have proposed an open, repeatable, and traceable method for translating evidence-based guideline narratives into implementable specifications of CDS services. Our approach, based on international standards such as CDS-Hooks and HL7 FHIR, enhances interoperability and potential adoption of CDS services across diverse healthcare systems.

Results: This methodology has been followed to create implementable specifications for 65 CDS services, automating CAREPATH consensus guideline consolidating guidance from 25 selected evidence-based guidelines. A total of 296 CDS rules have been formally defined, with input parameters defined as clinical concepts bound to FHIR resources and international code systems. Outputs include 346 well-defined CDS Cards, offering clear guidance for care plan activities and goal suggestions. These specifications have led to the

implementation of 65 CDS services integrated into the CAREPATH Adaptive Integrated Care Platform.

Discussion: Our methodology offers a systematic, replicable process for generating CDS specifications, ensuring consistency and reliability across implementation. By fostering collaboration between clinical expertise and technical proficiency, we enhance the quality and relevance of generated specifications. Clear traceability enables stakeholders to track the development process and ensure adherence to guideline recommendations.

KEYWORDS

clinical decision support, clinical guideline, automation, integrated care, multimorbidity, dementia, HL7 FHIR

1 Introduction

In the ever-evolving landscape of healthcare, the rising prevalence of multimorbidity combined with the complexity of medical knowledge poses significant challenges to clinical decision-making (1). Clinical guidelines, grounded in evidence-based practice, serve as essential tools for healthcare professionals in delivering optimal patient care (2). Nevertheless, the manual execution of these guidelines frequently leads to variations in practice, inefficiencies, and suboptimal outcomes, seemingly making the achievement of integrated care an overwhelming challenge (3, 4).

Integrated care is an organization-focused intervention that encompasses case-management, continuity of care, disease management, service integration, and multidisciplinary teamwork (5). It is designed to address the health and social needs of individuals living with multimorbidity, with the goal of reducing adverse healthcare outcomes, including potentially preventable hospitalizations (6). Older adults with multimorbidity can receive assistance in their own homes through Information and Communication Technologies (ICT) solutions. These solutions support them in their activities of daily living, help manage medical conditions and medications, and involve them in the healthcare process. Additionally, ICT solutions also improve physical activity and nutrition, reduce frailty, and facilitate health monitoring (7). While certain challenges remain to be addressed with these solutions, including concerns regarding data privacy and security threats, they hold significant potential for facilitating the transition from conventional medical practices to remote medicine (8, 9).

Computer-interoperable clinical guidelines play a crucial role in advancing such ICT solutions and digitizing healthcare (10). They enable the implementation of personalized clinical decision support (CDS) systems, aiding healthcare professionals in adhering to complex clinical protocols and facilitating guideline integration into daily practice. CDS systems integrate patient-specific data with evidence-based guidelines, providing real-time, personalized recommendations to healthcare providers. This integration holds great promise in streamlining clinical workflows, reducing errors, and ultimately enhancing patient outcomes (11, 12). Although CDS systems have undergone swift advancement since their initial implementation in the 1980s, their full adaptation in routine clinical practice has not yet been fully achieved for many reasons, such as ethical and legal issues, the intellectual challenge of creating knowledge, and technical dimensions of delivering CDS (13–15). Software engineers face challenges in understanding clinical guidelines due to a lack of medical expertise, which hampers their ability to automate CDS services, while clinicians without technical proficiency struggle to validate CDS implementations to ensure they align with guideline recommendations. The situation becomes more difficult when patients have multimorbidity conditions, because clinical guidelines are typically designed for individual conditions, and while they may address decision-making regarding other morbidities, they lack a systematic approach to identifying relationships between guidelines for different conditions (16).

The CAREPATH Project[1] aims to deliver a patient-centered integrated care platform to meet the needs of older patients with multimorbidity, including mild cognitive impairment (MCI) or mild dementia (MD) (17). Dementia and MCI are two of the most debilitating chronic conditions in older adults, affecting approximately 7.3 and 20% of this population (18), respectively, and leading to high-impact healthcare needs. Integrated solutions are necessary to manage this condition, especially when other chronic conditions coexist. Notably, pharmacological and non-pharmacological treatments for diseases such as heart failure or diabetes may differ in older patients with dementia compared to the general population. Developing a patient-centered integrated care platform is challenging, as the vast majority of clinical guidelines that would inform these tools typically focus on a single condition (19).

To address this challenge, Robbins et al. (20) presented clinical requirements addressing the needs of this patient group in the form of a reference, consensus clinical guideline to be used for the CAREPATH project. The development of the guideline was undertaken by a Clinical Reference Group (CRG) formed by CAREPATH project clinical partners based in Germany, Spain, Romania, and the UK. After a review of the literature to identify suitable clinical guidelines, 52 guidelines covering a range of chronic conditions, multimorbidity, and co-morbidity were assessed for quality using the AGREE II methodology (21). Based on this, 25 final guidelines were selected for examination, approval, or disapproval, grouping, and consolidation by the project CRG through a modified Delphi process (22). The final agreed guidance and actions were

1 CAREPATH Project Website, https://www.carepath.care/.

collated into the master narrative consensus guideline. The CAREPATH consensus clinical guideline provides advice, information, and actions in the following areas: overarching principles of management, MCI and dementia, physical exercise, nutrition and hydration, common use of drugs, coronary artery disease, heart failure, hypertension, diabetes, chronic kidney disease, chronic obstructive pulmonary disease (COPD), stroke, sarcopenia, frailty, and caregiver support.

CAREPATH aims to deliver integrated care solutions to multidisciplinary care teams, including health and social care providers, patients and their informal caregivers, enabling them to follow consensus guidelines in a personalized manner to create holistic care plans for older adults. The Adaptive Integrated Care Platform (AICP) is a clinician-facing application that allows healthcare professionals to review and update patient data retrieved from underlying Electronic Health Record (EHR) systems. It also enables them to assess personalized suggestions for editing the patient's care plan, such as setting clinical goals, adding or updating interventions (e.g., medications, lab orders, referrals, and patient interventions like self-monitoring activities, diet, and exercise). AICP is supported by two important components: the Technical and Semantic Interoperability Suite (TIS/SIS), which facilitates integration with EHR systems (23), and CDS services that process consensus-based guideline rules to deliver personalized care plan suggestions. Once the care plan is created, the Patient Empowerment Platform (PEP), which was developed with the involvement of patients, informal caregivers, and healthcare professionals, provides personalized assistance and guidance to patients (24). It sends reminders about care plan goals and activities, presents educational materials to reinforce treatment adherence, and collects feedback from patients via Patient Reported Outcome Measures (PROMs) to conduct geriatric assessments. Finally, the Home and Health Monitoring Platform (H/HMP) provides environment-aware services to continuously collect real-time data for early detection of onset and changes in functioning, autonomy, and underlying cognitive and physiological functions of patients.

This paper introduces a robust methodology for generating implementable specifications of CDS services, aimed at automating clinical guidelines. Through a collaborative co-creation landscape, we enable clinical experts and software engineers to jointly examine guidelines and develop human-readable CDS specifications. Our approach addresses the challenge of translating guideline suggestions into actionable guidance, bridging the gap between clinical expertise and technical implementation. Key strengths include a repeatable process, traceability, and emphasis on human-readable specifications, ensuring accessibility and alignment with evidence-based practices. By fostering interdisciplinary collaboration, our methodology empowers teams to create CDS services that effectively automate clinical guidelines while tailoring care plans to individual patient needs. Our approach is based on international standards, namely CDS-Hooks and HL7 FHIR, targeting to enhance the interoperability and potential adoption of CDS services across diverse healthcare systems.

2 Method

The methodology devised to implement clinical decision support services automating evidence-based clinical guidelines consists of four steps along with the two preliminary steps as depicted in Figure 1. The selection of best practice guidelines and the creation of consensus clinical guidelines are pre-requisites of this approach. They have already been presented in (20), hence their detailed description is out of the scope of this paper. The list of the selected best practice guidelines in CAREPATH is provided in (25–49). However, it should be noted that the methodology explained in this paper can be applied to any type and number of clinical guidelines, so the selection of guidelines is not crucial for the subsequent downstream process.

In the following subsections, we explain the details of the definition of flowchart-based rules, the definition of CDS rules in human-readable format, the definition of clinical concepts, and the preparation of CDS Hooks card templates.

2.1 Definition of flowchart-based rules

In the first step, the consensus guideline has been converted into flowchart-based rules, allowing integration into the digital platform for delivering care to dementia patients with multimorbidity. We attempted to formulate the sentences in the consensus guideline as flowchart rules, endeavoring to identify all clinical concepts. Close cooperation between technical personnel and CRG members was carried out to clearly assess the technical feasibility and clinical effectiveness of conversion of the narrative guideline into a flowchart. We have discussed and agreed on which parts of the consensus guideline can aid the clinicians if

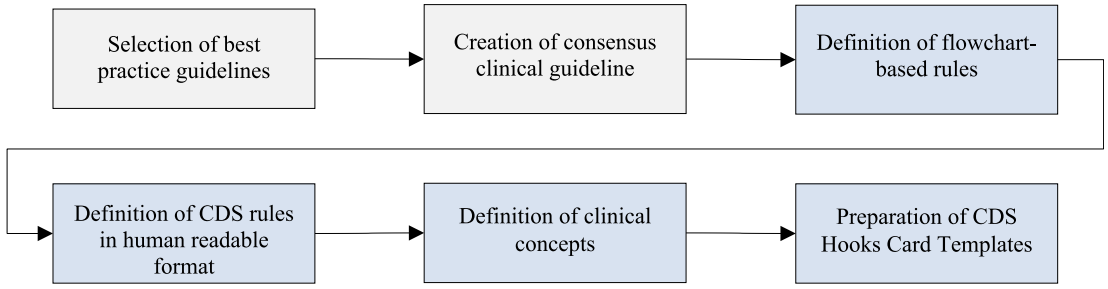

FIGURE 1
Overview of the methodology for transforming evidence-based clinical guidelines into implementable clinical decision support services.

automated as a clinical decision support service integrated into daily care practices. As a first step to create flowcharts with decision points to assess patient data, we have identified all the clinical concepts involved in textual guideline definitions. For each clinical concept, it was discussed with the CRG group whether it constitutes a diagnosis, an assessment to be conducted by the physician, a laboratory result, a medication, a clinical procedure, or a scored assessment to be performed. It was also determined whether the information would be retrieved from the EHRs of the patient, or whether it cannot be obtained from the EHR and needs assessment through physician facing CAREPATH tools, such as the AICP. Consequently, jointly agreed-upon parts of the consensus guideline have been converted into flowchart rules that pave the way for the implementation of clinical decision support services.

In our methodology, we utilized the Unified Modeling Language (UML) Activity Diagrams to draw the flowcharts. While activity diagrams are primarily used in the design phase of software engineering to describe system behavior as a workflow, researchers have also begun utilizing them for modeling clinical workflows (50–52). Activity diagrams enable us to graphically describe what clinical action needs to take place in which condition in an easy way. It also allows describing sequential and parallel processes. Activity diagrams consist of several concepts, such as activity, action, transition (control flow and object flow), decision node, swimlane and partition, each of which has a different graphical notation. In our approach, we only utilized the following concepts with the provided purpose of usage:

- Initial node: A circle representing the beginning of a workflow consisting of a set of actions or activities.
- Control flow: An arrow showing the sequence of workflow.
- Decision node: A diamond representing a test condition, such as "Has the patient met his/her blood pressure goal?." The control flow can only continue with one of the decision paths.
- Action/activity: A (rounded) rectangle representing an action from the consensus clinical guideline such as "Consider starting monotherapy with ACE inhibitors or Angiotensin II Receptor Blockers (ARBs) or Calcium channel blockers or Thiazide diuretics by also checking possible contraindications."
- Final node: An encircled circle representing the end of a workflow.

Figure 2 illustrates an example of a flowchart generated for the hypertension diagnosis procedure. If a patient has not been diagnosed hypertensive, they have not been sent home for diagnosis confirmation, and their systolic blood pressure (SBP) value is above 140 mmHg or diastolic blood pressure (DBP) value is above 90 mmHg, the guideline recommends short-term self-monitoring of blood pressure levels. It also recommends setting a follow-up appointment to confirm diagnosis after 2–4 weeks. If the SBP is between 130 and 139 mmHg or DBP is between 85 and 89 mmHg, it recommends categorizing patient's blood pressure as high-normal. If they are below 130 mmHg or 85 mmHg, respectively, it recommends normal categorization. On the other hand, if the patient has already been sent home, then based on the SBP and DBP values, patient's blood pressure can also be categorized as Grade 1, Grade 2, or Grade 3. In either case, the guideline recommends diagnosing the patient as hypertensive.

2.2 Definition of CDS rules in human readable format

In the second step, we developed directly implementable specifications for clinical decision support services to automate the consensus clinical guideline. For this purpose, we opted for the CDS Hooks formalism, which is a standard specification for clinical decision support services published by HL7.[2] It provides an API specification enabling synchronous, workflow-triggered CDS calls that return information and suggestions. The CDS Hooks specification describes a "hook"-based pattern for invoking decision support from within a clinician's workflow. User activity within the clinician's workflow triggers CDS hooks in real-time. When a triggering activity occurs, the CDS Client notifies each registered CDS service for the activity. These services must then provide near-real-time feedback about the triggering event. Each service receives basic details about the clinical workflow context (via the context parameter of the hook) along with any service-specific input data required (via the pre-fetch-template parameter).

In the CAREPATH context, this mechanism is utilized as follows (see Figure 3). CDSs in CAREPATH are employed to suggest personalized goals and interventions that can be put in a care plan based on the recommendations of clinical guidelines. AICP is responsible for calling the CDS services with important patient context data, crucial for personalizing suggestions. After presenting the suggestions to clinicians via user interfaces, the care plan of the patient can be created in a guided manner.

In CAREPATH, an HL7 FHIR-based interoperability approach is followed. All components utilize HL7 FHIR as a standard-based approach to represent patient data: the patient's EHRs retrieved from local systems by TIS/SIS are mapped to FHIR and stored in an open-source HL7 FHIR Repository, namely onFHIR.io (53), serving as a shared patient data repository. Data collected from the patient's home via home/health monitoring devices, such as vital signs, are stored as FHIR resources by H/HMP, and patient-collected data such as symptoms are represented as FHIR resources via PEP. AICP retrieves the relevant CDS input parameters from the FHIR repository as important patient context data and passes them to CDS services. In the CDS Hooks API, the patient data collected as FHIR data is passed as input to CDS services with the 'pre-fetch' parameter. The CAREPATH core data model conforms to HL7 FHIR Release 4, but the implemented architecture is not bound to this specific version. It can be easily adapted to accommodate later versions or modifications.

The response of CDS services can consist of textual recommendations communicated as information cards (which can be read and assessed by the clinician to create a care plan manually, such as adding medications based on the detailed guides about possible adverse reactions) or as directly reusable care plan components communicated as suggestion cards in conformance with the CDS Hooks API. In suggestion cards, the recommended goals and activities are represented as FHIR resources (such as MedicationRequest, Goal, Appointment resources) which can be used to constitute the care plan of the patient.

2 CDS Hooks Specifications, https://cds-hooks.hl7.org/.

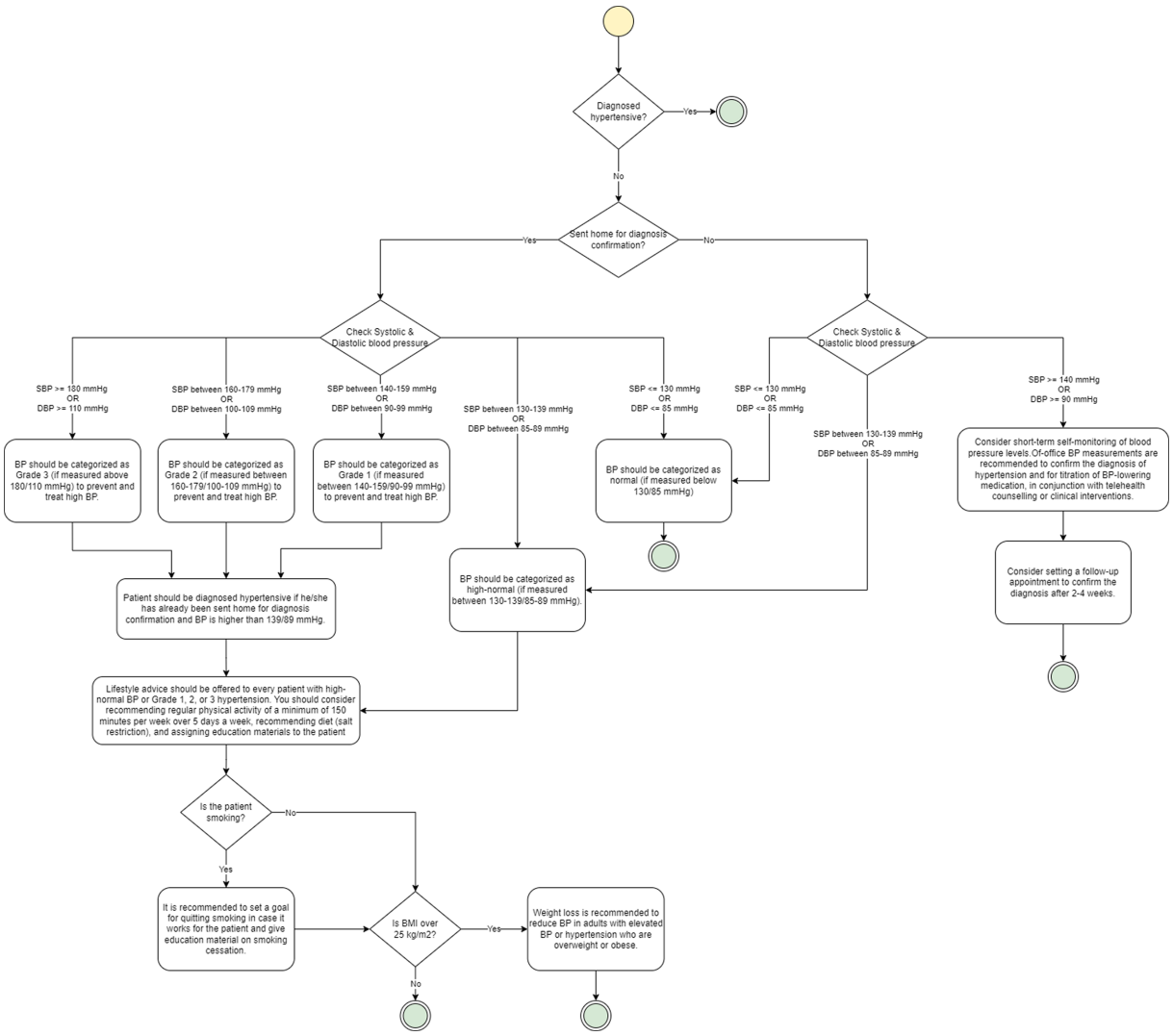

FIGURE 2
An example of a flowchart based on Hypertension guideline. The yellow circle represents the start node, while the green circles represent the end node. Diamonds are used to represent decision nodes, and rounded rectangles represent actions.

Based on the CDS Hooks standard, each CDS service can return any number of cards in response to the hook. Clinicians, as users, see these cards via AICP interfaces - one or more of each type – embedded in the workflow, and can interact with them as follows:

- **Information cards** provide text for the user to read. In our methodology, guidance from clinical guidelines, which may not be feasible or practical to automate but still provide crucial information to assist clinicians in creating individualized care plans, is represented as information cards. For example, in the Hypertension guideline, the guideline recommends discussing whether the patient is taking their medication as prescribed before considering changes to drug therapy, following the National Institute for Health and Care Excellence (NICE)'s guideline on medicines adherence (54). This guidance is presented as an information card and can be viewed by clinicians in a graphical user interface (UI) for reading and acting upon it. Example representations of information cards in CAREPATH AICP are illustrated in the Results section.
- **Suggestion cards** provide a specific recommendation for which the CDS Client renders a button that the user can click to accept. Clicking automatically populates the suggested change into the clinician's UI. In CAREPATH, CDS services can recommend adding certain care plan activities such as Referral Requests, Appointment Requests, and Lab Test Orders. These are represented as FHIR resources (as detailed in Section 2.4) and presented to the user with checkboxes via AICP. Clinicians can add them directly to the patient's care plan by clicking on the checkboxes next to these suggestions.

The flowcharts have been reviewed together with CRG members to determine the parts of the consensus guideline that should be presented as information cards or suggestion cards, in order to create a practical tool that can be easily utilized by clinicians as a part of their daily clinical workflow.

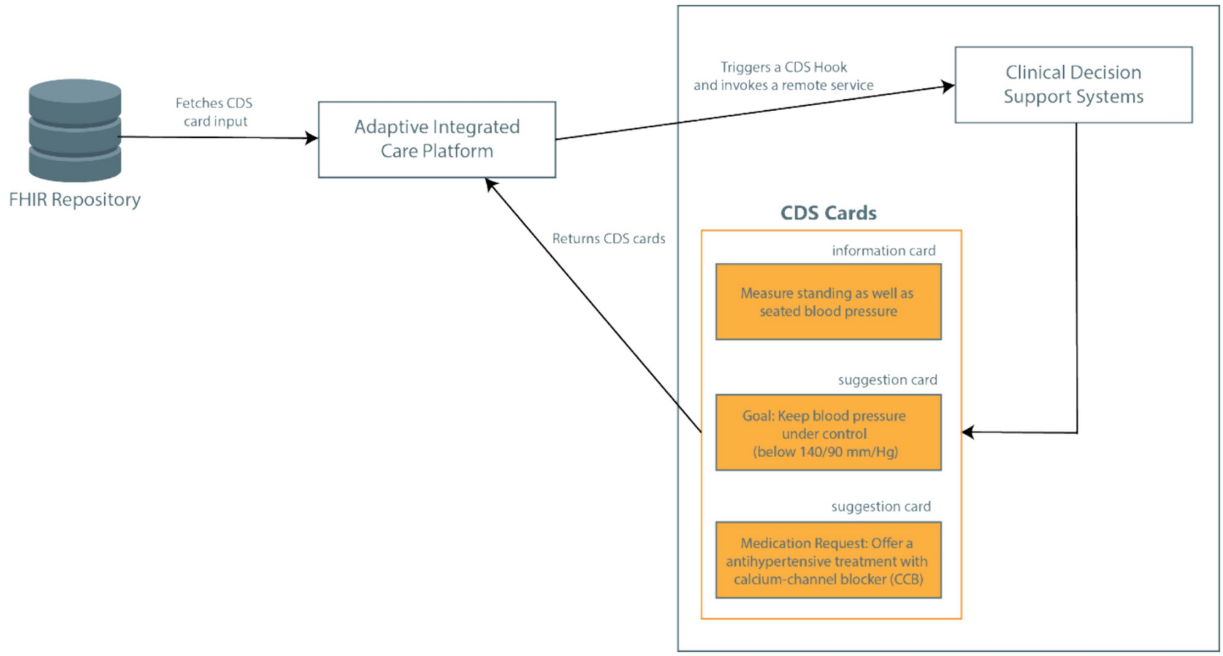

FIGURE 3
Use of CDS Hooks based services in CAREPATH architecture.

Rule ID	Context	Purpose	Trigger	Rule description	Input as Prefetch	Output as Cards
Hypertension 9	Patient is high-normal, or Hypertension Grade 1, 2, or 3	Lifestyle advice	This CDS will be triggered by AICP during Care Plan is being created or updated	IF the patient is categorized as [high-normal] OR [Grade 1 hypertension] OR [Grade 2 hypertension] OR [Grade 3 hypertension]	Observations: *Hypertension Grade	**Lifestyle advice for hypertensive patients (Card 7)** -Recommend Regular physical activity of a minimum of 150 minutes per week over 5 days a week -Recommend Diet (Salt Restriction) -Recommend assigning Education Material (Hypertension Education Material) to the patient
Hypertension 11	Patient is high-normal, or Hypertension Grade 1, 2, or 3	Goal Setting	This CDS will be triggered by AICP during Care Plan is being created or updated	IF the patient's (([SBP] is ≥160 mmHg AND [Age] >65 year) OR ([SBP] is between 140–159 mmHg AND ([Age]not >80 years)) AND does not have [Antihypertensive drugs]	Observations: *Hypertension Grade *BP measurement *Systolic Blood Pressure *Diastolic Blood Pressure Demographics *Age Medication *Antihypertensive drugs	**Start hypertension treatment to achieve blood pressure goals (Card 11)** -Recommend BP Goal BP GOAL should be set as between 130 and 140 mmHg, and diastolic BP (DPB) to <80 mmHg

FIGURE 4
A partial view of a CDS Hooks specification table, illustrating several CDS rules from Hypertension guideline.

We created a formal template to document CDS Hooks specifications for delivering the advice, information, and actions suggested for each area addressed by the consensus clinical guideline as depicted in Figure 4.

In this table, each flowchart rule identified in the first step is represented as a row. The columns of this template can be summarized as follows:

- Each rule is identified with a unique **identifier**. We begin with the section title of the consensus guideline and assign a unique number for each rule, such as 'Hypertension 1,' 'Hypertension 2,' 'Diabetes 1,' 'COPD 1' and so on.
- Each rule has a **context** attribute, which is mostly informative and describes the current state of the patient for which the rule will be applied.
- Each rule has a purpose. The **purpose** field is critical for CDS specifications. We have examined and categorized the purpose of the advice, information, and actions suggested by the consensus guideline into the following categories:
  - o Information
  - o Goal management
  - o Diagnosis
  - o Lifestyle advice (Nutritional intervention and Physical exercise)
  - o Drug treatment

- Adverse events and medication contraindications
- Symptom assessment
- Complication management
- Planning next visit

These purpose categories are utilized to group the suggestions, and separate CDS service implementations are done based on these categories. This facilitates presenting guidance from consensus guideline in a modular way in the user interfaces provided to clinicians. Different panels of the AICP pages are configured to be linked with different CDS service instances based on the purpose category, allowing clinicians to easily review the guidance provided by the consensus guideline.

- Each CDS rule has a **triggering condition**. Most of the time, for CDS rules automating clinical guideline suggestions, the triggering component is AICP. AICP calls the CDS services with the required input. Whenever the input parameters are updated from the user interface of AICP, the CDS services are triggered again.
- **Rule descriptions** are mainly retrieved from the consensus guideline and formalized to be easily converted into computer-interpretable rules. Each patient parameter represented as a clinical concept is enclosed within brackets (e.g., [SBP] designated for systolic blood pressure).
- Parameters represented as clinical concepts used in Rule descriptions within brackets are listed input parameters in the "**Input as prefetch**" column. These parameters need to be pre-fetched by the CDS client as FHIR resources from the FHIR repository and passed to the CDS services as an input. For this purpose, these parameters are first mapped to FHIR Resources, such as Condition, Observation, and Medication, and then represented as clinical concepts as explained in Section 2.3.
- Finally, the output of CDS rules is briefly described as CDS Hooks cards. Here in this table, only the titles, card numbers, and summaries are presented. As part of the CDS Hooks specification of CDS Services, there is a separate sheet where all the identified cards are clearly described, as presented in Section 2.4.

2.3 Definition of clinical concepts

Clear consensus on clinical concepts is a crucial step in CDS implementation for processing patient data to provide personalized suggestions. It is a step forward to create a common dictionary between clinical experts in CRG and technical experts who will implement CDS services. It is also essential to establish semantic interoperability with existing EHR systems to collect patient parameters in a machine-processable manner.

In this step, the parameters identified in rule descriptions and the "Input as prefetch" column are represented as clinical concept definitions (see Figure 5 for examples). Firstly, as CDS Hooks specifications require CDS parameters as HL7 FHIR resources, we have categorized clinical concepts as Condition, Observation, and Medication resources.

The second important step is to bind each clinical concept to a code from international code systems. Based on discussions with CRG members, conditions have been coded either with ICD-10 or SNOMED CT codes, with categorization as diagnoses or symptoms. Medications are uniformly coded with ATC codes, while lab tests represented as FHIR Observations are coded with LOINC codes. Additionally, the agreed-upon unit of the lab test result observation is specified in reference to UCUM.

Assessments to be carried out by clinicians via AICP interfaces are also represented as FHIR Observation resources. These are coded with LOINC or SNOMED CT whenever possible. In instances where a direct mapping to a code in international code systems such as LOINC and SNOMED CT is not feasible, local codes have been created to designate these observations. The data types of these assessments, represented as FHIR Observation resources, are usually specified either as boolean Yes/No values, or as a value-set. Value-sets define a set of codes drawn from one or more code systems as possible values of these assessment observations. For example, such a value-set for representing smoking status observation is presented in Figure 5, where a set of LOINC codes is selected to represent possible values of a smoking status observation.

Additionally, the possible sources of these parameters have been identified. Some can be directly extracted from the patient's EHR, while others require assessments during the visit, recorded via AICP. Some parameters can be retrieved from PEP, and others from H/HMP. This approach ensures that rule implementers can have a clear understanding of the clinical concepts to be processed by the CDS service implementation as parameters.

2.4 Preparation of CDS hooks card templates

After CDS rules are defined in a human-readable format, where the relevant clinical concepts are identified, the fourth step involves further detailing the specifications of CDS outputs identified in rule definitions. For this purpose, we have prepared CDS Hooks card templates. In successful responses, CDS Services respond with a 200 HTTP response containing an object that includes an array of cards.

Each of the cards identified in the Rules template is specified with all the details required in the CDS Hooks standard specification, as explained below. An example illustration of these in a user interface, such as AICP, is displayed in Figure 6.

- **Summary**: A short (usually one sentence) explanation of the suggestion, displayed in user interfaces as the title of the card.
- **Detail**: A detailed description sourced from the consensus clinical guideline. This description is displayed when the user clicks the arrow on the right side of the card title. It can be represented as plain text or in GitHub Flavored Markdown language.[3] This field is optional.
- **Source**: The primary source of guidance for the decision support represented by the card. In CAREPATH, we provide the exact section number and page number of the referenced clinical guideline (e.g., "Holistic patient centered CAREPATH best practice guideline, Chapter 12.2 [pp. 40]").

3 GitHub Flavored Markdown Spec, https://github.github.com/gfm/.

CONDITIONS (Diagnosis and Symptoms)					
Terms/concepts used in guidelines	**Coding System**	**Code**	**Designation**	**Category**	**Possible Source**
Angioneurotic oedema	ICD-10	T78.3	Angioneurotic oedema	diagnosis	EHR via TIS/SIS (Physicians will be able to confirm/amend via AICP)
	ICD-10	R60	Oedema, not elsewhere classified	diagnosis	EHR via TIS/SIS (Physicians will be able to confirm/amend via AICP)
Apnoea	ICD-10	R06.8	Other and unspecified abnormalities of breathing	symptom	EHR via TIS/SIS (Physicians will be able to confirm/amend via AICP)
Hyperkalaemia	ICD-10	E87.5	Hyperkalaemia	diagnosis	EHR via TIS/SIS (Physicians will be able to confirm/amend via AICP)
Hypertension	ICD-10	I10 - I16	Hypertensive diseases	diagnosis	EHR via TIS/SIS (Physicians will be able to confirm/amend via AICP)
OBSERVATIONS (Lab tests, Vital Signs)					
Terms/concepts used in guidelines	**Coding System**	**Code**	**Designation**	**Unit**	**Possible Source**
Albumin per 24 hours	LOINC	21059-1	Albumin [Mass/volume] in 24 hour Urine	mg/dL	EHR via TIS/SIS (Physicians will be able to confirm/amend via AICP)
BMI	LOINC	39156-5	Body mass index (BMI) [Ratio]	kg/m2	Height values retrieved from EHR via TIS/SIS and H/HMP Platform and PEP
Diastolic blood pressure	LOINC	8462-4	Diastolic blood pressure	mm[Hg]	From H/HMP, PEP , AICP, and EHR via TIS/SIS
Smoking status	LOINC	72166-2	Tobacco smoking status	(Smoking status values)	EHR via TIS/SIS (Physicians will be able to confirm/amend via AICP)
Systolic blood pressure	LOINC	8480-6	Systolic blood pressure	mm[Hg]	From H/HMP, PEP , AICP, and EHR via TIS/SIS
OBSERVATIONS (Assessments)					
Terms/concepts used in guidelines	**Coding System**	**Code**	**Designation**	**Unit**	**Possible Source**
HFrEF class	LOINC	88020-3	Functional capacity NYHA	(Functional capacity NYHA values)	AICP
Any high-grade sinoatrial or A-V block	http://kroniq.srdc.com.tr/fhir/CodeSystem/observation-code	high-grade-sinoatrial-or-av-block	High-grade sinoatrial or A-V block	Yes/No	AICP
MEDICATIONS					
Terms/concepts used in guidelines	**Coding System**	**Code**	**Designation**		**Possible Source**
ACE inhibitors	ATC	C09A	ACE inhibitors, plain		EHR via TIS/SIS (Physicians will be able to confirm/amend via AICP)
	ATC	C09B	ACE inhibitors, combinations		EHR via TIS/SIS (Physicians will be able to confirm/amend via AICP)
Alpha-blocker	ATC	C02CA	Alpha-adrenoreceptor antagonists		EHR via TIS/SIS (Physicians will be able to confirm/amend via AICP)
Antidepressants	ATC	N06A	Antidepressants		EHR via TIS/SIS (Physicians will be able to confirm/amend via AICP)
ARBs	ATC	C09C	Angiotensin II Receptor Blockers (ARBs), plain		EHR via TIS/SIS (Physicians will be able to confirm/amend via AICP)
	ATC	C09D	Angiotensin II Receptor Blockers (ARBs), combinations		EHR via TIS/SIS (Physicians will be able to confirm/amend via AICP)
VALUE SETS					
Value set name	**Coding System**	**Code**	**Designation**		
Smoking status values	LOINC	LA18976-3	Current every day smoker		
	LOINC	LA18977-1	Current some day smoker		
	LOINC	LA15920-4	Former smoker		
	LOINC	LA18978-9	Never smoker		
	LOINC	LA18979-7	Smoker, current status unknown		
	LOINC	LA18980-5	Unknown if ever smoked		
	LOINC	LA18981-3	Current Heavy tobacco smoker		
	LOINC	LA18982-1	Current Light tobacco smoker		
Functional capacity NYHA values	LOINC	LA28404-4	Class I		
	LOINC	LA28405-1	Class II		
	LOINC	LA28406-9	Class III		
	LOINC	LA28407-7	Class IV		

FIGURE 5
An excerpt from Hypertension clinical concepts table showcasing examples across different types.

- **Suggestions**: An array of suggestions that allows a service to recommend a set of changes in the context of the current activity (e.g., adjusting the dose of a currently prescribed medication for the medication-prescribe activity). Each suggestion can contain an array of Actions, each defining a suggested action. Within a suggestion, all actions are logically ANDed together, meaning that selecting a suggestion selects all the actions within it. If there are alternative suggestions, separate suggestions should be created as part of the suggestions array. Each suggestion must have a label summarizing the suggested actions.
  - o Each **Action** needs to have a type, which can be "create," "delete," or "update." In the CAREPATH context, "create" means that the suggested actions (such as referral, appointment, lab test request) will be added as care plan activities to the care plan; "delete" means removing an existing care plan activity from the care plan, and "update" means updating an existing care plan activity in the care plan.
  - o A human-readable description of the suggested action may be presented to the end-user, along with a description of the FHIR Resource that is suggested to be created, updated, or deleted. Therefore, each Action needs to have short title summarizing the suggested action. The presentation of suggestions and their corresponding actions within AICP is depicted in Figure 6.

If an Information card is suggested by the consensus guideline, then only the first three attributes (i.e., summary, detail, and source) are necessary. In other words, Information cards do not contain any

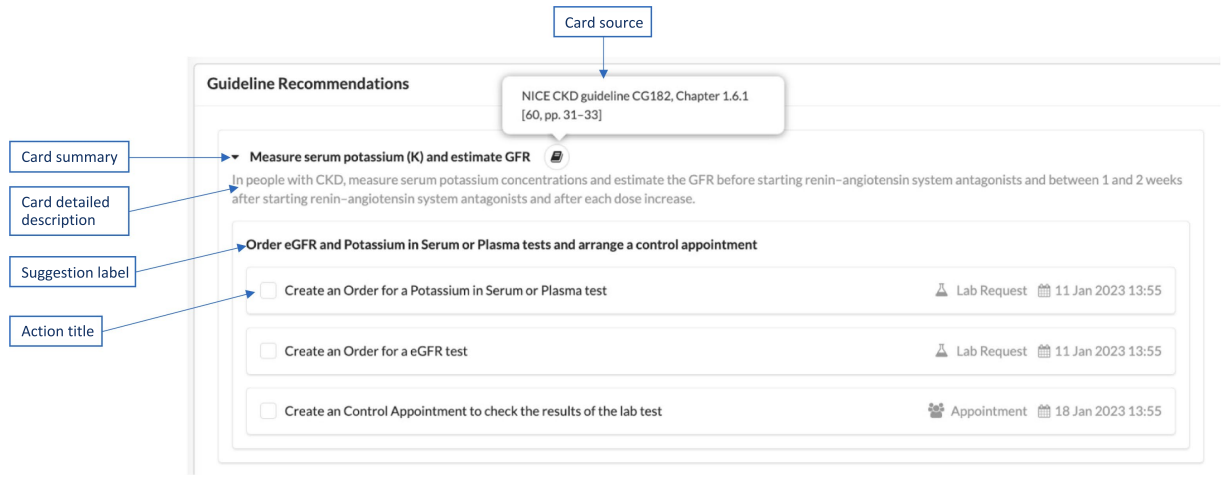

FIGURE 6
Presenting CDS Hooks cards in a user interface, such as AICP.

suggestions, as their purpose is only to provide some information. An example of an information card definition as a part CAREPATH CDS specifications is presented in Card 4 in Table 1.

Suggestion cards, on the other hand, always contain at least one Suggestion, which includes at least one Action. In CAREPATH, we have defined 7 types of actions, which are lab order, referral, appointment, patient activity, education material, goal, or medication. Within each Action, the exact FHIR Resource suggested to be added to the patient's care plan should be present. In CAREPATH, we use the following FHIR resources to represent consensus guideline suggestions as care plan components: *ServiceRequest*, *Appointment*, *CommunicationRequest*, *Goal*, and *MedicationRequest*. The details of different types of Actions alongside the FHIR resources used in them are explained below.

- **Lab Order Suggestions**: Consensus guidelines may suggest lab orders to be requested as part of the care plan. These are represented as the *ServiceRequest* resource in HL7 FHIR. An example of a lab order suggestion action within a suggestion card is presented in Action 2 of the card in Table 1. With the 'category' attribute of the *ServiceRequest* resource, we identify it as a lab request, referencing our local 'care-plan-activity-category' value set. The specific lab test requested is specified via the 'code' attribute of the *ServiceRequest* resource. In the example in Table 1, Action 2 of Card 21 suggests a lab test order for 'Creatinine [Mass/volume] in Serum or Plasma', indicated by the '2160–0' code from LOINC. Lab order categories are always indicated via a code from LOINC in CAREPATH. If there is guidance in the consensus guideline about when this lab test needs to be conducted, this is represented via the 'occuranceDateTime' attribute.
- **Referral Suggestions**: Consensus guidelines may suggest referrals to specialists as part of the care plan when a second opinion is needed. These are represented as *ServiceRequest* in HL7 FHIR.[4] The 'category' attribute of the *ServiceRequest* resource identifies it as a referral request, referencing our local 'care-plan-activity-category' value set. When guidance is available in the consensus guideline, the specialty of the practitioner to whom the referral is targeted is specified via the 'performerType' attribute of the *ServiceRequest* resource. For example, Action 3 in Table 2 indicates a referral to a cardiologist via the '175651000' code from SNOMED CT. Here, we always provide a code from the 'performer-role' value set defined by HL7.[5] If there is guidance in the consensus guideline about when this referral needs to be conducted, this is represented via the 'occuranceDateTime' attribute.
- **Appointment Suggestions**: Consensus guidelines may suggest appointments to be scheduled as part of the care plan. These appointments can be for regular care plan review visits, to check the effects of treatments, or to discuss the results of referrals. They are represented as *Appointment* resource in HL7 FHIR.[6] An example appointment action within a suggestion card is presented in Action 2 of the card in Table 2. The critical attributes are the appointment description and the proposed date, which is represented via the 'start' attribute.
- **Patient Activity Suggestions**: Consensus guidelines may suggest certain type of activities to be carried out by the patients as part of their care plans, such as physical exercises and self-measurement of vital signs. These are represented as *ServiceRequest* resources in HL7 FHIR. An example patient activity suggestion action within a suggestion card is presented in Action 1 of the card in Table 2. With the 'category' attribute of the *ServiceRequest* resource, we identify that it is a patient order, referencing our local 'care-plan-activity-category' value set. The specific activity type to be carried out is specified via the 'code' attribute of the *ServiceRequest* resource. In the example in Table 2, Action 1 suggests the patient to measure their blood

4 HL7 FHIR, Release 4, ServiceRequest Resource, http://hl7.org/fhir/r4/servicerequest.html.

5 HL7 FHIR, Release 4, Performer Role Value Set, https://build.fhir.org/valueset-performer-role.html.

6 HL7 FHIR, Release 4, Appointment Resource, http://hl7.org/fhir/r4/appointment.html.

TABLE 1 An example of an Information Card (Card 4) and a CDS card for arranging a follow-up visit to assess treatment effectiveness, containing Appointment and Lab Order actions (Card 21).

CARD 4				
Summary	Consider hypertension diagnosis with category Grade 1.			
Detailed description	BP should be categorized as normal (if measured below 130/85 mmHg), high-normal (130–139/85–89 mmHg), grade 1 (140–159/90–99 mmHg), grade 2 (160–179/100–109 mmHg) or grade 3 (≥ 180/110 mmHg) to prevent and treat high BP.			
Source	Holistic patient-centered CAREPATH best practice guideline, Chapter 12.2 [pp. 40]			
	Link to CAREPATH best practice guideline			
CARD 21				
Summary	Arrange a follow-up visit in 1 month.			
Detailed description	* Adults initiating a new or adjusted drug regimen for hypertension should have a follow-up evaluation of adherence and response to treatment at monthly intervals until control is achieved. * Renal function should be frequently assessed to detect possible increases in serum creatinine and reductions in eGFR as a result of BP-related reductions in renal perfusion.			
Source	Holistic patient centered CAREPATH best practice guideline, Chapter 12.3.1 [pp. 41 and 42]			
	Link to CAREPATH best practice guideline			
Suggestion 1				
label	Consider checking lab tests for eGFR and serum creatinine and setting a follow-up appointment within a month.			
ACTION 1	type	create		
	description	Consider setting a follow appointment after 1 month for follow-up evaluation of adherence and response to treatment.		
	resource	Appointment	extension	http://kroniq.srdc.com.tr/fhir/StructureDefinition/title | Follow appointment after 1 month for follow-up evaluation of adherence and response to treatment
			description	Follow appointment after 1 month for follow-up evaluation of adherence and response to treatment.
			status	proposed
			start	*{{Today + 1 month}}*
			specialty	–
ACTION 2	type	create		
	description	Consider ordering a serum creatinine test to assess renal function.		
	resource	Service Request	status	draft
			extension	http://kroniq.srdc.com.tr/fhir/StructureDefinition/title | Serum creatinine test
			intent	proposal
			occurranceDateTime	*{{Today + 1 month}}*
			category	http://www.kroniq.srdc.com.tr/fhir/care-plan-activity-category | lab-request | Lab Request
			code	LOINC | 2,160–0 | Creatinine [Mass/volume] in Serum or Plasma
			performerType	–
			performer	Patient
			text.status	generated
			text.div	Have serum creatinine before the control visit

pressure, indicated by the '85354–9' code from LOINC in the 'code' attribute. If there is guidance in the consensus guideline about when this activity needs to be conducted, this is represented via the 'occurranceDateTime' attribute. In this example, the patient is asked to measure their blood pressure twice a day.

- **Education Material Suggestions**: Consensus guidelines may suggest educational materials to be assigned to the patient as a part of the care plan. These are represented as *CommunicationRequest* resources in HL7 FHIR.[7] The payload attribute is utilized to refer to an online educational material that can be offered to the patient via the 'payload.contentAttachment.url' attribute. An example is

7 HL7 FHIR, Release 4, Communication Request Resource, http://hl7.org/fhir/r4/communicationrequest.html.

TABLE 2 An example of a CDS card for the management of resistant hypertension, containing Referral, Appointment and Patient Activity actions.

CARD 18				
Summary	Management of resistant hypertension.			
Detailed description	The recommended treatment strategy for resistant hypertension should include appropriate lifestyle measures and treatment with optimal or best-tolerated doses of three or more drugs, which should include a diuretic, typically an ACE inhibitor or an ARB, and a CCB. Secondary causes have to be ruled out when BP recommended treatment strategy fails to lower office systolic and diastolic BP values to <140 mmHg and/or < 90 mmHg, respectively, and the inadequate control of BP is confirmed by Ambulatory BP Monitoring or home BP monitoring in patients whose adherence to therapy has been confirmed.			
Source	Holistic patient-centered CAREPATH best practice guideline, Chapter 12.3.2 [pp. 42] and Chapter 12.2 [pp. 40]			
	Link to CAREPATH best practice guideline			
Suggestion 1				
label	Consider short-term self-monitoring of blood pressure levels to confirm inadequate control of BP.			
ACTION 1	type	create		
	description	Consider short-term self-monitoring of blood pressure levels to confirm inadequate control of BP.		
	resource	Service Request	status	draft
			extension	http://kroniq.srdc.com.tr/fhir/StructureDefinition/title | Self-monitoring of BP
			intent	proposal
			occuranceTiming	"start": *{{Today}}*, "end": *{{Today + 2 weeks}}*, "frequency": 2, "period": 1, "periodUnit": "d"
			category	http://www.kroniq.srdc.com.tr/fhir/care-plan-activity-category | patient-order | Patient Order
			authoredOn	*Automatically set to the date the CDS call is made*
			code	LOINC | 85,354–9 | Blood pressure panel
			performer	Patient
ACTION 2	type	create		
	description	Consider setting a follow appointment to confirm resistant hypertension after 2–4 weeks.		
	resource	Appointment	extension	http://kroniq.srdc.com.tr/fhir/StructureDefinition/title | Follow appointment to confirm resistant hypertension after 2–4 weeks
			description	Follow appointment to confirm resistant hypertension after 2–4 weeks.
			status	proposed
			start	*{{Today + 2 weeks}}*
			specialty	–
ACTION 3	type	create		
	description	Consider a Referral to Cardiologist for ruling out secondary causes.		
	resource	Service Request	status	draft
			extension	http://kroniq.srdc.com.tr/fhir/StructureDefinition/title | Referral to Cardiologist
			intent	proposal
			occuranceDateTime	*{{Today}}*
			category	http://www.kroniq.srdc.com.tr/fhir/care-plan-activity-category | referral | Patient referral to specialist
			authoredOn	*Automatically set to the date the CDS call is made*
			performerType	SNOMED | 175,651,000 | Cardiologist
			performer	–
			text.status	generated
			text.div	Referral to Cardiologist for ruling out secondary causes of resistant hypertension

TABLE 3 An example of a CDS card for offering lifestyle interventions for hypertensive patients, containing Communication Request actions.

CARD 7				
Summary	Offer Lifestyle interventions for hypertensive patients.			
Detailed description	Lifestyle advice should be offered to every patient with high-normal BP or Grade 1, 2, or 3 hypertension. Please check Diet Management and Exercise Planning pages for detailed diet and exercise plans to be added to the care plan of the patient.			
Source	Holistic patient centered CAREPATH best practice guideline, Chapter 12.3.1 [pp. 41]			
	Link to CAREPATH best practice guideline			
Suggestion 1				
label	Offer lifestyle advice and educational materials to hypertensive patients for healthy diet and physical activity.			
ACTION 1	type	create		
	description	Give education material on healthy diet.		
	resource	Communication Request	status	draft
			extension	http://kroniq.srdc.com.tr/fhir/StructureDefinition/title | Education material on healthy diet
			subject	Patient
			authoredOn	*Automatically set to the date the CDS call is made*
			payload.contentAttachment.language	en
			payload.contentAttachment.url	https://www.nhsinform.scot/healthy-living/food-and-nutrition
			payload.contentAttachment.title	Diet and nutrition - benefits of a balanced diet
ACTION 2	type	create		
	description	Give education material on physical activity for healthy living.		
	resource	Communication Request	status	draft
			extension	http://kroniq.srdc.com.tr/fhir/StructureDefinition/title | Education material on physical activity for healthy living
			subject	Patient
			authoredOn	*Automatically set to the date the CDS call is made*
			payload.contentAttachment.language	en
			payload.contentAttachment.url	https://www.nhsinform.scot/healthy-living/keeping-active
			payload.contentAttachment.title	Physical activity – health benefits of exercise

presented in Action 1 and 2 of the Lifestyle Interventions card shown in Table 3.

- **Goal Suggestions**: Consensus guidelines may suggest personalized goals to be assigned to the patient as part of the care plan. For example, in the diabetes section of the consensus guideline, personalized HbA1C, blood pressure, and lipid targets are suggested based on the patient's various parameters, such as glucose level, age, comorbidities, and recent lab test results. These goals are represented as *Goal* resources in HL7 FHIR.[8] The objective of the goal is indicated via the 'description.code' attribute, referencing international code systems. In the example presented in Table 4, the code '135840009' from SNOMED CT is used to specify that this is a 'Blood Pressure monitoring' goal. The specifics of the goal target are specified via the 'target' attribute, where the 'target.measure' attribute indicates that this is a goal for systolic blood pressure, referencing LOINC code '8480–6', with the target values indicated via the 'target.detailRange' attributes between 130 and 140 mmHg.
- **Medication Suggestions**: Consensus guidelines may suggest adding, updating the dose, or discontinuing a medication as part of the personalized care plan for the patient. For example, in the hypertension section of the consensus guideline, if the patient cannot achieve their blood pressure goals while already on dual medication, the consensus guideline suggests considering a triple combination of ACEi/ARB, CCB, and diuretic, while also checking for possible contraindications. These medication recommendations can be represented as *MedicationRequest* resources in HL7 FHIR.[9] In the example presented in Table 5, the first suggestion card recommends adding a beta-blocking agent. Other possible options can be added as additional alternative suggestion cards. The code "C07" from ATC is used to specify that the recommended drug is a beta-blocking agent.

8 HL7 FHIR, Release 4, Goal Resource, http://hl7.org/fhir/r4/goal.html.

9 HL7 FHIR, Release 4, Medication Request Resource, https://hl7.org/fhir/R4/medicationrequest.html.

TABLE 4 An example of a Goal suggestion CDS card.

CARD 11				
Summary	Systolic BP should be targeted to between 130 and 140 mmHg, and diastolic BP to <80 mmHg.			
Detailed description	The evidence supports the recommendation that multi-morbid older patients with cognitive impairment (>65 years, including patients over 80 years) should be offered BP-lowering treatment if their systolic BP is ≥160 mmHg. There is also justification to now recommend BP-lowering treatment for old patients (aged >65 but not >80 years) at a lower BP (i.e., grade 1 hypertension where systolic BP is between 140 and 159 mmHg). Systolic BP should be targeted to between 130 and 140 mmHg, and diastolic BP to <80 mmHg.			
Source	Holistic patient centered CAREPATH best practice guideline, Chapter 12.1 [pp. 40]			
	Link to CAREPATH best practice guideline			
Suggestion 1				
label	Keep blood pressure under control.			
ACTION 1	type	create		
	description	Keep systolic blood pressure under control (between 130 and 140 mm/Hg)		
	resource	Goal	lifecycleStatus	proposed
			meta.tag	http://kroniq.srdc.com.tr/fhir/CodeSystem/concept-id | GoalSystolicBP
			extension	http://kroniq.srdc.com.tr/fhir/StructureDefinition/title | Keep systolic blood pressure under control
			category	http://terminology.hl7.org/CodeSystem/goal-category | safety
			startDate	*Automatically set to the date the CDS call is made*
			description.text	Keep systolic blood pressure under control (between 130–140 mm/Hg)
			description.code	SNOMED | 135,840,009 | Blood Pressure monitoring (regime/therapy)
			target.dueDate	*{{Today + 3 months}}*
			target.measure	LOINC | 8,480–6 | Systolic blood pressure
			target.detailRange	low:130, high:140
ACTION 2	type	create		
	description	Keep diastolic blood pressure under control (below 80 mm/Hg)		
	resource	Goal	lifecycleStatus	proposed
			meta.tag	http://kroniq.srdc.com.tr/fhir/CodeSystem/concept-id | GoalDiastolicBP
			extension	http://kroniq.srdc.com.tr/fhir/StructureDefinition/title | Keep diastolic blood pressure under control
			category	http://terminology.hl7.org/CodeSystem/goal-category | safety
			startDate	*Automatically set to the date the CDS call is made*
			description.text	Keep diastolic blood pressure under control (below 80 mm/Hg)
			description.code	SNOMED | 135,840,009 | Blood Pressure monitoring (regime/therapy)
			target.dueDate	*{{Today + 3 months}}*
			target.measure	LOINC | 8,482–4 | Diastolic blood pressure
			target.detailRange	low:-, high:80

Possible side effects are presented as information cards, as depicted in Table 5. Considering that there could be too many options for the clinician to decide on, especially when considering possible side effects, it is also possible to represent medication recommendations as Information cards only. This enables the clinician to manually edit the medication plan via AICP after reviewing all the guidance provided. In CAREPATH, we have chosen to follow this approach to make the CDS specifications more concise.

3 Results

3.1 Output CDS rules and CDS hooks cards

Following the presented methodology, we analyzed the CAREPATH consensus clinical guideline, which provides advice, information, and actions in the following areas: overarching principles of management, mild cognitive impairment and dementia, physical exercise, nutrition and hydration, common use of drugs, coronary artery disease, heart

TABLE 5 An example of a Medication suggestion CDS Card (Card 31) and a possible side effect CDS card (Card 38).

CARD 31				
Summary	Consider triple combination of ACEi/ARB, beta-blocker, CCB and diuretic by also checking possible contraindications.			
Detailed description	For CAD patients who do not meet their BP goals on dual therapy, consider triple combination of ACEi/ARB, beta-blocker, CCB and diuretic by also checking possible contraindications.			
Source	Holistic patient-centered CAREPATH best practice guideline, Chapter 12.3.2 [pp. 42]			
	Link to CAREPATH best practice guideline			
Suggestion 1				
label	Consider adding Beta Blockers as a third therapy.			
ACTION 1	type	create		
	description	Consider prescribing Beta Blockers.		
	resource	MedicationRequest	lifecycleStatus	proposed
			description.text	Prescribe Beta Blocker as a part of triple therapy
			Medication.code	ATC | C07 | Beta Blocking Agents
CARD 38				
Summary	Compelling side effects for Beta-Blockers.			
Detailed description	Beta-blockers has compelling side effects for the patients with one of the following conditions: asthma or any high-grade sinoatrial or A-V block or bradycardia (heart rate < 60 beats per min).			
Source	Holistic patient centered CAREPATH best practice guideline, Chapter 12.3.2, Table 2			
	Link to CAREPATH best practice guideline			

failure, hypertension, diabetes, chronic kidney disease, COPD, stroke, sarcopenia, frailty, and caregiver support. We drew flowcharts, defined CDS rules and clinical concepts, and finally produced detailed implementable CDS-Hooks specifications for CDS services automating the following sections:

- Recommendations for the management of Mild dementia and mild cognitive impairment
- Recommendations for the management of Physical exercise
- Recommendations for the management of Nutrition and hydration
- Recommendations for the management of Commonly used drugs
- Recommendations for the management of Coronary artery disease
- Recommendations for the management of Heart failure
- Recommendations for the management of Hypertension
- Recommendations for the management of Diabetes
- Recommendations for the management of Chronic kidney disease
- Recommendations for the management of Chronic obstructive pulmonary disease
- Recommendations for the management of Stroke
- Recommendations for the management of Sarcopenia and frailty
- Recommendations for the management of Caregiver support

The full specifications are provided in the Supplementary material. In Tables 6, 7, we summarize the results of this process. The rules have been categorized under the following nine categories based on the purpose of recommendations:

1. **Diagnosis**: Guideline recommendations for diagnosing a patient's condition based on their current health parameters and status. For instance, hypertension guidelines recommend diagnosing hypertension if the patient has already undergone home diagnosis confirmation and their blood pressure remains higher than 139/89 mmHg.
2. **Lifestyle advice**: Guideline recommendations related to nutritional intervention, physical exercise, and smoking cessation.
3. **Goal management**: Guideline recommendations for assigning patients targets to achieve, such as maintaining systolic blood pressure between 130 and 140 mmHg or providing weight loss advice to adults with elevated blood pressure or hypertension who are overweight or obese.
4. **Drug treatment**: Guideline recommendations for initiating new medication therapy for newly diagnosed patients or adjusting existing medication therapy if disease progression is not controlled.
5. **Adverse events and medication contraindications**: Guideline recommendations for informing clinicians about possible adverse events and contraindications before starting a new medication therapy. For instance, the CAREPATH consensus clinical guideline recommends closely monitoring the impact of BP-lowering drugs on the well-being of the patient due to increased risk of adverse events (e.g., injurious falls) in older adults. When combination therapy is used, it suggests starting at the lowest available doses.
6. **Information and guidance about disease management**: Includes guidance for clinicians on important aspects of disease treatment associated with cognitive impairment and dementia, reminders about assessments needed before treatment planning and presenting useful information for sharing/discussion with patients and their caregivers. For example, "Before initiating pharmacological treatment for diabetes, the person's cardiovascular status and risk should be assessed to determine whether they have chronic heart failure" or "Keeping the environment at home safe to reduce the risk of falling and injury."
7. **Symptom recording**: For reminding clinicians to assess patient's specific symptoms at certain times or under certain conditions. For example, diabetes guidelines recommend assessing symptoms such as distress, disabilities, depression, anxiety, disordered eating, visual and hearing impairments, cognitive capacities, and other geriatric syndromes using a

TABLE 6 The number of rules defined for different categories in each section.

	HT	DM	COPD	MD& MCI	STR	S&F	CAD	HF	CKD	CUD	N&H	PE	CS
Diagnosis	12	3	9	3	1	1	3	4	4	–	–	–	–
Lifestyle advice	3	3	–	1	–	2	2	–	1	–	11	5	1
Goal management	2	5	–	–	1	–	–	–	1	–	–	–	–
Drug treatment	17	22	12	8	7	–	9	13	20	9	–	–	–
Adverse events and medication contraindications	17	–	–	4	–	–	–	–	–	–	–	–	–
Information and guidance about management	–	9	2	19	2	2	3	1	8	–	–	–	–
Symptom recording	–	1	1	–	–	–	–	–	–	–	–	–	–
Complication management and referrals	1	10	6	1	2	2	1	1	3	–	–	–	–
Planning next visit	–	2	–	2	–	–	–	–	1	–	–	–	–
TOTAL	**52**	**55**	**30**	**38**	**13**	**7**	**18**	**19**	**38**	**9**	**11**	**5**	**1**

The abbreviations used in the header refer to the following: HT, Hypertension; DM, Diabetes; COPD, Chronic Obstructive Pulmonary Disease; MD&MCI, Mild Dementia & Mild Cognitive Impairment; STR, Stroke; S&F, Sarcopenia & Frailty; CAD, Coronary Artery Disease; HF, Heart Failure; CKD, Chronic Kidney Disease; CUD, Commonly Used Drugs; N&H, Nutrition & Hydration; PE, Physical Exercise; CS, Caregiver Support.

Comprehensive Geriatric Assessment at the initial visit, at periodic intervals, and when there is a change in disease, treatment, or life circumstance, including caregivers and family members in this assessment.

8. **Complication management and referrals**: Recommendations for referring patients to other departments or specialists in case of suspected complications, emergencies, or when consultancy/expertise from another specialty is required. For example, in hypertension treatment, referral to a cardiologist is recommended to rule out secondary causes if recommended treatment strategies fail to lower blood pressure values. Additionally, referral to a respiratory disease specialist is recommended for diagnosing obstructive sleep apnea if the patient exhibits symptoms such as snoring, apnea, nocturia, nocturnal dyspnea, nighttime cardiovascular events, or resistant hypertension, along with daytime sleepiness. Moreover, a referral to emergency services is advised if the patient's clinic blood pressure exceeds 180/110 mmHg.
9. **Planning next visit**: For scheduling follow-up appointments to evaluate patient's adherence to care plan activities and their response to treatment.

Table 6 presents the number of rules defined for each section of the holistic guideline based on these categories. In total, 296 CDS rules have been defined. Among them, 117 (40%) are related to drug treatment, 46 (16%) to information and guidance about management, 40 (13%) to diagnosis, 29 (10%) to lifestyle advice, and 64 (21%) to other categories. No rules have been defined for drug treatment in the Sarcopenia & frailty, Nutrition & hydration, Physical exercise, and Caregiver support sections, because these guidelines do not directly address the treatment of specific diseases. Similarly, no rules related to diagnosis, complication management, and referral exist in the Commonly used drugs, Nutrition & hydration, Physical exercise, and Caregiver support sections. In goal management, guidelines for Hypertension, Diabetes, Stroke and Chronic kidney disease recommend setting targets for systolic blood pressure, diastolic blood pressure, weight, HbA1c, fasting glucose, LDL cholesterol, HDL cholesterol, Total cholesterol, and Hemoglobin. Since the CAREPATH study mainly focuses on multimorbidity management in the elderly with dementia, the largest number of rules for information and guidance about management has been defined in the Mild dementia & mild cognitive impairment section.

In each CDS rule, there exists one or more CDS Hooks card to achieve the specific objective of that rule. Table 7 shows the number of CDS Hooks cards defined for each section and the number of actions in those cards per action type. In the CAREPATH study, we defined 326 CDS Hooks cards to implement 296 CDS rules. Among them, the majority of the cards (191 out of 326, 59%) appeared in the Hypertension, Diabetes, COPD, and Chronic kidney disease sections, followed by 38 (12%) in MD & MCI and 31 (10%) in the Heart failure sections.

As explained in the Methodology section, a CDS Hooks card can be an information card, meaning that there is no action in it, or it can contain suggestions in which there exists at least one action. In Table 7, the number of information cards in each section is presented in the "Information & Medication contraindication" row. In hypertension, there exist 17 medication contraindication rules, which are modeled as information cards in CAREPATH. The rest of the rows in the table show the number of actions per type in the other CDS Hooks cards. Here, there is an additional type, autofill, which has not been explained in the methodology. In CAREPATH, autofill CDS Hooks cards are intended to present guideline recommendations suggesting diagnosis or assessment of a patient based on a recent measurement. For instance, hypertension guidelines recommend diagnosing Bradycardia

TABLE 7 The number of CDS Hooks cards defined for each section and the number of actions per type in the cards.

	HT	DM	COPD	MD& MCI	STR	S&F	CAD	HF	CKD	CUD	N&H	PE	CS
Cards	55	55	40	38	13	7	18	31	41	9	11	6	2
Information & Medication contraindication	33	21	12	27	3	4	3	6	11	1	9	5	2
Patient activity	2	1	–	–	–	–	1	–	–	–	–	1	–
Appointment	2	2	1	2	–	–	–	3	4	–	–	–	–
Referral	4	7	7	1	3	2	5	5	3	–	2	–	–
Education material	3	1	2	1	–	–	1	1	1	–	–	–	–
Goal	6	17	–	–	1	–	1	–	1	–	–	–	–
Lab request	4	6	3	3	–	–		10	17	–	–	–	–
Medication request	17	20	11	6	7	1	9	10	14	8	–	–	–
Autofill	4	–	8	–	–	–	–	6	3	–	–	–	–

if the patient's heart rate is less than 60 bpm, diagnosing Hyperkaliemia if the patient's potassium level is more than 5.5 mmol/L or considering severe left ventricular dysfunction if the patient's left ventricular ejection fraction is less than 40%.

3.2 Implementation of CDS engine

In CAREPATH, based on the CDS Service specifications presented in Section 2, software engineers have implemented the CDS services via a CDS Engine implementation in the Scala programming language. For each category in each section presented in Table 6, a CDS-Hooks-complaint REST endpoint has been implemented. For some categories that contain a considerable number of rules, such as drug treatment or information and guidance about management, multiple endpoints have been created. Consequently, a total of 65 CDS-Hooks-compliant REST endpoints have been implemented.

In CAREPATH, the patient data retrieved from EHRs, created via AICP, and collected from patients via H/HMP and PEP, are all represented in HL7 FHIR and maintained in a FHIR repository. Within the implementation of CDS-Hooks endpoints in Scala, the prefetch parameters have been expressed as FHIR queries, to retrieve the indicated patient input from a FHIR server, acting as the patient data store.

The CDS Hooks cards, represented as separate tables in the CDS specifications, have been defined as parametrized JSON files, using a template language, namely Mustache. These are instantiated for each patient by filling in the placeholders with patient-specific parameters by our CDS Engine. The CDS Logic, defined as rules in the CDS specifications, is implemented as rules defined via FHIR Path expressions, mapping retrieved input parameters to pre-defined CDS Hooks Template cards. The defined CDS Hooks cards and service definitions are available as open-source on GitHub[10].

In CAREPATH, we have preferred a Scala-based implementation. However, given the open specifications presented in Section 2, Supplementary material, and CDS-Hooks standard specifications, any other programming language could have been used to realize the implementation of these RESTful CDS services. Clear, open specifications mapping the clinical concepts to FHIR resources and international code systems, and rules defined based on these clinical concepts, enabled engineers who do not have clinical expertise to easily realize CDS implementations.

3.3 Usage of CDSs in a real-word environment

The Adaptive Integrated Care Platform (AICP) is one of the core components of the CAREPATH system, facilitating collaborative management of the care of multimorbid patients with mild dementia. It serves as the direct interface to care team members, allowing for the definition, updating, reconciling, and sharing of care plans, as well as the utilization of clinical decision support modules supporting these operations. It provides healthcare professionals with relevant information to guide decisions in an effective way, both during follow-up visits and in initial diagnosis processes. AICP has been implemented as a Web application providing an easy-to-navigate dashboard for care team members to view the basic medical history of the patient along with the care plan lifecycle history. The AICP care plan management graphical user interfaces have been designed to integrate the CDS services and to present the suggestions coming from CDS services in the best possible manner to facilitate care plan editing in the guidance of evidence-based clinical guidelines. The design was made with the involvement of healthcare professionals. First, the user requirements were collected through interviews conducted with them. Then, based on the user requirements, several mockups were drawn. These mockups were presented to healthcare professionals and their feedback was received. At the end of a few rounds, the final design emerged.

10 CAREPATH CDS Specifications, https://github.com/srdc/carepath-cds-specifications.

The input parameters of CDS services may be retrieved from the EHRs of the patient, including the patient's existing diagnosis, medications, and lab test results, from PEP for symptoms recorded, and from H/HMP for measurements retrieved from health devices. During the analysis of CDS services, we realized that some input parameters are clinical assessments which need to be carried out by the clinician during the visit with the patient. An example could be assessing "whether the patient's condition is stable or not."

AICP has been designed to provide a specific page for the management of each section, described in Section 3.1; e.g., Hypertension diagnosis/treatment, Diabetes diagnosis/treatment, CAD management etc., along with additional pages to support some common functionalities such as reviewing the current status of the patient (such as physical examination, review of lab results), providing overarching lifestyle and physical exercise recommendations, and reviewing the questionnaires assigned to the patient.

The care plan management pages have been divided into two main parts, as illustrated in Figures 7, 8. In Part A, the clinician is reminded about the important parameters that will affect personalized decisions about care plan goals and activity suggestions. These parameters have been identified in the third step of our methodology, which is the definition of clinical concepts. The values of these concepts are mostly retrieved from EHRs, and clinicians can amend them if necessary (e.g., manually adding new lab results). Clinicians can make new assessments, mostly for assessments that need to be carried out during that encounter. Figure 7 shows the first part of the Hypertension treatment page consisting of six different panels. In the first panel, the clinician examines the patient's latest systolic and diastolic blood pressure measurements as well as the number of falls since the last visit. The clinician can also record new values for those fields. Based on the latest systolic and diastolic blood pressure values, the guideline recommends categorizing the patient as Grade 1. In the second panel, the lab results of the patient are presented. It should be noted that these panels do not present the full medical summary of the patient. For each section, such as hypertension management, only the lab results, conditions, symptoms, assessments, etc. that are necessary for clinical assessment in the context of this section (that are listed in the clinical concepts table of the respective CDS services) are presented. In the third panel, comorbidities are shown. In the example, the CDS services automating hypertension guideline recommended CKD diagnosis, because the patient's eGFR value is less than 60 mL/min. In the fourth and fifth panels, assessments and symptoms are presented, respectively.

Based on the reviewed patient data and the provided clinical assessments in Part A, CDS services run in the background and provide personalized suggestions about what needs to be put in the care plan of the patient in Part B, such as goals (e.g., personalized systolic blood pressure, LDL cholesterol, HbA1c target), control appointments, lab test requests, referrals, medication requests, education materials, and patient orders (e.g., measuring blood pressure at home).

Figure 8 displays the implementation of Part B in the Hypertension treatment page, consisting of three panels. As explained in Section 2.4, in CAREPATH, we have chosen to represent medication recommendations as Information cards and enable the clinician to manually edit the medication plan. Therefore, in the Medication treatment panel, the medication-related guideline recommendations are presented under the medication list, and the clinician is provided with add, edit, and delete buttons to update the patient's medication treatment plan.

In the Goal Overview panel, the clinician can see the most recent systolic and diastolic blood pressure measurements of the patient in a chart view, observe the patient's adherence to the previous goals, and update the goals based on the guideline recommendations.

The guideline recommendations, selected by the CDS Engine based on the patient parameters provided in Part A, are presented at the end of the page. Clinicians can decide whether to add a suggested item to the care plan or not by clicking on the checkbox near it. If needed, they can edit their details (e.g., the date of a control appointment). In the example shown in Figure 8, the guideline recommended targeting systolic BP between 120 and 140 mmHg and diastolic BP below 80 mmHg for the patient who is under treatment. It also recommended arranging a follow-up visit in 1 month and ordering lab tests for eGFR and serum creatinine. Since the patient has resistant hypertension (because the patient did not meet his BP targets on triple therapy), the guideline also recommended short-term self-monitoring of blood pressure levels to confirm inadequate control of BP, setting a follow-up appointment to confirm resistant hypertension after 2–4 weeks, and a referral to cardiologist for ruling out secondary causes.

4 Discussion

This paper presents a methodology for generating implementable specifications for clinical decision support (CDS) services aimed at automating clinical guidelines. We have established a co-creation framework facilitating collaborative exploration of clinical guidelines by both clinical experts and software engineers. Through a systematic, traceable approach, our methodology enables the generation of open, human-readable CDS specifications. This open and traceable co-creation approach has especially helped us to address the challenges of automating multimorbidity guidelines. We have demonstrated that it is technically possible to consolidate suggestions from multiple conflicting guidelines and transform them into implementable specifications. We believe this methodology contributes to making healthcare more manageable for healthcare providers dealing with multiple chronic conditions and provides a practical example for future studies.

Understanding clinical guidelines poses a significant challenge for software engineers lacking medical expertise, hindering their ability to develop CDS services for automation (55). Conversely, clinicians without technical proficiency encounter difficulties in validating CDS implementations to ensure alignment with guideline recommendations. Our approach addresses these challenges by fostering interdisciplinary collaboration, allowing both groups to collectively translate clinical guideline suggestions into actionable directives for personalized care plan development.

Key strengths of our methodology include:

- Repeatable Process: Our methodology offers a systematic, replicable process for generating CDS specifications, ensuring consistency and reliability across implementations.

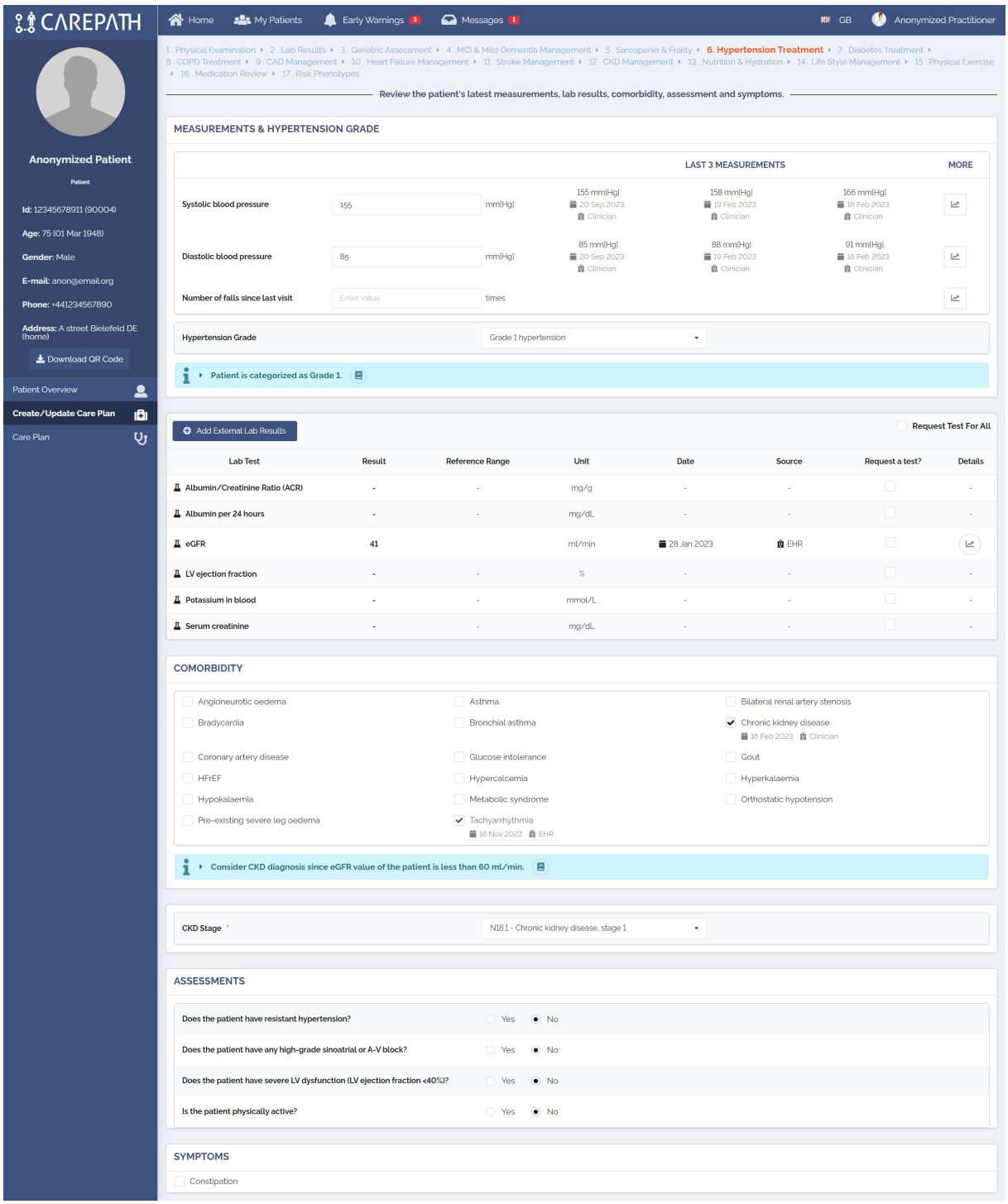

FIGURE 7
An example representation of clinical concepts identified during the definition of CDS rules in AICP pages.

- Co-Creation Landscape: By establishing a collaborative environment, we facilitate synergy between clinical expertise and technical proficiency, enhancing the quality and relevance of generated specifications.
- Traceability: Our approach provides clear traceability, enabling stakeholders to track the development process and ensure adherence to guideline recommendations.
- Human-Readable Specifications: We emphasize the creation of human-readable specifications, enhancing accessibility and facilitating comprehension for stakeholders across disciplines.
- Actionable Guidance: Our methodology translates clinical guideline suggestions into actionable guidance, enabling the creation of personalized care plans tailored to individual patient needs.

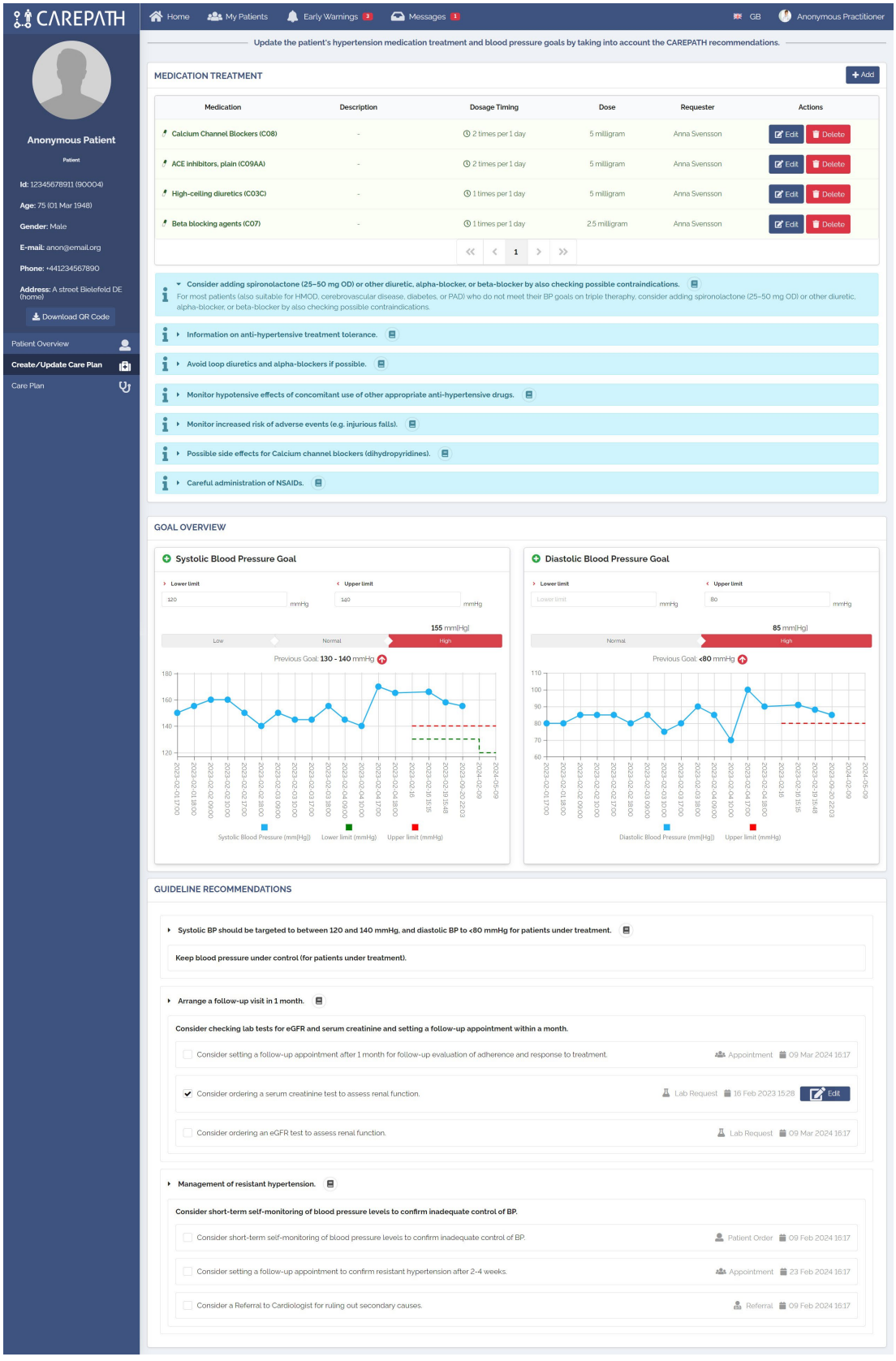

FIGURE 8
An example representation of CDS Hooks Cards in AICP interfaces.

By bridging the gap between clinical expertise and technical implementation, our methodology empowers interdisciplinary teams to develop CDS services that effectively automate clinical guidelines while ensuring alignment with evidence-based practices.

We have adopted a standardized approach guided by CDS Hooks Specifications, leveraging HL7 FHIR to define clinical concepts. Our methodology ensures clarity by precisely delineating the input/output parameters of CDS services in alignment with HL7 FHIR Resources, grounding clinical semantics within international code systems. This establishes a universal, shared lexicon—facilitating seamless communication between clinical and technical experts. Moreover, our clear specifications streamline the implementation of CDS services, as input parameters can be readily accessed from a FHIR repository via FHIR queries. By adhering to standards and facilitating easy mapping to FHIR-based implementations, our research enhances the interoperability and potential adoption of CDS services across diverse healthcare systems. This robust framework not only accelerates integration with external health IT systems but also paves the way for widespread implementation, thereby maximizing the impact of our research in clinical practice. In doing so, it complements prior studies facing challenges in disseminating and sharing knowledge artifacts for clinical decision support across different electronic health record platforms (56, 57).

CDS services for multimorbid older adults with MCI/MD need to address "whole-of-person" interventions to improve their quality of life (19), considering not only social issues but also physical and psychological difficulties (58). The CDS services implemented, following the methodology outlined in this paper, take a holistic approach to these patients, including specific healthcare conditions not typically found in guidelines, such as nutrition, exercise, frailty, and sarcopenia. Furthermore, they enable the entire healthcare team to participate in the care process using the same platform, considering not only patients' diseases but also environmental factors, caregiver support, quality of life, and psychosocial conditions.

The importance of patient privacy and data security in healthcare delivery necessitates careful planning and robust protection measures, particularly in highly automated workflows (59). Although the methodology outlined in this paper allows for the automation of clinical guidelines by producing implementable specifications for CDS services, it is limited to semi-automation, hence it does not provide a methodology for full automation. Healthcare professionals are still required to review CDS recommendations, make decisions, and exercise judgment at critical decision points in the workflow.

In future work, the usability, safety and technology acceptance of the CAREPATH ICT platform, including the developed tools and implemented CDS services, will be evaluated in a Technical Validation and Usability (TVU) study. This study will involve 16 patients with their informal caregivers and 16 healthcare professionals. Additionally, a Clinical Investigation (CI) involving over 200 patients will be conducted. These evaluations will take place in four European countries (Spain, Romania, Germany and the United Kingdom) over a period of 2 years.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

MG: Conceptualization, Methodology, Software, Writing – original draft. GL: Conceptualization, Methodology, Project administration, Software, Writing – original draft. AA: Software, Writing – original draft. OP: Software, Writing – review & editing. BA: Validation, Writing – review & editing. TA: Validation, Writing – review & editing. WS-B: Formal analysis, Investigation, Resources, Writing – review & editing. TR: Formal analysis, Investigation, Resources, Writing – review & editing. RA: Formal analysis, Investigation, Resources, Writing – review & editing. PA: Formal analysis, Investigation, Resources, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work presented in this paper was funded by the European Union's Horizon 2020 research and innovation program under grant agreement No 945169.

Acknowledgments

The authors would like to acknowledge the support of the CAREPATH consortium.

Conflict of interest

MG, GL, and AA are employed by Software Research & Development and Consultancy Corporation.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2024.1386689/full#supplementary-material

- CAD_CDSHooksSpecifications.xlsx
- CareGiverSupport_CDSHooksSpecifications.xlsx

- CKD_CDSHooksSpecifications.xlsx
- CommonlyUsedDrugs_CDSHooksSpecifications.xlsx
- COPD_CDSHooksSpecifications.xlsx
- Diabetes_CDSHooksSpecifications.xlsx
- HF_CDSHooksSpecifications.xlsx
- Hypertension_CDSHooksSpecifications.xlsx
- MCI_CDSHooksSpecifications.xlsx
- NH_CDSHooksSpecifications.xlsx
- PhysicalExercise_CDSHooksSpecifications.xlsx
- SarcopeniaFrailty_CDSHooksSpecifications.xlsx
- Stroke_CDSHooksSpecifications.xlsx

Additionally, the custom Value Set for listing care plan activity categories is provided in the following document.

- ValueSet_CarePlanActivityCategories.xlsx

References

1. Mercer SW, Smith SM, Wyke S, O'Dowd T, Watt GC. Multimorbidity in primary care: developing the research agenda. *Fam Pract*. (2009) 26:79–80. doi: 10.1093/fampra/cmp020

2. Lugtenberg M, Burgers JS, Westert GP. Effects of evidence-based clinical practice guidelines on quality of care: a systematic review. *BMJ Qual Saf*. (2009) 18:385–92. doi: 10.1136/qshc.2008.028043

3. Fischer F, Lange K, Klose K, Greiner W, Kraemer A. Barriers and strategies in guideline implementation—a scoping review In: S. Parthasarathy editor. *Healthcare*, vol. *4*: MDPI (2016). 36.

4. Lichtner G, Spies C, Jurth C, Bienert T, Mueller A, Kumpf O, et al. Automated monitoring of adherence to evidenced-based clinical guideline recommendations: design and implementation study. *J Med Internet Res*. (2023) 25:e41177. doi: 10.2196/41177

5. Goodwin N, Stein V, Amelung V. What is integrated care? *Handb Integr Care*. (2021) 3-25. doi: 10.1007/978-3-030-69262-9

6. Peart A, Barton C, Lewis V, Russell G. A state-of-the-art review of the experience of care coordination interventions for people living with multimorbidity. *J Clin Nurs*. (2020) 29:1445–56. doi: 10.1111/jocn.15206

7. Fares N, Sherratt RS, Elhajj IH. Directing and orienting ICT healthcare solutions to address the needs of the aging population In: A. Vitali editor. *Healthcare*, vol. *9*: MDPI (2021). 147.

8. Chen C, Ding S, Wang J. Digital health for aging populations. *Nat Med*. (2023) 29:1623–30. doi: 10.1038/s41591-023-02391-8

9. Gilbert S, Ricciardi F, Mehrali T, Patsakis C. Can we learn from an imagined ransomware attack on a hospital at home platform? *NPJ Digital Med*. (2024) 7:65. doi: 10.1038/s41746-024-01044-5

10. Riaño D, Peleg M, Ten Teije A. Ten years of knowledge representation for health care (2009–2018): topics, trends, and challenges. *Artif Intell Med*. (2019) 100:101713. doi: 10.1016/j.artmed.2019.101713

11. Kruse CS, Ehrbar N. Effects of computerized decision support systems on practitioner performance and patient outcomes: systematic review. *JMIR Med Inform*. (2020) 8:e17283. doi: 10.2196/17283

12. Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA*. (2005) 293:1223–38. doi: 10.1001/jama.293.10.1223

13. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digital Med*. (2020) 3:17. doi: 10.1038/s41746-020-0221-y

14. Sittig DF, Wright A, Osheroff JA, Middleton B, Teich JM, Ash JS, et al. Grand challenges in clinical decision support. *J Biomed Inform*. (2008) 41:387–92. doi: 10.1016/j.jbi.2007.09.003

15. Mayo AT, Myers CG, Bucuvalas JC, Feng S, Juliano CE. Supporting robust teamwork-bridging technology and organizational science. *N Engl J Med*. (2023) 388:2019–21. doi: 10.1056/NEJMp2300172

16. Pournik O., Ahmad B., Despotou G., Lim Choi Keung S. N., Mohamad Y., Gappa H., et al. (2022). CAREPATH methodology for development of computer interpretable, integrated clinical guidelines. Proceedings of the 10th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion, Lisbon, Portugal. (pp. 7–11).

17. Pournik O, Ahmad B, Lim Choi Keung SN, Khan O, Despotou G, Consoli A, et al. CAREPATH: developing digital integrated care solutions for multimorbid patients with dementia. *Stud Health Technol Inform*. (2022) 295:487–90. doi: 10.3233/SHTI220771

18. Bacigalupo I, Mayer F, Lacorte E, Di Pucchio A, Marzolini F, Canevelli M, et al. A systematic review and meta-analysis on the prevalence of dementia in Europe: estimates from the highest-quality studies adopting the DSM IV diagnostic criteria. *J Alzheimers Dis*. (2018) 66:1471–81. doi: 10.3233/JAD-180416

19. Lugtenberg M, Burgers JS, Clancy C, Westert GP, Schneider EC. Current guidelines have limited applicability to patients with comorbid conditions: a systematic analysis of evidence-based guidelines. *PLoS One*. (2011) 6:e25987. doi: 10.1371/journal.pone.0025987

20. Robbins T. D., Muthalagappan D., O'Connell B., Bhullar J., Hunt L. J., Kyrou I., et al. (2022). Protocol for creating a single, holistic and digitally implementable consensus clinical guideline for multiple multi-morbid conditions. Proceedings of the 10th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion, Lisbon, Portugal. (pp. 1–6).

21. Dans AL, Dans LF. Appraising a tool for guideline appraisal (the AGREE II instrument). *J Clin Epidemiol*. (2010) 63:1281–2. doi: 10.1016/j.jclinepi.2010.06.005

22. Custer RL, Scarcella JA, Stewart BR. The modified Delphi technique-a rotational modification. *Vocat Tech Educ*. (1999) 15:04521022. doi: 10.21061/jcte.v15i2.702

23. Pournik O, Ahmad B, Lim Choi Keung SN, Peake A, Rafid S, Tong C, et al. Interoperable E-health system using structural and semantic interoperability approaches in CAREPATH In: O Mantas, P Gallos, E Zoulias, A Hasman, MS Househ and M Charalampidou et al, editors. *Healthcare transformation with informatics and artificial intelligence*: IOS Press (2023). 608–11.

24. Gencturk M., Laleci Erturkmen G. B., Gappa H., Schmidt-Barzynski W., Steinhoff A., Abizanda P., et al. (2022). The design of a mobile platform providing personalized assistance to older multimorbid patients with mild dementia or mild cognitive impairment (MCI). Proceedings of the 10th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion, Lisbon, Portugal. (pp. 37–43).

25. Muth C, Blom JW, Smith SM, Johnell K, Gonzalez-Gonzalez AI, Nguyen TS, et al. Evidence supporting the best clinical management of patients with multimorbidity and polypharmacy: a systematic guideline review and expert consensus. *J Intern Med*. (2019) 285:272–88. doi: 10.1111/joim.12842

26. National Institute for Health and Care Excellence. Multimorbidity: clinical assessment and management. NICE guideline [NG56] (2016). Available at: https://www.nice.org.uk/guidance/ng56 (Accessed 2024-04-22)

27. Petersen RC, Lopez O, Armstrong MJ, Getchius TS, Ganguli M, Gloss D, et al. Practice guideline update summary: mild cognitive impairment: report of the guideline development, dissemination, and implementation Subcommittee of the American Academy of neurology. *Neurology*. (2018) 90:126–35. doi: 10.1212/WNL.0000000000004826

28. Atherton JJ, Sindone A, De Pasquale CG, Driscoll A, MacDonald PS, Hopper I, et al. National Heart Foundation of Australia and Cardiac Society of Australia and New Zealand: guidelines for the prevention, detection, and management of heart failure in Australia 2018. *Heart Lung Circ*. (2018) 27:1123–208. doi: 10.1016/j.hlc.2018.06.1042

29. McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Böhm M, et al. 2021 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure: developed by the task force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) with the special contribution of the heart failure association (HFA) of the ESC. *Eur Heart J*. (2021) 42:3599–726. doi: 10.1093/eurheartj/ehab368

30. Umemura S, Arima H, Arima S, Asayama K, Dohi Y, Hirooka Y, et al. The Japanese Society of Hypertension guidelines for the management of hypertension (JSH 2019). *Hypertens Res*. (2019) 42:1235–481. doi: 10.1038/s41440-019-0284-9

31. National Institute for Health and Care Excellence. Chronic obstructive pulmonary disease in over 16s: diagnosis and management. NICE guideline [NG115] (2019). Available at: https://www.nice.org.uk/guidance/ng115 (Accessed 2024-04-22)

32. Ismail Z, Black SE, Camicioli R, Chertkow H, Herrmann N, Laforce R Jr, et al. Recommendations of the 5th Canadian consensus conference on the diagnosis and treatment of dementia. *Alzheimers Dement*. (2020) 16:1182–95. doi: 10.1002/alz.12105

33. Frederiksen KS, Cooper C, Frisoni GB, Frölich L, Georges J, Kramberger MG, et al. A European academy of neurology guideline on medical management issues in dementia. *Eur J Neurol*. (2020) 27:1805–20. doi: 10.1111/ene.14412

34. National Institute for Health and Care Excellence. Dementia: assessment, management and support for people living with dementia and their carers. NICE guideline [NG97] (2018). Available at: https://www.nice.org.uk/guidance/ng97 (Accessed 2024-04-22)

35. World Health Organization. mhGAP Intervention Guide – Version 2.0: for mental, neurological and substance use disorders in non-specialized health settings (2019).

Available at: https://www.who.int/publications/i/item/9789241549790 (Accessed 2024-04-22)

36. Visseren FL, Mach F, Smulders YM, Carballo D, Koskinas KC, Bäck M, et al. 2021 ESC guidelines on cardiovascular disease prevention in clinical practice: developed by the task force for cardiovascular disease prevention in clinical practice with representatives of the European Society of Cardiology and 12 medical societies with the special contribution of the European Association of Preventive Cardiology (EAPC). *Eur J Prev Cardiol*. (2022) 29:5–115. doi: 10.1093/eurjpc/zwab154

37. Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Dennison Himmelfarb C, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association task force on clinical practice guidelines. *J Am Coll Cardiol*. (2018) 71:e127–248. doi: 10.1016/j.jacc.2017.11.006

38. Volkert D, Beck AM, Cederholm T, Cruz-Jentoft A, Goisser S, Hooper L, et al. ESPEN guideline on clinical nutrition and hydration in geriatrics. *Clin Nutr*. (2019) 38:10–47. doi: 10.1016/j.clnu.2018.05.024

39. Williams B, Mancia G, Spiering W, Agabiti Rosei E, Azizi M, Burnier M, et al. 2018 ESC/ESH guidelines for the management of arterial hypertension: the task force for the management of arterial hypertension of the European Society of Cardiology (ESC) and the European Society of Hypertension (ESH). *Eur Heart J*. (2018) 39:3021–104. doi: 10.1093/eurheartj/ehy339

40. Kleindorfer DO, Towfighi A, Chaturvedi S, Cockroft KM, Gutierrez J, Lombardi-Hill D, et al. 2021 guideline for the prevention of stroke in patients with stroke and transient ischemic attack: a guideline from the American Heart Association/American Stroke Association. *Stroke*. (2021) 52:e364–467. doi: 10.1161/STR.0000000000000375

41. Gladstone DJ, Lindsay MP, Douketis J, Smith EE, Dowlatshahi D, Wein T, et al. Canadian stroke best practice recommendations: secondary prevention of stroke update 2020. *Can J Neurol Sci*. (2022) 49:315–37. doi: 10.1017/cjn.2021.127

42. National Institute for Health and Care Excellence. Chronic heart failure in adults: diagnosis and management. NICE guideline [NG106]. (2018). Available at: https://www.nice.org.uk/guidance/ng106 (Accessed 2024-04-22)

43. American Diabetes Association. 12. Older adults: standards of medical Care in Diabetes—2021. *Diabetes Care*. (2021) 44:S168–79. doi: 10.2337/dc21-S012

44. National Institute for Health and Care Excellence. Type 2 diabetes in adults: management. NICE guideline [NG28]. (2015). Available at: https://www.nice.org.uk/guidance/ng28 (Accessed 2024-04-22)

45. Karter AJ, Warton EM, Lipska KJ, Ralston JD, Moffet HH, Jackson GG, et al. Development and validation of a tool to identify patients with type 2 diabetes at high risk of hypoglycemia-related emergency department or hospital use. *JAMA Intern Med*. (2017) 177:1461–70. doi: 10.1001/jamainternmed.2017.3844

46. National Institute for Health and Care Excellence. Diabetic foot problems: prevention and management. NICE guideline [NG19]. (2015). Available at: https://www.nice.org.uk/guidance/ng19 (Accessed 2024-04-22).

47. National Institute for Health and Care Excellence. Chronic kidney disease: assessment and management. NICE guideline [NG203]. (2021). Available at: https://www.nice.org.uk/guidance/ng203 (Accessed 2024-04-22).

48. Ketteler M, Block GA, Evenepoel P, Fukagawa M, Herzog CA, McCann L, et al. Executive summary of the 2017 KDIGO chronic kidney disease–mineral and bone disorder (CKD-MBD) guideline update: what's changed and why it matters. *Kidney Int*. (2017) 92:26–36. doi: 10.1016/j.kint.2017.04.006

49. Darryl Quarles L, Kendrick J. Management of hyperphosphatemia in adults with chronic kidney disease. (2024) Available at: https://www.uptodate.com/contents/management-of-hyperphosphatemia-in-adults-with-chronic-kidney-disease#H1687951240 (Accessed 2024-04-22)

50. Johnson KB, FitzHenry F. Case report: activity diagrams for integrating electronic prescribing tools into clinical workflow. *J Am Med Inform Assoc*. (2006) 13:391–5. doi: 10.1197/jamia.M2008

51. Spyrou S., Bamidis P., Pappas K., Maglaveras N. (2005). Extending UML activity diagrams for workflow modelling with clinical documents in regional health information systems. Connecting Medical Informatics and Bioinformatics: Proceedings of the 19th Medical Informatics Europe Conference (MIE2005), Geneva, Switzerland. (pp. 1160–1165).

52. Carvalho ECAD, Jayanti MK, Batilana AP, Kozan AM, Rodrigues MJ, Shah J, et al. Standardizing clinical trials workflow representation in UML for international site comparison. *PLoS One*. (2010) 5:e13893. doi: 10.1371/journal.pone.0013893

53. HL7 FHIR® based secure data repository. onFHIR.io. (2020) Available at https://onfhir.io/ (Accessed 2024-04-22).

54. National Institute for Health and Care Excellence. Medicines adherence: involving patients in decisions about prescribed medicines and supporting adherence. Clinical guideline [CG76]. (2009). Available at: https://www.nice.org.uk/guidance/cg76 (Accessed 2024-04-22).

55. Gooch P, Roudsari A. Computerization of workflows, guidelines, and care pathways: a review of implementation challenges for process-oriented health information systems. *J Am Med Inform Assoc*. (2011) 18:738–48. doi: 10.1136/amiajnl-2010-000033

56. Greenes RA, Bates DW, Kawamoto K, Middleton B, Osheroff J, Shahar Y. Clinical decision support models and frameworks: seeking to address research issues underlying implementation successes and failures. *J Biomed Inform*. (2018) 78:134–43. doi: 10.1016/j.jbi.2017.12.005

57. Laka M, Carter D, Milazzo A, Merlin T. Challenges and opportunities in implementing clinical decision support systems (CDSS) at scale: interviews with Australian policymakers. *Health Policy Technol*. (2022) 11:100652. doi: 10.1016/j.hlpt.2022.100652

58. Peart A, Lewis V, Barton C, Russell G. Healthcare professionals providing care coordination to people living with multimorbidity: an interpretative phenomenological analysis. *J Clin Nurs*. (2020) 29:2317–28. doi: 10.1111/jocn.15243

59. Zayas-Cabán T, Haque SN, Kemper N. Identifying opportunities for workflow automation in health care: lessons learned from other industries. *Appl Clin Inform*. (2021) 12:686–97. doi: 10.1055/s-0041-1731744

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 05 June 2024
DOI 10.3389/fmed.2024.1377209

OPEN ACCESS

EDITED BY
Oya Beyan,
University Hospital of Cologne, Germany

REVIEWED BY
Valentina Petkova,
Medical University of Sofia, Bulgaria
Stavros Pitoglou,
National Technical University of Athens, Greece

*CORRESPONDENCE
Katja Hoffmann
✉ katja.hoffmann@tu-dresden.de

RECEIVED 26 January 2024
ACCEPTED 08 May 2024
PUBLISHED 05 June 2024

CITATION
Hoffmann K, Nesterow I, Peng Y, Henke E, Barnett D, Klengel C, Gruhl M, Bartos M, Nüßler F, Gebler R, Grummt S, Seim A, Bathelt F, Reinecke I, Wolfien M, Weidner J and Sedlmayr M (2024) Streamlining intersectoral provision of real-world health data: a service platform for improved clinical research and patient care.
Front. Med. 11:1377209.
doi: 10.3389/fmed.2024.1377209

Streamlining intersectoral provision of real-world health data: a service platform for improved clinical research and patient care

Katja Hoffmann[1]*, Igor Nesterow[1], Yuan Peng[1], Elisa Henke[1], Daniela Barnett[2], Cigdem Klengel[1], Mirko Gruhl[1], Martin Bartos[3], Frank Nüßler[3], Richard Gebler[1], Sophia Grummt[1], Anne Seim[1], Franziska Bathelt[4], Ines Reinecke[2], Markus Wolfien[1,5], Jens Weidner[1] and Martin Sedlmayr[1]

[1]Faculty of Medicine Carl Gustav Carus, Institute for Medical Informatics and Biometry, Technische Universität Dresden, Dresden, Germany, [2]Data Integration Center, Center for Medical Informatics, University Hospital Carl Gustav Carus, Dresden, Germany, [3]Department of Informatics, Klinikum Chemnitz gGmbH, Chemnitz, Germany, [4]Thiem-Research GmbH, Cottbus, Germany, [5]Center for Scalable Data Analytics and Artificial Intelligence, Dresden, Germany

Introduction: Obtaining real-world data from routine clinical care is of growing interest for scientific research and personalized medicine. Despite the abundance of medical data across various facilities — including hospitals, outpatient clinics, and physician practices — the intersectoral exchange of information remains largely hindered due to differences in data structure, content, and adherence to data protection regulations. In response to this challenge, the Medical Informatics Initiative (MII) was launched in Germany, focusing initially on university hospitals to foster the exchange and utilization of real-world data through the development of standardized methods and tools, including the creation of a common core dataset. Our aim, as part of the Medical Informatics Research Hub in Saxony (MiHUBx), is to extend the MII concepts to non-university healthcare providers in a more seamless manner to enable the exchange of real-world data among intersectoral medical sites.

Methods: We investigated what services are needed to facilitate the provision of harmonized real-world data for cross-site research. On this basis, we designed a Service Platform Prototype that hosts services for data harmonization, adhering to the globally recognized Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) international standard communication format and the Observational Medical Outcomes Partnership (OMOP) common data model (CDM). Leveraging these standards, we implemented additional services facilitating data utilization, exchange and analysis. Throughout the development phase, we collaborated with an interdisciplinary team of experts from the fields of system administration, software engineering and technology acceptance to ensure that the solution is sustainable and reusable in the long term.

Results: We have developed the pre-built packages "ResearchData-to-FHIR," "FHIR-to-OMOP," and "Addons," which provide the services for data

harmonization and provision of project-related real-world data in both the FHIR MII Core dataset format (CDS) and the OMOP CDM format as well as utilization and a Service Platform Prototype to streamline data management and use.

Conclusion: Our development shows a possible approach to extend the MII concepts to non-university healthcare providers to enable cross-site research on real-world data. Our Service Platform Prototype can thus pave the way for intersectoral data sharing, federated analysis, and provision of SMART-on-FHIR applications to support clinical decision making.

KEYWORDS

service platform, intersectoral data sharing, health data space, real-world data, interoperability, data harmonization, research data infrastructure, secondary use of clinical data

1 Introduction

For scientific research, there is a high interest in using data from routine clinical care, so-called real-world data (RWD) (1–4). Although large amounts of RWD are available in various institutions, such as hospitals, outpatient clinics, and physician practices, the intersectoral data exchange between sites is hindered by their heterogeneity in terms of structure, content, and compliance with data protection regulations (2, 5). To address this challenge, the German Medical Informatics Initiative (MII) was launched in 2018, initially focusing on university hospitals to foster the exchange and utilization of RWD (6, 7). At that time, Data Integration Centers (DIC) were established at the medical sites of the university hospitals, and standardized solutions were developed for effective data use and exchange in both healthcare and research, with a focus on interoperability and data harmonization. For example, all university hospitals in the MII defined a dataset description, the MII Core dataset (CDS) (8) using the Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) international standard communication format (9). The FHIR MII CDS consists of basic modules (e.g., Person, Case, Diagnosis, Procedure, Laboratory Test Results, Medication), and extension modules (e.g., Oncology, Pathology Results, Molecular Genetics, Intensive Care) (8). This forms the foundation for cross-site data exchange and the integration of third-party applications via SMART-on-FHIR technology (10).

To facilitate the data exchange of RWD for scientific purposes, the HL7 International (11) and the Observational Health Data Sciences and Informatics (OHDSI) (12, 13) community have announced collaboration in 2021 (14). The OHDSI community develops the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) (15), as well as tools for data quality assessment and analysis (16). Common to all standardized data formats is the need to develop individual processes for the extraction, transformation and loading (ETL) of data from different data sources, which remains a major challenge (5). Nevertheless, hospitals have succeeded in overcoming these initial hurdles and making their own RWD available in a harmonized form for research (17, 18) and possibly also for patient care.

Since certain diseases, such as cardiovascular diseases, diabetes, allergies, and mental illnesses, are often not treated at university hospitals and there is insufficient data available, especially for rare diseases, it is essential to obtain RWD as well from non-university healthcare providers to ensure a comprehensive and diverse dataset for research studies. In order to make RWD available for research by non-university healthcare providers, the Medical Informatics Hub in Saxony (MiHUBx) was founded in 2021 (19). Among the project goals is to investigate whether the concepts of the MII can be transferred to non-university healthcare providers. As a first result within this project, Bathelt et al. (1) demonstrated the possibility to utilize an existing portable and standardized infrastructure from a university hospital setting and transferred it to non-university sites to support feasibility requests for participation in multicentre studies (1). However, the work of Bathelt et al. (1) was limited in terms of data availability, terminologies used, and harmonized data formats, so that RWD from non-university sites are still insufficiently available for research. Yet, due to limited human and economic resources and expertise in HL7 FHIR and OMOP CDM, it is hardly possible for non-university healthcare providers to develop the required services for data harmonization and provision in standardized data formats themselves. It is therefore necessary to provide the services in such a way that they can be conveniently deployed and easily used by data providers for different studies. In this paper, all tools for data harmonization, provision and management, such as databases, programs for ETL processes, analysis tools and other applications, are referred to as services.

The aim of this work is the development of pre-built packages that contain services and support the harmonization and provision of RWD in FHIR MII CDS and OMOP CDM format and thereby provide reuse potential for various projects. Another goal of this work is the development of a versatile and modular Service Platform Prototype that facilitates project administration, service management, data management and analysis. In this context, we focus on the following two research questions:

1. What services are needed to facilitate the provision of harmonized RWD for cross-site research?
2. How can the necessary services (from 1) be technically compiled so that they can be used by hospitals with few

resources and limited expertise in HL7 FHIR and OMOP CDM to make harmonized RWD accessible for research studies?

2 Materials and methods

2.1 Materials for data harmonization and provision

2.1.1 Compilation and integrated implementation of existing software resources

For the harmonization of RWD from source systems into the basic modules of the FHIR MII CDS version 1.0 (20), and into the OMOP CDM format (15), we have selected the following materials, as these are successfully used at the University Hospital Dresden. In addition, we investigated the applicability of other established tools. For this purpose, we used the following developed concepts of the MII and the software tools published by the MII consortium Medical Informatics in Research and Care in University Medicine (MIRACUM) (21) as MIRACOLIX Tools (22):

Clinical Data Repository: A Clinical Data Repository (CDR) is a database in which patient-centered healthcare data from various IT systems [e.g., electronic health record (EHR), laboratory system, biobank] are stored in a site-specific data model.

FHIR server BLAZE: A FHIR Server is a software solution that stores and manages FHIR resources. It acts as a bridge connecting healthcare applications and systems, allowing them to exchange patient information in a consistent and structured format or to answer population-wide aggregate queries quickly. The FHIR server BLAZE (23) was initially developed within the German Biobank Alliance project (24), aimed at high-throughput performance (25). BLAZE comes with a built-in feature to authenticate requests against an OpenID Connect provider.

MIRACUM FHIR Gateway: The MIRACUM FHIR Gateway (26) is a PostgreSQL database with a table for storing FHIR Resources, which are represented in JSON format. It serves as a temporary storage.

ETL process DWH-TO-FHIR: The ETL process DWH-TO-FHIR extracts research data, which is provided as database views, from the data warehouse (DWH) of the site, constructs FHIR resources according to the FHIR MII CDS structure definition v1.0 and loads them into a *MIRACUM FHIR Gateway*. The application DWH-TO-FHIR implemented Basic Authentication to enforce access controls to the FHIR server's resources. This involved sending a username and password in plain text over the network.

ETL process ROTATOR: The ROTATOR (fROm gaTewAy TO seRver) is an application that reads FHIR resources from the MIRACUM FHIR Gateway and loads them onto a FHIR Server, such as BLAZE.

OHDSI tools: The OHDSI tools are provided as Docker containers by Gruhl et al. (27). The basis of the OHDSI tools is the OMOP CDM PostgreSQL database, which is divided into the standardized OMOP CDM data tables v5.3 and the OMOP CDM standardized vocabularies (state February 2023), such as SNOMED CT (Systematized Nomenclature of Medicine–Clinical Terms), ICD-10-GM (International Classification of Diseases, German Modification) and OPS ("Operationen- und Prozedurenschlüssel," surgery and procedure key). Furthermore, the OHDSI tools by Gruhl et al. (27) includes additional dockerized tools for analyzing research data in OMOP CDM: (1) the R-based application ACHILLES that can be used for data characterization and visualization, (2) the web-based application ATLAS that can be used for cohort definition and scientific analysis, and (3) the R-based application Data Quality Dashboard (DQD) that can be used for data quality analysis.

ETL process FHIR-TO-OMOP: The ETL process FHIR-TO-OMOP extracts FHIR resources from a FHIR Server (e.g., the FHIR server BLAZE) or from the MIRACUM FHIR Gateway, transforms them into the standardized format of OMOP CDM and loads them into an OMOP CDM database (28, 29). The application provides HTTP Basic Authentication to enforce access controls to the FHIR resources of the FHIR server.

TRANSITION Database: The TRANSITION Database is a relational database that was originally developed for the semantic mapping of system-specific documented diagnoses to Orpha codes for rare diseases (30). It is deployed as a PostgreSQL database v14 (31), and offers tables with required terminology bindings (e.g., FHIR value sets), which can essentially support the semantic mapping of the RWD to required code systems. For example, the semantic mapping of the gender (e.g., value "1" for female) to a FHIR ValueSet (e.g., code "female" (32)) can be achieved via the TRANSITION Database. As an example of the database structure, the table for the semantic mapping of the vital parameters can be found as a csv file in the (**Supplementary File 2**).

Keycloak: Since medical data needs to be protected in an enhanced manner, the Keycloak server offers significant advantages (e.g., centralized Identity Management, Customizable Authentication Flows) for the medical domain, particularly in terms of securing sensitive patient data, ensuring compliance with healthcare regulations, and facilitating interoperability among diverse healthcare IT systems. The Authentication Server Keycloak (33) can be used as an OpenID (34), OAuth 2.0, or SAML Connect provider to validate requests to the FHIR endpoints.

2.1.2 Original contributions–expanding with custom software additions

We analyzed the already established data provision pipeline at University Hospital Dresden by consulting experts in the fields of data integration, provision, protection and security. From this, we derived a generic process for data harmonization and provision. Subsequently, we identified missing materials, which we have developed and adapted. These are technically described below and further explained in the "3 Results" section.

Research Data Repository: In order to be able to provide the ETL process DWH-TO-FHIR to other sites, a database structure is required which represents medical data and follows the logic of the basic modules of the FHIR MII CDS specification v1.0 (20). For this reason, we developed the Research Data Repository (RDR), deployed as a PostgreSQL database v14 (31). This repository encompasses tables for various FHIR resource types, like *Patient*, *Encounter*, *Condition*, *Observation*, *Procedure*, *Medication*, and *MedicationAdministration*, and is used for the cross-project and project-related storage of research data from routine clinical care (RWD).

Structural Mapping Guideline MII CDS: To reduce implementation time and to facilitate the specific data mapping, which must be done at each site, the Structural Mapping Guideline MII CDS is developed (**Supplementary File 1**).

The ETL process RDR-to-FHIR: To streamline the ETL process to provide RWD in FHIR MII CDS format, the sub-processes DWH-to-FHIR and ROTATOR were merged into the ETL process RDR-to-FHIR. RDR-to-FHIR loads project-related RWD from the RDR, constructs FHIR resources according to the FHIR MII CDS specification v1.0 and loads directly into the FHIR Server without buffering via the MIRACUM FHIR Gateway. To facilitate process execution, a RESTful API client was implemented to send data to the FHIR API. In addition, to enhance security beyond Basic Authentication in the ETL process DWH-TO-FHIR, we have implemented the authentication framework OAuth 2.0 for accessing resources of the FHIR server BLAZE. This approach uses bearer tokens created by the authentication server Keycloak.

ETL process FHIR-TO-OMOP: To enable OAuth 2.0 authentication against the FHIR server BLAZE, we added this authentication method to this application.

2.2 Pre-built packages for data harmonization and provision

We have developed pre-built packages, e.g., for the transformation of RWD to FHIR, to facilitate creating, running and connecting the multiple services (see section "2.1 Materials for data harmonization and provision"). In close collaboration with an interdisciplinary team of software developers, database engineers as well as infrastructure, security, and usability experts, we defined specific applications and composed the services to easy-to-install, pre-configurable installation packages based on Docker v24 (35) and Docker Compose v2 (36). We made our decisions based on previous experience and preliminary works. For testing purposes, we provided a test dataset.

To facilitate data and system administration, we have included the following additional materials in our packages, which can be used optionally.

pgAdmin: The open-source administration and management tool pgAdmin4 (37) provides a user-friendly web interface for the efficient administration of PostgreSQL databases used for RDR, TRANSITION database, Keycloak, and OMOP CDM.

Portainer: The open source container administration tool Portainer community edition (38) simplifies the management of containerized applications that are used to provide all the services described in the section "2.1 Materials for data harmonization and provision." Portainer provides a web interface for interacting with Docker containers (35) to create, manage, and deploy containers without the need for extensive knowledge of Docker commands.

2.3 Service Platform Prototype

To support research projects and system administration, we have developed a web-based Service Platform Prototype that facilitates project-related data provision for research projects. Based on experience with the design of user interfaces for interactive systems, an initial Low-Fidelity Prototype of the service platform was created to serve as a basis for discussion for a joint, interdisciplinary Graphical User Interface (GUI) design. By incorporating feedback from usability experts, this Low-Fidelity Prototype was iteratively transformed into a High-Fidelity Prototype. To this end, regular joint meetings were held with software developers and usability experts to ensure that the interaction principles for design solutions in accordance with ISO 9241-110 (39) were taken into account and implemented in the service platform interface (task appropriateness, self-descriptiveness, conformity to expectations, learnability, controllability, robustness against errors, user retention).

Our Service Platform Prototype was developed using Quarkus v3.4.3 (40). The frontend provides a form-based interface for users to specify the services and configurations needed for their project. The backend, managing the service deployment, is responsible for the instantiation of Docker containers, network setup, configuration management, and error handling. Angular v16.2.8 (41) was used for the frontend, complemented by Angular Bootstrap v16.0.0 (42). Communication between the microservices (e.g., to manage projects and Docker containers) is orchestrated by the Java Docker API v3.3.3 (43). Communication between the services is facilitated by RabbitMQ (44), a message broker that ensures reliable and real-time messaging.

3 Results

3.1 A generic data harmonization and provision process to foster data availability

We developed a generic process for the project-related provision of research data, which is shown in **Figure 1** and described below.

The data provision process begins with a project-independent and site-specific ETL pre-process CDR-TO-RDR [cf. **Figure 1** (Step 1)] that extracts RWD from the CDR, transforms it with the help of the *Structural Mapping Guideline* and the *TRANSITION Database* to the data structure of the *Research Data Repository (RDR)* and loads it into a *project-independent* instance of the *RDR*.

According to the MII concepts, the organizational starting point for the provision of data for a research project is a request from the data user. After the technical and legal feasibility check, the required data for the study is requested from the *project-independent RDR* instance and pseudonymized or anonymized in the subsequent process step to ensure that the research data does not allow any conclusions to be drawn about the identity of a patient [**Figure 1** (Step 2)].

To provide the project-related data in FHIR MII CDS format, instances of *RDR-TO-FHIR* and the *FHIR server BLAZE* are created and the ETL process is executed [**Figure 1** (Step 3)]. To provide the project-related data in the OMOP CDM format, instances of *FHIR-TO-OMOP* and the *OHDSI tools* (cf. OHDSI tools) are created and the ETL process is executed [**Figure 1** (Step 4)].

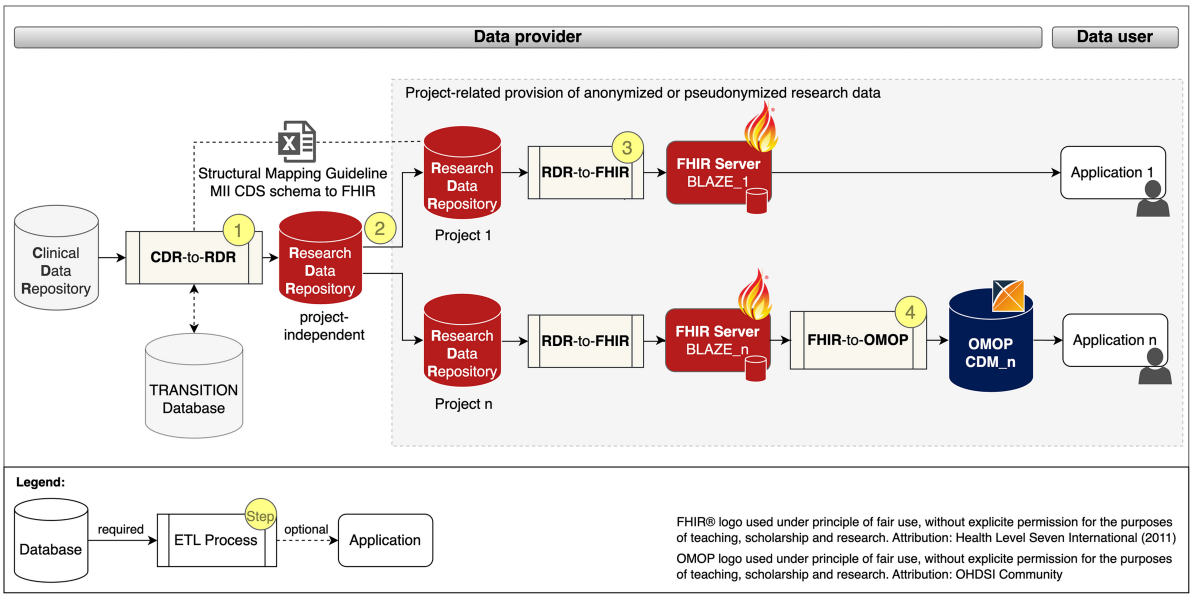

FIGURE 1
Project-related provision of research data for providers and users. Fast Healthcare Interoperability Resources (FHIR); Observational Medical Outcomes Partnership (OMOP); common data model (CDM); extract transform load (ETL); Clinical Data Repository (CDR); Research Data Repository (RDR): Medical Informatics Initiative (MII); Core dataset (CDS).

3.2 Pre-built packages to facilitate interoperability on the fly

We developed the following three pre-build packages: (1) ResearchData-TO-FHIR package for converting RWD to FHIR resources (cf. red bordered box in Figure 2), (2) FHIR-TO-OMOP package for converting the FHIR resources to the OMOP CDM format (cf. blue bordered box in Figure 2), and (3) the Addons package for deployment of optional services to simplify database, container, and security management (cf. gray bordered box in Figure 2). Figure 2 illustrates the composition of our implemented services and the data flows.

ResearchData-TO-FHIR package: The ResearchData-TO-FHIR package v2.2.0 provides a Docker compose specification that allows to automatically retrieve the images of the *RDR*, the *FHIR server BLAZE*, and the *RDR-to-FHIR* (cf. section "2.1 Materials for data harmonization and provision" and red bordered box in Figure 3), creates Docker containers as instances of the images that can be used to provide project-related research data in FHIR MII CDS format. For testing purposes, the package provides a test dataset. The package also includes the Structural Mapping Guideline (cf. Research Data Repository). By adjusting environment variables, the installation is customizable.

FHIR-TO-OMOP package: The FHIR-TO-OMOP package v1.1.0 (35) provides a Docker compose specification that allows to automatically retrieve the images of the *OHDSI tools* and the *FHIR-TO-OMOP* (cf. section "2.1 Materials for data harmonization and provision" and blue bordered box in Figure 3), creates Docker containers as instances of the images that can be used to provide project-related research data in OMOP CDM format. Through environment variables, the installation is customizable and the synthetic dataset from the ResearchData-TO-FHIR package can be used for testing purposes.

Addons package: The Addons package v1.0.0 (36) provides Docker compose specifications that allow to optionally deploy the PostgreSQL administration platform *pgAdmin*, the *TRANSITION database*, the container administration tool *Portainer* and the Authentication server *Keycloak* with a demo configuration for protecting the FHIR server Blaze (cf. ResearchData-TO-FHIR package).

The source code of the ResearchData-TO-FHIR package v2.2.1 (45), the FHIR-TO-OMOP package v1.1.0 (46) and the Addons package v1.0.0 (47) includes instructions for installation and usage as well as for further developments.

3.3 Service Platform Prototype to enable an easy to use and modular infrastructure

While the pre-built packages and their services, which are integrated into the Service Platform Prototype, were described in 3.2, the Service Platform Prototype v1.0.0 (48) itself is described below.

We implemented two key services: (1) the Project Management Service and (2) the Container Management Service. The Project Management Service allows researchers to create and manage their research projects. The Container Management Service automates the installation, operation and administration of services for each research project, such as the launch of ETL-processes and the display of technical details, such as unique identifiers, Docker names, assigned communication ports or operational status, to inform users about the current activity and availability of services.

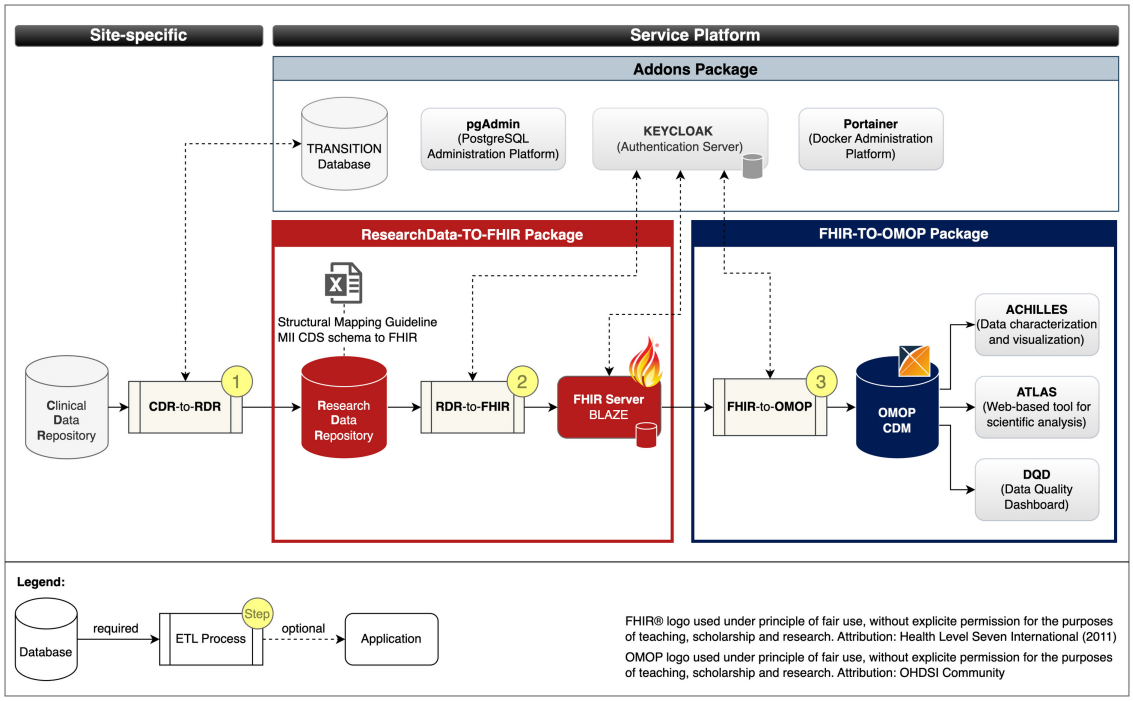

FIGURE 2
Services and data flows. Fast Healthcare Interoperability Resources (FHIR); Observational Medical Outcomes Partnership (OMOP); common data model (CDM); extract transform load (ETL); Clinical Data Repository (CDR); Research Data Repository (RDR): Medical Informatics Initiative (MII); Core dataset (CDS).

The frontend layout contains a navigation bar (Figure 3A), a sidebar (Figure 3B) and a main content area (Figures 3C, D). The navigation bar offers links to subpages that provide services and further information on database administration (i.e., via pgAdmin), authentication (i.e., Keycloak) and container administration (i.e., Portainer). The research projects are listed in the sidebar and new projects can be created via clicking on the respective button. The main content area is divided horizontally. The upper area displays the project-related services and offers functions for use and management, such as starting and stopping the Docker containers, starting the ETL processes, accessing the web frontend of the services and receiving further information (Figure 3C). The project-related containers are listed in the lower area, where further technical details are displayed, e.g., information on Docker containers, images, communication ports and operating status (Figure 3D).

The demo server, as an instance of the Service Platform Prototype, is available at https://tu-dresden.de/med/demoserver. This repository (48) also contains the developer documentation, including initial installation instructions.

4 Discussion

Our aim was to determine which services and processes are required to enable the provision of RWD for cross-site research and patient care, and how the services can be made available and usable to non-university healthcare providers with limited resources in a low-threshold manner. In particular, we determined which services are required for the harmonization of healthcare data into HL7 FHIR standard-based MII CDS format for intersectoral exchange, and to the OMOP CDM format for national or even international cross-site research. From this, we derived a process for the project-related provision of RWD based on the already established data provision pipelines of the university sites (21) in Germany. On this basis, we developed the pre-built packages *ResearchData-TO-FHIR*, *FHIR-TO-OMOP*, and *Addons* for data harmonization and provision to facilitate interoperability on the fly. To streamline the use of the particular services, especially for non-university healthcare providers, we developed a versatile and modular Service Platform Prototype that demonstrates the administration of research projects based on RWD.

To the best of our knowledge, our work is the first that shows how services for data harmonization, provision, and analytics can be provided to non-university healthcare providers in a low-threshold manner. The proposed pre-built Packages and Service Platform Prototype streamline the process of setting up research project environments and reduce the time and technical expertise required to provide RWD for research studies and feasibility inquiries, such as those conducted by the German Portal for Medical Research (Deutsches Forschungsdatenportal für Gesundheit, FDPG) (18, 49). Even though admission to the FDPG is currently only possible for sites that participate in the MII (mainly universities), our work is also highly relevant for non-university service providers, as feasibility studies can also be carried out using the OHDSI tools. We also consider it likely that our work may be of use to the European Health Data Space (EHDS) (50)

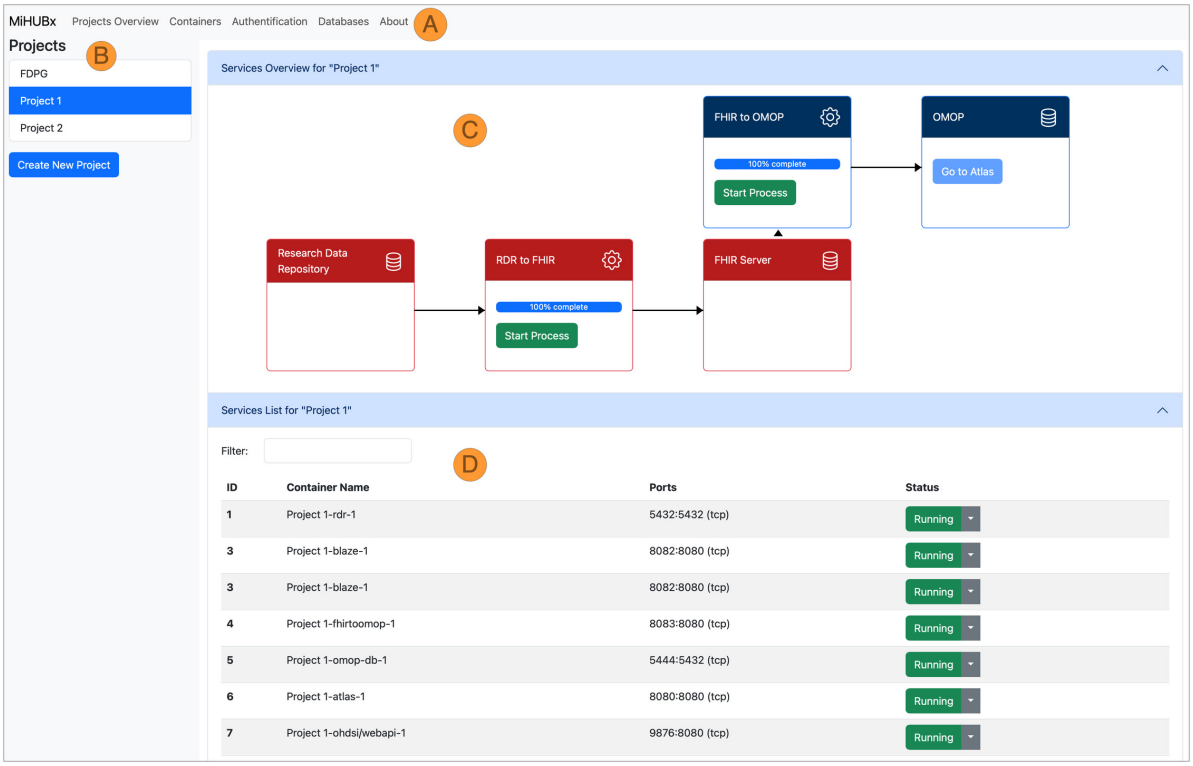

FIGURE 3
Schematic representation of the Service Platform Prototype. Navigation Bar **(A)**, Sidebar **(B)**, Main Content Area divided in Service Overview **(C)** and Service List **(D)**, Fast Healthcare Interoperability Resources (FHIR); Observational Medical Outcomes Partnership (OMOP); Research Data Repository (RDR); Identifier (ID); Medical Informatics Hub in Saxony (MiHUBx).

in the future, as our platform can be extended with appropriate services for data harmonization.

Although our work is a big step toward intersectoral provision of RWD, it has some limitations due to the following factors:

(1) Healthcare providers use site-specific clinical information systems to store RWD, which have limited standardized interfaces for the data exchange. Therefore, the RWD from possibly multiple data sources must be harmonized into the data format of our provided Research Data Repository in a pre-processing step by the sites themselves, which can be a challenge depending on the type and scope of data storage, the available human and economic resources and the knowledge of medical informatics. The legally binding interface specifications currently developed and established in Germany, such as the HL7 FHIR standard-based *Information technology systems in hospitals* (Informationstechnische Systeme in Krankenhäusern, ISIK) (51) and *Medical Information Objects* (Medizinische Informationsobjekte, MIO) (52), could provide a remedy here, as services based on these standards could be developed and made available via our platform to automatically convert the RWD into the FHIR MII CDS format for research purposes.

(2) The FHIR MII CDS v1.0, established as the standard for data exchange within the MII, presents certain limitations, particularly in its dataset specifications for specific medical fields like oncology and ophthalmology. Although plans are in place to refine profiles for oncology in the upcoming FHIR MII CDS v2, there remains a need to develop more appropriate dataset specifications for specific medical domains. These improvements are crucial for effectively incorporating such specifications into similar service platforms, especially in key areas like *Observations*, *Imaging Studies*, *Diagnostic Reports*, *Procedures*, and *Medication Administrations*.

(3) In addition, the services/applications cannot yet be installed "out of the box" via the frontend. In order to further minimize the technical hurdle, this is an important goal for the future.

Despite the limitations, our pre-built packages together with the Service Platform Prototype can already be used to provide data for specific research projects in a time-saving manner. We believe that our research represents a significant contribution in research data management, offering an efficient, user-friendly, and reproducible way to establish project-specific data provision pipelines. This may make our work interesting not only for non-university service providers, but also for university sites. In addition, our Service Platform Prototype can serve as a foundation for third-party applications, e.g., based on SMART-on-FHIR, which can be used not only for research but also for patient care.

Next, we will implement the Service System Platform at two non-university hospitals in Germany as part of a pilot study. In order to achieve a high level of acceptance among end users, we will test its functionality and usability. As part of the roll-out, an accompanying acceptance analysis in the form of an observation protocol and in-depth face-to-face interviews will be conducted to ensure the solution is usable and sustainable. For this purpose, we collaborate closely with experts from the fields of technology acceptance and usability. Thanks to our heterogeneous teams, the

interdisciplinary perspective can have a supportive effect in order to strengthen user-friendliness and thereby actual usage of the Service System Platform.

5 Conclusion

In conclusion, the developed Service Platform Prototype together with the pre-built packages represent an essential step forward in managing and facilitating medical research studies, with a focus on data harmonization, and collaborative effectiveness.

Data availability statement

The original contributions presented in this study are included in this article/**Supplementary material**, further inquiries can be directed to the corresponding author.

Author contributions

KH: Conceptualization, Investigation, Methodology, Project administration, Software, Visualization, Writing – original draft, Writing – review & editing. IN: Conceptualization, Methodology, Project administration, Software, Writing – review & editing. YP: Resources, Software, Validation, Writing – review & editing. EH: Resources, Software, Writing – review & editing. DB: Resources, Software, Writing – review & editing. CK: Software, Writing – review & editing. MG: Resources, Writing – review & editing. MB: Validation, Writing – review & editing. FN: Writing – review & editing. RG: Writing – review & editing. SG: Resources, Writing – review & editing. AS: Resources, Writing – review & editing. FB: Conceptualization, Funding acquisition, Resources, Writing – review & editing. IR: Resources, Writing – review & editing. MW: Supervision, Writing – review & editing. JW: Writing – review & editing. MS: Funding acquisition, Supervision, Writing – review & editing.

Funding

The authors declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the German Federal Ministry of Education and Research (https://www.bmbf.de/bmbf/en/) within the project "Medical Informatics Hub in Saxony (MiHUBx)." Grant numbers 01ZZ2101A (Dresden) and 01ZZ2101E (Chemnitz). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We would like to express our gratitude to medRxiv–the preprint server for Health Sciences–for providing a platform for the dissemination of our research findings prior to formal peer review (53). The accessibility and rapid dissemination that medRxiv provides has been instrumental in promoting scientific discourse and the exchange of ideas in our field. We recognize the valuable role of preprint servers in fostering collaboration and accelerating scientific discovery.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2024.1377209/full#supplementary-material

SUPPLEMENTARY FILE 1
Structural Mapping Guideline MII CDS schema to FHIR (Excel).

SUPPLEMENTARY FILE 2
Example table for the semantic mapping of vital parameter: observation_vital_parameter_loinc_code.csv (CSV).

References

1. Bathelt F, Reinecke I, Peng Y, Henke E, Weidner J, Bartos M, et al. Opportunities of digital infrastructures for disease management–exemplified on COVID-19-related change in diagnosis counts for diabetes-related eye diseases. *Nutrients.* (2022) 14:2016. doi: 10.3390/nu14102016

2. Liu F, Panagiotakos D. Real-world data: A brief review of the methods, applications, challenges and opportunities. *BMC Med Res Methodol.* (2022) 22:287. doi: 10.1186/s12874-022-01768-6

3. Orsini LS, Berger M, Crown W, Daniel G, Eichler HG, Goettsch W, et al. Improving transparency to build trust in real-world secondary data studies for hypothesis testing–why, what, and how: Recommendations and a road map from the real-world evidence transparency initiative. *Value Health.* (2020) 23:1128–36.

4. Kalra D. The importance of real-world data to precision medicine. *Pers Med.* (2019) 16:79–82.

5. Wolfien M, Ahmadi N, Fitzer K, Grummt S, Heine KL, Jung IC, et al. Ten topics to get started in medical informatics research. *J Med Internet Res.* (2023) 25:e45948. doi: 10.2196/45948

6. Gehring S, Eulenfeld R. German medical informatics initiative: Unlocking data for research and health care. *Methods Inf Med.* (2018) 57:e46–9.

7. Semler S, Wissing F, Heyder R. German medical informatics initiative: A national approach to integrating health data from patient care and medical research. *Methods Inf Med.* (2018) 57:e50–6.

8. Medical Informatics Initiative. *The medical informatics initiative's core data set*. Mainz: Medical Informatics Initiative (2023).

9. HL7 FHIR. *Health level seven international fast healthcare interoperability resources (HL7 FHIR)*. (2023). Available online at: http://hl7.org/fhir/ (accessed December 15, 2023).

10. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: A standards-based, interoperable apps platform for electronic health records. *Journal of the Am Med Inf Assoc.* (2016) 23:899–908.

11. HL7 International. 2024 Available online at: https://www.hl7.org (accessed January 13, 2024).

12. OHDSI. 2024 Available online at: https://www.ohdsi.org (accessed January 13, 2024).

13. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational health data sciences and informatics (OHDSI): Opportunities for observational researchers. *Stud Health Technol Inform.* (2015) 216:574–8.

14. OHDSI. *HL7 international and OHDSI announce collaboration to provide single common data model for sharing information in clinical care and observational research*. (2024). Available online at: https://www.ohdsi.org/ohdsi-hl7-collaboration/ (accessed January 13, 2024).

15. OHDSI. *Standardized data: The OMOP common data model*. (2024). Available online at: https://www.ohdsi.org/data-standardization/ (accessed December 15, 2023).

16. OHDSI. *Software tools*. (2024). Available online at: https://www.ohdsi.org/software-tools/ (accessed January 13, 2024).

17. Prokosch HU, Bahls T, Bialke M, Eils J, Fegeler C, Gruendner J, et al. editors. The COVID-19 data exchange platform of the German university medicine. In: Séroussi B, Weber P, Dhombres F, Grouin C, Liebe JD, Pelayo S, et al. editors. *Studies in health technology and informatics*. Amsterdam: IOS Press (2022).

18. Prokosch HU, Gebhardt M, Gruendner J, Kleinert P, Buckow K, Rosenau L, et al. editors. Towards a national portal for medical research data (FDPG): Vision, status, and lessons learned. In: Hägglund M, Blusi M, Bonacina S, Nilsson L, Cort Madsen I, Pelayo S, et al. editors. *Studies in health technology and informatics*. Amsterdam: IOS Press (2023). doi: 10.3233/SHTI230124

19. Deutsche Zentrum für Luft und Raumfahrt. *MiHUBx: Ein digitales ökosystem für forschung, diagnostik und therapie*. (2023). Available online at: https://www.gesundheitsforschung-bmbf.de/de/mihubx-ein-digitales-okosystem-fur-forschung-diagnostik-und-therapie-13054.php (accessed December 15, 2023).

20. Medizininformatik-Initiative. *Basismodule des kerndatensatzes der MII*. (2023). Available online at: https://www.medizininformatik-initiative.de/de/basismodule-des-kerndatensatzes-der-mii (accessed December 15, 2023).

21. Prokosch HU, Acker T, Bernarding J, Binder H, Boeker M, Boerries M, et al. MIRACUM: Medical informatics in research and care in university medicine: A large data sharing network to enhance translational research and medical care. *Methods Inf Med.* (2018) 57:e82–91.

22. Miracum. *MIRACOLIX tools*. (2024). Available online at: https://www.miracum.org/das-konsortium/datenintegrationszentren/miracolix-tools/ (accessed January 15, 2024).

23. Kiel A. *Blaze FHIR® server with internal, fast CQL engine*. (2023). Available online at: https://samply.github.io/blaze/ (accessed December 15, 2023).

24. Schüttler C, Prokosch HU, Hummel M, Lablans M, Kroll B, Engels C, et al. The journey to establishing an IT-infrastructure within the German biobank alliance. *PLoS One.* (2021) 16:e0257632. doi: 10.1371/journal.pone.0257632

25. Gruendner J, Deppenwiese N, Folz M, Köhler T, Kroll B, Prokosch HU, et al. The architecture of a feasibility query portal for distributed COVID-19 fast healthcare interoperability resources (FHIR) patient data repositories: Design and implementation study. *JMIR Med Inform.* (2022) 10:e36709. doi: 10.2196/36709

26. GitHub. *FHIR gateway*. (2024). Available online at: https://github.com/miracum/fhir-gateway (accessed January 15, 2024).

27. Gruhl M, Reinecke I, Sedlmayr M. Specification and distribution of vocabularies among consortial partners. *Stud Health Technol Inform.* (2020) 270:1393–4. doi: 10.3233/SHTI200458

28. Peng Y, Henke E, Reinecke I, Zoch M, Sedlmayr M, Bathelt F. An ETL-process design for data harmonization to participate in international research with German real-world data based on FHIR and OMOP CDM. *Int J Med Inform.* (2023) 169:104925. doi: 10.1016/j.ijmedinf.2022.104925

29. Henke E, Peng Y, Reinecke I, Zoch M, Sedlmayr M, Bathelt F. An ETL-process design for incremental loading German real-world data based on FHIR and OMOP CDM: Algorithm development and validation. *JMIR Med Inform.* (2023) 11:e47310. doi: 10.2196/47310

30. Kümmel M, Reinecke I, Gruhl M, Bathelt F, Sedlmayr M. *Transition database for a harmonized mapping of German patient data to the OMOP CDM – The 'German TDB'*. (2020). Available online at: https://www.ohdsi.org/wp-content/uploads/2020/05/13_Poster.pdf (accessed November 15, 2023).

31. PostgreSQL. *PostgreSQL: The world's most advanced open source relational database*. (2023). Available online at: https://www.postgresql.org (accessed December 16, 2023).

32. HL7. *FHIR core R4, administrative gender*. (2023). Available online at: https://hl7.org/fhir/r4/valueset-administrative-gender.html (accessed January 3, 2023).

33. Keycloak. *Open source identity and access management*. (2024). Available online at: https://www.keycloak.org (accessed April 19, 2024).

34. OpenID. *What is OpenID connect*. San Ramon, CA: OpenID Foundation (2023).

35. docker. *Make better, secure software from the start*. (2024). Available online at: https://www.docker.com (accessed December 16, 2023).

36. docker. *Docker compose*. (2023). Available online at: https://docs.docker.com/compose/ (accessed January 19, 2023).

37. pgAdmin. *Open source administration and development platform for PostgreSQL*. (2023). Available online at: https://www.pgadmin.org (accessed December 29, 2023).

38. Portainer.io. *Container management software for kubernetes and docker*. (2023). Available online at: https://www.portainer.io (accessed December 29, 2023).

39. Geis T, Tesch G. *Basiswissen usability und user experience: Aus– und weiterbildung zum UXQB certified professional for usability and user experience (CPUX) – foundation level (CPUX-F). 2., überarbeitete und aktualisierte Auflage*. Heidelberg: dpunkt.verlag (2023). 312 p.

40. Quarkus. *A Kubernetes native java stack tailored for OpenJDK HotSpot and GraalVM, crafted from the best of breed Java libraries and standards*. (2023). Available online at: https://quarkus.io (accessed December 16, 2023).

41. Angular. *The web development framework for building the future*. (2023). Available online at: https://angular.io (accessed December 16, 2023).

42. Bootstrap widgets. *The angular way*. (2024). Available online at: https://ng-bootstrap.github.io/ (accessed January 09, 2024).

43. gitHub. *Java docker API client*. (2024). Available online at: https://github.com/docker-java/docker-java (accessed January 08, 2024).

44. RabbitMQ. *Message broker*. (2023). Available online at: https://www.rabbitmq.com (accessed December 16, 2023).

45. Mihubx. *ResearchData-TO-FHIR package*. (2024). Available online at: https://gitlab.ukdd.de/pub/mihubx/fhir-to-omop-install-package (accessed May 14, 2024)

46. Mihubx. *FHIR-TO-OMOP package*. (2024). Available online at: https://gitlab.ukdd.de/mihubx/fhir-to-omop-install-pkg (accessed May 14, 2024)

47. Mihubx. *Addons package*. (2024). Available online at: https://gitlab.ukdd.de/pub/mihubx/addons-install-pkg (accessed May 14, 2024)

48. Mihubx. *MiHUBx service platform prototype*. (2024). Available online at: https://gitlab.ukdd.de/pub/mihubx/mihubx-service-portal (accessed May 14, 2024).

49. Das Deutsche Forschungsportal für Gesundheit [FDPG]. 2024. Available online at: https://forschen-fuer-gesundheit.de (accessed January 07, 2024).

50. Horgan D, Hajduch M, Vrana M, Soderberg J, Hughes N, Omar MI, et al. European health data space–an opportunity now to grasp the future of data-driven healthcare. *Healthcare.* (2022) 10:1629. doi: 10.3390/healthcare10091629

51. gematik. *ISiK – Informationstechnische systeme in krankenhäusern*. (2024). Available online at: https://fachportal.gematik.de/informationen-fuer/isik (accessed January 07, 2024).

52. KBV *MIO – Medizinische Informationsobjekte*. (2024). Available online at: https://mio.kbv.de/site/mio (accessed April 19, 2024).

53. Hoffmann K, Nesterow I, Peng Y, Henke E, Barnett D, Klengel C, et al. Streamlining intersectoral provision of real-world health data: A service platform for improved clinical research and patient care. *medRxiv* [Preprint]. (2024):doi: 10.1101/2024.01.29.24301922

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 27 June 2024
DOI 10.3389/fmed.2024.1348442

OPEN ACCESS

EDITED BY
Oya Beyan,
University Hospital of Cologne, Germany

REVIEWED BY
Adamantios Koumpis,
University Hospital of Cologne, Germany
Wilson Tumuhimbise,
Mbarara University of Science and Technology, Uganda
Zeinab Gholamnia Shirvani,
Babol University of Medical Sciences, Iran

*CORRESPONDENCE
Reem S. AlOmar
✉ rsomar@iau.edu.sa

RECEIVED 02 December 2023
ACCEPTED 05 June 2024
PUBLISHED 27 June 2024

CITATION
Aljerian NA, Alharbi AA, AlOmar RS, Binhotan MS, Alghamdi HA, Arafat MS, Aldhabib A and Alabdulaali MK (2024) Showcasing the Saudi e-referral system experience: the epidemiology and pattern of referrals utilising nationwide secondary data. *Front. Med.* 11:1348442. doi: 10.3389/fmed.2024.1348442

Showcasing the Saudi e-referral system experience: the epidemiology and pattern of referrals utilising nationwide secondary data

Nawfal A. Aljerian[1,2], Abdullah A. Alharbi[3], Reem S. AlOmar[4]*, Meshary S. Binhotan[5,6], Hani A. Alghamdi[7], Mohammed S. Arafat[1], Abdulrahman Aldhabib[1] and Mohammed K. Alabdulaali[8]

[1]Medical Referrals Centre, Ministry of Health, Riyadh, Saudi Arabia, [2]Emergency Medicine Department, King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia, [3]Family and Community Medicine Department, Faculty of Medicine, Jazan University, Jazan, Saudi Arabia, [4]Department of Family and Community Medicine, College of Medicine, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia, [5]Emergency Medical Services Department, College of Applied Medical Sciences, King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia, [6]King Abdullah International Medical Research Centre, Riyadh, Saudi Arabia, [7]Department of Family and Community Medicine, College of Medicine, King Saud University, Riyadh, Saudi Arabia, [8]Ministry of Health, Riyadh, Saudi Arabia

Introduction: Referrals are an integral part of any healthcare system. In the Kingdom of Saudi Arabia (KSA) an electronic referral (e-referral) system known as the Saudi Medical Appointments and Referrals Centre (SMARC) began formally functioning in 2019. This study aims to showcase the Saudi experience of the e-referral system and explore the epidemiology of referrals nationally.

Methods: This retrospective descriptive study utilised secondary collected data between 2020 and 2021 from the SMARC system. Cross tabulations with significance testing and colour-coded maps were used to highlight the patterns across all regions.

Results: The study analysed over 600,000 referral requests. The mean age of patients was 40.70 ± 24.66 years. Males had a higher number of referrals (55.43%). Referrals in 2021 were higher than those in 2020 (56.21%). Both the Autumn and Winter seasons had the highest number of referrals (27.09% and 27.43%, respectively). The Surgical specialty followed by Medicine had the highest referrals (26.07% and 22.27%, respectively). Life-saving referrals in the Central region were more than double those in other regions (14.56%). Emergency referrals were also highest in the Southern regions (44.06%). The Central and Eastern regions had higher referrals due to unavailable sub-speciality (68.86% and 67.93%, respectively). The Southern regions had higher referrals due to both unavailable machine and unavailable beds (18.44% and 6.24%, respectively).

Conclusion: This study shows a unique system in which referrals are between secondary, tertiary, and specialised care. It also highlights areas of improvement for equitable resource allocation and specialised care in slightly problematic areas as well as the use of population density in future planning.

KEYWORDS

epidemiology, e-referral, health policy, e-health systems, public health

Introduction

Digital health is transforming healthcare into real-time, individualised care, enhancing diagnosis, treatment, and patient empowerment (1). It provides opportunities beyond conventional healthcare for prevention, early illness detection, and chronic disease management (2). However, literature shows mixed results of this digital transformation across different countries (3). In new medicine, digital technologies can reinforce best practices like electronic referrals (4).

E-referrals are critical for providing quality healthcare. Efficient referral systems promote collaboration across all levels of care (5). Referral system success depends on many factors including patient barriers, resources, technology, and patient behaviour (6).

Saudi Arabia has recently undergone significant healthcare reforms and system changes as part of the National Transformation Programme launched in 2015 under Saudi Vision 2030. This aims to provide equitable, high-quality healthcare for all through innovations such as a robust digital health infrastructure (7–9). One key component of the digital health transformation is the establishment of a national electronic referral system known as the Saudi Medical Appointments and Referrals Centre (SMARC). SMARC facilitates referrals between healthcare facilities across all levels of care in the Kingdom. It utilises a Unified System of Medical Referrals (USMR) to receive and coordinate referral requests nationally through a centralised platform (10).

Whilst Saudi Arabia has made significant progress in implementing digital health, few studies have evaluated the impacts and effectiveness of these efforts. One study found preparedness amongst Saudi facilities for adapting to Vision 2030 changes was varied (11). Understanding patterns and utilisation of the new e-referral system across regions can provide insights into its performance and areas needing improvement.

The e-referral system within the Kingdom of Saudi Arabia (KSA), previously known as Ehalati, faced challenges when initially launched in 2012 including fragmented systems across hospitals, lack of integration between public and private facilities, and inadequate expertise in digital health solutions. The information technology platform at that time lacked features like artificial intelligence, robust data analytics, and interoperability, making centralised data management difficult. With many hospitals relying on their own individual platforms, doctors often depended on informal referral networks to coordinate care. However, aligned with Vision 2030, the centralised electronic referral system was revamped and fully reimplemented in 2019 as the SMARC (12).

Since 2019, substantial improvements have been made with Ministry of Health (MoH) support to transition to a unified, national e-referral platform. Targeted training programmes were implemented to build digital health capabilities across facilities. The SMARC system leverages advanced health information technologies like artificial intelligence and predictive analytics for improved care coordination. By standardising the e-referral platform and workflows across all public and private hospitals, SMARC addressed fragmentation and seamlessly integrates referrals digitally. With all governmental and majority of private healthcare facilities now connected to the centralised system, SMARC facilitates efficient nation-wide referral management and represents a major milestone in the digital transformation of Saudi Arabia's health sector.

Other countries have also implemented effective digital health systems, such as Catalonia's electronic health information exchange which has been a European leader since 2009. This system enabled critical health data sharing during the COVID-19 pandemic (13). Additionally, the European Health Data Space (EHDS) promotes individuals' electronic health data access and use for research and public benefit (14). Global digital health initiatives like Catalonia's and the EHDS exemplify how digital systems can improve health outcomes and research. Lessons from these efforts can inform Saudi Arabia's digital health advancements under Vision 2030.

Saudi Arabia currently serves a population of almost 34 million through a combination of public and private facilities across 13 administrative regions. As part of Vision 2030 reforms, the healthcare system is being upgraded to boost quality, efficiency and value through integrating public and private sectors. This includes establishing five new business units to manage the 13 healthcare regions alongside national insurance companies, overseen by the MoH and new insurance centres (15–17).

This study is the first to showcase the KSA's nationwide referral patterns using routine data from the new SMARC e-referral system. Examining referral epidemiology and trends will provide insights into the system's effectiveness and inform future optimisations to enhance its impact as a key digital health initiative under Vision 2030.

Materials and methods

Setting and data source

Under the new healthcare transformation adopted by the MoH, the 13 administrative areas will be pooled into five BUs as follows; Asir, Jazan, and Najran in the Southern BU; Aljouf, Hail, Northern Border and Tabuk in the Northern BU; Riyadh and Alqassim in the Central BU; Makkah, Medina, and Albaha in the Western BU, and the Eastern administrative area in the Eastern BU (15).

All hospitals have a designated coordination department usually known as the Office of Coordination and Eligibility for Treatment (OCET), which has access to the USMR. The OCET receives a referral request from the treating physician which is then uploaded to the USMR. Depending on a patient's medical condition, the referral request is uploaded as either lifesaving, emergency, or routine. These three types of referrals are categorised by SMARC to facilitate the referral process, and to timely secure acceptance to patients who are in most need.

For emergency and routine referrals, the OCET has the privilege to choose up to three hospitals that can potentially offer the needed service at the same region, or alternatively the USMR will automatically choose three appropriate hospitals. The SMARC system has built-in timeframes for referral requests to be accepted, depending on the urgency. For emergency referrals, hospitals have 72 h to accept the request, whilst routine referrals have a 14-day timeframe. If the initially chosen hospitals reject the request within the allotted timeframe, the request is sent to additional hospitals for consideration. If no hospital has accepted the referral request once the timeframe elapses, the case is escalated to the SMARC medical referral management team. They will find an appropriate alternative from the pool of public and private hospitals, searching both within the same region and in other regions if needed. Importantly, whilst awaiting

referral acceptance, patients continue receiving necessary healthcare management at the sending hospital to ensure stability until the transfer is arranged.

To expedite the referral request for life-threatening cases, SMARC offers a 24-h lifesaving hotline (1937) in which any treating physician can call directly. The call is answered by a SMARC lifesaving agent and directed to an on-call medical consultant for review and acceptance. If the request is accepted as a lifesaving by the on-call consultant, the treating physician through the OCET, will upload the request to the USMR along with the acceptance code and the name of the receiving hospital. Treating physicians who requested emergency referrals can also call and use this service when patients' health conditions deteriorate whilst waiting the emergency referral acceptance.

Additionally, SMARC oversees referrals for Saudi patients seeking to return to the KSA for treatment. For these cases, Saudi Embassies abroad have access to the USMR and may initiate a referral request. Figure 1 describes the process of the referral requests acceptance.

This study utilised routinely collected secondary data extracted from the SMARC e-referral system database between 2020 and 2021. Permission was obtained to access and analyse this de-identified dataset for research purposes, which was provided by the SMARC team after obtaining necessary approvals. The informed consent was waived given the retrospective nature of this study which relied solely on anonymized secondary data. The dataset was checked for completeness and consistency, and any incomplete or inconsistent records were removed prior to analysis. No personally identifiable information was included to maintain patient confidentiality. The variables included in the dataset are described in the Measurements section.

Study design

This retrospective study utilised secondary routinely collected data provided by the SMARC e-referral system. The dataset includes all referral requests submitted through the SMARC system nationally in 2020 and 2021, with no exclusion criteria applied.

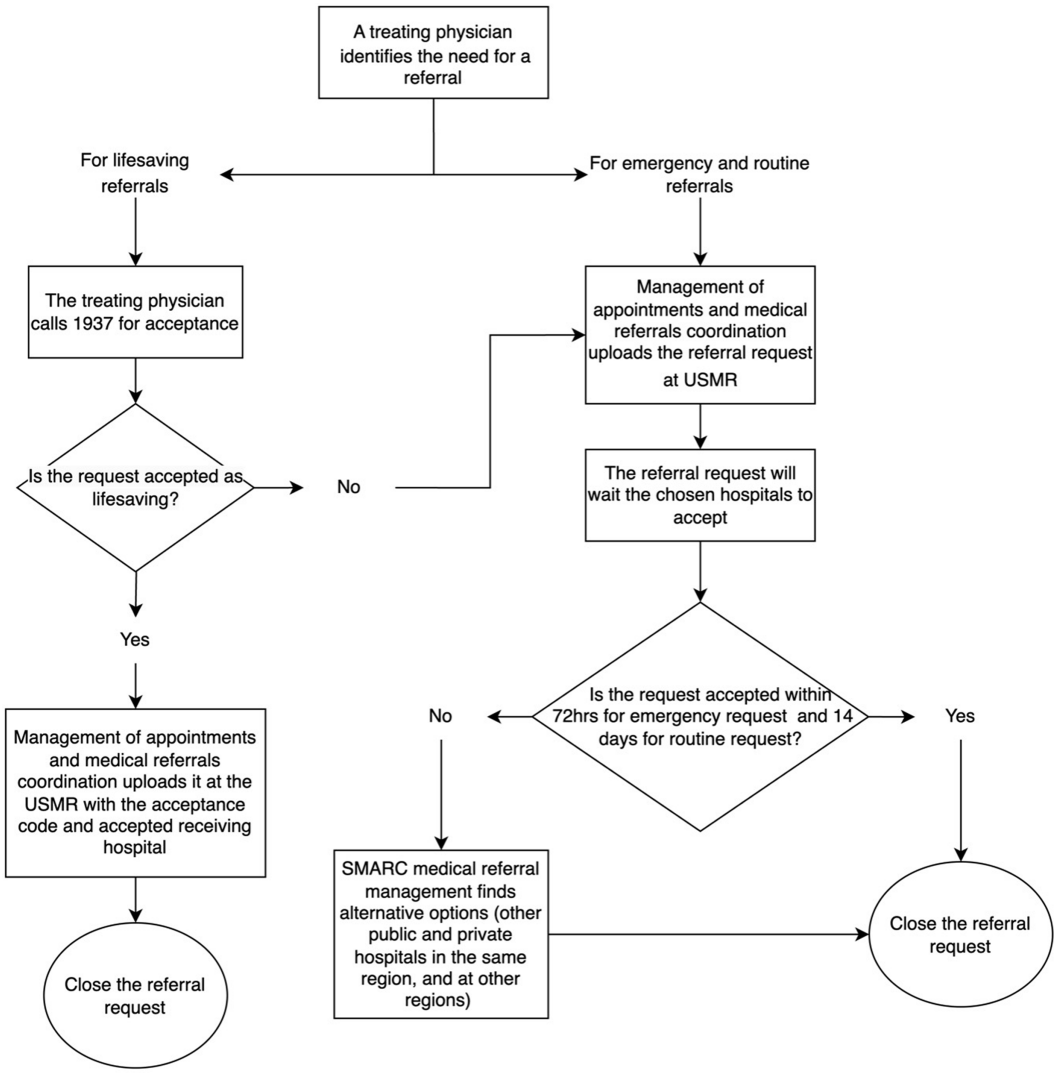

FIGURE 1
The process of referral acceptance across Saudi regions.

Ethical considerations

Both the MoH and Imam Abdulrahman Bin Faisal University institutional review boards have approved this study (23-77-E) and (IRB-2023-01-305). Standard precautions were taken to protect the confidentiality and privacy of patients' data involved.

Measurements

The dataset includes variables on sex, age, date of referral (month and year), type of referral (e.g., lifesaving, routine), bed type (e.g., ward bed or burn bed), reason for referral (e.g., unavailable speciality or unavailability of a specialised physician), medical speciality requesting the referral (e.g., medicine or surgery), region of referral request according to the five business units of the New Model of Care as well as according to the entire 13 administrative regions of the country.

Statistical analysis

To answer the objectives of the study, cross tabulations of explanatory variables according to the five BUs were performed, and tests of significance through Chi-squared tests and ANOVA tests were computed where appropriate. All analyses were run using the Stata statistical software version 16 (18). To further study the distribution of referral requests across the 13 administrative areas, colour-coded maps were drawn in ArcGIS (GIS software) version 10.0 (19), according to the percentage of referrals of each area.

Results

Sociodemographic characteristics of patients

Table 1 presents the sociodemographic characteristics of all patients. The total number of patients was 671,672 with an average age of 40.70 ± 24.66 years. Over 55% of referrals were for males. Non-Saudi's made up 15.11% of the total referrals. Most referral requests originated from the Western BU (34.99%), and the least originated from the Eastern BU (11.02%). Referrals were higher in 2021 compared to 2020 (56.21% and 43.79%, respectively).

TABLE 1 Sociodemographic characteristics of patients with referral requests.

Characteristics	Total (%) 671,672 (100.00)
Age (μ, SD)	36.88 (23.40)
Gender	
Males	372,308 (55.43)
Females	299,364 (44.57)
Nationality	
Non-Saudi	101,474 (15.11)
Saudi	570,198 (84.89)
Region (BUs)	
Central	101,793 (15.16)
Eastern	74,018 (11.02)
Western	235,020 (34.99)
Northern	118,212 (17.60)
Southern	142,629 (21.23)
Year	
2020	294,114 (43.79)
2021	377,558 (56.21)

Pattern of referrals across months

Upon examining the overall monthly pattern of referrals in Figure 2, both years of 2020 and 2021 have commenced with a high percentage of referrals, dipping to their lowest levels in April 2020 and May 2021. In 2020, the chart displays an initial decline in medical case referrals between February and March. From the beginning of April onwards, there is a significant graduate increase in referrals through to the year end. In contrast, 2021 displays a less consistent trend with more fluctuations and a significant increase in referrals in March, June, August, and December. Comparatively, referral rates from both years meet by the end of the year, indicating an expected new standard for medical referrals has developed.

Patterns across medical specialties

Figure 3 shows the pattern of medical specialties requesting referrals. Patients with referrals pertaining to internal medicine were the most common, reaching over a quarter of all requests (27.74%). Followed by general surgery and cardiac surgery (25.23% and 9.63%). The least common referral requests were for anaesthesia (0.03%).

Sociodemographic variables and region of referral request

Associations between sociodemographic variables and the region of referral request are presented in Table 2. Referrals originating from the Western and Southern regions were for patients who were relatively older than those from other regions (average age 42.38 and 41.63 years). Males dominated referrals from the Southern regions (59.18%), whereas for females, referrals were similarly high for both the Central and the Eastern regions. Requests for non-Saudis was highest in the Western region and lowest in the Eastern region (19.46% and 9.06%, respectively). All associations were significant at the 0.05 level.

Referral characteristics and region of referral request

The Central region had the highest number of referrals due to life saving events (14.56%), whereas the Northern region had the lowest

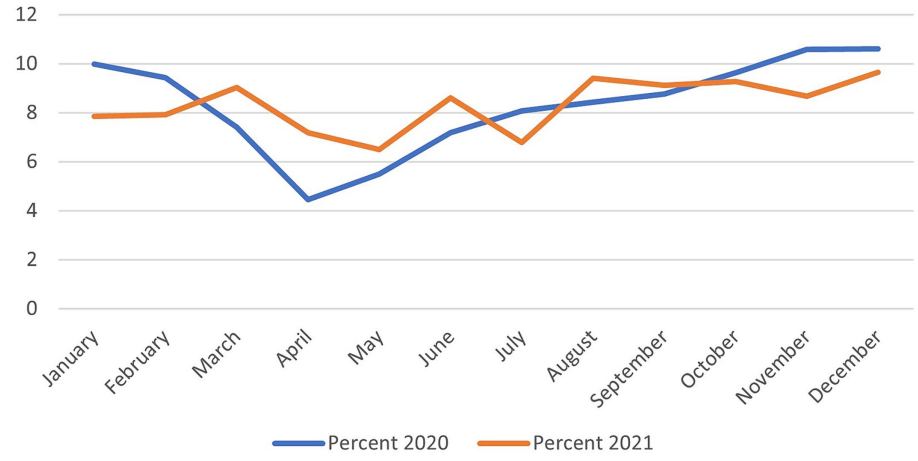

FIGURE 2
Pattern of monthly e-referrals for the years 2020 and 2021 across the Kingdom of Saudi Arabia.

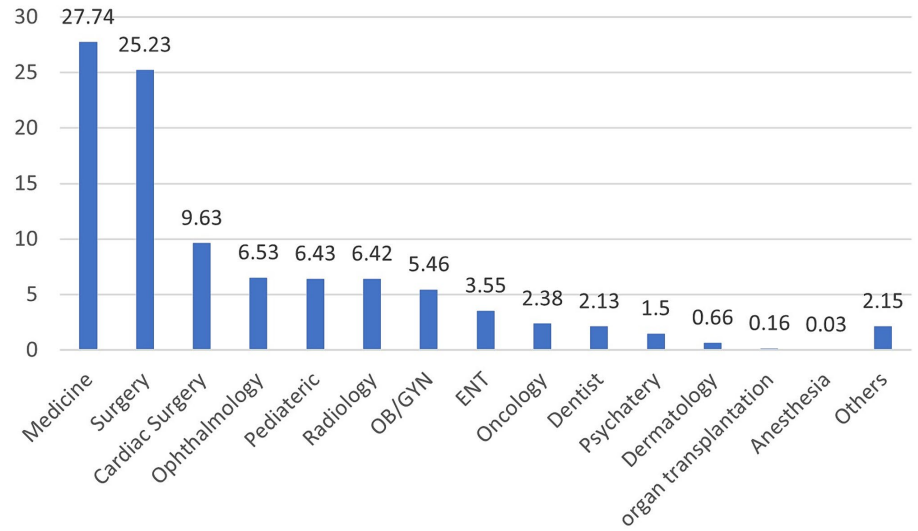

FIGURE 3
The e-referral requests by medical speciality for 2020 and 2021 across the Kingdom of Saudi Arabia.

at only 2.32%. For routine outpatient referrals, those originating from the Northern region were the highest reaching 57.39%. Emergency related referrals were most common in the Southern region and least common in the Eastern region (44.06% and 25.94%, respectively). As for dialysis, no referrals were registered from the Eastern region, whereas the Western region had 93 requests.

Also, 44.85% of referrals for ward beds were found to be in the Southern region. The Western region had the highest requests for isolation beds (5.63%). Requests for ICU beds was highest in the Central region (9.36%), whereas for CCU beds it was highest in the Western region (3.26%). As for PICU and NICU beds, they were highest in the Northern and Western regions, respectively.

As for reasons for referral, unavailable subspeciality was the most common reason and was highest in the Central region followed by the Eastern region (68.86% and 67.93%, respectively). The unavailability of a specialised physician was mostly reported in the Northern region (24.03%). The Southern region mostly reported the unavailability of a machine and the unavailability of a bed compared to all other regions (18.44% and 6.24%, respectively). Referrals due to social reasons were most commonly reported in the Western region, whilst there were 213 referrals due to a royal order from the Eastern region. Referrals due to injuries were only reported in the Southern region, whereas referrals due to health crises were highly reported in the Western region. All associations were significant at the 0.05 level (Table 3).

Total referral requests and referrals received by administrative areas

Both the Eastern and Makkah administrative areas were in the highest quintile with requests beyond 10.49% for both. However, Hail,

TABLE 2 Associations between sociodemographic variables and region of referral request.

Characteristics	Central 101,793 (15.16)	Eastern 74,018 (11.02)	Western 235,020 (34.99)	Northern 118,212 (17.60)	Southern 142,629 (21.23)
Age (μ, SD)	36.38 (22.46)	36.58 (23.06)	38.79 (23.48)	34.75 (23.55)	36.00 (23.77)
P-value	<0.001				
Gender					
Males	52,851 (51.92)	38,423 (51.91)	134,196(57.10)	62,546 (52.91)	84,408 (59.18)
Females	48,942 (48.08)	35,595 (48.09)	100,824 (42.90)	55,666(47.09)	58,221(40.82)
P-value	<0.001				
Nationality					
Non-Saudi	12,734 (12.51)	6,707(9.06)	45,737(19.46)	12,182 (10.31)	24,114 (16.91)
Saudi	89,059 (87.49)	67,311(90.94)	189,283 (80.54)	106,030 (89.69)	118,515(83.09)
P-value	<0.001				

Tabuk, and Najran administrative areas were within the lowest quintiles (Figure 4).

As for receiver areas, both Riyadh and Makkah were within the highest quintile both reaching above 12.80% of the total requests received. Whereas, Hail, Najran and the Northern areas were amongst the lowest (Figure 5).

Discussion

This study is the first to present the current status of the Saudi e-referral system. It also explored the patterns of e-referrals across the country utilising routinely collected data stored by the SMARC system. Patterns of referrals have been enormously studied worldwide (20–23). However, making clear comparisons are likely to be difficult due to differences between countries in, for example, local contexts and health care systems (24). Also, patterns of referrals in the current literature were mostly limited to primary healthcare referrals (20–23). This contrasts with the SMARC system in the KSA, which is concerned with secondary, tertiary, and specialised levels of care only. Also, this analysis of the Saudi e-referral system provides the first empirical evidence of inequalities across the different BUs. Previous studies in the KSA have shown that there are discrepancies in the quality of treatment provided to COVID-19 patients amongst the five different BUs at the outbreak's onset (15). Other several noteworthy observations can be drawn from the findings of this study.

Sex variations in e-referrals suggest the presence of disparities in healthcare-seeking patterns. Higher referrals amongst males are reflected in the higher proportion of males compared to females as shown in the 2022 census (25). However, sex variations were observed in referrals in other countries including America and Canada (23, 26, 27).

The observed discrepancy in the ratio of Saudi/non-Saudi patients may be ascribed to the inherent characteristics of the healthcare system and expatriates' situation in the KSA. Expatriates are primarily in the country for work purposes and are obligated to be medically fit in order to have a work visa. Also, since employers are required to provide health insurances for their foreign employees, most of them attain health services from private hospitals. This is despite the fact that free healthcare is provided to all citizens regardless of nationality in MoH facilities especially during the COVID-19 pandemic (28).

Regional variations in referral patterns likely stem from differences in healthcare resources and infrastructure. The uneven distribution of healthcare services across and within regions is well-established (29–33). Disparities between the five new business units in Saudi Arabia have been noted in prior studies on quality indicators for COVID-19 patients (15, 34, 35). Our analysis provides further evidence of disproportionality amongst the BUs regarding referral initiation and receipt.

The Eastern region and Makkah initiated the highest total number of referral requests. Contributing factors could include their greater population density (25) and regional health system capacities.

However, when examining referral request rates per 10,000 population, the Northern and Albaha regions were actually highest (Supplementary Table 1). This suggests medical resource limitations, also reflected when grouped into their respective Northern and Western BUs. Conversely, the Eastern region had the second lowest rate *per capita* despite having the most total referrals, highlighting the need to consider population density in resource allocation.

Riyadh and Makkah had the most referral requests. As the country's major healthcare hubs with advanced facilities and specialties (36), these regions likely attract more referrals due to advanced medical capabilities. Similar regional differences have been observed elsewhere globally (26). Further research into the distribution of health system resources, such as workforce and facilities, is needed to fully explain the variations in medical referrals across Saudi Arabia's regions.

Internal medicine emerged as the most commonly referring speciality highlighting the prevalence of chronic diseases and cardiovascular conditions in the population (37). Surgical related specialities followed which may be due to shortage of surgical staff, particularly in surgical sub-specialities. Patients and referring physicians often prefer and trust specialised centres, further driving the demand for surgical services (38). Comparatively, in Canada, dermatology was one of the top referred specialties, whereas in the KSA, dermatology related referrals were low (39).

One notable finding is the high number of referrals due to unavailable subspecialties, which is particularly high in the Central and Eastern regions. This may be an indication of a shortage of certain

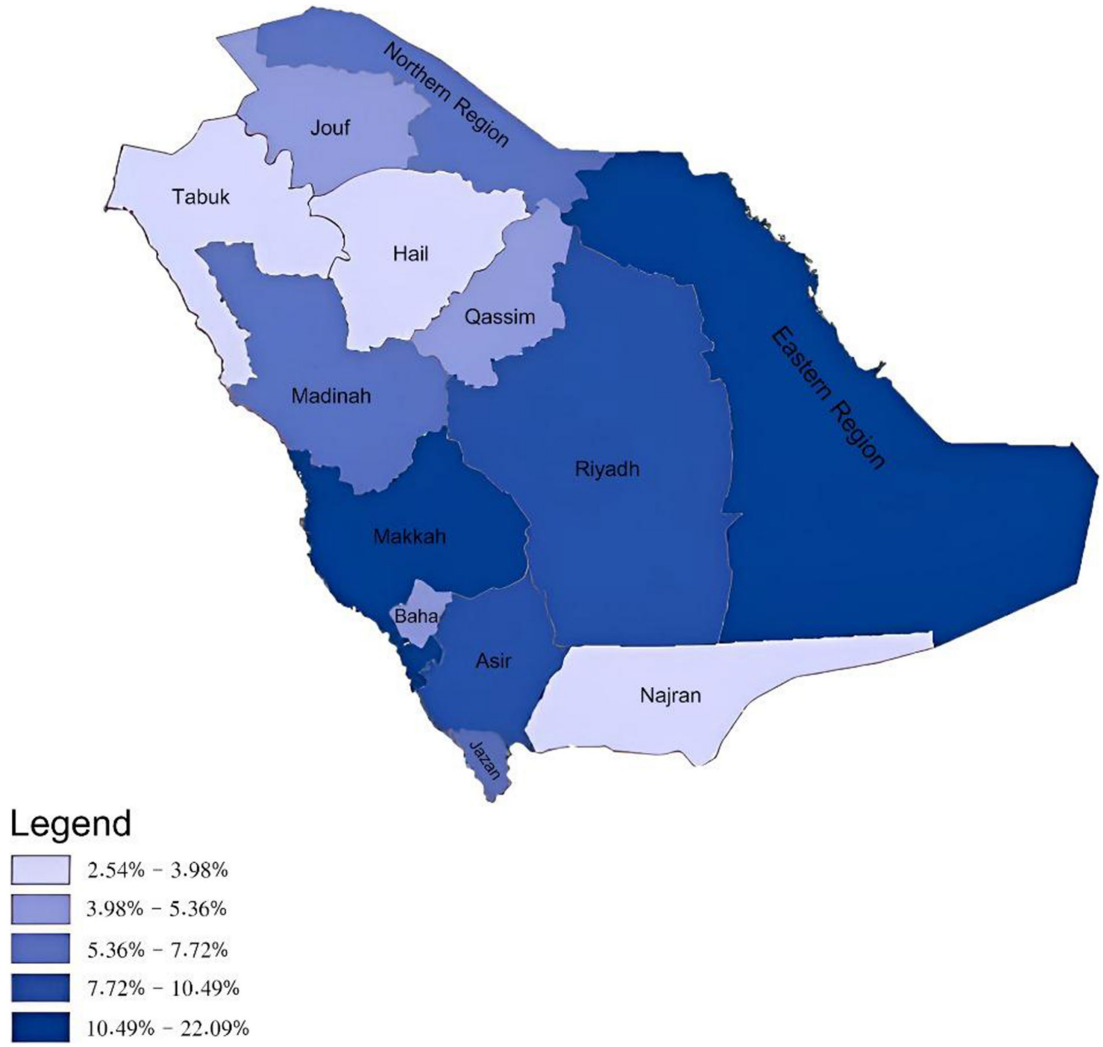

FIGURE 4
Total e-referral requests sent according to the 13 administrative areas of the Kingdom of Saudi Arabia.

sub-specialities, where despite the fact that these areas are home to two of the main cities of the country namely Riyadh and Dammam, which both include excellent healthcare services and high quality of care, these cities are surrounded by smaller towns with hospitals equipped with lesser specialised staff referring to those main cities. It may also indicate that healthcare staff within those areas pursue a high-quality of care (40). Also, the Northern region stands out with a significant proportion of referrals attributed to the unavailability of a specialised physician, indicating a potential need for improved access to specialised care in that region. In contrast, the Southern region reports a higher frequency of referrals due to the unavailability of a machine and bed, indicating infrastructure-related challenges. Additionally, it is worth noting that referrals due to injuries were exclusively reported in the Southern region. This suggests that the Southern region may have a higher prevalence of injury-related incidents or a greater need for specialised care for injuries compared to other regions. The Western region shows a higher occurrence of referrals due to social reasons, potentially reflecting the influence of social and cultural factors on healthcare-seeking behaviour.

Discrepancies in bed types across regions indicates potential differences in healthcare needs and allocation of resources. The Southern region exhibits a relatively higher number of referrals for ward beds, this may be attributed to the concentration of general hospitals or specialised facilities within that particular geographic location. The assignment of distinct bed categorisations, such as burn beds and isolation beds, may be a result of various factors, such as the prevalence of illnesses in a specific geographical region, the demand for specialised medical treatments, and the demographic attributes of the populace.

Lastly, referral rates are influenced by national and international incidents. In 2020, COVID-19 pandemic and its consequences including the lockdown could explain the low referral rate in 2020 compared to 2021. The drop in referrals during the pandemic has also been seen in different settings such as emergency departments, and heart diseases in countries including Italy and the United Kingdom (41, 42). However, this is the first study to observe the influence on a national level. With the implementation of the new regional healthcare transformations under the 2030 Vision and the merging of the existing 13 regions into five BUs; this research promises to encourage greater

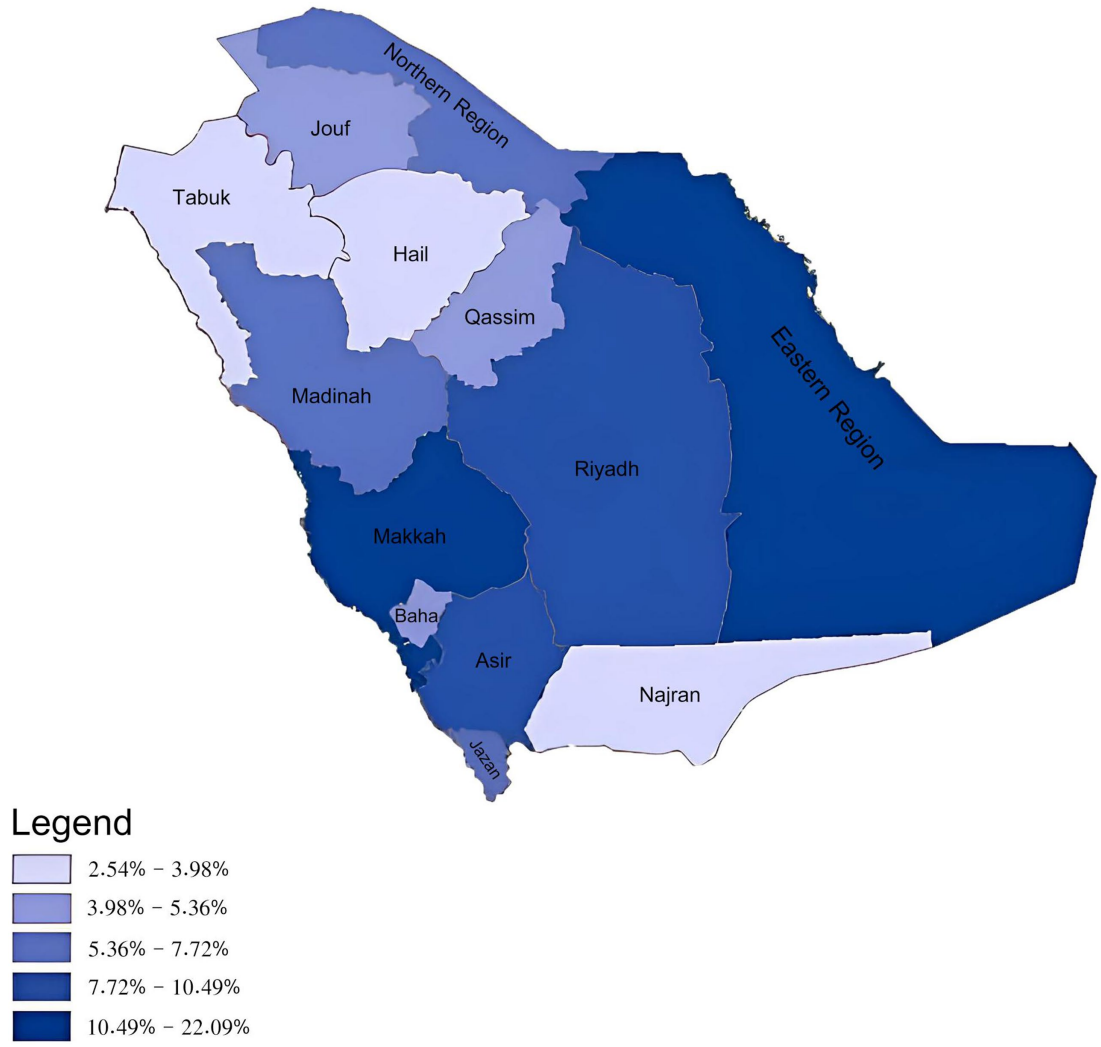

FIGURE 5
Total e-referral requests received according to the 13 administrative areas of the Kingdom of Saudi Arabia.

dedication to increasing the outstanding quality and equitable distribution of healthcare services.

Current findings show a momentary view of national referral patterns during a two-year time frame. These results provide an opportunity for improvement in terms of equity in resource allocation as well as enhancement of specialised care especially in problematic areas highlighted here. Furthermore, the use of this nationwide secondary data enabled us to explore the patterns of e-referrals across the country. However, certain limitations should be addressed. The reliance on secondary data obtained from the e-referral system limits the scope of variables examined. Also, the absence of similar studies in the wider literature makes direct comparisons challenging. Additionally, the study's focus on referral patterns may overlook other important aspects of healthcare, such as primary care utilisation or patient outcomes. Future research should address these limitations to provide a more comprehensive understanding of healthcare utilisation and effectiveness.

Additionally, several key insights for healthcare systems worldwide can be drawn from the evolution of SMARC. First, a unified e-health platform not only enhances service quality and efficiency but also improves access, conserves resources and eliminates service redundancies. Second, centralised tracking allows effective monitoring of health outcomes and resource utilisation, which aids in the identification of strengths and weaknesses within the system. Finally, this integrated approach increases strategic resource distribution, informs health policy and advances academic research, leading to greater optimization of healthcare delivery (43, 44). Investments in e-health and digital health provide economic benefits as well through streamlining operations and reducing administrative burdens, both of which are achieved by automating processes and minimising the need for in-person consultations. Digital health technologies improve diagnostic and treatment accuracy, improving patient outcomes and reducing medical errors; and extend service reach, particularly in underserved areas, maximising resource utilisation. These benefits contribute to an overall improvement in the efficiency of the healthcare system, ultimately leading to lower costs over the long-term (43, 44).

TABLE 3 The e-referral characteristics and region of referral requests in 2020 and 2021 across the Kingdom of Saudi Arabia.

Characteristic	Total 671,672 (100.00)	Central 101,793 (15.16)	Eastern 74,018 (11.02)	Western 235,020 (34.99)	Northern 118,212 (17.60)	Southern 142,629 (21.23)
Referral types						
Life saving	47,315 (7.04)	14,820 (14.56)	3,793 (5.12)	16,516 (7.03)	2,748 (2.32)	9,438 (6.62)
Routine OPD	317,484 (47.27)	51,944)51.03(	42,333 (57.19)	100,378 (42.71)	67,846 (57.39)	54,983 (38.55)
Routine inpatient	85,955 (12.80)	7,997 (7.86)	8,693 (11.74)	38,150 (16.23)	15,760 (13.33)	15,355 (10.77)
ER	220,802 (32.87)	27,030 (26.55)	19,199 (25.94)	79,883 (33.99)	31,854 (26.95)	62,836 (44.06)
Dialysis	116 (0.02)	2 (0.00)	0 (0.00)	93 (0.04)	4 (0.00)	17 (0.01)
P-value		<0.001				
Bed type						
OPD no bed	316,152 (47.07)	51,691 (50.78)	42,209 (57.03)	99,938 (42.52)	67,654 (57.23)	54,660 (38.32)
Ward	242,731 (36.14)	33,239 (32.65)	22,552 (30.47)	87,203 (37.10)	35,762 (30.25)	63,975 (44.85)
Burning bed	630 (0.09)	97 (0.10)	65 (0.09)	254 (0.11)	71 (0.06)	143 (0.10)
Isolation bed	27,067 (04.03)	1,742 (1.71)	2,557 (3.45)	13,225 (5.63)	3,024 (2.56)	6,519 (4.57)
ICU	47,217 (07.03)	9,528 (9.36)	4,247 (5.74)	19,505 (8.30)	4,825 (4.08)	9,112 (6.39)
CCU	18,603 (02.77)	2,360 (2.32)	1,094 (1.48)	7,653 (3.26)	3,067 (2.59)	4,429 (3.11)
PICU	8,102 (01.21)	1,392 (1.37)	659 (0.89)	2,702 (1.15)	1,664 (1.41)	1,685 (1.18)
NICU	11,170 (01.66)	1,744 (1.71)	635 (0.86)	4,540 (1.93)	2,145 (1.81)	2,106 (1.48)
P-value		<0.001				
Reason of referral						
Unavailable subspecialty	413,619 (61.38)	70,300 (68.86)	50,429 (67.93)	144,614 (61.33)	68,582(57.82)	79,694 (55.68)
Unavailable physician	114,882 (17.05)	17,736 (17.35)	13,233 (17.78)	35,724 (15.17)	28,482 (24.03)	19,706 (13.76)
Unavailable machine	89,790 (13.33)	10,932 (10.74)	5,938 (7.95)	29,420 (12.49)	17,080 (14.40)	26,361 (18.44)
Unavailable bed	24,305 (03.61)	1,817 (01.77)	1,956 (02.62)	10,057 (04.27)	1,537 (01.29)	8,921 (06.24)
Social reason	1,662 (0.25)	95 (0.09)	328 (0.44)	1,111 (0.47)	100 (0.08)	28 (0.02)
Health crisis	27,414 (04.06)	1,209 (01.17)	2,233 (02.99)	14,678 (06.24)	2,785 (02.34)	6,509 (04.55)
P-value		<0.001				

Conclusion

This study examined the underlying mechanism of an important telehealth tool, namely, e-referrals. Certain patterns were observed which included higher referrals for males, as well as in internal medicine and surgical related specialities, and unavailable subspeciality being the most commonly reported reason for referrals. We recommend the use of population density in the future planning of resource allocation and specialised care.

Data availability statement

The data analysed in this study is subject to the following licenses/restrictions: the data was requested from the Ministry of Health. Any researcher can also request this data provided they have an ethical approval. Requests to access these datasets should be directed to https://www.moh.gov.sa.

Ethics statement

Ethical approval for the study was obtained from the Ministry of Health's Institutional Review Board (23-77-E) and Imam Abdulrahman Bin Faisal's Institutional Review Board (IRB-2023-01-305). The informed consent was waived given the retrospective nature of this study which relied solely on anonymized secondary data.

Author contributions

NA: Conceptualization, Data curation, Supervision, Writing – review & editing. AAlh: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. RA: Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. MB: Investigation, Validation, Visualization, Writing – original draft, Writing – review & editing. HA: Project administration, Validation, Visualization,

Writing – original draft, Writing – review & editing. MAr: Conceptualization, Validation, Visualization, Writing – review & editing. AAld: Conceptualization, Validation, Visualization, Writing – review & editing. MAl: Conceptualization, Project administration, Validation, Visualization, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

The authors would like to thank all personnel involved in maintaining the SMARC database. The authors would also like to thank Maps Development Est. (MapDev Est.) (Member of Zaki Farsi Group) for kindly providing updated maps of the Kingdom of Saudi Arabia used in this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2024.1348442/full#supplementary-material

References

1. European Commission. *Communication from the commission to the European Parliament, the council, the European economic and social committee and the Committee of the Regions Youth Opportunities Initiative*. Belgium: European Commission Brussels (2011).

2. Awad A, Trenfield SJ, Pollard TD, Ong JJ, Elbadawi M, McCoubrey LE, et al. Connected healthcare: improving patient care using digital health technologies. *Adv Drug Deliv Rev*. (2021) 178:113958. doi: 10.1016/j.addr.2021.113958

3. Baltaxe E, Czypionka T, Kraus M, Reiss M, Askildsen JE, Grenkovic R, et al. Digital health transformation of integrated care in Europe: overarching analysis of 17 integrated care programs. *J Med Internet Res*. (2019) 21:e14956. doi: 10.2196/14956

4. Huebner L-A, Mohammed HT, Menezes R. *Using digital health to Support Best Practices: Impact of MRI Ordering Guidelines Embedded Within an Electronic Referral Solution*. Amsterdam, The Netherlands: ITCH (2019).

5. Seyed-Nezhad M, Ahmadi B, Akbari-Sari A. Factors affecting the successful implementation of the referral system: a scoping review. *J Family Med Prim Care*. (2021) 10:4364–75. doi: 10.4103/jfmpc.jfmpc_514_21

6. Nguru K, Ireri L. Challenges influencing proper implementation of quality health care referral system in Kaloleni sub-county, Kilifi County in Kenya. *Int Emerg Nurs*. (2022) 62:101169. doi: 10.1016/j.ienj.2022.101169

7. Yusuf N. Private and public healthcare in Saudi Arabia: future challenges. *Int J Bus Econ Dev*. (2014) 2:114–8.

8. Almalki M, FitzGerald G, Clark M. Health care system in Saudi Arabia: an overview. *East Mediterr Health J*. (2011) 17:784, 2011–2793. doi: 10.26719/2011.17.10.784

9. al-Kahtani N, Alrawiai S, al-Zahrani BM, Abumadini RA, Aljaffary A, Hariri B, et al. Digital health transformation in Saudi Arabia: a cross-sectional analysis using healthcare information and management systems society' digital health indicators. *Digit Health*. (2022) 8:205520762211177. doi: 10.1177/20552076221117742

10. Aljerian NA, Alharbi AA, Alghamdi HA, Binhotan MS, AlOmar RS, Alsultan AK, et al. External Vs Internal e-Referrals: Results from a Nationwide Epidemiological Study Utilizing Secondary Collected Data. *Risk Manag Healthc Policy*. (2024) 739–51. doi: 10.2147/RMHP.S453042

11. Alharbi MF. An analysis of the Saudi health-care system's readiness to change in the context of the Saudi National Health-care Plan in vision 2030. *Int J Health Sci*. (2018) 12:83.

12. MoH. Referral program (Ehalati): Ministry of Health; (2017). Available at: https://www.moh.gov.sa/en/Ministry/Structure/Programs/Referral/Pages/default.aspx.

13. Pérez Sust P, Solans O, Fajardo JC, Medina Peralta M, Rodenas P, Gabaldà J, et al. Turning the crisis into an opportunity: digital health strategies deployed during the COVID-19 outbreak. *JMIR Public Health Surveill*. (2020) 6:e19106. doi: 10.2196/19106

14. The European Council. European health data space: Council and parliament strike deal (2024). Available at: https://www.consilium.europa.eu/en/press/press-releases/2024/03/15/european-health-data-space-council-and-parliament-strike-provisional-deal/#:~:text=The%20aim%20of%20the%20EHDS,of%20the%20European%20Health%20Union.

15. Alharbi AA, Alqassim AY, Muaddi MA, Alghamdi SS. Regional differences in COVID-19 mortality rates in the Kingdom of Saudi Arabia: a simulation of the new model of care. *Cureus*. (2021) 13:e20797. doi: 10.7759/cureus.20797

16. Rahman R. The privatization of health care system in Saudi Arabia. *Health Serv Insights*. (2020) 13:117863292093449. doi: 10.1177/1178632920934497

17. MoH. *Health sector: Transformation strategy*. Saudi Arabia: MoH (2019).

18. StataCorp. *Stata statistical software: Release 16*. College Station, TX: StataCorp LLC (2019).

19. ESRI. *ArcGIS desktop: Release 10*. Redlands, CA: Environmental Systems Research Institute (2011).

20. Forrest CB, Majeed A, Weiner JP, Carroll K, Bindman AB. Comparison of specialty referral rates in the United Kingdom and the United States: retrospective cohort analysis. *BMJ*. (2002) 325:370–1. doi: 10.1136/bmj.325.7360.370

21. Shadd J, Ryan BL, Maddocks H, Thind A. Patterns of referral in a Canadian primary care electronic health record database: retrospective cross-sectional analysis. *Inform Prim Care*. (2011) 19:217–23. doi: 10.14236/jhi.v19i4.816

22. Ringberg U, Fleten N, Deraas TS, Hasvold T, Førde O. High referral rates to secondary care by general practitioners in Norway are associated with GPs' gender and specialist qualifications in family medicine, a study of 4350 consultations. *BMC Health Serv Res*. (2013) 13:147. doi: 10.1186/1472-6963-13-147

23. Liddy C, Singh J, Kelly R, Dahrouge S, Taljaard M, Younger J. What is the impact of primary care model type on specialist referral rates? A cross-sectional study. *BMC Fam Pract*. (2014) 15:22. doi: 10.1186/1471-2296-15-22

24. Liddy C, Arbab-Tafti S, Moroz I, Keely E. Primary care physician referral patterns in Ontario, Canada: a descriptive analysis of self-reported referral data. *BMC Fam Pract*. (2017) 18:1–8. doi: 10.1186/s12875-017-0654-9

25. GAS. General Authority for Statistics: Saudi census (2023) Available at: https://portal.saudicensus.sa/portal.

26. Reimer AP, Schiltz N, Koroukian SM, Madigan EA. National incidence of medical transfer: patient characteristics and regional variation. *J Health Hum Serv Adm*. (2016) 38:509–28. doi: 10.1177/107937391603800404

27. Tandjung R, Morell S, Hanhart A, Haefeli A, Valeri F, Rosemann T, et al. Referral determinants in Swiss primary care with a special focus on managed care. *PLoS One*. (2017) 12:e0186307. doi: 10.1371/journal.pone.0186307

28. Algaissi AA, Alharbi NK, Hassanain M, Hashem AM. Preparedness and response to COVID-19 in Saudi Arabia: building on MERS experience. *J Infect Public Health*. (2020) 13:834–8. doi: 10.1016/j.jiph.2020.04.016

29. Wunsch H, Angus DC, Harrison DA, Collange O, Fowler R, Hoste EA, et al. Variation in critical care services across North America and Western Europe. *Crit Care Med*. (2008) 36:2787–e8. doi: 10.1097/CCM.0b013e318186aec8

30. Alharbi AA, Alqumaizi KI, Bin Hussain I, AlHarbi NS, Alqahtani A, Alzawad W, et al. Hospital length of stay and related factors for COVID-19 inpatients among the four

southern regions under the proposed southern business unit of Saudi Arabia. *J Multidiscip Healthc*. (2022) 15:825–36. doi: 10.2147/JMDH.S362625

31. Steinwachs DM, Hughes RG. *Health services research: Scope and significance*. (2011). In: Hughes RG, editor. Patient Safety and Quality: An Evidence-Based Handbook for Nurses. Rockville (MD): Agency for Healthcare Research and Quality (US).

32. Corallo AN, Croxford R, Goodman DC, Bryan EL, Srivastava D, Stukel TA. A systematic review of medical practice variation in OECD countries. *Health Policy*. (2014) 114:5–14. doi: 10.1016/j.healthpol.2013.08.002

33. Kwan A, Daniels B, Saria V, Satyanarayana S, Subbaraman R, McDowell A, et al. Variations in the quality of tuberculosis care in urban India: a cross-sectional, standardized patient study in two cities. *PLoS Med*. (2018) 15:e1002653. doi: 10.1371/journal.pmed.1002653

34. Alharbi AA, Alqassim AY, Alharbi AA, Gosadi IM, Aqeeli AA, Muaddi MA, et al. Variations in length of stay of inpatients with COVID-19: a nationwide test of the new model of care under vision 2030 in Saudi Arabia. *Saudi J Biol Sci*. (2021) 28:6631–8. doi: 10.1016/j.sjbs.2021.07.040

35. Alharbi AA, Alqassim AY, Gosadi IM, Aqeeli AA, Muaddi MA, Makeen AM, et al. Regional differences in COVID-19 ICU admission rates in the Kingdom of Saudi Arabia: a simulation of the new model of care under vision 2030. *J Infect Public Health*. (2021) 14:717–23. doi: 10.1016/j.jiph.2021.04.012

36. Ministry of Health. Statistical Yearbook (2024) Available at: https://www.moh.gov.sa/en/Ministry/Statistics/book/Documents/Statistical-Yearbook-2022.pdf

37. Alzahrani MS, Alharthi YS, Aljamal JK, Alarfaj AA, Vennu V, Noweir MD. National and regional rates of chronic diseases and all-cause mortality in Saudi Arabia-analysis of the 2018 household health survey data. *Int J Environ Res Public Health*. (2023) 20:254. doi: 10.3390/ijerph20075254

38. Yahanda AT, Lafaro KJ, Spolverato G, Pawlik TM. A systematic review of the factors that patients use to choose their surgeon. *World J Surg*. (2016) 40:45–55. doi: 10.1007/s00268-015-3246-7

39. Thanh NX, Wanke M, McGeachy L. Wait time from primary to specialty care: a trend analysis from Edmonton, Canada. *Healthc Policy*. (2013) 8:35–44. doi: 10.12927/hcpol.2013.23375

40. MoH. *Statistical yearbook*. Saudi Arabia: Ministry of Health (2021).

41. Bellan M, Gavelli F, Hayden E, Patrucco F, Soddu D, Pedrinelli AR, et al. Pattern of emergency department referral during the COVID-19 outbreak in Italy. *Panminerva Med*. (2021) 63:478–81. doi: 10.23736/S0031-0808.20.04000-8

42. Sugand K, Aframian A, Park C, Sarraf KM. Impact of COVID-19 on acute trauma and orthopaedic referrals and surgery in the UK during the first wave of the pandemic: a multicentre observational study from the COVID emergency-related trauma and orthopaedics (COVERT) collaborative. *BMJ Open*. (2022) 12:e054919. doi: 10.1136/bmjopen-2021-054919

43. Hillestad R, Bigelow J, Bower A, Girosi F, Meili R, Scoville R, et al. Can electronic medical record systems transform health care? potential health benefits, savings, and costs. *Health Aff*. (2020) 24:1105. doi: 10.1377/hlthaff.24.5.1103

44. Olu O, Muneene D, Bataringaya JE, Nahimana M-R, Ba H, Turgeon Y, et al. How can digital health technologies contribute to sustainable attainment of universal health coverage in Africa? A perspective. *Front Public Health*. (2019) 7:341. doi: 10.3389/fpubh.2019.00341

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 16 July 2024
DOI 10.3389/fmed.2024.1411013

OPEN ACCESS

EDITED BY
Gokce Banu Laleci Erturkmen,
Software Research and Development Consulting, Türkiye

REVIEWED BY
Adamantios Koumpis,
University Hospital of Cologne, Germany
Remzi Celebi,
Maastricht University Institute of Data Science, Netherlands

*CORRESPONDENCE
Toomas Klementi
✉ toomas.klementi@taltech.ee

RECEIVED 02 April 2024
ACCEPTED 19 June 2024
PUBLISHED 16 July 2024

CITATION
Klementi T, Piho G and Ross P (2024) A reference architecture for personal health data spaces using decentralized content-addressable storage networks. *Front. Med.* 11:1411013. doi: 10.3389/fmed.2024.1411013

A reference architecture for personal health data spaces using decentralized content-addressable storage networks

Toomas Klementi[1]*, Gunnar Piho[1] and Peeter Ross[2,3]

[1]Department of Software Science, Tallinn University of Technology (TalTech), Tallinn, Estonia, [2]Department of Health Technologies, TalTech, Tallinn, Estonia, [3]Research Department, East Tallinn Central Hospital, Tallinn, Estonia

Introduction: This paper addresses the dilemmas of accessibility, comprehensiveness, and ownership related to health data. To resolve these dilemmas, we propose and justify a novel, globally scalable reference architecture for a Personal Health Data Space (PHDS). This architecture leverages decentralized content-addressable storage (DCAS) networks, ensuring that the data subject retains complete control and ownership of their personal health data. In today's globalized world, where people are increasingly mobile for work and leisure, healthcare is transitioning from episodic symptom-based treatment toward continuity of care. The main aims of this are patient engagement, illness prevention, and active and healthy longevity. This shift, along with the secondary use of health data for societal benefit, has intensified the challenges associated with health data accessibility, comprehensiveness, and ownership.

Method: The study is structured around four health data use case scenarios from the Estonian National Health Information System (EHIS): primary medical use, medical emergency use, secondary use, and personal use. We analyze these use cases from the perspectives of accessibility, comprehensiveness, and ownership. Additionally, we examine the security, privacy, and interoperability aspects of health data.

Results: The proposed architectural solution allows individuals to consolidate all their health data into a unified Personal Health Record (PHR). This data can come from various healthcare institutions, mobile applications, medical devices for home use, and personal health notes.

Discussions: The comprehensive PHR can then be shared with healthcare providers in a semantically interoperable manner, regardless of their location or the information systems they use. Furthermore, individuals maintain the autonomy to share, sell, or donate their anonymous or pseudonymous health data for secondary use with different systems worldwide. The proposed reference architecture aligns with the principles of the European Health Data Space (EHDS) initiative, enhancing health data management by providing a secure, cost-effective, and sustainable solution.

KEYWORDS

health data accessibility, comprehensiveness, and ownership dilemmas, primary and secondary use, a reference architecture for global health data space, decentralized content-addressable storage (DCAS) networks, semantic interoperability, European Health Data Space (EHDS)

1 Introduction

Health data encompasses information about an individual's or a population's health conditions, health outcomes, and quality of life (1). They include clinical, environmental, socioeconomic, and behavioral data relevant to health and wellness (2). Healthcare digitalization, when combined with accurate and high-quality health data, presents opportunities for delivering enhanced health and wellness-related services at reduced costs (3). However, health data introduces significant risks, as alone or combined with other data, it can reveal personal health status (4). The risk of revealing health status may reduce the willingness of individuals to participate in certain care processes, e.g., in mental health (5, 6) or drug abuse treatment. Health data leakage can also lead to discrimination against individuals by employers, insurers, or banks (7, 8).

The primary use of health data for diagnosis, treatment, and rehabilitation expects that pertinent information about a person's health is shared accurately and promptly with relevant parties, facilitating coordinated decision-making across all care settings (9). Beyond primary use, health data is utilized for secondary purposes (10) by various stakeholders, including policymakers, public health officials, researchers, physicians, the public, and industry (11). Routine clinical data is considered highly valuable (12) for advancing healthcare objectives and improving overall health outcomes.

Despite the value of routine clinical data collected during healthcare provision, significant portions of health data remain underutilized (13) due to the unstructured nature of the data and privacy and interoperability concerns. Moreover, the integration of medical data from various health data sources—Electronic Health Records (EHRs), medical devices for home use, innovative health and welfare applications, and health notes by patients—is beneficial in both primary and secondary use (14). However, the challenges related to data security, privacy, accessibility, comprehensiveness, and interoperability (15) result in the underutilization of data integration. We formulate these challenges as the following three dilemmas.

The dilemma of accessibility: The conflict between the desire for the accessibility of health data and the need to safeguard sensitive personal information.

This dilemma encapsulates the contradiction between ensuring data FAIR accessibility (16) and protecting sensitive personal information (17). A vast dataset with valuable routine health data is available worldwide, and broad and open access to this information is essential to maximize its benefits for society and citizens (18). However, given the delicate nature of personal data, there's an increasingly pressing need to fortify access controls. This presents a notable contradiction, as the pursuit of widespread health data FAIR accessibility clashes with the imperative to protect personal information (19).

The dilemma of comprehensiveness: The challenge to reconcile the need for the comprehensiveness of health data with their current fragmented nature (20).

Currently, a person's health data are preserved in different service providers' data repositories in provider-specific formats, preventing the gathering of a holistic representation of the individual's health record (21). Using the complete personal health records of a person, modern machine learning and AI methods can be used to gain a comprehensive picture of their health status (22). This would enable a transition from episodic, symptom-based treatment to continuous health monitoring and personal integrated care pathways, aiming to prevent diseases or diagnose them as early as possible. However, various factors prevent consolidating an individual's health data into a single, unified repository, including challenges related to semantic interoperability, diverse legal and ethical hurdles, and elevated risks of data leakage. As stated in research from 2018 (23), we still do not have a unified interoperability approach to cope with the semantic heterogeneity of health data. A review from 2019 concludes that no big-data analytics will happen without optimized data sharing and reuse, which we still lack despite different interoperability standards in the medical domain (24). Similar semantic interoperability-related challenges will be highlighted in the papers published in 2024 (25, 26).

The dilemma of ownership: The discrepancy between the data owner's rights to ownership and the practical inability to exercise those rights.

The presented statement highlights a dual dilemma. First, whether data and information can be considered property remains unresolved (27, 28). Second, the significant challenges associated with data ownership need to be addressed. While this paper refrains from definitively answering the first question, the authors generally favor an affirmative stance. Regardless of the stance on data ownership, prevailing legislation (29) ensures specific rights for the data subject concerning the information collected about them. Generally, in the EU, the processing of health data is prohibited unless there is a lawful basis under Article 6 of the GDPR and one of the exceptions mentioned in Article 9 is met (e.g., consent, contract, legal obligation, vital interests, public tasks, and legitimate interest). This legal framework ensures that individuals maintain control over their health data, emphasizing the importance of informed consent and transparency in processing such data (30). In reality, however, the practical exercise of these rights faces challenges, as data is preserved in third-party servers beyond the physical control of the data subject. In most countries, laws governing medical records place responsibility for storing health data on healthcare providers. These regulations are based on the healthcare provision legislation and do not need to be discussed in the context of this article.

Even the contemporary regional or national digital health platforms (DHPs) like the Estonian National Health Information System (EHIS) cannot resolve these dilemmas. First, as such systems are data processors according to the GDPR, they must process, protect, and secure data accordingly. Therefore, accessing data for secondary purposes is difficult due to complex content management and the need for de-identification (anonymization and pseudonymization) (31). Second, in such systems, the dilemma of data comprehensiveness has not been solved because of the international mobility of citizens. To solve this, the DHP must be pan-European or worldwide, or there is a need for an interoperability solution for the federation of national health systems. This is likely impossible and impractical as such systems are too complex to develop and operate. The third challenge

involves the data ownership dilemma. Within the intricate infrastructure of national or regional DHPs where data may be stored either in the cloud or on local servers, individuals do not know the whereabouts of their data. More critically, they might be unaware of who has access to their data and for what purposes it is being used. This situation further complicates individuals' ability to exercise their legal rights, leaving them powerless and disconnected from their health data.

In addition, the solution used in Estonia, which has 1.3 million citizens, may not be scalable in larger countries or, for instance, on a pan-European scale due to development and operation costs and data security and privacy challenges. One of the issues in such extensive DHP systems is health data concentration (32), which may be tempting for attackers because, in the event of a successful attack, it is possible to obtain the health data of many people. Between 2009 and 2022, there were 5,150 healthcare data breaches, resulting in the impermissible disclosure of 382,262,109 healthcare records in total (33). In 2021 alone, there were 686 HIPAA rule breaches affecting 500 or more health records, and the Accellion FTA Hack alone exposed the health information of at least 3.51 million individuals, making it the worst year for healthcare data breaches (34).

The more concentrated the data, the higher the costs for security; any breach could have severe consequences for individuals' privacy and well-being. Moreover, the dominance of a few entities in controlling health data raises questions about data ownership and control and the risks for data monopoly. Additionally, there are worries about the impact on healthcare innovation. A concentrated health data environment may hinder the development of diverse and competitive solutions, limiting the ability of small players to enter the market. Striking a balance between centralized and decentralized approaches, and prioritizing privacy and competition, is crucial in addressing the health data concentration issue. Policymakers, healthcare providers, and technology companies must collaborate on patient privacy, promote fair competition, and foster innovation in the health data ecosystem.

We propose and evaluate a reference architecture for a Personal Health Data Space based on DCAS networks (Figure 1). The focus of this paper is twofold. The first objective is to outline the typical use cases of health data for primary and secondary use based on existing health information systems (AS-IS) and to explain these systems' inability to resolve the three dilemmas. The second objective is to envisage an innovative DCAS network-based reference architecture for health data management (TO-BE), analyze its properties from the accessibility, comprehensiveness, and ownership dilemma perspectives, and evaluate security, data protection, scalability, and other aspects of the proposed solution under the typical primary and secondary use case scenarios.

The EHIS covers all Estonian residents and is one of the best digital health platforms (35). The Estonian model, operational since 2008 (36), provides valuable experiences that can be extrapolated for broader application. Our research utilizes four common health data use cases from the EHIS. Through this exploration, we shed light on issues and challenges associated with preserving health data within analogous unified national health data repositories. Our analysis underscores the need for cohesive solutions at the EU level, facilitating the seamless exchange of health data across institutional and national borders. Our discussion operates within the framework outlined by the GDPR (29) and the EHDS (37). This involves managing citizens' health data responsibly, ensuring data privacy, and enabling the reusability of health data for societal benefit. We posit that such a system establishes the groundwork for a fair data economy (38), wherein enterprises, especially small and medium-sized enterprises (SMEs), can engage in an innovative business landscape for intelligent health solutions. Simultaneously, citizens gain control over the utilization of their health data and actively participate in a just compensation mechanism, ensuring the equitable distribution of profits generated from innovative solutions based on their data.

The suggested reference architecture is in harmony with the fundamental principles of the European Health Data Space (EHDS) regulation proposal (Figure 2), significantly improving health data management by ensuring security, cost-efficiency, and sustainability. This architecture guarantees individuals' ownership and complete control over their health information while enabling semantic interoperability with existing hospital, regional, and national systems and respecting privacy and data protection laws. Through this solution, people have the opportunity to amalgamate their health information from diverse sources—various healthcare institutions, mobile applications, medical devices for home use, and personal health notes—into a single, integrated Personal Health Record [PHR; (39)]. This all-encompassing PHR can be shared with healthcare professionals, independent of the healthcare provider's location or the type of information system in use. Moreover, this solution empowers individuals to share their de-identified (anonymous or pseudonymous) health data for secondary use for the benefit of society according to explicit legal consent.

The rest of the paper is organized as follows: Section 2 delves into four health data use case scenarios based on the EHIS—primary medical use, medical emergency use, secondary use, and personal use. These EHIS scenarios are then examined through accessibility, comprehensiveness, and ownership to advocate the need for health data management based on DCAS network technology. Section 3 proposes the reference architecture to resolve health data accessibility, comprehensiveness, and ownership dilemmas through preserving semantically interoperable PHRs in DCAS networks. Section 4 evaluates and assesses the critical attributes of the proposed architecture. Section 5 compares the solutions with similar existing ones and examines their integration with existing health information systems and alignment with the EHDS initiative (37).

2 Methods

We adhere to the Design Science (DS) methodology (40), Figure 3, encompassing three steps: (1) investigating a problem, (2) designing a solution (treatment design), and (3) evaluating the solution's effectiveness in addressing the problem (treatment validation). While treatment implementation is not part of DS but is part of the engineering cycle, the figure shows treatment implementation to demonstrate the place and role of the prototype solution in our study.

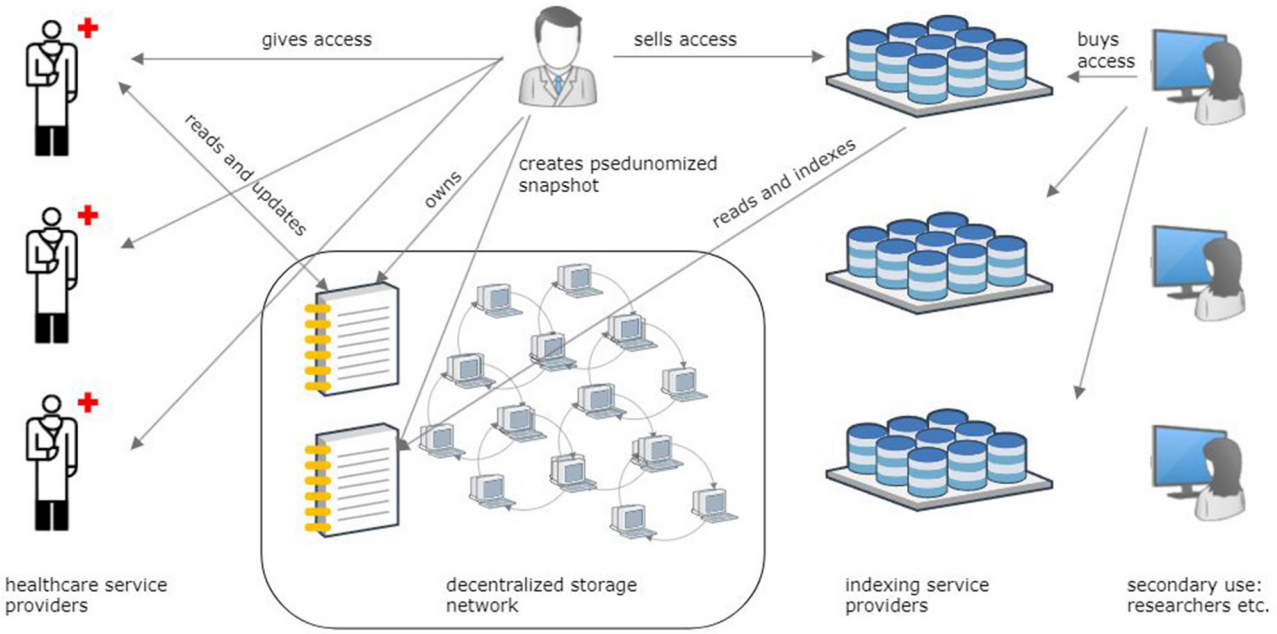

FIGURE 1
Overview of the reference architecture for storing personal health records in a decentralized content-addressable storage network and sharing health data for primary and secondary purposes.

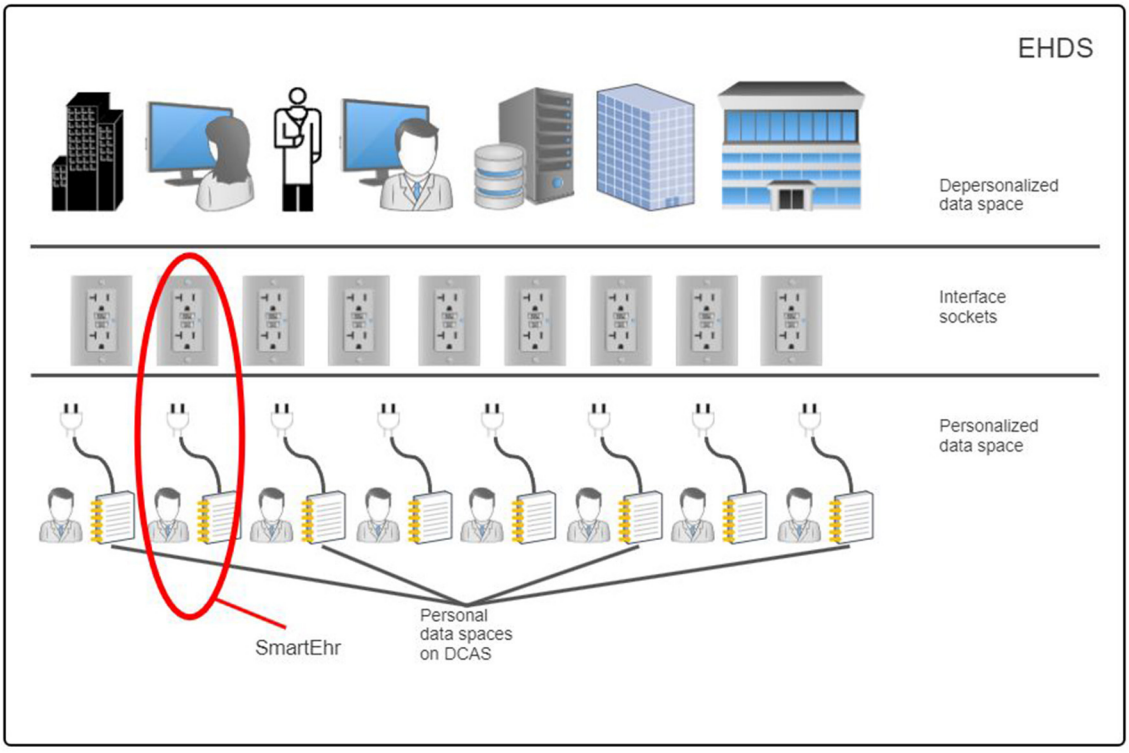

FIGURE 2
The personal data space in the decentralized content-addressable storage network is valuable for existing hospital, regional, and national health information systems for secure and sustainable retention of personal health data and to support semantic interoperability in data exchange.

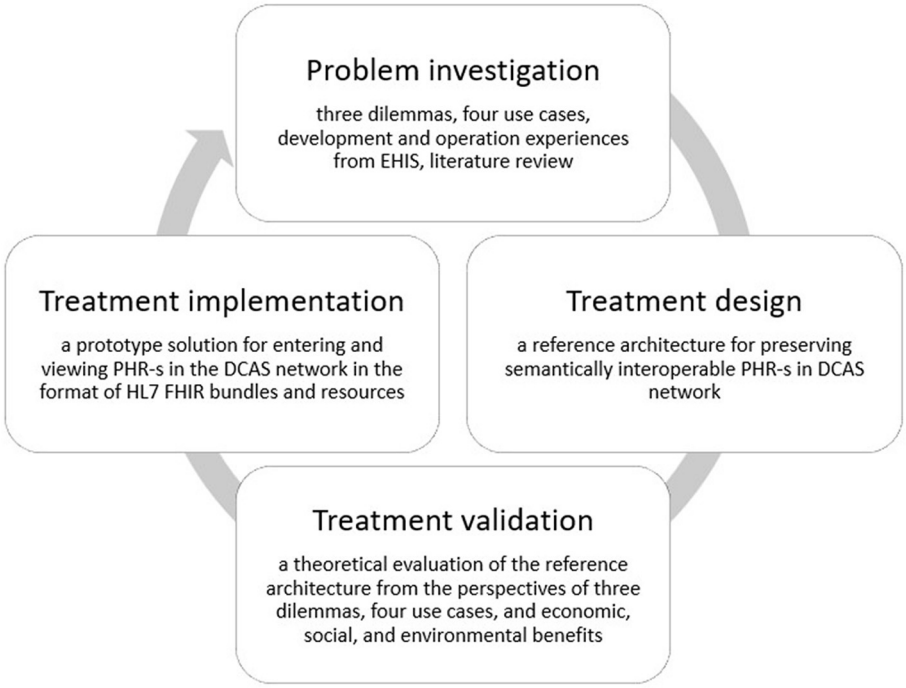

FIGURE 3
The design science methodology used in the development of the proposed reference architecture.

We articulate the problem through three dilemmas: data accessibility, data comprehensiveness, and data ownership (Section 1). Our analysis is based on a literature review and experiences in EHIS operation and handling. We first describe four use cases (this section, Section 2) based on EHIS operation and explain, based on these use cases, why even national systems like the EHIS fail to address the three dilemmas. As a solution (Section 3), we propose keeping the master copy of the PHR of each person's health record on the DCAS network under the complete control and ownership of the data subject. We will then show (Section 4) how the proposed solution will effectively address the three formulated dilemmas when utilizing the same four use case scenarios and explain how the proposed system supports seamless and coherent interoperability with the existing hospital, regional, and national information systems and data registers.

The Estonian Health Information System (EHIS, Figure 4) is a central national DHP through which health service providers, such as doctors, nurses, midwives, physiotherapists, and other healthcare professionals, can exchange data and see health data entered by other healthcare professionals about a patient. The EHIS consists of (1) central national databases, e.g., EHR, Prescription Centre, and Picture Archiving and Communication System (PACS); (2) digital health services built on the existing e-government infrastructure, e.g., digital prescription, e-referral, e-consultation, and e-ambulance; and (3) digital decision support systems and cross-sectoral services exploiting nationwide databases, e.g., drug-drug interaction database, clinical decision support system (DSS) for primary care, patient summary. The EHIS provides secure, robust, and reliable internet-based data exchange services for healthcare providers and natural persons. Healthcare service providers must, according to law, transfer specific, defined, structured, and standardized data to the EHIS. Data exchange between healthcare providers and the EHIS is ensured by implementing international standards, such as HL7 CDA and LOINC. Persons can access the EHIS through the Health Portal (41) (available in Estonian, English, and Russian).

In the case of the EHIS and the Health Portal, it is important to note their inseparable connection to other e-government services and tools in Estonia. The EHIS relies on a comprehensive information technology base infrastructure developed at the national level and is a central electronic database where residents' health history is recorded from birth to death. Technically, the health information system has been implemented on top of the state infrastructure solutions [ID card and mobile ID, (42), X-tee, (43), etc.] that most Estonians use extensively. The system is successfully connected to other information technology solutions offered to Estonian citizens, making it convenient for all users. According to the United Nation's E-Government survey, Estonia ranks very high in the E-Government Development Index (44), which might explain peoples' positive attitude toward the Health Portal.

2.1 Medical primary use case

A healthcare institution's internal and external information systems and databases are used in the daily work of doctors, nurses, and other healthcare professionals. Electronic Medical Records (EMR) and other Clinical Information Systems are the central in-house clinical information systems. For patient management, healthcare professionals primarily use the EMR. In Estonia, most clinical processes in healthcare institutions have been digitized.

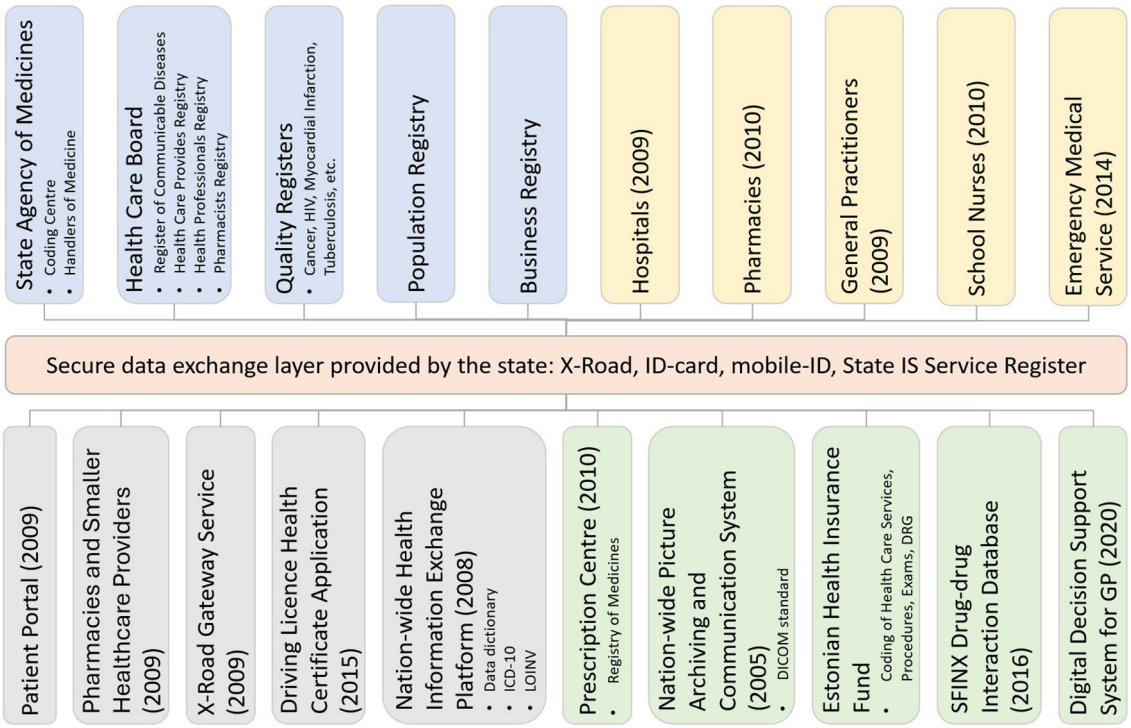

FIGURE 4
The Estonian National Health Information System architecture. The coloring schema is as follows: orange—central government infrastructure services; blue—national registers; yellow—integrated health service providers; gray—services that either use or provide services to EHIS; green—the central services of EHIS. The year shown in brackets indicates the year of deployment.

Still, paper-based documents have not disappeared entirely, e.g., intensive care spreadsheets, hospital internal orders, nurses' notes. The integrated EMR seamlessly communicates with external information systems if the person has been treated in another healthcare institution in Estonia, a healthcare worker wants to see previous data, or a doctor needs some central services, such as clinical DSS or e-consultation. If the person has been imaged or lab tests have been performed in other institutions, the EMR can query and retrieve relevant images from the nationwide PACS or receive lab test results from another EMR or EHR system. One very convenient service is a digital prescription: the doctor issues a prescription in the EMR, which uploads the digital prescription to the central prescription center after making several queries from national databases, e.g., to find out the reimbursement rate given to the specific patient. Since all digital documents used in healthcare in Estonia are linked to a person's unique personal code, the patient can go to any pharmacy and show their ID code. The pharmacist will immediately see all prescriptions issued for the patient and dispense the appropriate medicine to the patient. E-referral, e-consultation, and other digital health services follow similar principles. Documents completed in the healthcare institution, examination reports, or test results are converted by the EMR into a standard data exchange form and sent to the EHIS, where they are parsed and kept in different repositories. This allows clinical systems to compose either a time series based on data collected in the EHIS from various healthcare institutions, e.g., the dynamics of lab test results over time, or a standard Patient Summary (45). The benefits of a centrally developed, integrated, secure, internet-based, standard-following DHP such as the EHIS are related to data availability, sharing, and security. The medical professional gets a complete overview of the patient's contacts in the healthcare system and their content.

2.2 Medical emergency use case

The work of ambulance and emergency medicine departments has been digitalized in Estonia. Paramedics use tablet devices with specially designed e-ambulance software to enter data. E-ambulance and emergency medicine software are integrated with the EHIS (Figure 4). This way, the paramedic can see the patient's previous health data at the scene. The data available to paramedics is not limited to the text or diagnoses; previous medical imaging reports and electrocardiograms (ECG) can also be viewed. The ambulance can use the software to transmit critical information about the patient to the hospital before the patient arrives.

2.3 Secondary use case

Unfortunately, health data secondary use for public health, clinical research, medical claims management, or the pharmaceutical industry does not yet benefit significantly from the

EHIS. In the EHIS, secure data exchange between various clinical parties is resolved well, but ensuring data quality still has issues and challenges. Although various international classifications and terminologies are in use, their use is insufficient, and medical records still contain a lot of free text. This forces the National Institute for Health Development (NIHD), responsible for public health in Estonia, to collect data separately through the information systems they developed. This causes data duplication and discrepancies.

Firstly, the NIHD collects most of its data through its internet portal, a legally mandated data entry system for healthcare providers to report to the NIHD. This portal, in combination with other government data collection systems, e.g., the EHIS, can be seen as a redundant system and duplicate data entry. The data NIHD collects is often available in other systems, but due to the gaps in data quality and interoperability, it cannot be automatically transferred to the NIHD databases. Secondly, data entered directly into NIHD systems and cleansed for better quality is not shared back in an interoperable way to clinical/administrative healthcare systems. This limits the value of the NIHD's data and analytics, as it cannot contribute to the general quality enhancement of clinical and administrative decision-making processes in hospitals.

The same trend of data being collected in separate information systems can be observed in the case of randomized clinical trials conducted by pharmaceutical companies. However, new registries, such as the Breast Cancer Screening Registry, have been started, which query data directly from the EHIS. Still, systemic weaknesses in cross-sectoral and cross-institutional regulation, coordination, and clinical data standardization limit the secondary use of health data. This creates a need for manual data processing and culminates in inefficient information handling and systems development.

Hospitals often use several software applications for administrative data when automated integration with medical systems is not in place. Frequently, manual data entry is needed for reporting and statistics. In most hospitals, the raw data is electronic but manually transferred for reporting and statistics. Additionally, regulations on the health information system, prescription system, reimbursement system, public health reporting system, or vertical registries (cancer, HIV, tuberculosis, myocardial infarction, etc.) are not always harmonized, or the clinical information classes are defined too generally to be usable practically. Therefore, each responsible agency, specialty, or sector develops its terminologies and data structures independently. This leads to point-to-point solutions, lessens system interoperability, and ultimately increases manual data processing and complicates software development.

2.4 Personal primary use case

In the Health Portal (Figure 5) of the EHIS, a person can see their health and medical data and may perform several activities. This data has been collected according to how the person's treating physician or healthcare institution sent them to the EHIS in a standardized way. A person can submit declarations of intent, appoint a representative, perform actions on their behalf and on behalf of the person represented, and view the medical invoices submitted by healthcare institutions to the Estonian Health Insurance Fund about their medical treatment. All prescriptions in Estonia are in digital form, and a person can see the issued prescriptions and their status in the portal.

All residents can access their data to determine their consent for specific health data sections. This means the patient can restrict access to certain documents, medical records, and all personal data in health information systems. Access restrictions can be imposed on one individual document or all information contained in the EHIS. From the point of view of data security and privacy, it is essential to note that a person can monitor all activity logs in the Health Portal, i.e., see which medical professional has requested their data and when and what document was viewed.

2.5 EHIS from the perspectives of the three dilemmas

The Estonian National Health Information System (EHIS) is a pioneer in digitizing healthcare on the national level. However, the system faces significant challenges related to the dilemmas of accessibility, comprehensiveness, and ownership.

1. *Accessibility*: The EHIS fails to resolve the accessibility dilemma as it lacks features for secondary data usage, as previously mentioned. Consequently, the initial aspect of the dilemma, necessitating data access, remains unresolved. Moreover, the EHIS falls short in ensuring comprehensive protection of personal data, as its measures aimed at limiting access are reactive rather than preventive. While data owners can detect unauthorized access, they cannot preemptively exclude it.
2. *Comprehensiveness*: The EHIS fails to resolve the dilemma of comprehensiveness primarily because, at the global level, it operates as an isolated data silo. Moreover, even at the local scale, the EHIS does not provide a holistic perspective of an individual's health profile. Research suggests that patient data stored within healthcare facilities tends to be more accurate and thorough than EHIS data (46). Additionally, the exclusion of patient-generated data, such as lifestyle information and data collected from wearable devices, further restricts the system's capacity to offer the complete picture. Consequently, the EHIS merely presents a simplified and partial representation of the data, contradicting its initial aspirations for comprehensiveness.
3. *Ownership*: The EHIS fails to resolve the ownership dilemma, as the institution managing the data retains physical control. While the data subject possesses certain rights, such as the ability to restrict access to specific data and monitor the audit trail of data usage, the managing institution remains the de facto owner of the data. This scenario resembles feudal land ownership relations, where the land belongs to the landlord, and the peasant has limited rights to utilize part of it for personal use.

To surmount these challenges, a different approach is needed—one that embraces decentralized technology to enhance system agility, incorporates patient-generated and -entered health data to ensure data comprehensiveness, and empowers patients with preemptive and complete control over their health information. Such a system would facilitate seamless cross-border health data exchange, support the integration of innovative health

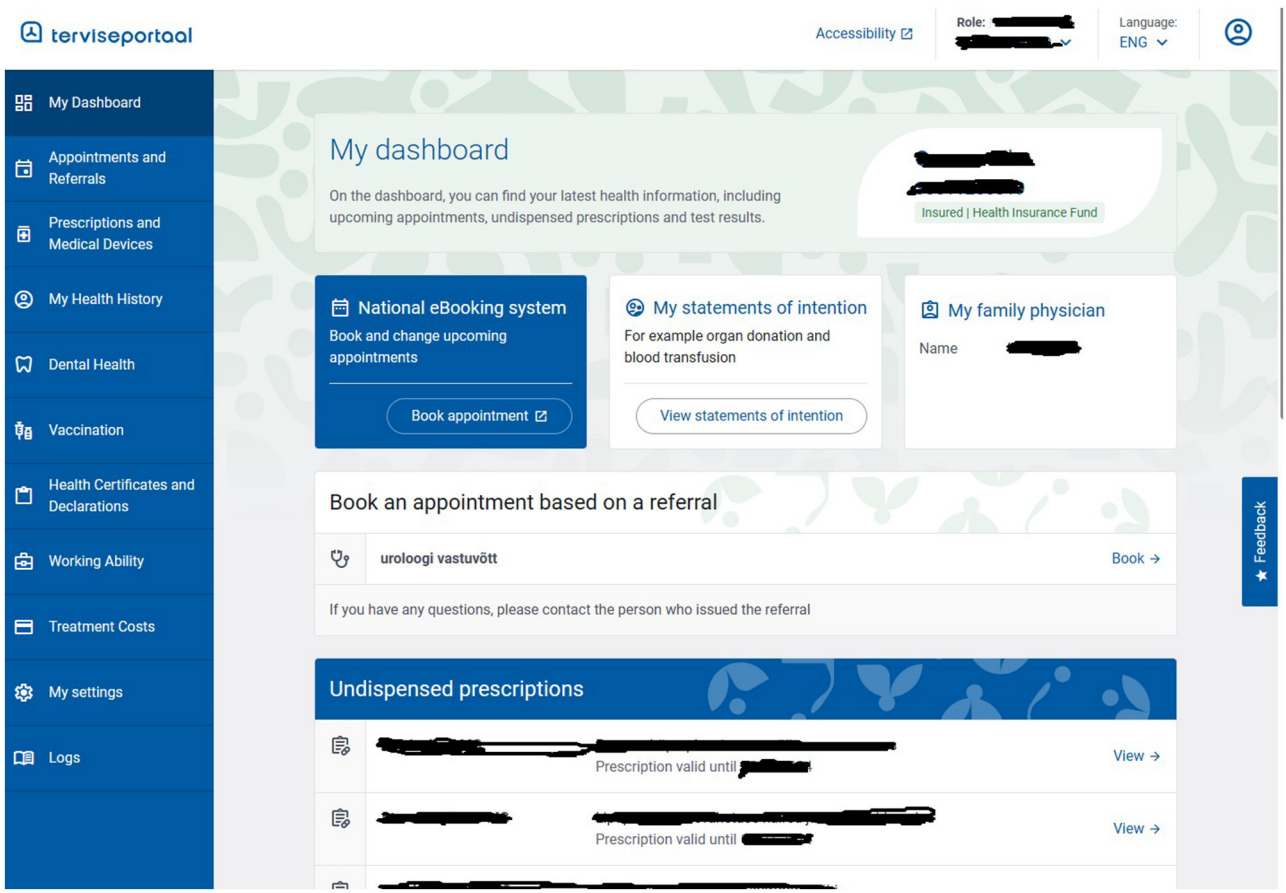

FIGURE 5
A screenshot from the Health Portal of the Estonian National Health Information System.

technologies, and streamline consent management for secondary data use.

3 A reference architecture for personal health records

3.1 An overview of the architecture and fundamentals

The proposed architectural solution to solve the three dilemmas is based on the novel decentralized content-addressable storage (DCAS) network technology (Figure 1). We first analyze data management risks to grasp the principles by which DCAS networks operate.

By aggregating all health data in one place and keeping it in a hospital, regional, or national health information system, the risk of data management increases due to a single point of failure, attractiveness to attackers, the complexity of security management, difficulties in access management, and the complexity of regulatory requirements. The opposite also applies—splitting a large dataset into smaller components reduces the risk of managing each component and the whole. Continuing this iterative data volume-reducing process leads to a scenario where the risk linked to an individual tiny data fragment approaches zero, and the implementation of intricate and costly security measures becomes superfluous.

DCAS networks operate on a similar principle. They are peer-to-peer networks wherein nodes run open-source software designed to store an enormous quantity of tiny data fragments. When some data, such as a file or a document, is to be stored in such a network, the data is first split into data fragments of a few kilobytes each. These fragments are then distributed across various nodes according to the network protocol. Each fragment represents an insignificant fraction of the complete dataset, making it feasible to distribute them between nodes without jeopardizing the privacy of the entire dataset. As the anonymity of DCAS network nodes is part of the DCAS protocols, the trustworthiness of the node operators is not imperative for secure data storage within the network, as individual data fragments are not informative. In addition, no node knows to which dataset the fragment belongs, the location of nodes, or the nodes where the remaining data fractions are stored.

Conceptually, a DCAS network resembles a paper shredder (Figure 6), cutting a classified document into tiny strips, none of which divulge the document's contents. Unlike a physical shredder, a software-based implementation can reconstruct the original document from its shredded components. This reversal process

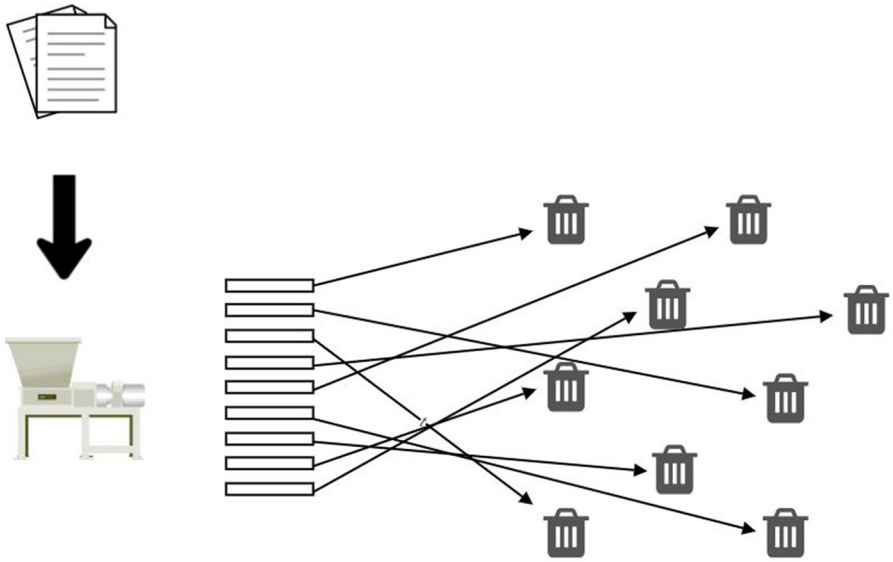

FIGURE 6
The decentralized content-addressable storage as an electronic paper shredder.

merely necessitates knowledge of the root hash of the original document, which a data owner must only keep to themselves. Here and in the future, a data owner means a person who keeps their data on a DCAS network and, if necessary, shares that data for primary or secondary use.

In the following, we provide concise overviews of the fundamental characteristics of a DCAS network. While Ethereum Swarm (47) has inspired these descriptions, they are formulated broadly enough to apply to any implementation of a DCAS network.

Content addressing. In contemporary internet architecture, location-based addressing is widely employed. The typical structure of a web URL consists of several components: the server name, which is substituted by the IP address during the name resolution process, the name of the sought-after resource, and the path to the directory where the desired resource is situated. This method of addressing is called location-based addressing, as the resource's address signifies its physical location.

In contrast, content-based addressing is not based on the location of a resource but highlights its content (48). Content-based addressing, in many respects, is more intuitive than location-based addressing. When searching for a specific resource, its content is of primary importance rather than its physical location. This can be observed in everyday activities like shopping. In a store, individuals are not concerned with the product's precise shelf but are interested in milk or bread, irrespective of their spatial arrangement.

DCAS networks use content-based addressing. Each network node has an overlay address, a randomly generated integer. To avoid duplication of overlay addresses, large, 256-bit integers are used. The Kademlia distance (49) between two network nodes is an integer obtained by the exclusive logical addition (XOR) of overlay addresses of nodes. For instance, the Kademlia distance between overlay addresses 5 (0101) and 6 (0110) is 3 (0011). The Kademlia distance has all the fundamental characteristics of distance, including non-negativity, symmetry, the zero value of a node's distance from itself, and triangle inequality. In the DCAS network, each shard of information is stored on the node whose Kademlia distance is closest to the shard's hash value. The hashes of the shards are arranged into a Merkle tree (50), which is stored in the DCAS network following the same information-splitting protocol. The hash value of the Merkle tree's root serves as the address of stored data.

When retrieving data from the DCAS network, the process is reversed. Specifically, the network protocol implemented in the node software locates the node closest to the given hash value and finds its underlying address (IP address in the context of the internet). Subsequently, a request is sent to this identified node to access the desired data. Content addressability serves as a supplementary measure to ensure data integrity. This is achieved by enabling the consumer to verify the content of the downloaded data by calculating its hash value and comparing it to its address, thus confirming that the data has not been altered.

Decentralization. Firstly, let's delve into some terminological considerations. The term "decentralized" is frequently employed to convey the idea of a system comprising numerous smaller, independent entities. To illustrate, a "decentralized data network" is commonly understood as a federation of diverse data sources, each independently comprehensive within localized boundaries (51). This implies that information about a specific subject is internally cohesive within these local confines. While these data sources may lack global completeness by not encompassing all available information about a particular topic, they wield control over the data within their purview.

However, this paper adopts a more stringent definition for "decentralized", signifying a system where data lacks completeness even locally, information stored on individual nodes is indecipherable, and no governing body exists locally or globally. The absence of a governing body within the DCAS

network means that independent node operators individually determine all decisions, including joining or leaving the network. At the same time, the network protocol incentivizes each node to make decisions that contribute to the network's objectives.

Redundancy. Network decentralization refers to the absence of a governing authority body within the network (52). Consequently, network nodes can disconnect from the network at any given moment. Upon leaving, these nodes take with them the data shards they have been storing. This presents a significant challenge, as restoring the data that these shards were part of is impossible. Naturally, such a situation is unacceptable, necessitating the implementation of redundancy to prevent data loss.

One potential approach to address redundancy involves storing each piece of data not only on the node closest to it based on Kademlia distance, but also on a set of nodes that belong to a specific neighborhood of responsibility surrounding the closest node (53). Since overlay addresses are randomly assigned to the network nodes, and the Kademlia distance has nothing to do with geographical dimensions, network nodes belonging to this neighborhood are typically dispersed worldwide under the management of different operators. Based on network size and its rate of churn, a sufficiently large radius of the neighborhood can be chosen, ensuring that the loss of a single piece of data resulting from the departure of the node storing it is close to zero (54).

The outlined redundancy method represents just one approach to guarantee data redundancy. Alternatively, more efficient techniques like Erasure Coding (55) may be used. Despite distinct algorithms, the objective remains to ensure data preservation within the network when nodes exit the network.

Immutability and de-duplication. Content addressability leads directly to the immutability of the data (56). This is due to using hash values as addresses, where any change in the content of the data results in a change in its address. Consequently, the altered data becomes a new addressable entity for the network, while the previous version remains accessible at the earlier address. Therefore, the DCAS networks inherently retain the version history of any data modifications.

As described, the data is typically fragmented into tiny pieces stored independently as individual entities within the DCAS network. Likely, only a particular portion of these pieces will be modified when changes occur in the data. Those pieces that remain unaltered continue to exist online at their former addresses. Thus, DCAS networks efficiently maintain the version history of the dataset, ensuring that only one copy of the data exists within the network, excluding the copies required for redundancy.

Mutable address space. Content addressability has numerous advantages (57). As previously mentioned, content addressability results in data immutability, as any modification to the data corresponds to a change in its address. However, there are cases where it becomes essential to store mutable data at a designated address. To accommodate this need, each user in a DCAS network is allocated a personal mutable address space. This dedicated space allows users to manage and modify data within specific addresses without conflicting with the immutability constraints associated with content addressability.

Incentives. Decentralized networks' successful emergence and sustainability rely on establishing a precise and robust incentivization mechanism (58). This mechanism must adequately motivate network operators to bear the costs associated with providing services and is typically facilitated through compensation from the users of the network services. However, the absence of a central governing authority poses a challenge in orchestrating this compensation process.

Adopting a compensation mechanism built on blockchain and smart contracts is imperative to achieve incentives in complete network decentralization (59). Within such systems, it is feasible to use crypto tokens for payment. Ethereum Swarm, which operates on the BZZ crypto tokens (60), is an example of a decentralized compensation mechanism implementation. Alternative compensation mechanisms have also been implemented. However, any method reliant on traditional fiat currency necessitates the involvement of an intermediary body, compromising the network's decentralization.

3.2 Core application

The core application (Figure 7) is open-source software that operates on the user's device, serving as a personal portal to health data. This application primarily aims to present a person's health data stored online in a DCAS network in a user-friendly manner. Additionally, it enables persons to perform various tasks such as annotating, searching, filtering, and sorting health information. Furthermore, it establishes data communication with the DCAS network using an abstraction layer, which ensures independence from the implementation of a specific DCAS network. Moreover, the core application employs software layers to incorporate protocols and standards commonly used in healthcare to facilitate interoperability. The core application's functionality can be expanded by integrating separate downloadable modules.

The following subsections present a detailed description for each component of the core application.

3.2.1 Root hash management

The root hash granting access to the data should be known to the data owner exclusively. This hash value plays a crucial role in granting access to the data; therefore, the data owner must thoroughly protect it. In the unfortunate event of losing the hash value, retrieving access to the data becomes impossible. Consequently, the method employed for storing the hash value must incorporate safeguards to prevent both unauthorized access and accidental loss; therefore, safeguarding and securing this hash value is a primary responsibility of the core application.

Whenever a modification is made to the data, the hash value is updated to reflect the changes. The new hash value permits access to the modified data, while the previous hash value represents the prior version of the data. Preserving the entire version history of the data within the core application may not be feasible due to practical limitations. A possible approach is to include the address to the previous data version within the data itself. This enables the core application to retain the whole history of data amendments.

In addition, it is essential to consider the possibility of the stored hash value being unavailable due to, e.g., the data owner's device

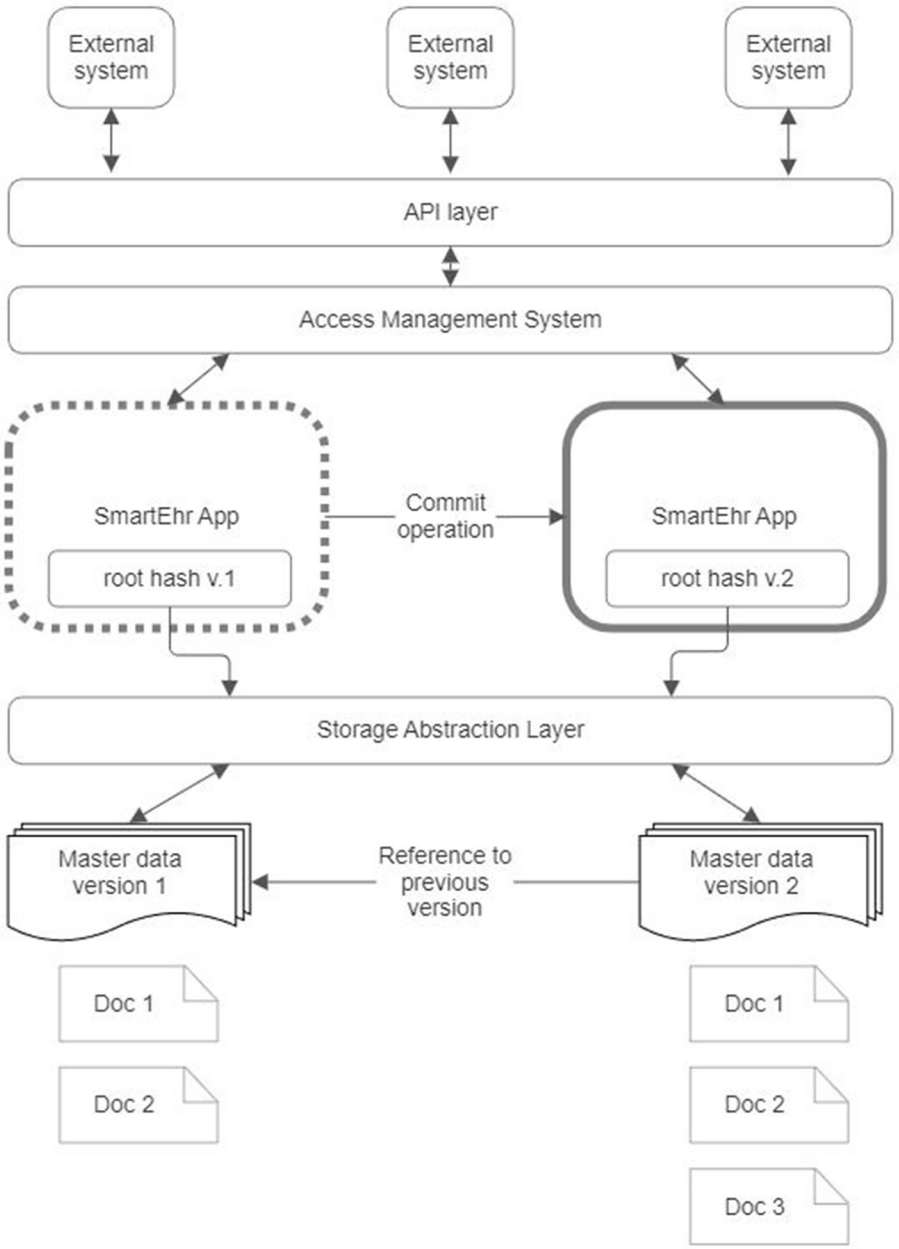

FIGURE 7
Core application architecture of the proposed reference architecture for personal health record.

being lost. In this case, storage of a constantly changing hash value in a recoverable manner poses a significant challenge. A plausible alternative involves storing the changing root hash within the DCAS network. This is where the personal mutable address space of the DCAS network proves valuable. By storing the encrypted hash value within the user's private mutable address space, the core application simplifies its task to retaining the constant address where the current root hash resides.

This constant value facilitates the implementation of secret sharing algorithms, like Shamir's Secret Sharing (61), to effectively mitigate the risk of data loss. This secret-sharing framework mathematically divides the constant address where the current root hash resides into multiple shares (Figure 8). Each share is then stored separately in the main applications of the data owner's closest relatives so that only one share is retained by one relative. This secret-sharing mechanism ensures that the address can be recovered by gathering a sufficient number of shares that meet or exceed the predetermined threshold. Conversely, it is impossible to reveal the secret address if the number of shares is below that threshold.

The solution above also provides a means to safeguard data against unauthorized modification. In this approach, recording the hash value of the modified data is exclusively permitted in the owner's mutable address space. Consequently, any alteration to the

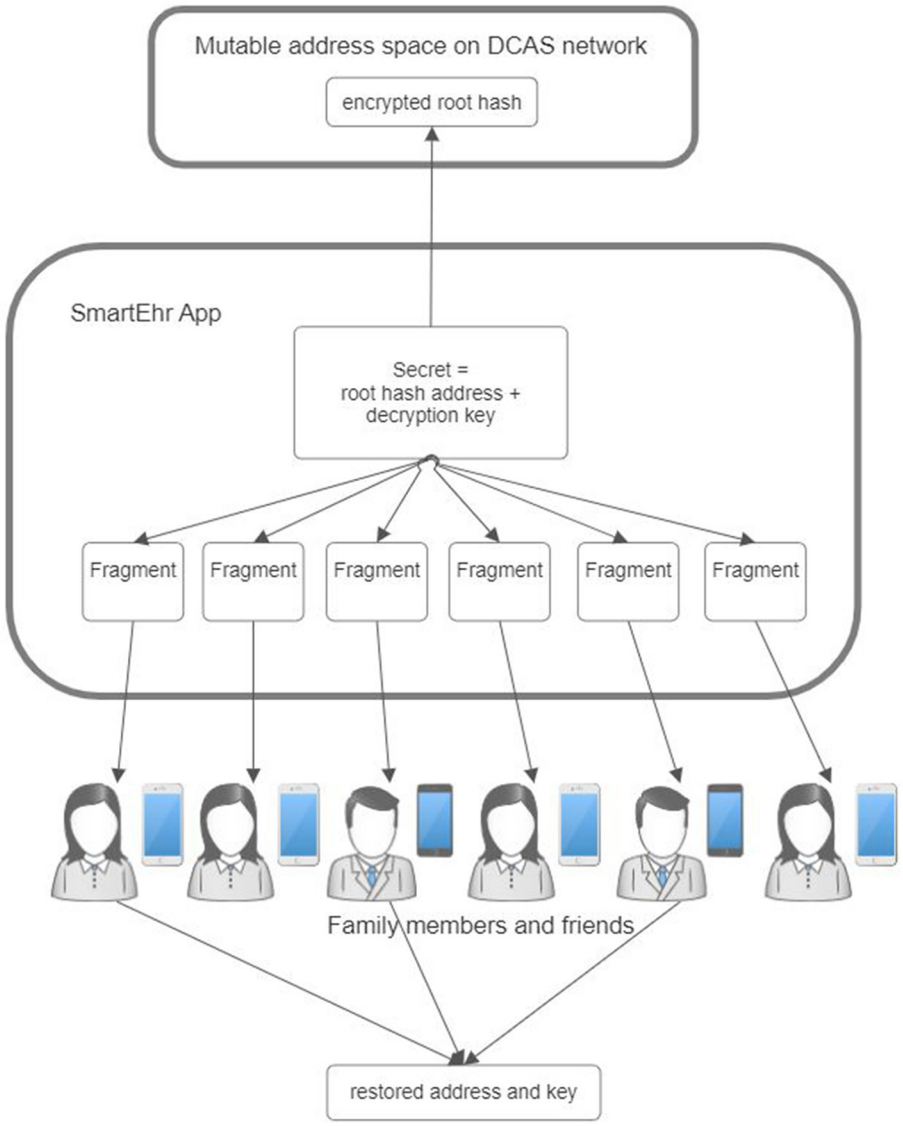

FIGURE 8
A root hash management in the proposed reference architecture for personal health record.

data necessitates the owner's approval by storing the revised hash of modified data. This confirmation process can be likened to the commit operation commonly employed in databases. Without such confirmation, any changes to the data are lost.

3.2.2 Storage abstraction layer

The Storage Abstraction Layer (SAL) is an intermediary component, facilitating communication between the core application and the DCAS network. This intermediary layer ensures the core application's independence from the specific implementation details of the storage network. It aligns with the principle of dependency inversion commonly employed in software development. Incorporating an intermediate layer such as SAL, the core application can remain unaffected if replacement of the layer becomes necessary. The core application solely requires functionality related to the reading and writing of data, while SAL effectively manages all other intricacies.

3.2.3 Content handlers

Numerous standards exist to represent health data, including various HL7 standards and versions, OpenEHR (62), ISO 13606 (63), and ContSys (64). It is desirable for the core application to not be restricted solely to clinical data but to offer the capability of managing diverse information about an individual's health and general lifestyle. As this data can be generated by various devices from different manufacturers, they might exhibit disparate formats and employ distinct data models. Content handlers in the core application are to handle this multitude of data models effectively. These autonomous software modules adhere to the dependency inversion principle, akin to the Storage Abstraction Layer.

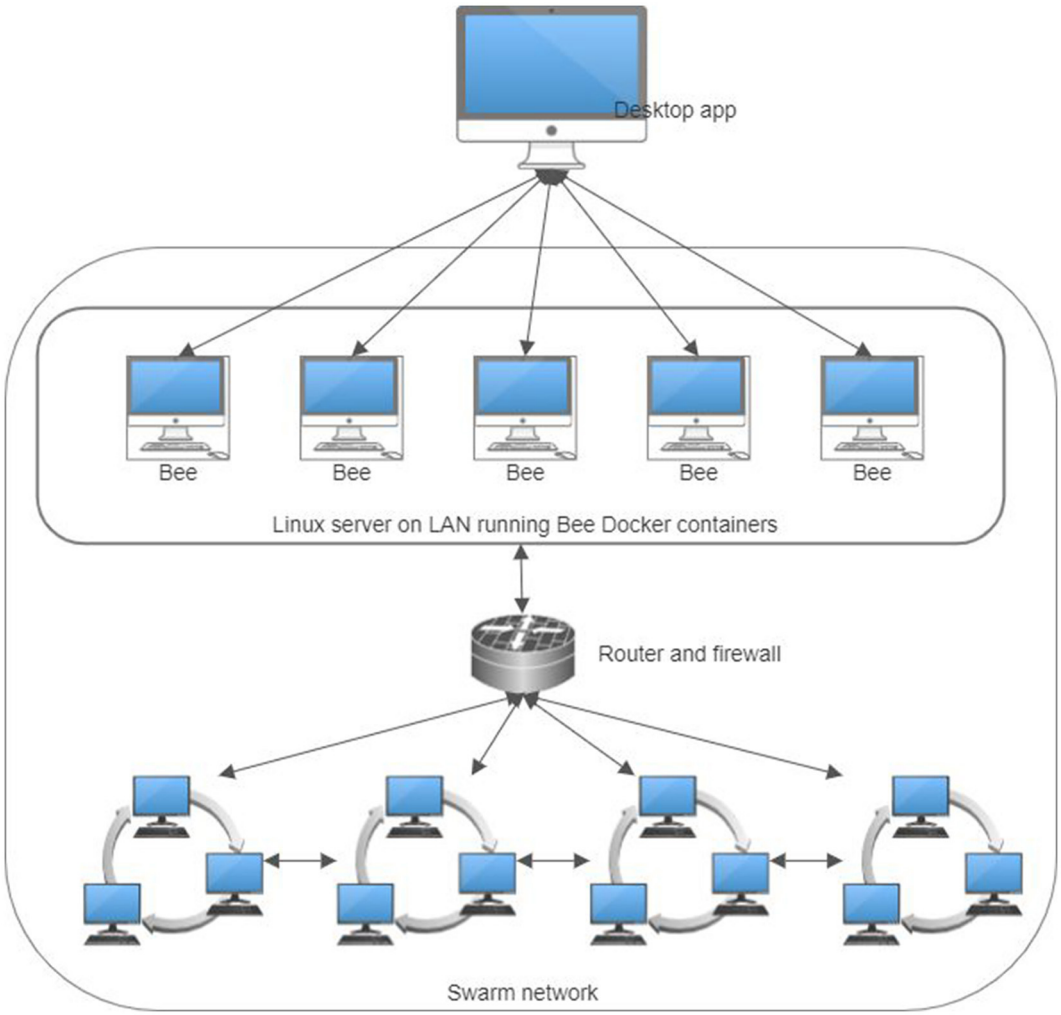

FIGURE 9
The infrastructure of the practical experiments for storing personal health records in a decentralized content-addressable storage network.

Incorporating these content handlers into the core application does not necessitate any modifications to the core application itself. The data should be presented online in a self-descriptive manner, enabling the bootloader to select the appropriate content handler for processing.

3.2.4 Interoperability layers

The purpose of the interoperability layers is to facilitate the integration of the core application with external information systems. A key objective of these layers is to enable healthcare providers to access patient data in the context of primary and secondary use. As previously mentioned, one data-sharing approach involves disclosing the data address (its root hash). Nevertheless, a preferable alternative is to grant data access via an application programming interface (API), such as FHIR, which preserves the confidentiality of the root hash. It is reasonable to use federated [on-the-fly adaptation according to the third-party data exchange protocol; (65–67)] rather than integrated (based on a standard data format) or unified (based on a common standard) interoperability (68) to achieve flexible and adaptable interoperability across hospital, regional, and national health information systems.

3.2.5 Extension modules

Extension modules are plugins that serve the purpose of augmenting the existing capabilities of the core application. These supplementary features encompass various enhancements, such as integrating diverse wearable devices into the core application and incorporating various algorithms enabling individuals to supervise and assess their health-related behaviors. It is vital to note that these extension modules obtain access to individuals' PHR through the core application, while concurrently enabling other system components to harness the data they generate.

4 Evaluation of the proposed architecture

4.1 Practical experiments

Practical experiments were conducted to validate the viability of the proposed reference architecture. Due to the sensitivity surrounding medical data and the constraints imposed by legal regulations, obtaining medical data for testing poses significant challenges. Instead, we used the Synthea package (69) to generate synthetic health data. Synthea is an open-source data generator renowned for producing realistic medical history data for synthetic patients, encompassing various healthcare scenarios. It allows for the creation of datasets of any desired magnitude. For our experiment, a dataset comprising 1,000 synthetic persons was generated.

As Synthea generates data in the format of FHIR bundle resources, we selected this data format for our experiment. However, it's important to note that our choice of FHIR format does not necessarily imply its superiority in DCAS networks. Ultimately, Resource Description Framework (RDF) and personal knowledge graphs offer more flexible solutions. Since FHIR is also concerned with developing RDF (70) and other concentrated and thin data formats [e.g., FHIR Shorthand (71)], we are likely not far from the desired and practical results to support federated semantic interoperability with a third-party hospital, regional and national healthcare systems, and innovative welfare applications.

We opted for Ethereum Swarm (47) as our DCAS network for several compelling reasons:

1. Full decentralization: Ethereum Swarm operates without a central authority, ensuring a decentralized ecosystem.
2. Robust incentivization: The network boasts a robust mechanism encouraging participation and contribution.
3. Ideal for small data storage: Ethereum Swarm is well-suited for efficiently storing small data fragments, such as FHIR resources.
4. Open-source nature: Ethereum Swarm is open-source and fosters transparency and collaborative development.

The Swarm network comprises independent nodes running the Bee software (72), compatible with both Linux and Windows systems. For our setup, we have chosen Ubuntu Linux as our operating environment. Despite its modest resource requirements, Bee performs optimally with an SSD hard drive and a fast network connection, handling network traffic efficiently.

The software development environment for this project was Microsoft Visual Studio 2022. The FHIR bundles generated were dissected into individual resources and stored in an SQL Server database to facilitate ease of manipulation. Subsequently, each resource was uploaded to the Swarm network as a distinct entity, uniquely addressed with a hash key. A patient's resource index was stored separately as an FHIR bundle resource, incorporating multiple FHIR Reference resources. The .NET task-based asynchronous pattern (TAP) enhanced query efficiency. A dedicated program in C# was designed to upload the generated FHIR resources. This involved strategically alternating queries between five Bee Docker container nodes and executing 40 simultaneous POST requests in parallel for each, optimizing the uploading process (Figure 9). Parallel queries were similarly employed for data downloads. Due to Swarm's massively parallel protocol, which sends simultaneous requests to numerous network nodes for data chunks, the overall user experience was comparable to, if not better than, traditional web browsing. A screenshot of the experimental app showing a list of the generated FHIR resources stored on the Ethereum Swarm live network is shown in Figure 10.

4.2 Medical primary use case

Relying on utilizing DCAS networks to preserve Personal Health Records, the proposed reference architecture (Figure 1) integrates with existing hospital, regional, and national health information systems seamlessly and in a semantically interoperable manner (Figure 2). This architecture features a person-owned application (Figure 7) that operates on the person's device. This application is responsible for securely storing the root hash of the person's health data and facilitating the reading and writing of data within the DCAS network.

In the primary use scenario, a person (data owner) can share data with a healthcare provider by disclosing the root hash of their data (Figure 1). Once the healthcare service provider completes the necessary edits and saves the additions to PHR, a new data version and the corresponding new hash value are generated. The service provider relays the updated value to the data owner, who securely stores it via their application. The healthcare service provider should not retain the original or the revised root hash.

Alternatively, data sharing can occur without disclosing the root hash. One possible method is utilizing a standardized API, such as HL7 FHIR (73), integrated within the data owner's application. However, in such cases, additional measures must be developed to uphold the integrity and reliability of the shared data (74–76).

In medical data, the integrity of information holds paramount importance. A key strategy to ensure data reliability involves the digital signing of entries by the respective contributors. In this context, the data's trustworthiness hinges on the trustworthiness of data entry. Beyond signing the added or modified part of the data, an additional layer of security can be established if the healthcare provider signs the data they enter and the root hash of the entire dataset as it was presented to the healthcare provider during the medical treatment or service provided.

The data subject can conceal specific portions of their data by restricting access for particular healthcare providers. This concealment involves generating a new version of the data, accompanied by a corresponding alteration in the root hash, as elucidated earlier. Significantly, the de-duplication feature outlined earlier clarifies that creating a partially concealed data set does not involve duplicating the entire dataset. Instead, it only stores the modified data fragments in the DCAS network.

When a service provider adds an entry and signs it, they essentially endorse the data they contributed and the entire dataset as it was presented to them. This ensures a comprehensive and signed record of the data collection, offering a transparent snapshot of the information available to the service provider at the time of data entry.

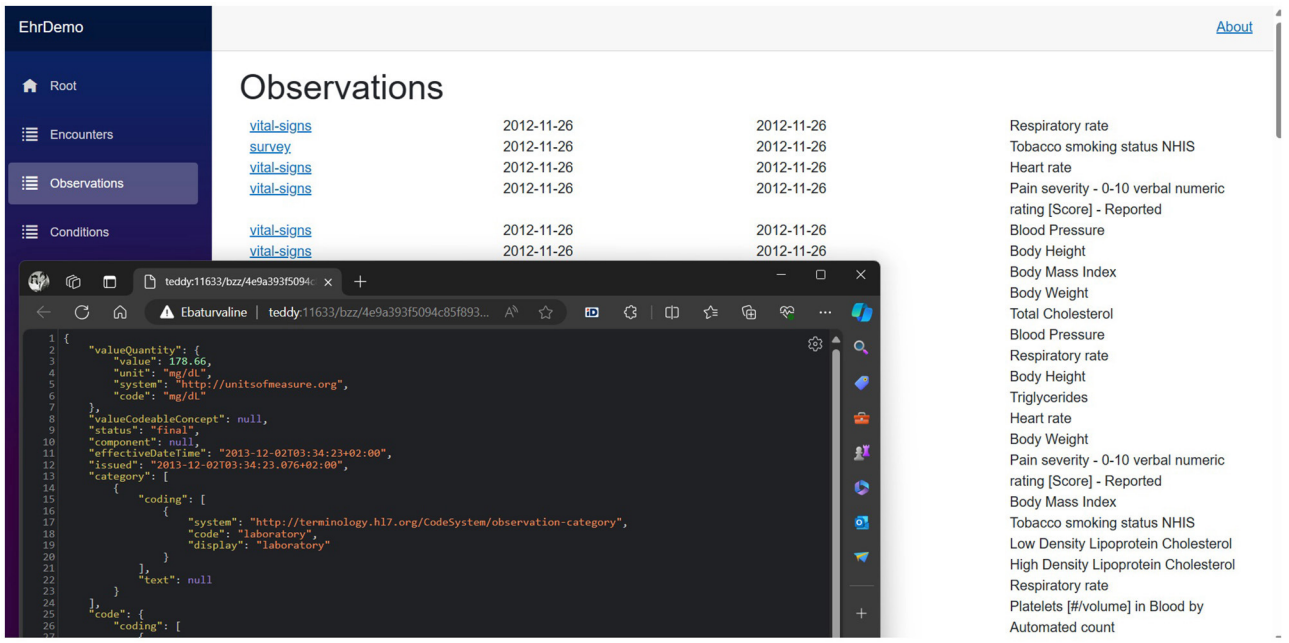

FIGURE 10
A screenshot of the proof-of-concept app shows a list of FHIR observation resources, with one open in the browser.

4.3 Medical emergency use case

The proposed architecture offers a simple solution for emergency access to an individual's health data. For this, a distinct data subset must be created encompassing vitally important information, such as data about chronic conditions and ailments, medications, allergies, and other related details. These particular data entities form a specialized subset within the comprehensive health data and are endowed with a unique address within a DCAS network, enabling global accessibility. Individuals should consistently carry the reference to this subset, either in digital format stored on a microchip or physically embodied as a QR code on a wearable tag or implemented through alternative means. In a medical emergency, medical personnel can retrieve the most critical health data of the individual by scanning the aforementioned QR code or reading it from the microchip. This method allows access only to the depersonalized subset of health data encompassing vital information during emergencies, while protecting the identity and other PHR data.

4.4 Secondary use case

For secondary use (Figure 1), the Personal Health Record must be de-identified (31) to make anonymized or pseudonymized data versions. This process involves the removal of any information that could lead to the identification of the subject, while preserving the reliability of the data. To achieve this, a third party trusted by all stakeholders plays a crucial role. Whether a national institution or a purpose-built organization, this entity verifies the data subject's identity. Subsequently, it validates and removes all signatures associated with the data and appends its own signature to the dataset. This signature proves the reliability of the de-identified data, now derived from the trustworthiness of the third party that signed the data. Through this multifaceted approach, data de-identification not only preserves data subjects' privacy but also ensures the integrity and credibility of the de-identified dataset.

This de-identified dataset is stored within the DCAS network as a separate entity, assigning a new address (hash) to the data. The person may share (possibly for compensation) this hash with third parties interested in utilizing the data for secondary purposes. In real life, the transfer of data from the person to the end user would probably not take place directly but through a data intermediary who aggregates the data of multiple persons and prepares them as a comprehensive data registry for the end-consumers for data analysis.

4.5 Personal primary use case

In the context of the DCAS network architecture, the personal primary use case focuses on empowering individuals with complete control over their health data. By leveraging DCAS technology, individuals can manage, share, and protect their health data more effectively, fostering a more personalized and secure healthcare experience.

The cornerstone of the personal primary use case is the individual's ability to consolidate and control their health data through a unified Personal Health Record (PHR). This PHR aggregates information from various healthcare providers, mobile applications, home medical devices, and personal health notes. As the data owner, the individual retains exclusive access to the

root hash, ensuring that they control who can access their data and under what circumstances. This control extends to updating, annotating, and managing their health data directly through a user-friendly core application.

One of the critical features of the proposed architecture is its emphasis on semantic interoperability. The PHR can be shared with healthcare providers across regions and systems, ensuring that the data is meaningful and useful regardless of the recipient's technology. This particularly benefits individuals who travel frequently or receive care from multiple providers. Sharing the root hash or utilizing standardized APIs, individuals can grant healthcare professionals access to their up-to-date health records, facilitating informed and timely medical decisions.

The architecture empowers individuals by enhancing transparency and ownership of their health data. Users can monitor all access to their health records. This transparency builds trust in the system and encourages individuals to engage more actively in their healthcare management. The ownership aspect is particularly transformative as it shifts the control of health data from institutions to individuals, enabling them to decide how their data is used and shared.

In addition to primary use, the architecture supports the secondary use of health data while maintaining privacy. Individuals can anonymize or pseudonymize their data and share it for research or commercial purposes. This contributes to societal health benefits and opens up opportunities for individuals to receive compensation for their data. The trusted third-party intermediary ensures that de-identified data remains credible and secure, facilitating its use in various secondary applications.

Integrating Artificial Intelligence (AI) and Machine Learning (ML) algorithms into the DCAS-based reference architecture adds a significant layer of personalization and precision to healthcare management. These technologies can analyze the comprehensive health data stored in the PHR to generate tailored lifestyle and healthcare recommendations. For instance, AI and ML can propose dietary adjustments, exercise plans, or preventive measures based on the individual's health history, genetic information, and real-time data from wearable devices. However, it is crucial to maintain a clear distinction between the recommendations provided to individuals and those given to healthcare professionals. Suggestions for personal use should focus on lifestyle and preventive care, empowering individuals to make informed decisions about their health. In contrast, recommendations for doctors should assist in clinical decision-making, ensuring they have accurate and relevant information to provide the best possible care. This separation is vital to prevent confusion and ensure that clinical advice remains in the domain of qualified healthcare providers.

4.6 Resolving the three dilemmas

The dilemma of accessibility is resolved by partitioning the entire personal health data space (Figure 2) in a DCAS network under the complete control and ownership of a data-owning person into distinct non-intersecting sub-spaces of identifiable and de-identified (anonymized or pseudonymized) health data. Identifiable personal health data stored within the former is exclusively controlled by their data owners (data subjects). As long as the root hash of the data remains secret and known solely to the owner, no other party, except those that the owner has explicitly shared the root hash with, has even a theoretical chance of accessing this data. Conversely, the data owner can generate numerous de-identified health data copies with minimal risk of re-identifying the data owner. These copies can be freely shared for secondary use.

The dilemma of comprehensiveness is resolved by consolidating a person's health data from multiple healthcare institutions, portable health devices, health-related applications, and other sources into a complete Personal Health Record (PHR). Since this comprehensive PHR remains under the exclusive physical control of the owner (data subject), the concentration of data does not increase the data leakage risks, as in the event of a successful attack, only one person's data can leak. A master copy of PHR data is used only in cases of initial use of data by sharing this data only with healthcare professionals from desired healthcare facilities regardless of region or national affiliation.

In addition, the ownership dilemma is resolved by storing personal health data within DCAS networks, where access requires the owner's root hash. The network's decentralization ensures that access is exclusively granted to the owner without intermediaries, e.g., without system administrators of hospital, regional, or national information systems. Consequently, the owner can manage their data much like any other private property, though they must acknowledge specific distinctive characteristics inherent to data compared to physical assets.

5 Analysis and discussions

5.1 Related works

The proposed DCAS-based architecture for personal health data presents an innovative approach to data management, emphasizing user control and data sharing. It resolves three critical health data challenges: accessibility, comprehensiveness, and ownership. In light of these challenges, we outline several initiatives that tackle similar issues.

MyData global (77) is a community advocating for human-centric data management, emphasizing data portability, interoperability, and user empowerment. They declare that they "*help people and organizations to benefit from personal data in a human-centric way.*" MyData aims to transform the data economy by ensuring individuals have more control over their data and can share it between services.

The International Data Space (IDS) (78) promotes data ownership through its data sovereignty principles, ensuring providers retain control over their data. This framework supports ownership rights across various industries, including healthcare. However, implementing ownership principles within IDS depends on the specific use cases and sectors involved.

Mediceus (79) ensures data ownership by providing a user-centric platform where individuals control their health data. Users can manage and share their data securely, maintaining ownership and control. While similar to DCAS in focusing on health data, Mediceus uses a more centralized approach to data management.

MIDATA's cooperative (80) model ensures that users are co-owners of their health data. This model prioritizes user interests and provides ownership rights through consent-based data sharing. Users have significant control over their data, although the cooperative model requires active participation and trust in its management.

Solid project (81) empowers users with ownership of their data by storing it in Pods (personal data spaces) managed by pod providers. Users can decide who accesses their data and revoke access anytime, ensuring solid data ownership. However, the ownership model is broader and not exclusive.

While these projects address issues related to accessibility, comprehensiveness, and ownership, they fall short of providing a holistic solution to all three.

5.2 Interoperability and privacy aspects

As illustrated in Figure 2, according to the proposed reference architecture, every citizen has a personal data space on the DCAS network, where health data as a PHR is preserved under the person's ownership and complete control. A detailed explanation of how health data is represented as PHRs on the DCAS network is beyond the scope of this document. However, we are working toward a unified clinical data model, formalized as RDF-based Knowledge Graphs, which supports ContSys ontology and federated semantic interoperability (66, 67, 82–94).

RDF is the standard data interchange model on the Web (95). An FHIR observation resource represented as RDF triplets is illustrated in Figure 11.

Traditionally, the RDF specification employs URIs to represent resources. However, within the realm of DCAS networks, an intriguing prospect arises: substituting URIs with hash values. Such an approach could alleviate numerous issues inherent in URIs, including collisions (distinct resources have the same URL) and aliases (multiple URLs refer to the same resource). By comparing URIs symbol by symbol, a match would unequivocally denote the same resource, eliminating ambiguity. Thanks to the deduplication feature of DCAS networks, it is ensured that a resource cannot possess disparate URIs.

Moreover, the immutable nature of addresses in DCAS guarantees that the meaning associated with any DCAS address-based URI remains constant over time. Unlike URLs on the internet, changes in ownership, and potential unavailability, the hash values (content addresses) of resources on a DCAS network remain unchangeable. This could pave the way for a new version of the internet, aligning closely with Tim Berners-Lee's vision of the Giant Global Graph (96).

We wish to underscore some considerations concerning data de-identification. Firstly, standard FHIR resources conventionally reference the treating physician and the data owner, typically the patient. While usually needed in API requests, this reference becomes redundant when storing data as Personal Health Records in the Personal Knowledge Graph. A more efficient approach involves preserving all demographic data in a distinct data subgraph. An affiliation to the owner is implicitly established by graph connectivity, obviating the explicit need for references to the subject within the resources. This omission of direct references to the data subject streamlines the pseudonymization process, requiring only the sharing of the address of the subgraph housing clinical data. Other identifiable data, such as the treating physician's name and their working institution, can also be separated by preserving them in a separate sub-graph, thus further strengthening the mechanisms for protecting personal data.

Under ordinary circumstances, the root hash of personal data is known exclusively to the data owner. While the owner may share it for primary use by medical service providers, it is conceivable to design protocols facilitating data sharing without divulging the hash. However, for secondary use, a prerequisite is the pseudonymization of the data. This involves creating a pseudonymized copy by expunging all references to individuals, institutions, locations, etc., retaining only essential clinical data. Additionally, all dates within the dataset could be rendered relative to the owner's birthdate. To fortify re-identification control, the hash of the pseudonymized dataset may be integrated into the original dataset, ensuring that only the original owner can reverse the pseudonymization process.

5.3 Compatibility with the European Health Data Space

The proposed reference architecture seamlessly aligns with and fully embraces the core principles of the European Health Data Space (EHDS) initiative, offering several valuable enhancements. The following outlines and provides commentary on enhancements resulting directly from the DCAS network characteristics or the proposed reference architecture.

Data security. The EHDS advocates for the availability of PHR data via access points established by member states. However, such access points entail heightened data leakage risks. In contrast, the proposed reference architecture employs a DCAS network for storing personal data, significantly mitigating such risks. By decentralizing data access, any potential breach would, at worst, result in the leakage of only one person's data without any impact on the security of others. This minimizes the vulnerability associated with centralized databases, where a breach could compromise millions of individuals' data.

The protocol design achieves data security in a DCAS network. Each network node stores the data locally as a key-value pair. The value corresponds to a distinct data fragment, while the key signifies its address (hash value). Individual fragments are encrypted utilizing distinct keys, rendering the data incomprehensible to the node. Consequently, the network nodes lack access to meaningful information regarding the content of the stored data. Moreover, the network routing protocol ensures that the transmission source of a particular data fragment holds no implications regarding its ownership. In other words, the recipient node remains unaware of whether the sender serves as the original data source or simply functions as an intermediary forwarder. Collectively, these measures signify that network nodes possess no discernible knowledge regarding the content or the rightful owner of the stored data. Consequently, the risk of data leakage becomes virtually negligible within such a system.

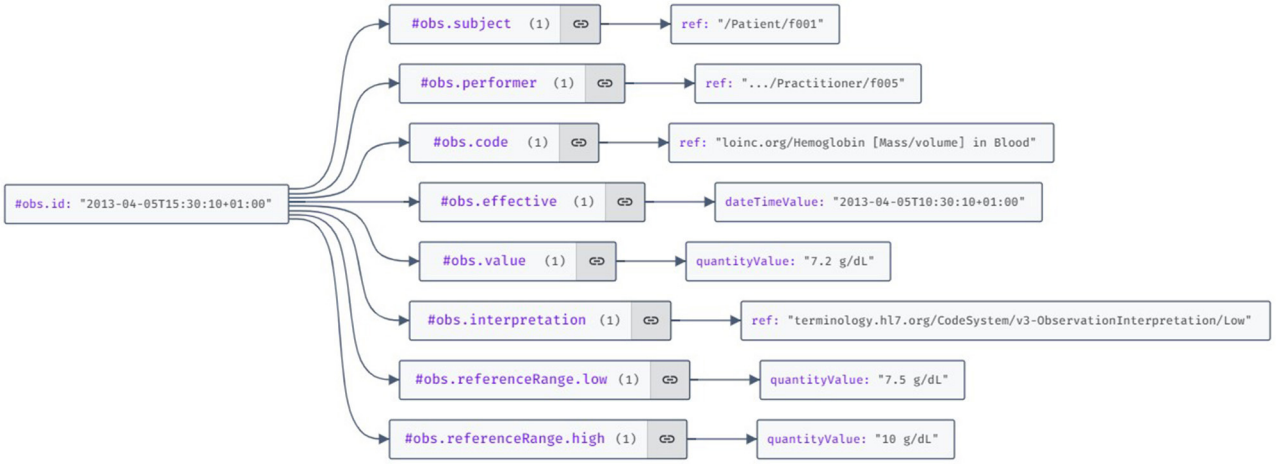

FIGURE 11
Visualization of the HL7 FHIR (Fast Healthcare Interoperability Resources) observation resource example using the JSON Crack features.

The inherently distributed nature of the DCAS network renders it challenging to launch cyber-attacks against it successfully. The absence of a single point of failure confers a substantial advantage, as the network remains unaffected even if specific nodes are compromised due to such attacks. Thus, in theory, the proposed architecture exhibits exceptional resilience against cyber threats.

Cost efficiency. Retaining personal data within DCAS networks external to the EHDS infrastructure generates substantial cost reductions for the entire system. This cost-effectiveness stems from two key factors: First, the absence of concentrated personal data in the system eliminates the need for extensive security measures associated with centralized storage and data-sharing protocols. Consequently, the security mechanisms implemented are notably more economical. Second, the utilization of DCAS networks predominantly leverages existing IT infrastructure. This strategic approach significantly diminishes the initial investments required to implement the entire solution and the ongoing expenses essential for its maintenance. The result is a streamlined, cost-effective system that aligns with contemporary economic considerations while ensuring enhanced data security.

Eliminating single points of failure. Another vulnerability of storing personal health data in a centralized repository lies in a single point of failure. In centralized repositories, the imperative becomes ensuring regular backups, consequently escalating the overall system cost. In contrast, in DCAS networks, each data point is dispersed across multiple nodes according to the built-in redundancy measures, eliminating the data loss risks associated with a centralized repository. This inherent resilience safeguards against potential data loss and obviates the need for recurrent and resource-intensive backup procedures. Opting for DCAS networks enhances data security and presents a cost-efficient alternative by eradicating the expenses of mitigating the risks of a single point of failure.

Simplicity. Eliminating the need to store personal data within central repositories simplifies the system considerably. Typically, an escalation in the complexity of information systems correlates with an augmented security risk, as a more intricate structure expands the potential attack surface (97). A simplified system streamlines operational aspects and inherently mitigates security risks. The logic is straightforward: the less intricate the system, the more manageable and controllable potential security risks become. Simplicity, in this context, acts as a strategic ally, making the system more dependable (98) and security management simpler. Simplicity enhances the system's efficiency and bolsters its security.

Reducing ecological impact. Managing health data for hundreds of millions of individuals in centralized systems demands substantial resources, encompassing hardware, energy, and labor, resulting in a notable ecological footprint. A centralized system's infrastructure, by its very nature, has enormous environmental impact. In contrast, DCAS networks utilize resources more efficiently. Operating predominantly on existing infrastructure, they demand relatively few additional resources. Consequently, the ecological footprint of such a decentralized solution is markedly smaller. Utilizing DCAS networks, we enhance the operational efficiency of health data management along with environmental sustainability by making informed choices to minimize the overall ecological impact of health data management systems.

Empowering data ownership. The core strategic objective of the EHDS is that of data owners maintaining absolute control over their data. When personal data resides on third-party servers, achieving data owner control becomes challenging. However, adopting DCAS networks establishes a paradigm where data owners have complete and exclusive control over their data. Furthermore, the authority to decide on data sharing rests solely with the owner, reinforcing the realization of the stipulated strategic goal. By embracing DCAS networks, we align with the EU's vision of robust data ownership and establish a framework that empowers individuals with unequivocal access control, ensuring the integrity and privacy of their data per EU strategic objectives.

Data integrity and version control. In DCAS networks, utilizing hash values as data addresses guarantees data integrity. Users can compute and compare the hash value with the original data address. A congruence between the two assures

the downloader that the downloaded data has not been altered. Furthermore, content addressability introduces an automatic versioning mechanisms' any alteration to the data results in assigning a new address reflective of the modified content. Simultaneously, the prior version of the data persists at its original address. This inherent version control facilitates the preservation of the data modification history. Notably, this characteristic empowers the creation of diverse sub-branches within the data, a useful feature for scenarios requiring selective information disclosure. Subsequently, these branches can be seamlessly amalgamated into a cohesive whole when needed.

Data preservation. Given the absence of a central control mechanism, the primary concern within a DCAS network is the preservation of stored data. Volunteers, the main operators of DCAS network nodes, may depart from the network independently. To mitigate the risk of data loss, the network must incorporate effective preservation mechanisms. One such mechanism involves providing rewards to network node operators, which incentivizes them to keep their network nodes online. Additionally, data preservation is facilitated by redundancy, wherein data is distributed across multiple network nodes. Consequently, the departure of a single node does not result in data loss. Ensuring an expansive network size, minimizing the likelihood of node departure, and maintaining sufficient data redundancy make it possible to minimize the probability of data loss to nearly negligible levels.

Re-centralization poses a significant risk to decentralized data networks, referring to accumulating a significant proportion of the network nodes under the control of a single operator. This consolidation empowers the operator to disrupt or halt the network's functionality. To avert this potential threat, the network must attain a substantial scale to render the concentration of a majority of network nodes under the oversight of a single operator unfeasible, both from a technical and financial standpoint. Ensuring a sizable network diminishes the likelihood of re-centralization, safeguarding the network's integrity and resilience.

Data quality enhancement. The reference architecture we propose substantially improves data quality. By storing PHR in a single logical location in a unified and coherent manner, issues arising from incomplete or conflicting information can be mitigated by the data owner's validation. Furthermore, the inherent characteristics of DCAS networks automatically guarantee data integrity and facilitate the preservation of a full version history.

Comprehensiveness. Storing a PHR in a unified location under the data owner's complete control resolves the prevalent issue of fragmented and incomplete data. Such data completeness effectively tackles the drawbacks associated with the secondary use of health data, which often necessitates gathering data from disparate service providers and increases the data privacy risks associated with secondary use.

Global scalability. DCAS networks operate using the Kademlia metric, eliminating the geographical dimension. For redundancy purposes, each data chunk is stored on all nodes belonging to a Kademlia neighborhood. It is important to recognize that within the Kademlia metric, nodes belonging to the same neighborhood may be widely dispersed geographically. In light of this, since each node only stores a small portion of the data, the question of where the data is stored in a geographical sense becomes meaningless. Ultimately, the data is stored simultaneously nowhere and everywhere.

Data de-duplication. Within the network, only one logical copy of identical data exists at any given time. This becomes particularly evident when dealing with large, immutable data entities (e.g., images, videos). Even if these entities are included in multiple data sets, such as in the pseudonymization process, only a single logical copy is present within the network. Thus, there is no need for redundant copies of these large data entities; a mere reference to them is sufficient.

5.4 Future work

This paper concludes the first part of our research by proposing the reference architecture for resolving health data accessibility, comprehensiveness, and ownership dilemmas by preserving semantically interoperable PHRs in DCAS networks. We have sketched the ideas (99) and submitted the technical solution as an EU patent application (100). Still, we have only proposed a technical solution. The proposed architecture's social, organizational, and legal aspects and applicability in real-life primary and secondary cases are for future study. The same is related to formal and real-life-based evaluation of the properties of DCAS networks in medical, medical emergency, secondary, and private use cases. Therefore, most of the research topics we proposed in Klementi et al. (99) are still to be studied and analyzed. Those topics are as follows:

- *Data model*—currently, we have only preliminary ideas of how the data in PHR in a DCAS network should be preserved; therefore, a data model that supports federated semantic interoperability with the existing and future developed hospital, regional, and national systems and also supports various data communication protocols (e.g., HL7 v.2.7, CDA or FHIR), reference models (e.g., HL7 RIM or openEHR RM), classifiers (SNOMED, ICD, LOINC or their different versions), languages (e.g., English, Estonian) as well as structured and unstructured data must be designed and implemented.
- *Data quality*—the mechanisms must be implemented for how the data is validated technically and clinically before being preserved in PHR in a DCAS network.
- *Data interoperability*—our research group is related to the development of TermX,[1] a platform for developing healthcare terminology and interoperability and other federated semantic interoperability-related development activities (66, 67, 90, 91, 101).
- *Primary use*—together with physicians, we are designing primary use-case studies to combine real-world clinical and patient-entered data in the treatment of selected diseases, e.g., cardiovascular and prostate diseases.

1 https://termx.org/

- *Secondary use*—we are designing different real-world secondary use cases related to clinical trials, public health, medical statistics, care efficiency, quality, etc.
- *Data security and privacy*—one of the directions here is to design a technical and organizational solution for health data de-identification so that the de-identified data is reliable for secondary use; another direction is to design and conduct proper real-world evidence-based experiments to justify these properties in primary and secondary use-cases.
- *Data integrity and transparency*—although data integrity and transparency arise from DCAS properties, we have to justify these in real-world evidence-based experiments during primary and secondary use.
- *Linked data*—the potential role of a DCAS network as the foundation for the Giant Global Graph (by Tim Berners-Lee) is an interesting related research topic.

6 Conclusion

The reuse of health data presents a significant challenge that currently lacks an effective solution. This article delves into the issue through the lenses of accessibility, completeness, and ownership. To address these challenges, we propose a novel, globally scalable architecture for a personal health data space based on decentralized content-addressable networks. It ensures that data subjects retain complete and exclusive control over their data, while enabling them to share it with third parties as they see fit.

To illustrate the problems, we present four use cases from the Estonian e-health system, demonstrating how the current methods fail to effectively address the three dilemmas. Following this, we analyze how the proposed new strategy resolves these issues.

The proposed architecture presents a notable departure from previous approaches to health data management and introduces a paradigm shift in the manner in which data storage is conceived. Therefore, it is expected that society will require a significant period of adjustment. Consequently, the feasibility of implementing the described solution in the immediate future appears remote. Nonetheless, it remains imperative for societal discourse to acclimate to emerging technological possibilities and navigate alongside them.

By providing enhanced control, interoperability, security, and transparency, the proposed solution has the potential to fundamentally transform how individuals interact with their health data. It empowers individuals to take an active role in their healthcare journey, fostering a more patient-centric and secure healthcare environment.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

TK: Conceptualization, Investigation, Software, Writing – original draft, Writing – review & editing. GP: Funding acquisition, Investigation, Methodology, Resources, Software, Supervision, Writing – original draft, Writing – review & editing. PR: Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research has been supported by the ICT Programme of the European Union through the European Social Fund and the IT Academy Research Measures (102).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Esnault C, Rollot M, Guilmin P, Zucker JD. Qluster: an easy-to-implement generic workflow for robust clustering of health data. *Front Artif Intell.* (2023) 5:1055294. doi: 10.3389/frai.2022.1055294

2. Xiang D, Cai W. Privacy protection and secondary use of health data: strategies and methods. *BioMed Res Int.* (2021) 2021:6967166. doi: 10.1155/2021/6967166

3. Kokshagina O. Managing shifts to value-based healthcare and value digitalization as a multi-level dynamic capability development process. *Technol Forecast Soc Change.* (2021) 172:121072. doi: 10.1016/j.techfore.2021.121072

4. Schäfke-Zell W. Revisiting the definition of health data in the age of digitalized health care. *Int. Data Priv. Law.* (2022) 12:33–43. doi: 10.1093/idpl/ipab025

5. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. *J Med Internet Res.* (2021) 23:e24387. doi: 10.2196/24387

6. Huckvale K, Nicholas J, Torous J, Larsen ME. Smartphone apps for the treatment of mental health conditions: status and considerations. *Curr Opin Psychol.* (2020) 36:65–70. doi: 10.1016/j.copsyc.2020.04.008

7. Prince AE, Schwarcz D. Proxy discrimination in the age of artificial intelligence and big data. *Iowa L Rev*. (2019) 105:1257.

8. Zindorf NE. Discrimination in the 21st century: protecting the privacy of genetic information in employment and insurance. *Tulsa LJ*. (2000) 36:703–26.

9. Slawomirski L, Oderkirk J. *Digital Technology: Making Better Use of Health Data. New Health Technologies: Managing Access, Value and Sustainability*. Paris: OECD Publishing (2017). p. 185. doi: 10.1787/9789264266438-9-en

10. PWC. *Transforming Healthcare Through Secondary Use of Health Data*. London: PriceWaterhouseCoopers (2009).

11. Hackl WO, Ammenwerth E. SPIRIT: systematic planning of intelligent reuse of integrated clinical routine data: a conceptual best-practice framework and procedure model. *Methods Inf Med*. (2016) 55:114–24. doi: 10.3414/ME15-01-0045

12. Wade TD. Refining gold from existing data. *Curr Opin Aller Clin Immunol*. (2014) 14:181. doi: 10.1097/ACI.0000000000000051

13. Zurynski Y, Smith CL, Vedovi A, Ellis LA, Knaggs G, Meulenbroeks I, et al. *Mapping the Learning Health System: a Scoping Review of Current Evidence*. Sydney: Australian Institute of Health Innovation and the NHMRC Partnership Centre for Health System Sustainability (2020).

14. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. *J Big Data*. (2019) 6:1–25. doi: 10.1186/s40537-019-0217-0

15. Martinez-Garcia M, Hernández-Lemus E. Data integration challenges for machine learning in precision medicine. *Front Med*. (2022) 8:784455. doi: 10.3389/fmed.2021.784455

16. Hulsen T, Friedeckỳ D, Renz H, Melis E, Vermeersch P, Fernandez-Calle P. From big data to better patient outcomes. *Clin Chem Lab Med*. (2023) 61:580–6. doi: 10.1515/cclm-2022-1096

17. Clayton EW, Embí PJ, Malin BA. Dobbs and the future of health data privacy for patients and healthcare organizations. *J Am Med Informat Assoc*. (2023) 30:155–60. doi: 10.1093/jamia/ocac155

18. Jamshidi M, Moztarzadeh O, Jamshidi A, Abdelgawad A, El-Baz AS, Hauer L. Future of drug discovery: the synergy of edge computing, internet of medical things, and deep learning. *Fut Internet*. (2023) 15:142. doi: 10.3390/fi15040142

19. Williamson SM, Prybutok V. Balancing privacy and progress: a review of privacy challenges, systemic oversight, and patient perceptions in AI-driven healthcare. *Appl Sci*. (2024) 14:675. doi: 10.3390/app14020675

20. Li H, Li C, Wang J, Yang A, Ma Z, Zhang Z, et al. Review on security of federated learning and its application in healthcare. *Fut. Generat. Comput. Syst*. (2023) 144:271–90. doi: 10.1016/j.future.2023.02.021

21. Getzen E, Ungar L, Mowery D, Jiang X, Long Q. Mining for equitable health: assessing the impact of missing data in electronic health records. *J Biomed Informat*. (2023) 139:104269. doi: 10.1016/j.jbi.2022.104269

22. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*. (2019) 17:1–9. doi: 10.1186/s12916-019-1426-2

23. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J Biomed Health Informat*. (2018) 22:1589–604. doi: 10.1109/JBHI.2017.2767063

24. Gansel X, Mary M, van Belkum A. Semantic data interoperability, digital medicine, and e-health in infectious disease management: a review. *Eur J Clin Microbiol Infect Dis*. (2019) 38:1023–34. doi: 10.1007/s10096-019-03501-6

25. Amar F, April A, Abran A. Electronic health record and semantic issues using fast healthcare interoperability resources: systematic mapping review. *J Med Internet Res*. (2024) 26:e45209. doi: 10.2196/45209

26. Fennelly O, Moroney D, Doyle M, Eustace-Cook J, Hughes M. Key interoperability factors for patient portals and electronic health records: a scoping review. *Int J Med Informat*. (2024) 2024:105335. doi: 10.1016/j.ijmedinf.2023.105335

27. Liddell K, Simon DA, Lucassen A. Patient data ownership: who owns your health? *J Law Biosci*. (2021) 8:lsab023. doi: 10.1093/jlb/lsab023

28. Martani A, Geneviève LD, Elger B, Wangmo T. "It's not something you can take in your hands". Swiss experts' perspectives on health data ownership: an interview-based study. *Br Med J Open*. (2021) 11:e045717. doi: 10.1136/bmjopen-2020-045717

29. EU. 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) (Text with EEA relevance). *Off J Eur Un*. (2016) 119:1–88.

30. Kahn SD, Terry SF. Who owns (or controls) health data? *Sci Data*. (2024) 11:156. doi: 10.1038/s41597-024-02982-1

31. Kovačević A, Bašaragin B, Milošević N, Nenadić G. De-identification of clinical free text using natural language processing: a systematic review of current approaches. *Artif Intell Med*. (2024) 53:102845. doi: 10.1016/j.artmed.2024.102845

32. Graef I. When data evolves into market power: Data concentration and data abuse under competition law. In: Moore M, Tambini D, editors. *Digital Dominance: The power of Google, Amazon, Facebook, and Apple*. Oxford University Press (2018). p. 72–97.

33. Murray-Watson R. Healthcare data breach statistics. *HIPAA J*. (2021). Available online at: https://www.hipaajournal.com/healthcare-data-breach-statistics/

34. Alder S. Largest healthcare data breaches of 2021. *HIPAA J*. (2021). Available online at: https://www.hipaajournal.com/largest-healthcare-data-breaches-of-2021/

35. Ardielli E. Implementation of eHealth applications by primary care physicians in the European Union member states. *Int J Electr Healthc*. (2021) 11:378–398. doi: 10.1504/IJEH.2021.117829

36. Metsallik J, Ross P, Draheim D, Piho G. Ten years of the e-health system in Estonia. In: Rutle A, Lamo Y, MacCaull W, Iovino L, editors, *CEUR Workshop Proceedings. Vol. 2336. 3rd International Workshop on (Meta)Modelling for Healthcare Systems (MMHS)* (2018). p. 6–15. Available online at: ceur/ws.org/Vol/2336/MMHS2018_invited.pdf (accessed June 27, 2024).

37. Directorate-General for Health and Food Safety. *A European Health Data Space: Harnessing the Power of Health Data for People, Patients and Innovation*. European Commission. (2022). Available online at: https://health.ec.europa.eu/document/download/17c7065c/c432-445f/9b27/8ccf283581bc_en?filename=com_2022/196_en.pdf (accessed January 28, 2024).

38. Rantanen MM, Koskinen J. Humans of the European data economy ecosystem-What do they demand from a fair data economy? In: *Human-Centric Computing in a Data-Driven Society: 14th IFIP TC 9 International Conference on Human Choice and Computers, HCC14 2020, Tokyo, Japan, September 9–11, 2020, Proceedings 14*. Berlin: Springer (2020). p. 327–39.

39. Kahn JS, Aulakh V, Bosworth A. What it takes: characteristics of the ideal personal health record. *Health Affairs* (2009) 28:369–76. doi: 10.1377/hlthaff.28.2.369

40. Wieringa RJ. *Design Science Methodology for Information Systems and Software Engineering*. Berlin: Springer (2014).

41. Health and Welfare Information Systems Centre. *The Health Portal*. (2024). Available online at: https://www.terviseportaal.ee/en/ (accessed March 17, 2024).

42. Information Systems Authority. *Electronic Identity eID*. (2024). Available online at: https://www.ria.ee/en/state-information-system/electronic-identity-eid-and-trust-services/electronic-identity-eid (accessed March 17, 2024).

43. Information Systems Authority. *Data Exchange Layer X-TEE*. (2024). Available online at: https://www.ria.ee/en/state-information-system/data-exchange-platforms/data-exchange-layer-x-tee (accessed March 17, 2024).

44. United Nations Department of Economica and Social Affairs. *E-Government Survey 2022. The Future of Digital Government*. (2024). Available online at: https://desapublications.un.org/sites/default/files/publications/2022-09/Web%20version%20E-Government%202022.pdf (accessed June 27, 2024).

45. Blobel B, et al. The international patient summary standard and the extensibility requirement. In: *pHealth 2020: Proceedings of the 17th International Conference on Wearable Micro and Nano Technologies for Personalized Health. Vol. 273*. Amsterdam: IOS Press (2020). p. 54.

46. Bertl M, Kankainen KJI, Piho G, Draheim D, Ross P. Evaluation of data quality in the Estonia national health information system for digital decision support. In: *Proceedings of the 3rd International Health Data Workshop*. Leicester (2023). p. 13.

47. Swarm. *Swarm Is a Decentralised Storage and Communication System for a Sovereign Digital Society*. (2022). Available online at: ethswarm.org (accessed March 24, 2022).

48. Doan TV, Psaras Y, Ott J, Bajpai V. Toward decentralized cloud storage with IPFS: opportunities, challenges, and future considerations. *IEEE Internet Comput*. (2022) 26:7–15. doi: 10.1109/MIC.2022.3209804

49. Maymounkov P, Mazieres D. Kademlia: a peer-to-peer information system based on the XOR metric. In: *International Workshop on Peer-to-Peer Systems*. Berlin: Springer (2002). p. 53–65.

50. Merkle RC. A digital signature based on a conventional encryption function. In: *Conference on the Theory and Application of Cryptographic Techniques*. Berlin: Springer (1987). p. 369–78.

51. Gray JN. An approach to decentralized computer systems. *IEEE Trans Softw Eng*. (1986) 6:684–92.

52. Rashid A, Siddique MJ. Smart contracts integration between blockchain and Internet of Things: opportunities and challenges. In: *2019 2nd International Conference on Advancements in Computational Sciences (ICACS)*. Lahore: IEEE (2019). p. 1–9.

53. Marandi A, Sehat H, Lucani DE, Mousavifar S, Jacobsen RH. Network coding-based data storage and retrieval for kademlia. In: *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*. Helsinki: IEEE (2021). p. 1–7.

54. Bustamante FE, Qiao Y. Designing less-structured P2P systems for the expected high churn. *IEEE/ACM Trans. Netw*. (2008) 16:617–27. doi: 10.1109/TNET.2007.903986

55. Balaji S, Krishnan MN, Vajha M, Ramkumar V, Sasidharan B, Kumar PV. Erasure coding for distributed storage: an overview. *Sci China Inform Sci*. (2018) 61:1–45. doi: 10.1007/s11432-018-9482-6

56. Ungureanu C, Atkin B, Aranya A, Gokhale S, Rago S, Calkowski G, et al. HydraFS: a high-throughput file system for the HYDRAstor content-addressable storage system. *FAST*. (2010) 10:225–39.

57. Hinsen K. The magic of content-addressable storage. *Comput Sci Eng*. (2020) 22:113–9. doi: 10.1109/MCSE.2019.2949441

58. Lakhani VH, Jehl L, Hendriksen R, Estrada-Galiñanes V. Fair incentivization of bandwidth sharing in decentralized storage networks. In: *2022 IEEE 42nd International Conference on Distributed Computing Systems Workshops (ICDCSW)*. Bologna: IEEE (2022). p. 39–44.

59. Chen Y, Richter JI, Patel PC. Decentralized governance of digital platforms. *J Manag*. (2021) 47:1305–37.

60. Dyson SF. Blockchain investigations-beyond the "money". *J Br Blockchain Assoc*. (2019) 2:6. doi: 10.31585/jbba-2-2-(6)2019

61. Shamir A. How to share a secret. *Communications of the ACM*. (1979) 22:612–3.

62. Kalra D, Beale T, Heard S. The openEHR foundation. *Stud Health Technol Informat*. (2005) 115:153–73.

63. ISO. *ISO 13606-1:2019 Health Informatics—Electronic Health Record Communication—Part 1: Reference Model*. Geneva: International Organization for Standardization (2019).

64. ISO. *13940:2015 Health Informatics—System of Concepts to Support Continuity of Care*. Geneva: International Organization for Standardization (2015).

65. Klementi T, Kankainen KJI, Piho G, Ross P. Prospective research topics towards preserving electronic health records in decentralised content-addressable storage networks. In: *HEDA@ Petri Nets*. Bergen (2022). p. 14.

66. Randmaa R, Bossenko I, Klementi T, Piho G, Ross P. Evaluating business meta-models for semantic interoperability with FHIR resources. In: *HEDA-2022: the International Health Data Workshop, June 19–24, 2022*. Bergen: CEURAT (2022). p. 14.

67. Sõerd T, Kankainen K, Piho G, Klementi T, Ross P. Towards specification of medical processes according to international standards and semantic interoperability needs. In: *MODELSWARD* (2023). p. 160–7.

68. Chen D, Doumeingts G, Vernadat F. Architectures for enterprise integration and interoperability: past, present and future. *Comput Indus*. (2008) 59:647–59. doi: 10.1016/j.compind.2007.12.016

69. Synthea. *Synthea Is a Synthetic Patient Population Simulator*. (2023). Available online at: https://github.com/synthetichealth/synthea (accessed March 09, 2023).

70. Prud'hommeaux E, Collins J, Booth D, Peterson KJ, Solbrig HR, Jiang G. Development of a FHIR RDF data transformation and validation framework and its evaluation. *J Biomed Informat*. (2021) 117:103755. doi: 10.1016/j.jbi.2021.103755

71. Shivers J, Amlung J, Ratanaprayul N, Rhodes B, Biondich P. Enhancing narrative clinical guidance with computer-readable artifacts: authoring FHIR implementation guides based on WHO recommendations. *J Biomed Informat*. (2021) 122:103891. doi: 10.1016/j.jbi.2021.103891

72. Swarm. *Bee Is the Software Run By Swarm Network Nodes*. (2023). Available online at: https://github.com/ethersphere/bee (accessed March 09, 2023).

73. HL7. *FHIR Is a Standard for Health Care Data Exchange, Published by HL7®*. (2022). Available online at: http://hl7.org/fhir/ (accessed January 28, 2024).

74. Kask M, Piho G, Ross P. Systematic literature review of methods for maintaining data integrity. In: *Advances in Model and Data Engineering in the Digitalization Era: MEDI 2021 International Workshops: DETECT, SIAS, CSMML, BIOC, HEDA, Tallinn, Estonia, June 21–23, 2021, Proceedings 10*. Berlin: Springer (2021). p. 259–68.

75. Kask M, Klementi T, Piho G, Ross P. Maintaining data integrity in electronic health records with hyperledger fabric. In: *The 3rd International Workshop on Health Data Co-located with STAF 2023, 18–21 July*. Leicester (2023). p. 1–17.

76. Kask M, Klementi T, Piho G, Ross P. Preserving decentralized EHR-s integrity. In: *Telehealth Ecosystems in Practice*. Amsterdam: IOS Press (2023). p. 296–7.

77. MyData Global. *Empowering Individuals by Improving Their Right to Self-determination Regarding Their Personal Data*. (2024). Available online at: https://mydata.org/

78. International Data Space Association. *The Future of the Data Economy Is Here*. International Data Spaces Association (2024). Available online at: https://internationaldataspaces.org/

79. Mediceus. *Your Data, Your Health, Your Choice*. Mediceus (2024). Available online at: https://www.mediceus.pt/

80. Midata Cooperative. *My Data—Our Health*. MiData Cooperative (2024). Available online at: https://www.midata.coop/en/home/

81. Solid. *Your Data, Your Choice. Advancing Web Standards to Empower People*. Solid (2024). Available online at: https://solidproject.org/

82. Piho G, Roost M, Perkins D, Tepandi J. Towards archetypes-based software development. In: Sobh T, Elleithy K, editors. *Innovations in Computing Sciences and Software Engineering*. Dordrecht: Springer (2010). p. 561–6.

83. Piho G, Tepandi J, Parman M, Perkins D. From archetypes-based domain model of clinical laboratory to LIMS software. In: *The 33rd International Convention MIPRO*. New York, NY: IEEE (2010). p. 1179–84.

84. Piho G, Tepandi J, Roost M. Domain analysis with archetype patterns based Zachman Framework for enterprise architecture. In: *2010 International Symposium on Information Technology. Vol. 3*. New York, NY: IEEE (2010). p. 1351–6.

85. Piho G, Tepandi J, Roost M. Evaluation of the archetypes based development. In: *Databases and Information Systems VI*. Amsterdam: IOS Press (2011). p. 283–95.

86. Piho G, Tepandi J, Roost M, Parman M, Puusep V. From archetypes based domain model via requirements to software: exemplified by LIMS Software Factory. In: *2011 Proceedings of the 34th International Convention MIPRO*. New York, NY: IEEE (2011). p. 570–5.

87. Piho G. *Archetypes Based Techniques for Development of Domains, Requirements and Software: Towards LIMS Software Factory*. Tallinn University of Technology (2011). Available online at: digi.lib.ttu.ee/i/?636 (accessed June 27, 2024).

88. Piho G, Tepandi J, Roost M. Archetypes based techniques for modelling of business domains, requirements and software. In: *Information Modelling and Knowledge Bases XXIII*. Amsterdam: IOS Press (2012). p. 219–38.

89. Piho G, Tepandi J, Parman M. Towards LIMS (laboratory information management systems) software in global context. In: *2012 Proceedings of the 35th International Convention MIPRO*. New York, NY: IEEE (2012). p. 721–6.

90. Piho G, Tepandi J, Thompson D, Tammer T, Parman M, Puusep V. Archetypes based meta-modeling towards evolutionary, dependable and interoperable healthcare information systems. *Proc Comput Sci*. (2014) 37:457–64. doi: 10.1016/j.procs.2014.08.069

91. Piho G, Tepandi J, Thompson D, Woerner A, Parman M. Business archetypes and archetype patterns from the HL7 RIM and openEHR RM perspectives: towards interoperability and evolution of healthcare models and software systems. *Proc Comput Sci*. (2015) 63:553–60. doi: 10.1016/j.procs.2015.08.384

92. Kankainen KJI. Usages of the ContSys standard: a position paper. In: Bellatreche L, Chernishev G, Corral A, Ouchani S, Vain J, editors. *Advances in Model and Data Engineering in the Digitalisation Era, MEDI 2021, 21–23 June 2021, Tallinn, Estonia. Vol. 1481 of Communications in Computer and Information Science*. Dordrecht; Heidelberg; New York; London: Springer Nature (2021). p. 314–24.

93. Kankainen K, Klementi T, Piho G, Ross P. Using SNOMED CT as a semantic model for controlled natural language guided capture of clinical data. In: *HEDA@ Petri Nets* (2022). p. 1–13.

94. Kankainen K. Usages of the ContSys standard: A position paper. In: *Advances in Model and Data Engineering in the Digitalization Era: MEDI 2021 International Workshops: DETECT, SIAS, CSMML, BIOC, HEDA, Tallinn, Estonia, June 21–23, 2021, Proceedings 10*. Berlin: Springer (2021). p. 314–24.

95. W3C. *Resource Description Framework*. (2024). Available online at: https://www.w3.org/RDF/ (accessed March 09, 2024).

96. Hendler J, Berners-Lee T. From the Semantic Web to social machines: a research challenge for AI on the World Wide Web. *Artif Intell*. (2010) 174:156–61. doi: 10.1016/j.artint.2009.11.010

97. Zhan Y, Ahmad SF, Irshad M, Al-Razgan M, Awwad EM, Ali YA, et al. Investigating the role of Cybersecurity's perceived threats in the adoption of health information systems. *Heliyon*. (2024) 10. doi: 10.1016/j.heliyon.2023.e22947

98. Avizienis A, Laprie JC, Randell B. *Fundamental Concepts of Dependability. Department of Computing Science Technical Report Series* (2001).

99. Klementi T, Kankainen KJI, Piho G, Ross P. Prospective Research Topics towards Preserving Electronic Health Records in Decentralised Content-Addressable Storage Networks. In: *HEDA@ Petri Nets. Proceedings of The International Health Data Workshop co-located with 10th International Conference on Petrinets (Petri Nets 2022)* Bergen (2022).

100. Klementi T, Piho G. *Method and System for Managing Data Using Decentralized Content-Addressable Storage Networks* (2024). Submitted Patent. European Patent Office, priority number EP24166173.5.

101. Bossenko I, Linna K, Piho G, Ross P. Migration from HL7 clinical document architecture (CDA) to fast health interoperability resources (FHIR) in infectious disease information system of Estonia. In: *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*. Tallinn (2023). p. 882–5.

102. Estonian Research Information System. *IT Academy Research Support Measures Programme for 2018–2022: Artificial Intelligence and Machine Learning; Data Science and Big Data; Robots-People Collaboration and the Internet of Things in Industry Processes*. Estonian Research Council (2023).

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 30 July 2024
DOI 10.3389/fmed.2024.1393123

OPEN ACCESS

EDITED BY
Mauro Giacomini,
University of Genoa, Italy

REVIEWED BY
Giovanna Nicora,
University of Pavia, Italy
Daniele Pala,
University of Pavia, Italy
Simone Rancati, University of Pavia, in collaboration with reviewer DP
Lorenzo Peracchio, University of Pavia, in collaboration with reviewer GN

*CORRESPONDENCE
Tuncay Namli
✉ tuncay@srdc.com.tr

RECEIVED 28 February 2024
ACCEPTED 16 July 2024
PUBLISHED 30 July 2024

CITATION
Namli T, Sınacı AA, Gönül S, Herguido CR, Garcia-Canadilla P, Muñoz AM, Esteve AV and Ertürkmen GBL (2024) A scalable and transparent data pipeline for AI-enabled health data ecosystems.
Front. Med. 11:1393123.
doi: 10.3389/fmed.2024.1393123

A scalable and transparent data pipeline for AI-enabled health data ecosystems

Tuncay Namli[1]*, Ali Anıl Sınacı[1], Suat Gönül[1], Cristina Ruiz Herguido[2], Patricia Garcia-Canadilla[2], Adriana Modrego Muñoz[2], Arnau Valls Esteve[2] and Gökçe Banu Laleci Ertürkmen[1]

[1]SRDC Software Research Development and Consultancy A. Ş., Ankara, Turkey, [2]Fundacio Sant Joan De Deu, Barcelona, Spain

Introduction: Transparency and traceability are essential for establishing trustworthy artificial intelligence (AI). The lack of transparency in the data preparation process is a significant obstacle in developing reliable AI systems which can lead to issues related to reproducibility, debugging AI models, bias and fairness, and compliance and regulation. We introduce a formal data preparation pipeline specification to improve upon the manual and error-prone data extraction processes used in AI and data analytics applications, with a focus on traceability.

Methods: We propose a declarative language to define the extraction of AI-ready datasets from health data adhering to a common data model, particularly those conforming to HL7 Fast Healthcare Interoperability Resources (FHIR). We utilize the FHIR profiling to develop a common data model tailored to an AI use case to enable the explicit declaration of the needed information such as phenotype and AI feature definitions. In our pipeline model, we convert complex, high-dimensional electronic health records data represented with irregular time series sampling to a flat structure by defining a target population, feature groups and final datasets. Our design considers the requirements of various AI use cases from different projects which lead to implementation of many feature types exhibiting intricate temporal relations.

Results: We implement a scalable and high-performant feature repository to execute the data preparation pipeline definitions. This software not only ensures reliable, fault-tolerant distributed processing to produce AI-ready datasets and their metadata including many statistics alongside, but also serve as a pluggable component of a decision support application based on a trained AI model during online prediction to automatically prepare feature values of individual entities. We deployed and tested the proposed methodology and the implementation in three different research projects. We present the developed FHIR profiles as a common data model, feature group definitions and feature definitions within a data preparation pipeline while training an AI model for "predicting complications after cardiac surgeries".

Discussion: Through the implementation across various pilot use cases, it has been demonstrated that our framework possesses the necessary breadth and flexibility to define a diverse array of features, each tailored to specific temporal and contextual criteria.

KEYWORDS

artificial intelligence, dataset, harmonization, transparency, FHIR, interoperability, health data spaces

1 Introduction

1.1 Background and objectives

Transparency, and traceability are considered among the key requirements for trustworthy artificial intelligence (AI) by the AI-Act which will be governing the use of AI solutions in the EU (1). Lack of transparency in the data preparation process, i.e., the difficulty in tracking and understanding the transformations and manipulations that the data undergoes before being used for training is a major issue for building trustworthy AI solutions (2).

Today's AI models depend on a complex, iterative process involving extensive communication among medical professionals, data scientists, and database administrators. Medical experts outline the specific data needed for the AI project to data scientists and AI developers, who then pass these requirements to database administrators. These administrators are responsible for retrieving the relevant data from existing sources, such as Electronic Health Records (EHR), based on the defined variables. Typically, this procedure is manually carried out by database administrators, resulting in time-consuming, labor-intensive tasks that lack transparency and traceability. Data scientists check the accuracy and relevance of the data, while medical professionals evaluate the performance of the AI model trained using this data. This prone-to-error and laborious back-and-forth continues until there is a mutual understanding and satisfaction with the data prepared for the AI application. This lack of transparency can lead to several issues:

- Reproducibility: Without knowing the exact steps taken during data preparation, it becomes difficult to reproduce the same dataset or validate the results obtained from the AI model. This also hampers the ability to effectively train AI models across several sites via a federated learning architecture. For example, researchers may decide to exclude data from certain patients with specific phenotypes (e.g., having a condition like epilepsy). Even if they document this exclusion by indicating the name of the disease, the lack of clear coding for exclusion criteria can still pose a problem due to insufficient transparency. Medical terminologies (e.g., ICD-10 codes for diagnosis) are used to indicate phenotypes, and the usage of these medical concepts can vary among different healthcare settings. Even if the same terminology is used, the practical definition of the phenotype can differ between two healthcare settings. Therefore, when an AI model is deployed in a different setting, it is crucial that the phenotype definitions are clear and transparent. This allows for proper configuration and customization of data mapping or preprocessing steps according to how medical terminologies are used in that specific setting.
- Debugging: When unexpected results occur during model training or inference, it can be challenging to identify the root cause without knowing how the training data was prepared.
- Bias and Fairness: Data transformations and preprocessing steps can inadvertently introduce biases into the dataset, leading to biased AI models. During training data extraction and data cleaning step, decisions on how to handle missing values can introduce biases. For example, if data for certain racial groups is more likely to have missing values, imputing these with overall mean values might not accurately reflect the health status of these groups. Then suppose that an AI model trained and deployed to predict health risks for a diverse patient population where this data cleaning step not documented transparently. Because the model was trained on a biased dataset, it may not perform well for these underrepresented groups. Due to this unidentified bias, the model may perform well on the majority population but poorly on minorities which may exacerbate existing health disparities. Without traceability, it's challenging to detect and mitigate these biases.
- Compliance and Regulation: In regulated domains such as healthcare, there are regulatory requirements for documenting data processing steps for transparency and auditability purposes.

In this paper, we propose a formal data preparation pipeline specification to overcome the limitations of the manual and error-prone data extraction processes for AI and data analytics applications, while also addressing the issue of traceability. We introduce a declarative JSON-based language designed for specifying how to extract data from datasets that adhere to Common Data Models, specifically those compatible with HL7 Fast Healthcare Interoperability Resources (FHIR) standards (3), as part of a pipeline process to prepare data for AI. Our goal is to enhance the scalability, transparency, and reproducibility of AI applications by streamlining the data extraction phase and clearly separating medical knowledge from data engineering knowledge. Our approach also endeavors to offer a practical methodology for realizing the objectives outlined in the European Health Data Space (EHDS) legislation (4) which aims to facilitate health data portability and foster the development of a unified market for health data for secondary use purposes. The transparent and declarative model used to define the data preparation pipeline

supports the aggregation of health data from diverse sources, thereby making them accessible for clinical research.

Our objective is to establish a data preparation pipeline originating from Electronic Health Record (EHR) sources to generate AI-ready training datasets. This task presents several challenges due to the inherent complexity of EHRs, rendering them unsuitable for direct use as feature vectors in training AI models (5–7). EHR data is structured in intricate, nested, high-dimensional models with diverse data types often linked to external domain-specific terminologies and code systems, enhancing the semantic understanding of data entities. However, this structure doesn't align directly with the flat feature vectors expected by AI methods, typically represented as normalized, domain-agnostic value sets. Moreover, EHR data records unevenly distributed clinical events, resulting in irregularly sampled sparse time series data, further complicated by the presence of missing values.

Converting EHR data into feature vectors suitable for AI methods necessitates multiple steps, involving various decisions. These decisions encompass selecting domain-specific codes from international code systems to determine which data entities from EHR to include, identifying necessary temporal joins and aggregations, determining the resampling strategy for longitudinal EHR data, specifying transformations for normalization, unit conversion, and harmonization, and devising approaches to handle missing data. The design of our declarative data preparation pipeline definition has been guided by these challenges. It is crafted to transparently define each step of the transformation pipeline as a sequence of data processing and transformation actions in a standardized manner.

Our data preparation pipeline model is technology agnostic; it provides a machine processable definition of the pipeline steps. In this paper, to demonstrate the effectiveness of the proposed methodology, we also briefly describe our implementation of an engine, called "onfhir-feast", that processes this machine processable pipeline definition to extract AI-ready datasets from EHR sources. Implemented as a high-performance distributed engine, we showcase its ability to efficiently extract datasets for various use cases. This domain-specific, technology-agnostic language establishes a standardized approach for a variety of stakeholders, including data scientists, health data owners, and AI or clinical decision support service vendors, to collaborate and develop AI-based solutions or conduct research studies. This framework enables scalability and reproducibility, ensuring that solutions and studies can be effectively implemented and replicated across different healthcare settings.

1.2 Related research

One of the pioneering initiatives to enable observational research on top of EHRs is OHDSI (8). OHDSI offers the OMOP Common Data Model (CDM) (9), which standardizes the structure and content of observational data with the support of a standardized vocabulary. Additionally, OHDSI provides a suite of open-source tools, including ATLAS for designing and executing observational research studies, and ACHILLES for characterizing and visualizing source data. While OMOP CDM serves as a solid foundation, it requires extension and specialization to cater to the needs of domain specific research studies, such as cancer research (10) and medical imaging (11). In OMOP CDM approach, it is not possible to document these extensions and customizations in a machine processable and traceable manner. Our approach addresses this gap by introducing a FHIR-based CDM, which meticulously documents all customizations in a machine-processable manner via the profiling methodology. Although OHDSI's open-source tools facilitate population queries and dataset extraction, this process is not documented in a machine processable manner which diminishes the end-to-end transparency, and traceability of the dataset preparation process. This deficiency hampers reproducibility and auditability, key requirements of AI-Act.

There have been a number of efforts in the literature to flatten the hierarchical EHR data to create AI-ready tabular datasets. Fiddle (6) provides an open-source generic preprocessing pipeline implementation for extracting structured data from the EHR data with three distinct steps, namely for pre-filtering, transforming and post-filtering. As HL7 FHIR is widely supported by numerous health care institutions and vendors of clinical information systems, several efforts focused on flattening EHR data represented as FHIR resources for extracting AI friendly data sets. Liu et al. (12) utilized the FHIR Bulk Data API to create population-level exports from clinical systems, into a file format often referred to as "Flat-FHIR' represented in NDJSON-based data format. FHIR-DHP (5) proposes a generic data harmonization pipeline (DHP) that is composed of data exchange, mapping, and export operations to transform EHR data to FHIR standard first, then to a relational database format, and exporting the data to a custom flattened JSON format as an AI-friendly format. FhirExtinguisher (13) has extended the FHIR Search API with an additional projection layer using FHIRPath, to build a tool for transforming FHIR resources into tabular data. Pathling (14) proposes an extended FHIR Analytics API, as a specialization of the FHIR API that focuses on providing functionality useful for health data analytics applications, namely: importing bulk FHIR data, execution of aggregation-based queries across a data set, searching via FHIRPath queries for cohort selection and extracting datasets to create custom data extracts for input into other tools and workflow. Although these efforts provided a generic methodology (5, 12) and/or extended API specification and implementation (13, 14) to flatten EHR data as tabular data sets, they do not provide a declarative model to formally define the data preparation pipeline. Our technology agnostic approach complements these, by providing an additional level of abstraction for enabling transparency, traceability, and reproducibility of AI methods. It should be noted that these efforts such as (13, 14) can be utilized to implement a transformation engine implementing the declarative data preparation pipeline definition proposed in this paper.

2 Materials and methods

2.1 Common data modeling for AI use case

A pivotal component of our methodology involves the construction of Common Data Models (CDMs) tailored for AI

applications. In our methodology, the CDMs are meticulously built utilizing the HL7 Fast Healthcare Interoperability Resources (FHIR) standard, leveraging the FHIR profiling technique. FHIR profiling is the process of defining or constraining FHIR resources to address specific requirements. This involves customizing FHIR's generic, standardized resources to create more precise models that cater to particular use cases, workflows, or data exchange scenarios within healthcare applications. These profiles dictate how FHIR resources are used, including the elements they must contain, the cardinality of these elements (e.g., optional, mandatory, repeating), and value sets or data types for each element.

In healthcare, AI applications are generally built for specific use cases, and the data requirements, so called variables needed for executing the AI model or visualizing the results, are declared within those use cases. One of the primary benefits of utilizing a CDM with HL7 FHIR profiling is the creation of a customized standard data model that is specifically tailored to the unique requirements of each AI use case. By defining a machine processable CDM that precisely aligns with the specific data elements, structures, and terminologies relevant to the use case, we ensure that the AI system is built upon a solid foundation of accurate and relevant data.

Utilizing a CDM defined by the HL7 FHIR standard is the establishment of a standardized interface for querying and accessing health records. This standardization not only simplifies the integration of disparate health information systems but also ensures that AI algorithms can access the necessary data in a consistent and reliable manner. By facilitating a uniform method to search and retrieve health records, we significantly reduce the complexity and variability often encountered in health data, thus enabling more efficient data processing and analysis.

The use of HL7 FHIR in defining our CDM enables the explicit declaration of information critical to the AI use case, such as phenotype and AI feature definitions. This is achieved by referring to inherent FHIR structures and standardized medical terminologies through the value set references. With this approach, an AI use case transparently declares its information of interest. As a result, our methodology not only enhances the semantic interoperability of health data but also ensures that the AI systems have access to a rich and semantically coherent dataset. This level of specificity and clarity in data representation is essential for the development of AI algorithms that are both effective and reliable in clinical settings.

As a common data model for a specific analytic or AI use case, we propose to provide a FHIR Implementation Guide including the following machine processable FHIR based definitions.

- A FHIR CapabilityStatement defining the list of related FHIR resource types needed for this use case as well as references to search parameters to be used to search related data for each resource type.
- A list of StructureDefinition resources defining syntactic and semantic customizations and restrictions representing a category of health events or facts needed for the use case.
- A list of ValueSet and/or CodeSystem resources defining the relevant concepts from terminology systems for restricting certain elements value sets or define information of interest.
- The adoption of a Common Data Model based on the HL7 FHIR standard, tailored through the FHIR profiling approach, offers significant advantages for the development of AI in healthcare. It provides a standardized, customizable, and semantically rich framework for accessing and processing health records, thereby laying the groundwork for scalable and transparent AI solutions in healthcare.

2.2 Declarative model to define the data preparation pipeline

We propose an end-to-end data preparation pipeline that begins with clinical data sources, such as Electronic Health Records (EHRs) and supplies AI systems with training datasets. This pipeline can also be utilized to run intelligent clinical applications and decision support services built based on AI models readily on EHRs by seamlessly retrieving the input parameters.

By utilizing the Common Data Model built upon HL7 FHIR, we establish a standardized interface for accessing source data effortlessly. Our goal is to create a transparent pipeline utilizing this FHIR interface to generate a dataset optimized for AI applications. However, this presents a challenge: converting the nested, hierarchical data model of FHIR into a tabular or time series format compatible with mainstream AI frameworks such as TensorFlow (15), Pythorch (16) or Scikit-learn (17).

EHR data is intricate, featuring high dimensionality, irregular time series sampling, and a variety of data types with diverse representations of clinical events. Converting this complex EHR data into flat feature vectors that align with Machine Learning (ML) techniques poses several challenges.

- First and foremost, performing temporal joins and aggregations over EHR data is necessary to derive features or outcome variables in a dataset. These derived features will become columns in the tabular format expected by AI frameworks. This process also requires tailoring to the specific requirements of each use case. For instance, consider a scenario involving EHR data where a particular lab result, such as creatinine, is represented as a FHIR Observation type resource according to the FHIR standard. In a specific use case, like predicting complications after cardiac surgeries, various creatinine results may be relevant. These could include the creatinine level before surgery, the first creatinine result within 24 h after surgery, the latest creatinine result, the average of all results, and the difference between the first and last creatinine results. Each of these aspects needs to be defined as separate features specific to the given use case scenario. Thus, it's essential to establish a method for defining how these use case-specific tabular feature sets can be extracted from the hierarchical, relational FHIR-based model for each unique use case scenario.
- Frequently, transformations are required to adjust the scale or discretize the numeric values found in EHR data, such as laboratory values. This is necessary to create normalized features that align with the expectations of ML models. Additionally, numeric values expressed in various units may

need to be converted to a specified unit for the sake of harmonization.

- In EHR data, clinical events are logged as they occur within the clinical workflow, leading to irregular sampling of time series events, which differs from the regular sampling expected by ML methods. Consequently, it's essential to establish strategies for resampling longitudinal EHR data to meet the requirements of specific use cases.

Each of these steps requires numerous decisions from data scientists in the data preparation process. Transparency within these decisions is vital for data transparency, as they heavily impact the characteristics, and quality of the resulting dataset. To ensure clarity in defining these steps or decisions, we introduce a declarative model aimed at precisely defining each step of the transformation in the pipeline from EHR data to AI-ready feature sets.

The HL7 FHIR API offers a standardized query language, included within the FHIR API's search interaction, enabling the querying of health data. Additionally, there's another language known as FHIRPath (18), designed for processing and navigating FHIR content. Our declarative model leverages these FHIR Query and FHIRPath statements to transparently define health datasets as a series of data processing and transformation steps in a standardized manner. Through this declarative model, transformation steps can be precisely defined and executed to prepare training, validation, or test datasets for AI. Moreover, it can also facilitate the preparation of features for executing decision support models during online prediction.

In the process of designing our declarative model, we aimed to recognize the steps typically taken by data scientists or research groups when creating a dataset through conventional methods, which often involve coding in Python and/or SQL. We endeavored to devise a practical approach to achieve the same using FHIR constructs. The following steps have been identified and form the primary sections of our declarative model:

- Definition of target population: This step entails identifying the target cohort by declaratively specifying the characteristics of the data entities that will comprise the target population for a specific use case, utilizing inclusion and exclusion criteria. In certain scenarios, entities eligible for inclusion in the dataset may be limited to specific time periods. This step also allows for defining these eligible time periods tailored to the specific use case requirements.
- Definition of feature groups: In this step, we define an intermediate result set, as a group of base features, that can be retrieved from EHR and can be utilized in the next step to calculate the final set of features required by the use case. At this step it is also possible to define transformations such as unit conversions to create harmonized data sets.
- Definition of final datasets: This step includes defining the individual features based on the base features identified in feature groups. At this step, we first define the rules for resampling of longitudinal health data, and also define anchor time points that are important for the use case. Following this, we declaratively specify how final features in the dataset can be calculated based on the base features, and anchor timepoints, through a set of temporal and contextual constraints, aggregations, and transformations.

TABLE 1 An example definition of target population for a simple use case; "Patients diagnosed with Parkinson."

```
{
"url": "https://aiccelerate.eu/cohorts/pilot2/
parkinson_cohort",
"name": "parkinson_cohort",
"title": "Patients diagnosed with Parkinson",
"description": "Patients diagnosed with Parkinson
(ICD-10 G20 code)",
"version": "0.1",
"date": "2022-04-21T00:00:00",
"fhirVersion": "4.0.1",
"publisher": "AICCELERATE WP1 Team (SRDC Corp.)",
"entityType": ["Patient"],
"eligibilityCriteria": [
   {
      "fhirSearch": "?",
      "description": "All patients with a parkinson
diagnosis (ICD-10 G20)",
      "filter": [
        {
          "resourceType": "Condition",
          "fhirSearch": "?code = http://hl7.org/fhir/
sid/icd-10{\T1\textbar} G20&patient = {{Patient}}",
          "entities": ["Condition.subject"],
          "eventTime": "Condition.onsetDateTime"
        }
     ]
   }
 ]
```

In the subsequent sections, we explore the intricacies of the methodologies and processes involved, elucidating the benefits and functionalities of the suggested approach and solution. Through the application of the proposed language, we offer exemplary definitions to demonstrate the adaptability of our method across a diverse range of use cases.

2.2.1 Definition of target population

The initial phase in preparing the dataset involves defining the target cohort, which entails specifying the characteristics or phenotype of the entities (for instance, patients) designated as the target population for the current use case, and whose information will be incorporated into the dataset. In this context, we adhere to the definition provided by OHDSI, which describes a cohort as "a set of persons who satisfy one or more inclusion criteria for a duration of time" (19).

In our approach, a single population definition is engineered for versatility across various use cases, thereby allowing it to be a distinct, reusable component within different dataset definitions. For illustration, Table 1 presents a population definition tailored for

datasets focusing on Parkinson's disease patients. Each construct, starting with the population definition, is initiated with metadata elements such as title, description, version, and a canonical URL, to provide a comprehensive overview of the definition. To establish a population definition effectively, it is essential first to identify the specific entities comprising the population. Within the FHIR framework, there are distinct resource types—such as Patient, Practitioner, and Organization—designed to represent individuals (e.g., patients or healthcare practitioners) or entities (e.g., organizations), along with their foundational information. All ancillary resources link back to these primary resources to delineate their interrelations. For instance, a lab result, denoted by a FHIR Observation resource, specifies its associated patient through a reference type element that points to the pertinent FHIR Patient resource.

As depicted in Figure 1, our methodology employs the names of FHIR resource types, such as "Patient," to denote that our target population primarily consists of patients, reflecting the common practice in health data analytics. Unlike OHDSI, our approach expands the definition of the target population to include not only individual entities like patients or practitioners but also conceptually broader categories such as encounters or episodes of care. These categories encompass both the individual involved and specific events, such as a hospital visit or a surgical care episode. This broader categorization allows us to leverage the relationships established in FHIR between resources like FHIR Encounter or EpisodeOfCare and other FHIR resources (for example, medications administered during a hospital stay). By doing so, we facilitate precise grouping of data based on distinct criteria, thereby enhancing the clarity and utility of the data for health analytics.

The subsequent step involves defining the eligibility criteria, which detail the characteristics or phenotypes of the entities in question. Our framework supports the definition of multiple eligibility criteria, recognizing that a single entity may exhibit different characteristics based on varying representations of underlying facts. For instance, patients with diminished kidney function might be identified through FHIR Condition resources that implicitly diagnose with specific ICD-10 codes, or through eGFR measurements depicted by FHIR Observation resources where the value falls within a certain range. This flexibility also accommodates the inclusion of various sub-cohorts within the dataset. The process of defining eligibility criteria begins with a FHIR query statement targeting the base entity type, which in our scenario is the FHIR Patient resource. At this example, we impose no limitations on demographic information such as age, gender, or ethnicity, which are typically included in the FHIR Patient resource type. The criteria are further refined through additional filter definitions applied to other FHIR resource types. For example, to isolate patients diagnosed with Parkinson's disease, we employ a FHIR query on FHIR Condition resources. FHIR facilitates a universal search mechanism via RESTful API, providing a comprehensive list of search parameters for each resource type. These parameters enable queries on FHIR resources using filters based on coded, numeric, Boolean, textual, temporal, or relational information, including references among resources. In our methodology, we utilize these FHIR search statements to delineate a specific result set. In population definitions, these queries serve to filter entities that satisfy at least one condition specified by the query. Specifically, we target patients who have at least one FHIR Condition resource coded with the ICD-10 code "G20" for Parkinson's diagnosis. Each filter explicitly states the search parameter linking the population to that resource type (e.g., patient = {{Patient}} indicates the 'patient' parameter linked to our population's Patient entities) and includes a FHIRPath expression that specifies the path for entity identifiers (e.g., "Condition.subject"). For more intricate scenarios, additional filters on other resource types can be defined to specify further characteristics required for an entity to be considered eligible for the cohort. Moreover, FHIRPath expressions allow for the imposition of additional conditions on the result set for each filter, addressing constraints that the standard FHIR query mechanism may not accommodate.

In certain cases, entities qualify for inclusion in a cohort only during specific time frames or across multiple intervals. This means that the relevance of an entity to a use case hinges on its state within these designated periods. For instance, in the context of constructing a dataset for analyzing or predicting the progression of Parkinson's disease, our interest is confined to the period following a patient's Parkinson diagnosis. By utilizing a FHIRPath expression to mark the event time, we can determine the precise moment each entity enters the cohort, which for our example is the onset time of Parkinson, as recorded in the FHIR Condition resource. While it's also possible to define an exit time when an entity no longer meets the cohort criteria, this aspect is not utilized in our example scenario. Consider a use case aimed at examining patient outcomes in relation to a specific medication regimen over time. Here, the start and end times of medication administration, as documented in FHIR MedicationRequest resources, could serve as the markers for entering and exiting the cohort, respectively. Given that medication prescriptions are often renewed, multiple resources may document medication use for distinct periods. In such instances, it becomes necessary to identify multiple eligibility periods for patients. Our methodology allows for the specification of a minimum time gap between consecutive eligibility periods. For example, setting a 15-day minimum gap implies that if the interval between two prescriptions is less than 15 days, they are considered part of the same usage period. This approach enables the precise delineation of eligibility periods in alignment with clinical guidelines or practices. As we will detail in forthcoming sections, these eligibility periods—particularly the defined entry and exit times—are critical for the sampling of data used in creating training or validation datasets. They also play a pivotal role in the development of other features that hinge on these specific temporal markers.

The defined entry and exit times within population criteria are instrumental when establishing criteria based on the temporal relationship between two health events. An illustrative scenario, as discussed in the Book of OHDSI, involves identifying "patients who initiate ACE inhibitors monotherapy as first-line treatments for hypertension." In such a case, one might set up a filter on the Condition resource to search for a hypertension diagnosis, utilizing the diagnosis or onset date as the event time. Subsequently, an additional filter could be applied to MedicationRequest or MedicationStatement resources. This filter would search for a specific set of ATC codes corresponding to ACE inhibitors, incorporating an extra condition. This condition, defined using a FHIRPath expression, would stipulate that the temporal gap

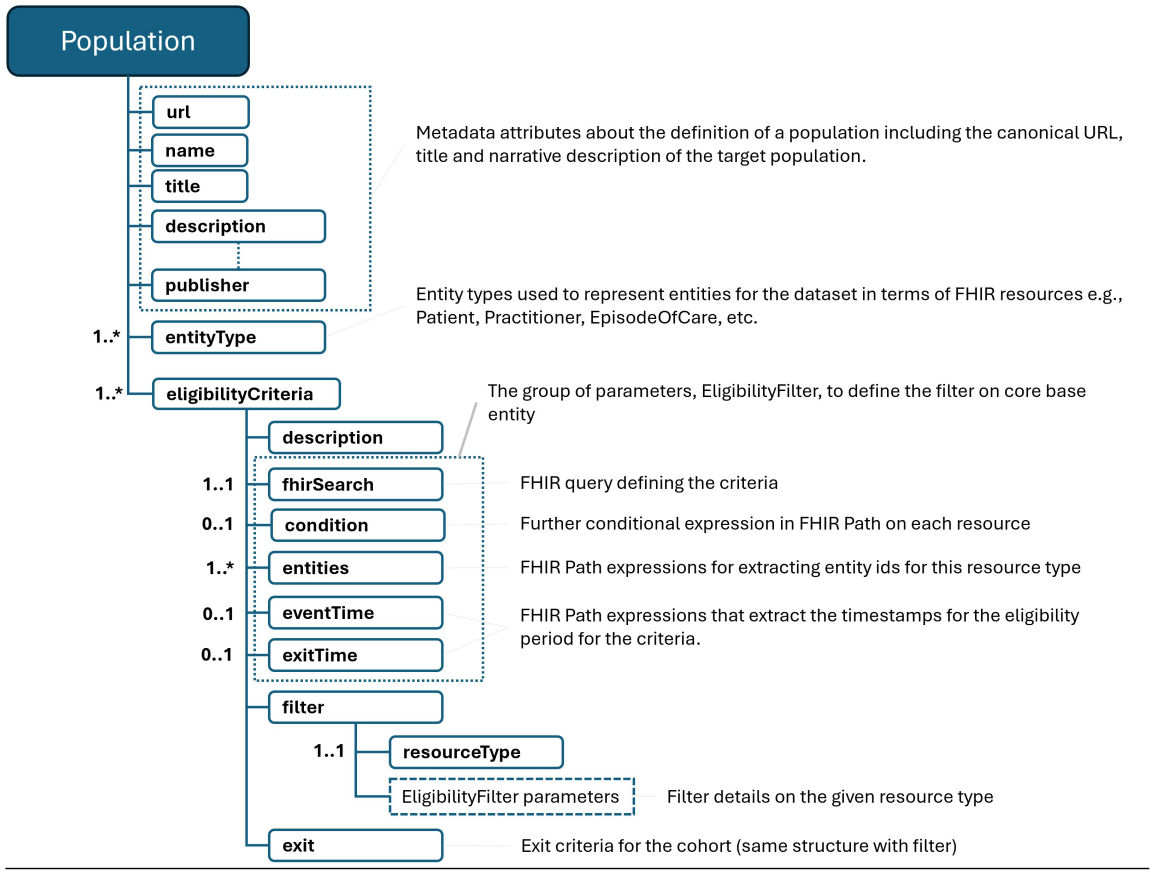

FIGURE 1
Population definition schema. "*" gives the cardinality of corresponding element and means it is an array and 0 or more cardinality.

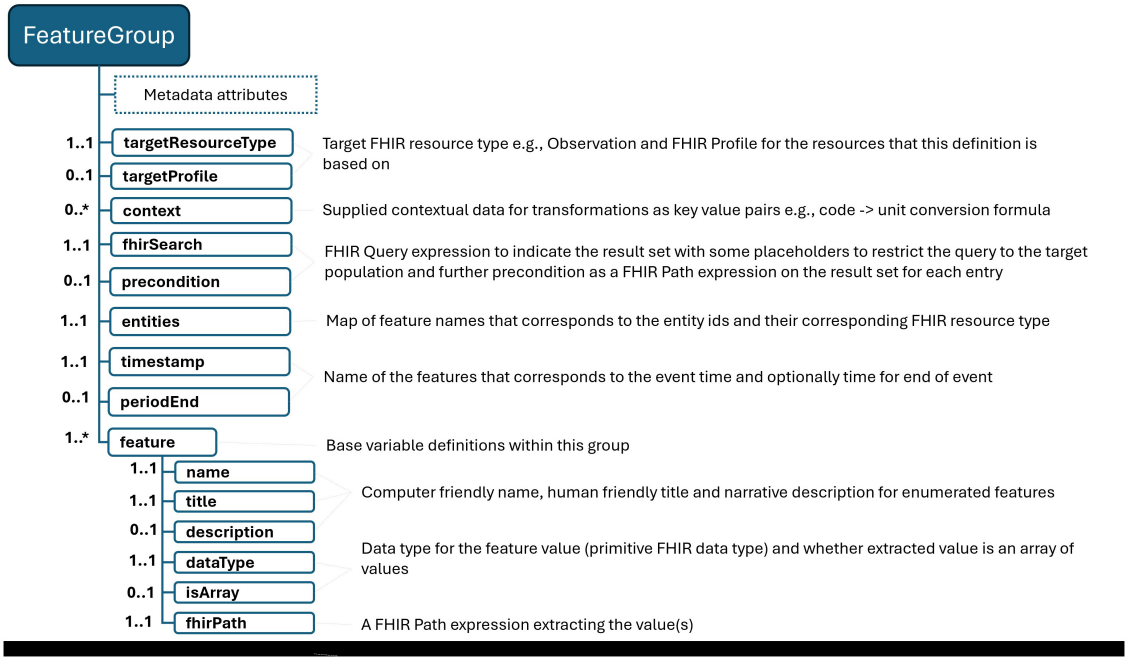

FIGURE 2
FeatureGroup definition schema. "*" gives the cardinality of corresponding element and means it is an array and 0 or more cardinality.

TABLE 2 A sample Feature Group definition to retrieve blood pressure measurements for the specific population.

```
{
    "url": "https://aiccelerate.eu/feature-groups/
pilot1/bloodpressure",
    "name": "bloodpressure",
    "title": "Blood Pressure Measurement",
    "description": "Represent a blood pressure
measurement including systolic and diastolic",
    "version": "0.1",
    "date": "2022-09-07",
    "fhirVersion": "4.0.1",
    "publisher": "AICCELERATE WP1 Team (SRDC
Corp.)",
    "targetResourceType": "Observation",
    "targetProfile":
"http://hl7.org/fhir/StructureDefinition/bp",
    "fhirSearch":
"?patient = {{Patient}}&category = http:
//terminology.hl7.org/CodeSystem/observation-
category{\T1\textbar}
vital-signs&code = http://loinc.org{\T1\textbar}
85354-9",
"entities": {
      "pid": "Patient"
    },
    "timestamp": "time",
    "feature": [
     {
       "name": "pid",
       "title": "Patient identifier",
       "description": "Patient identifier",
       "dataType": "id",
       "fhirPath": "Observation.subject"
      },      {
       "name": "time",
       "title": "Observation time",
       "description": "Time of measurement",
       "dataType": "dateTime",
       "fhirPath": "Observation.effectiveDateTime"
      },
      {
       "name": "systolic",
       "title": "Systolic BP",
       "description": "Systolic BP value",
       "dataType": "decimal",
       "fhirPath":
"Observation.component.where(code.coding.exists
(system = 'http://loinc.org' and
code = '8480-6')).first().valueQuantity.value"
      },
```

(Continued)

TABLE 2 (Continued)

```
      {
       "name": "diastolic",
       "title": "Diastolic BP",
       "description": "Diastolic BP value",
       "dataType": "decimal",
"fhirPath":
"Observation.component.where(code.coding.exists
(system = 'http://loinc.org' and
code = '8462-4')).first().valueQuantity.value"
      }
   ]
}
```

between the hypertension diagnosis and the initiation of ACE inhibitor therapy must be at least 365 days. This method enables the precise definition of eligibility criteria that hinge on the chronological sequencing of health-related events.

In addition to inclusion criteria, certain use cases necessitate the establishment of exclusion criteria. Continuing with the aforementioned example, the criterion "with no history of prior hypertension treatment" mandates verifying the absence of any hypertension treatment prior to the identified ACE inhibitors monotherapy, subsequently excluding those patients from the population. This is achieved through the same mechanism of filter definitions, which, when designated as exclusions, allow for the identification and removal of such cases. Entities for which at least one resource meets the FHIR query and the specified condition are thus excluded from the population. This method enables the precise tailoring of the population by omitting entities that do not meet the defined criteria.

2.2.2 Definition of feature groups

With an understanding of the necessary features and outcome variables, the following step involves identifying the specific FHIR resources required to compute these variables, ensuring access via the FHIR API while adhering to the agreed-upon common data model. To facilitate the reuse of these definitions across varying scenarios and dataset constructs, we introduce a concept known as a "feature group." Figure 2 briefly summarizes the definition schema. This construct allows for the delineation of result sets tailored to specific needs. For instance, as depicted in Table 2, one can establish a feature group aimed at gathering blood pressure readings for the targeted population, subsequently isolating systolic and diastolic values as base features for subsequent analyses while calculating other features. Essentially, feature group definitions articulate a FHIR result set—stemming from a particular FHIR query—alongside the specific data points to be extracted from this set.

Similar to defining a population, we employ FHIR search statements to outline the desired result set, specifying both the type of FHIR resource and the expected target FHIR profile to ensure the resulting resources conform accordingly. In the given example, we opt for the FHIR Blood Pressure profile, which mandates the use of the LOINC code 85354-9 to identify blood pressure measurement records specifically. This code is utilized as a filter within the

search statement. Additionally, the relevant FHIR reference or ID type search parameter is paired with an entity placeholder (e.g., patient = {{Patient}}). This approach signifies that our request is exclusively for records pertaining to patients within the defined population, ensuring that the data collected is directly relevant to our study's subjects.

We proceed to identify the base variables to be extracted from the result set determined by the FHIR search statements. This compilation should encompass potential identifiers for the entities involved, and, where applicable, associated time information that elucidates the timing of the clinical event or fact in question. FHIR resources, akin to other clinical record models, are capable of representing health-related facts or events in three distinct categories: (i) time-independent information, such as demographics provided by the FHIR Patient resource or family health conditions outlined in the FHIR FamilyMemberHistory resource; (ii) events/facts associated with a specific time point, like the onset date of a chronic condition detailed in the FHIR Condition resource or a particular laboratory result specified in the FHIR Observation resource; and (iii) events/facts pertinent to a defined time period, for instance, the duration of medication use indicated by the FHIR MedicationStatement resource. In the construction of these definitions, it is crucial to map entity identifiers to their corresponding entity types. For instance, in our scenario, we designate "pid" as the identifier for patients. Additionally, temporal variables—such as the timestamp of the clinical event/fact or the start and end times for events/facts spanning a period—must be articulated for the feature groups, except those involving time-independent information. In our case, we specify that the variable "time," representing the moment of blood pressure observation, will serve as the timestamp for the data in question.

Illustrated by our example, each variable is accompanied by metadata including its name, description, and data type (aligned with FHIR data types), along with a FHIR Path expression that specifies the method for extracting information from the result set. Beyond mere extraction, FHIR Path can be employed for data transformations or calculations. For instance, in situations where your common data model does not limit the units for a particular laboratory result or if there are several unit options, FHIR Path expressions can be used to convert numeric values from various units into a standardized unit, facilitating data harmonization. Similarly, these expressions can be applied to rescale or discretize numeric values, aiding in data normalization. Our approach allows for the inclusion of such contextual data within the definition itself, providing formulas for unit conversion, thresholds for clinical measurements, etc. This enables the use of FHIR Path expressions for performing the requisite calculations. By integrating contextual data and its metadata within the dataset definition and keeping it separate from the scripts, we adhere to our principles of transparency and readability. Additionally, this method enhances the configurability and reusability of the definitions.

2.2.3 Definition of dataset

We introduce the concept of "feature set"; similar to other constructs within our framework, its definition begins with essential metadata that provides a verbal description of the dataset to be prepared with respect to the feature set definition. The definition model is illustrated in Figures 3, 4. We outline a strategy for resampling the longitudinal health data, which often displays characteristics of sparsity and irregular sampling intervals, with various variables being recorded at disparate frequencies. The pivotal decision here involves determining the sampling time points for each entity, essentially deciding what each row in the dataset represents. This decision is intricately linked to the specific analytics or AI use case envisioned for the dataset. Current practices in the literature, employed by data scientists and researchers, offer several approaches for this:

- Selecting the start or end times of specific health events as sampling points. For example, utilizing the discharge time from a hospital as the sampling point for a dataset aimed at predicting hospital readmission.
- Segmenting a period to establish sampling points based on the frequency of the most regularly recorded data. An instance of this would be dividing the time from the end of surgery until discharge into 8-h intervals for a dataset intended to predict the length of stay following cardiac surgeries.
- Dividing a period while also incorporating outcome events into consideration. For example, segmenting the duration of an Intensive Care Unit (ICU) stay into 5-min intervals, but also using the occurrence time of sepsis as an additional sampling point and adjusting the time windows accordingly. This approach aims to predict sepsis during ICU stays by analyzing vital signs and other frequent measurements, ensuring snapshots of each patient are taken at 5, 10, 15 min, etc., prior to the observation of sepsis.

These strategies enable the creation of datasets that reflect the dynamics of patient health status over time, tailored to the specific analytical or predictive needs of the use case.

Within our framework, we've integrated a mechanism to streamline the definition of sampling strategies, as illustrated in Table 3 under the "referenceTimePoints" section. This mechanism allows users to specify the methodology for determining sampling time points in a structured manner:

- Method: The "method" element specifies the chosen methodology for sampling. For methods that require dividing a period into sub-periods, we leverage the eligibility period calculated for each entity based on the population definition. For instance, in a scenario where a patient's eligibility period is delineated by the time span from the end of their first surgery to their discharge from the hospital, specifying a period (e.g., 1 h) means this duration will be segmented into 1-h intervals.
- Outcome Events: If the determination of time points also takes into account certain outcome events, these are specified by referencing one or more FeatureGroup definitions. In the given example, a FeatureGroup that provides data on complications is utilized for this purpose.
- Configuration: Users can further refine the strategy by setting a time offset to define the exact sampling point relative to an event, as well as a minimum gap between two outcome events for them to be considered distinct outcomes. In the context of predicting post-operative complications, the example specifies that two complications must occur at least 8 h apart. Additionally, it stipulates that the initial sampling

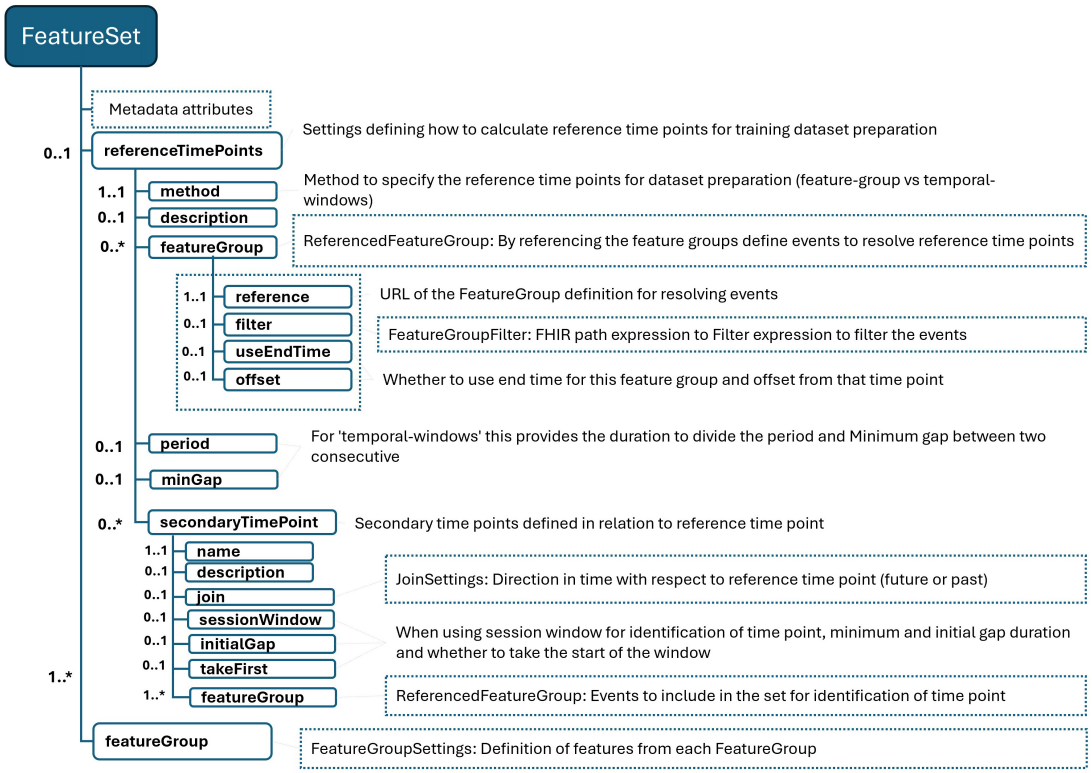

FIGURE 3
FeatureSet definition schema. "*" gives the cardinality of corresponding element and means it is an array and 0 or more cardinality.

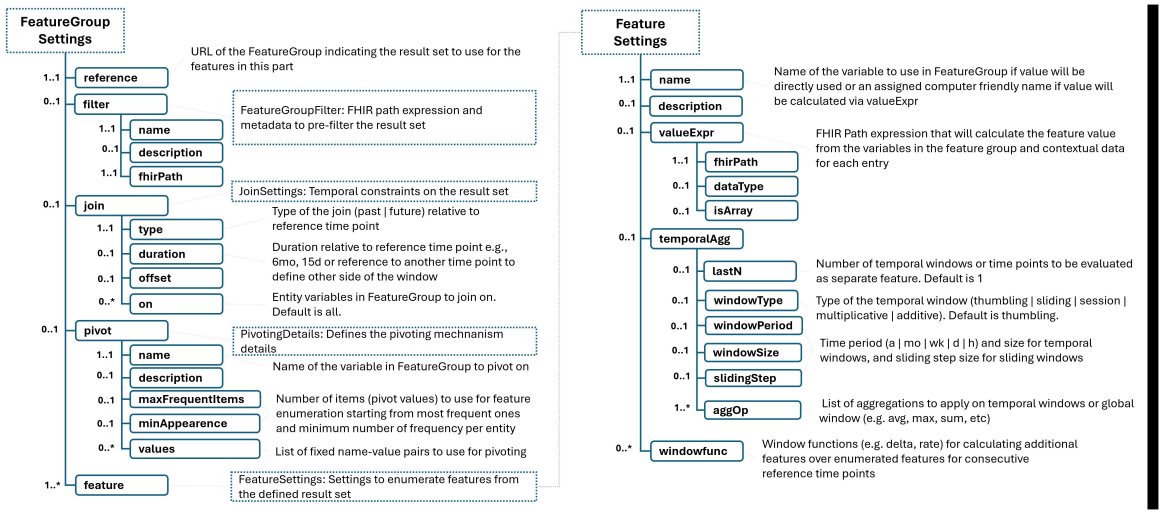

FIGURE 4
Remaining of FeatureSet schema. "*" gives the cardinality of corresponding element and means it is an array and 0 or more cardinality.

point should be set 1 h before the occurrence of the earliest complication.

This flexible mechanism enables precise configuration of sampling strategies, tailoring the dataset to capture clinically relevant events and periods. By incorporating both fixed intervals and event-driven sampling points, researchers can create datasets that more accurately reflect the complexities of patient care trajectories, enhancing the potential for insightful analysis and predictive modeling.

In addition to primary sampling points, certain scenarios benefit from the delineation of secondary time points, which correspond to significant health events within the patient's care continuum. These secondary time points are defined in relation

TABLE 3 A sample FeatureSet definition—Defining sampling time points and other time points.


```
{
    "url": "https://aiccelerate.eu/feature-sets/
pilot1_hsjd_complications",
    "name": "pilot1_hsjd_complications",
"title": "Feature set for AICCELERATE Pilot 1 for
predicting complications after the surgery",
    "description": "Feature set for AICCELERATE
Pilot 1 for predicting complications after the
surgery",
    "version": "0.1",
    "date": "2022-04-21",
    "fhirVersion": "4.0.1",
    "publisher": "AICCELERATE WP1 Team (SRDC
Corp.)",
    "referenceTimePoints": {
      "method": "temporal-windows",
      "description": "The time period between the end
of first surgical operation and the discharge time is
divided into 1h periods. However if patient has
complications then these event times are considered
as anchor points and reference time points are
calculated accordingly. Two complications with less
than 8 h are assumed same complication so no
reference time point is picked within this period.
Enumeration for reference time points start from 1 h
before the complications",
      "useEndTime": true,
      "period": "1h",
      "minGap": "8h",
      "offset": "1h",
      "featureGroup": [{
        "reference":"https:
//aiccelerate.eu/feature-groups/pilot1/complication"
      }],
      "secondaryTimePoint": [
      {
        "name": "lastSurgeryTime",
        "description": "Time of the latest main
surgery performed in the episode",
        "join": { "type": "past" },
        "featureGroup": [
        {
          "reference":"https://aiccelerate.eu/
feature-groups/pilot1/surgeryEncounter",
          "useEndTime": true,
          "filter": {
            "name": "isMainSurgery",
            "description": "If procedure is main
cardiac surgery",
            "fhirPath": "category = '394603008'"
        }
      }
   ]}
]},
.....
```


TABLE 4 A part of sample FeatureSet definition – Defining features from medication data for predicting progression to Advanced Parkinson Disease.


```
{
   "reference":"https:
//aiccelerate.eu/feature-groups/pilot2/medication",
   "join": {
   "type": "past",
   "duration": "3mo"
   },
   "feature": [
     {
      "name": "hasBenzodiazepinesRecently",
      "description": "Whether patient has
benzodiazepines or not within this period. ATC Code:
under N05CD",
      "valueExpr": {
      "fhirPath": "atcCode.startsWith('N05CD')",
"dataType": "boolean"
    },
"temporalAgg": [{
      "aggOp": ["any"]
    }]
   },
...
```


to the primary sampling points, offering a nuanced timeline that captures critical clinical milestones. In the provided example of Table 3, the secondary time point "lastSurgeryTime" is identified as the time marking the end of the patient's last surgery, as indicated by the relevant feature group that records surgery encounters. This point is determined to be the closest, yet prior, instance to the established primary sampling time point. To ensure the significance of each identified event, users have the flexibility to specify a minimum interval that should exist between two consecutive events. Furthermore, the framework allows users to select specific events (e.g., first, last, second to last) to serve as these secondary time points. Secondary time points, along with primary sampling points, play a crucial role in defining the temporal context for data analysis and feature extraction. For instance, in the example, the "lastSurgeryTime" serves as a pivotal reference for calculating features such as the number of hours elapsed since the most recent surgery at each sampling point. This approach allows for the inclusion of dynamic, temporally relevant information in the dataset, enhancing the precision of subsequent analyses and the development of predictive models that accurately reflect patient trajectories and outcomes.

The process of transforming raw health data into meaningful dataset features involves defining a set of temporal and contextual constraints, aggregations, and transformations based on the information provided by related feature groups. Each feature group encapsulates a category of health events (e.g., lab results, diagnoses, surgeries) along with the base facts of these events, the entity they are related to, and the time or period of the event. To convert

TABLE 5 A part of sample FeatureSet definition—Enumerating features from frequent SPO2 measurements in hospital after cardiac surgeries for predicting complications.

```
{
      "reference": "https:
//aiccelerate.eu/feature-groups/pilot1/vitalsign",
      "filter": {
       "name": "spo2",
       "fhirPath": "code = '2708-6'"
      },
      "feature": [
        {
         "name": "value",
         "description": "Aggregation of last (2,4 and
8)-h time windows for SPO2 measurements",
         "temporalAgg": [
           {
              "lastN": 3,
              "windowPeriod": "h",
              "windowSize": 2,
              "extending": "multiplicative",
              "aggOp": ["stddev", "avg", "max", "min",
"median", "kurtosis", "skewness"]
           }
         ],
         "windowFunc": ["delta"]
      },
      {
        "name": "value",
        "description": "Aggregations of last 3 1-h
time windows for body SPO2 measurements",
        "temporalAgg": [
          {
              "lastN": 3,
              "windowPeriod": "h",
              "windowSize": 1,
              "aggOp": ["stddev", "avg", "max", "min",
"median", "kurtosis", "skewness"]
          }
        ],
        "windowFunc": ["delta"]
      }
    ]
   },
...
```

these facts into actionable variables, it's essential to establish clear temporal relationships between the facts represented by the selected feature groups and the predefined anchor time points. For instance, as illustrated in Table 4, when defining features, one may only want to consider medication usage data from the most recent three months. This decision impacts the definitions of features within the dataset, such as a Boolean feature indicating recent benzodiazepine use by a patient. This "recency" is calculated in relation to the main sampling time point for each record, ensuring that the feature reflects current or recent medication use. The language designed for this purpose allows users to define temporal constraints with ease, specifying periods in relation to defined time points either by indicating a duration that looks forward (future) or backward (past) in time. This flexibility can include optional offsets or can be bounded between two specific time points. For example, to focus on diagnoses made after a patient's Parkinson's diagnosis, one could define a temporal period that spans from the time of the patient's eligibility to the sampling time point. This approach facilitates the generation of features that are not only relevant to the patient's current health state but also temporally aligned with the objectives of the study or analysis. It allows for the creation of datasets that can more accurately model health outcomes by incorporating the timing and sequence of health events in relation to significant clinical milestones.

In the process of defining features for a dataset, it's not only possible to apply temporal constraints to health event data, but you can also impose contextual constraints to further refine the data included in your analysis. This is accomplished by specifying filters on the data represented by a feature group. These filters are expressed using FHIRPath expressions, which allow for precise selection of data based on specific criteria. Table 5 shows an example of such a contextual constraint filtering complication data, where the result set is based on FHIR AdverseEvent resources. By applying a filter using the corresponding SNOMED-CT code, one can specifically target unexpected ICU admission events. This method ensures that the dataset only includes relevant adverse events, thereby enhancing the specificity and relevance of the analysis.

The language introduces a "pivoting mechanism" for efficiently handling scenarios where it's necessary to generate a standardized set of features across multiple concepts within the same category, such as laboratory test results. This mechanism is particularly useful for cases where analysts wish to extract a common suite of statistical measures (e.g., the latest, average, minimum, and maximum values) for a variety of tests or measurements that are relevant to their specific use case. The first step involves selecting a base variable from the FeatureGroup definition that will serve as the pivot. This could be, for example, the LOINC code for a laboratory test, which uniquely identifies the type of lab test being conducted. Users can then specify a list of values and corresponding labels for this pivot variable. These values could be specific LOINC codes for lab tests that are of particular interest in the use case. If the exact tests of interest are known ahead of time, they can be explicitly listed in the model. If the specific items of interest are not predetermined, the model allows users to define criteria for automatically selecting these pivot variables based on the data. For instance, one might specify that features should be enumerated for the 20 most frequently occurring lab tests in the dataset, provided that each of these tests appears in the records of at least 100 patients. This pivoting mechanism simplifies the process of generating a consistent set of features across multiple data points or concepts, which is particularly valuable when dealing with large and complex datasets. It ensures that analysts can focus on analyzing the most

relevant and frequently occurring data points without manually defining features for each possible variable.

In the process of defining features for a dataset, there are two primary methods to derive feature values from the underlying data represented by feature groups: direct use of base variable values and calculation through FHIR Path expressions.

- Direct Use of Base Variables: A feature can be directly based on the value of a base variable that has been defined within the related feature group. This approach is straightforward and involves using the raw value of a data point as a feature in the dataset. Table 5 shows an example using the "value" variable from a feature group that represents vital sign information.
- Calculation Through FHIRPath Expressions: Alternatively, features can be derived by applying FHIRPath expressions to calculate values from the data records within each feature group. This method allows for more complex transformations of the data. As shown in Table 4, an example illustrates how medication usage data, identified by ATC codes in the medication usage feature group, can be transformed into a feature indicating whether the patient is using a medication from the benzodiazepine group. This involves interpreting the ATC codes using FHIRPath expressions to identify specific medication classes and then summarizing this information into a binary feature (e.g., benzodiazepine usage: yes/no).

For addressing the challenges of data harmonization, especially when dealing with disparate measurement units, scales, or categorization needs stemming from different calibration standards of medical devices or varied clinical practices, the model introduces a mechanism for specifying and applying contextual information. The model facilitates this through a dedicated section within FeatureGroup or FeatureSet definitions, designed for the transparent declaration of contextual parameters. These parameters can encompass a wide array of transformational instructions, such as:

- Conversion formulas for standardizing units of measurement (e.g., converting temperature from Fahrenheit to Celsius or blood pressure readings from mmHg to kPa).
- Rescaling instructions for numerical values to align with a common scale or range, enhancing comparability.
- Categorization criteria based on clinical thresholds or norms, enabling the transformation of continuous data into discrete categories that reflect clinical significance (e.g., defining hypertension stages based on blood pressure readings).
- Terminology mappings, which are crucial for harmonizing data coded in different clinical terminologies or classification systems, such as mapping between different coding systems for diagnoses or medications (e.g., ICD to SNOMED-CT).

The proposed language provides a convenient method for generating multiple features from a single value by leveraging a combination of aggregation operators and temporal windowing strategies. This approach allows for the extraction of rich, time-sensitive insights from health data, particularly useful for variables that are measured repeatedly over time, such as vital signs or lab results. The key aspects of this feature include:

- Aggregation Operators: Users can apply a variety of standard aggregation functions, such as standard deviation, average, and maximum, to a set of data points. These functions are akin to those found in SQL and data processing frameworks like Apache Spark (20), ensuring familiarity and ease of use for those with a background in data science.
- Temporal Windowing: The language supports several types of temporal windows, including tumbling, extending, session, and sliding windows. This flexibility allows users to analyze data over specified periods in a manner that best suits their analytical or predictive needs. For example, users can look at the last 3 1-h windows or extend their analysis over longer periods, such as 2, 4, and 8 h, to observe trends or changes over time.
- Configuration Flexibility: Parameters such as the number of windows, window size, extension factor, or sliding step duration can be easily adjusted. This configurability enables users to tailor their analysis to specific requirements or hypotheses about the data.
- Extension Capability: While the language comes with a set of predefined aggregation operators, it is designed to be extensible. Implementors can introduce additional operators as needed, enhancing the language's applicability to a wide range of scenarios and datasets.
- Delta and Rate of Change: Beyond simple aggregations, the language supports operators for calculating changes between consecutive temporal windows, such as the delta or rate of change. This feature can be particularly insightful for tracking the progression or improvement of a patient's condition over time, offering a dynamic view of health status that static measurements cannot provide.

As exemplified in Table 5 and Figure 5, by applying these techniques to SpO2 (oxygen saturation) measurements, users can generate a comprehensive set of features that describe not just the current state but also the variability and trends of a patient's oxygen levels over time. Such detailed feature sets can significantly enhance the predictive power of analytical models, enabling more nuanced and accurate assessments of patient health and outcomes.

We introduce a systematic naming convention within the language to ensure that each feature generated through its advanced aggregation and temporal windowing capabilities receives a unique and descriptive name. This naming scheme is crucial for maintaining clarity and ease of reference when dealing with a potentially large number of features. The components of this naming scheme include:

- FeatureGroup Names: The base name derived from the FeatureGroup, which categorizes the health event or data type, e.g., "vitalsign".
- Filters: The specific aspect or measurement within the FeatureGroup, such as "spo2" for oxygen saturation levels.
- Join Expression and Temporal Aggregation Window: Indicators such as "l2" and "w1h" specify the temporal context of the feature, with "l2" denoting the second last window and "w1h" specifying a window period of 1 h.

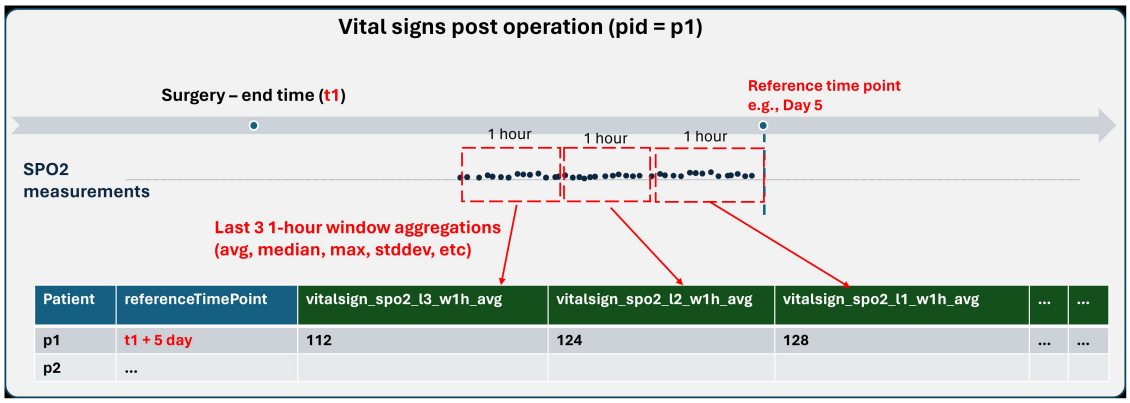

FIGURE 5
Population definition schema.

- Aggregation and Window Function Operators: The operation applied to the data, such as "avg" for average, clearly indicates the type of statistical measure calculated for the feature.

As illustrated in Figure 5, "vitalsign_spo2_l2_w1h_avg", illustrates how these elements combine to form a feature name that is both informative and concise. This feature name indicates that it represents the average oxygen saturation ("spo2") values from the "vitalsign" feature group, calculated over the second last 1-h window ("l2_w1h"). This structured approach to naming ensures that each feature's purpose and derivation are immediately apparent, facilitating easier analysis and interpretation of the data. It also aids in the automated processing of features, as the naming convention provides clear and consistent cues about the nature and temporal dynamics of the data encapsulated by each feature.

3 Results

3.1 Implementation: a feature repository for health data

We have developed a software, onfhir-feast, capable of processing declarative data preparation pipeline definitions in a high-performance distributed manner. This software enables two key functionalities: (1) Batch extraction of training or validation datasets from an integrated FHIR compliant data source and (2) Calculation of features for entities (e.g., patients) to support online prediction services integrated into the production environment as part of an AI-based decision support solution.

onfhir-feast is aligned with the emerging concept of a feature store, which is integral to AI pipelines. In the realm of machine learning, a feature store serves as a platform dedicated to managing and providing access to both historical and real-time feature data (21). It facilitates the creation of precise datasets at particular time points using historical feature data. Consistent with this definition, onfhir-feast manages Population, FeatureGroup, and FeatureSet definitions to provide a REST API for configuring or triggering a dataset extraction pipeline, enabling access to the dataset or real-time features for online predictions by leveraging these definitions.

A service built on an EHR system that provides data access via HL7 FHIR is typically optimized for patient-centric applications with user interfaces. However, when it comes to population-centric queries, especially in the context of AI pipelines, performance issues may arise due to the large volume of data involved. To address this challenge and prevent excessive workload on FHIR endpoints, onfhir-feast is designed akin to a health data warehouse. In this setup, only relevant data is synchronized periodically, typically at intervals such as every hour or every day, based on the Population and FeatureGroup definitions in the platform. Consequently, only FHIR resources updated since the last synchronization will be queried, resulting in a reduced workload on the system ensuring optimal performance and efficient utilization of resources.

In this synchronization process, the Population definitions take precedence. Entity identifiers of the resulting entities identified in each batch, based on these Population definitions, are stored in a specific population table within the configured time-series data repository. Subsequently, each FeatureGroup definition referenced in the activated dataset definitions is executed for the related population identified up to that point in time. This approach ensures that the platform only synchronizes the necessary data for the identified population.

The result sets of FeatureGroup executions are likewise stored in FeatureGroup-specific tables in the time-series data repository. Importantly, Population and FeatureGroup definitions can be reused across different dataset definitions. The platform manages this seamlessly to ensure that it never queries and processes the same FHIR resource more than once, thereby optimizing efficiency and resource utilization.

With this synchronization mechanism, onfhir-feast acts as a data warehouse similar to having an OMOP database populated with data pipelines mapping EHR system data. But in our case, users are more flexible to design their own tables, in other words Feature Groups, tailored to their use cases when needed.

Additionally, onfhir-feast offers an API to asynchronously trigger dataset extraction for preparing training or validation datasets. Users can choose to utilize all available data or specify a particular period, such as extracting a training dataset from data recorded in the previous year. Similarly, periodic dataset preparations can be scheduled and configured to support AI

model retraining scenarios. When such an extraction is triggered, the platform initiates the synchronization phase, which updates with the new data on the integrated system until the last synchronization point and populates the related tables in the time-series data repository. Subsequently, the relevant FeatureSet definition is executed on the loaded data from those tables for the identified entities within the population to prepare the dataset. The resulting datasets are then stored in the integrated "Offline Feature Repository."

Throughout the execution of this process, the client has the option to inquire about the status of the process via the REST API. Upon completion, automatically generated metadata of the dataset is stored and made accessible to the client. This metadata includes a list of features and outcome variables, along with their descriptions, basic statistics (such as the number of missing values, maximum, minimum, and average for numeric values, cardinalities, and value sets for nominal features, etc.).

The platform leverages Apache Spark (20), Akka (22) and Apache Software Foundation (23) frameworks to ensure reliable, fault-tolerant distributed processing for handling parallel FHIR queries during population identification and synchronization phases, as well as processing the result set for dataset preparation. It also provides extension mechanism to support the usage of different type of databases or persistency mechanisms as integrated repositories (such as time series data repository, Offline and Online Repository). Currently, PostgreSQL based repositories and file system-based repositories for storing data in Apache Parquet format are supported.

The platform can also serve as a component of a decision support application integrated with a trained AI model, tasked with preparing features for individual entities for online prediction. To facilitate this, a corresponding synchronous operation is provided as part of the REST API. During this process, the same dataset definitions, comprising the bundle of FeatureSet, Population, and FeatureGroup definitions, are executed. However, this time, they are applied to a single entity (e.g., a patient) to calculate the same feature list in a consistent manner. The process triggers the synchronization phase solely for that entity, considering data updated after the last synchronization time for the target population if the patient is included in the population. Subsequently, the FeatureSet definition is executed on the obtained data to prepare the features for online prediction.

3.2 Case studies

The proposed methodology and the implemented platform have been deployed and tested in 3 research projects supported within the EU Horizon 2020 and Horizon Europe frameworks namely, AICCELERATE (24), DataTools4Heart (25) and AI4HF (26) projects as part of several pilot studies.

Table 6 presents the list of pilot studies and use cases where the described methodology is followed and a range of dataset definitions are provided. In all these projects, the onfhir-feast platform is deployed locally on the data provider's data center to extract training and/or test datasets. For instance, in AICCELERATE, Pilot 2 involves utilizing datasets extracted from various data providers for cross-validation of AI models developed locally. Additionally, onfhir-feast serves as an integral component of the resulting solution for online prediction. In DataTools4Heart and AI4HF projects, onfhir-feast is incorporated into federated learning platforms to extract harmonized datasets from diverse data providers.

TABLE 6 Case studies where the methodology and the platform are used for preparing training/validation datasets.

Case study	# of data provider	# of variables
AICCELERATE		
Pilot 1—Patient Flow Management and Surgical Units	2	
1.1 Dataset for predicting complications after cardiac surgeries	1	916
1.2 Dataset for predicting length of stay (LoS) for patients after cardiac surgeries	1	583
1.3 Dataset for predicting duration of surgery, ICU stay and LoS before surgery	2	88
Pilot 2—Parkinson's Disease Digital Care Pathway	2	
2.1 Dataset for predicting progression to advanced Parkinson stage and predicting Mild Cognitive Impairment in Parkinson patients	2	402
Pilot 3—Palliative and chronic pediatric service delivery & patient workflow	3	
3.1 Dataset for clustering pediatric palliative patients into risk groups	1	117
3.2 Dataset for predicting tumor relapse after cancer treatment in pediatric patients	1	549
3.3. Dataset for predicting time needed for preparation (time to surgery) to a surgery	1	48
DataTools4Heart and AI4HF*		
Pilot 1—Medication prescription in patients with acute heart failure and chronic kidney disease or hyperkalaemia	9	
1.1 Dataset for analysing prescription patterns and clinical outcomes in terms of HF and CKD	9	604
Pilot 2—Risk score for acute HF in the emergency department	9	
2.1 Dataset for predicting on (HF/CV)-rehospitalization, cardiovascular event or mortality within 7-, 30-, 90-, 180-days, 1-, 3- and 5-year follow-up.	9	162
Pilot 3—Referral pathways for patients with HF	9	
3.1 Dataset for predicting the right specialty at the first time right to refer the patients for an in-hospital and general practitioners referral support model.	9	268

*DataTools4Heart and AI4HF projects are under development at the time of writing this manuscript. The numbers might change as the projects may evolve.

3.2.1 Example case study—predicting complications after cardiac surgeries

To illustrate the methodology and results achieved using the solution, we will now provide the details of the data preparation

TABLE 7 Feature group definitions and relation to CDM for the use case.

FHIR resource type	AICCELERATE CDM FHIR profile	Feature group definitions
Patient	AIC-Patient: Patient demographics - gender, birthdate → Set as mandatory	Patient_demographics (pid, gender, birthDate)
EpisodeOf Care	AIC-OperationEpisode: Surgical episode of care indicating the period from admission to discharge - type → Bind to a valueset for episode types to distinguish surgical episodes - diagnosis → Set as mandatory to identify pre-operative diagnosis for surgery	Episodes (pid, episodeId, time, endTime, preOpDiagnoses, comorbidtyDiagnoses)
Encounter	AIC-OperationEpisodeEncounter: Encounters related to surgical workflow. - type → Bind to a ValueSet with SNOMED-CT codes to distinguish ICU stays, ward stays, operation encounters	Icuorwardstay (pid, episodeId, encounterId, startTime, endTime, type, location, duration)
		surgeryEncounter (pid, episodeId, encounterId, startTime, endTime, category, priority, location, duration)
Condition	AIC-Condition: Diagnosis records for patients - code → Bind to ICD-10-CM value set	Condition (pid, encounterId, onsetDate, icd10Code)
Procedure	AIC-SurgeryPhaseDetails: Record to provide details of the main procedure performed in surgery. - category → Identify a fixed SNOMED-CT code to distinguish such records - code → Bind to ICD-10-PCS value set for surgery codes	Surgeries: Details of the surgery (pid, episodeId, encounterId, startTime, endTime, isMainSurgery, ccsCategory, mainProcedureCode, bodySite,duration, aristotleScore, stsScore, rachs1Score, extubationStatus, defibrillationStatus, minTemparature, cecTime, clampTime, arrestTime)
	AIC-ProcedureRelatedWithSurgicalWorkflow: Other related procedures performed in surgery—code → Bind to a ValueSet for interested procedure codes in SNOMED-CT for cardiac surgeries e.g. extracorporeal circulation procedure (cec), vascular clamp, extubating, defibrillation, etc.	
Medication Administration	AIC-MedicationAdministration: Record indicating an administered medication within surgical workflow in the hospital. - medication → Bind to ATC codes	Medications (pid, episodeId, time, atcCode, atcCategory, dose, doseUnit)
Observation	AIC-LabResultWithinSurgicalWorkflow: Record providing a related lab result - code → Bind to LOINC codes for lab results and provide a ValueSet to declare the interested lab tests for the use case	Lab (pid, episodeId, encounterId, time, code, value, unit, interpretation)
	Vitalsigns: A set FHIR standard profiles representing vital sign measurements e.g. body weight, temperature, SPO2, blood pressure, etc. - Fixed LOINC codes and units for each vital sign	Vitalsign (pid, time, code, value)
		bloodpressure (pid, time, systolic, diastolic)
AdverseEvent	AIC-ComplicationAfterOperation: Record indicating an adverse event after surgical operation. - event → Bind to a ValueSet including SNOMED-CT codes listing interested complications occur after cardiac surgeries including unexpected ICU admission	Complication (pid, episodeId, encounterId, time, code)

pipeline for one of the use cases within the AICCELERATE project's pilot 1 study. This particular use case revolves around predicting complications, specifically unexpected ICU admissions following cardiac surgeries and specific diagnostic procedures.

For this study, the target cohort is defined as the surgical episodes of patients who have undergone at least one cardiothoracic surgery or diagnostic procedure, such as cardiac catheterization or cardiac electrophysiology. These eligibility criteria are defined using a Population definition, which filters the FHIR EpisodeOfCare and Encounter resources based on the service type of encounter, utilizing the corresponding SNOMED-CT codes.

Within these episodes of care, which encompass the period from hospital admission to discharge, various types of encounters occur, including surgical encounters, ward stays, intensive care unit (ICU) stays, and pre-surgery visits. The study utilizes diagnostic data and basic patient demographic information from the pre-surgery phase. Additionally, it incorporates details of surgical or

diagnostic procedures performed, including specific interventions such as intubation, defibrillation, and hypothermic circulatory arrest, which may influence post-operative complications. Furthermore, intraoperative observations and assessments, such as minimum temperature and related surgery risk scores, are included in the study. For the post-operative phase, a specific set of lab results and frequent vital sign measurements obtained during ICU or ward stays are primarily utilized for prediction purposes. For instance, the data provider's dataset includes vital signs recorded at 5-min intervals for most of the time until discharge. To calculate outcome variables, a list of explicit complication data, including a range of post-operative complications and unexpected ICU admission events, is employed.

Table 7 illustrates the HL7 FHIR-based common data model and feature group definitions provided for the use case, along with their relationships. For instance, the FeatureGroup definitions "icuOrWardStay" and "surgeryEncounter" are dependent on the model described by AIC-OperationEpisodeEncounter, which customizes the FHIR Encounter resource model. On the other hand, the definition named "surgeries" relies on two profiles: one customizing the record representing the main surgical procedure and the other representing additional procedures performed in relation to cardiac surgeries. The table also details the primary customizations or restrictions applied to the standard resource model for each defined profile, as well as the parameters extracted from those records within the FeatureGroup definitions.

List of features and outcome variables that are designed for this use case in collaboration with clinicians and data scientists are provided in supplementary material as Supplementary Table 1. Related definitions are available open source at https://github.com/aiccelerate/data-extraction-suite/blob/main/definitions/pilot1-hsjd/.

Within this pilot study, the data is provided by the project partner Sant Joan de Deu hospital by getting data exports from corresponding EHR database tables in CSV format. For the transformation of data in CSV files into HL7 FHIR resources, the open source toFHIR platform (27, 28) is used as data integration platform, and onFHIR.io (29) is utilized as the secure health data repository.

The onfhir-feast tool, along with the data integration platform, is deployed on a server for demonstration and piloting purposes. Utilizing the toFHIR tool and corresponding mapping definitions, retrospective data provided in CSV format are transformed into FHIR resources compatible with the CDM for the specified use case. These FHIR resources are then stored in the onFHIR.io repository. Table 8 provides an overview of the data size, indicating the number of FHIR resources created as a result of the mappings. Following this, a batch dataset extraction job is initiated on onfhir-feast using the designed dataset preparation pipeline definition to create the dataset for training and testing of AI models. Moreover, the setup serves as an integral part of the prediction service served to healthcare professionals wrapping the trained AI model. The prediction service and UI component utilize onfhir-feast APIs to retrieve features for a patient within a surgical episode. Subsequently, this information is utilized for online prediction of complications for that patient.

We conducted a basic performance test using the same setup on a single personal computer (Lenovo ThinkPad) equipped with an 11th Gen Intel(R) Core (TM) i7-11800H processor running at 2.30GHz. The test was carried out within a controlled Docker environment featuring 8 CPU cores and 16GB of RAM. Initially, we executed the synchronization job independently, as the synchronization phase relies on the performance of the FHIR server to respond to queries. Subsequently, the dataset preparation job was performed, taking approximately 164 min to complete. The resulting dataset comprises 916 variables and 141,805 entries, covering 1,022 surgical episodes belonging to 906 patients. Furthermore, the metadata generation for this dataset, including basic statistics, required approximately 2 min. The API for retrieving features for a patient within a surgical episode at any chosen time demonstrated an average response time of 37 s.

TABLE 8 Number of FHIR resources created by mapping raw data and used in dataset creation.

FHIR resource	# of relevant resources	Details
Patient	906	
Episode of Care	1,022	
Encounter	4,581	Surgical encounters: 1,197 ICU stays: 783 Ward stays: 1294
Condition	2,310	
Medication administration	121,188	
Procedure	2,210	1,197 surgery 1,013 other procedures records
Observation	6,972,703	6,853,917 vital sign records 76,191 lab result records 1,108 others 41,487 blood pressure records
Adverse event	565	

4 Discussion

4.1 Principal findings

In this paper, we have introduced a declarative data preparation pipeline definition language designed to transparently outline each stage of the transformation process from EHR data to AI-ready feature sets. This framework ensures traceability by providing a clear depiction of the transformation and pre-processing operations applied to the data, from its retrieval from EHRs to its delivery to AI models for training.

Through our implementation in our pilot studies, we have demonstrated that, the framework is extensive enough for defining diverse set of features with different temporal and contextual criteria. In the realm of applying machine learning to electronic health record (EHR) data, researchers frequently resort to readily extracted, manually chosen obvious features due to the time-intensive nature of more thorough preprocessing methods (6). The proposed dataset definition language enables researchers to easily enumerate features with different representations and

temporal context using FHIR Path expressions, temporal windows, aggregation operators and window functions. Furthermore, as the important part of the definitions are parametrized, users have the chance to generate different versions of the datasets with different configurations which helps them to search for optimal or suitable solution with the underlying data. This capability is invaluable for researchers seeking the most effective or appropriate analytical models based on the available data, enabling a more dynamic and exploratory approach to data analysis.

Reproducibility poses a common challenge in AI research, with healthcare presenting a particularly pronounced instance of this issue. The limited availability of publicly accessible medical datasets serves as one indication of this challenge (30). While promoting increased data sharing is crucial, establishing reusable and standardized definitions for key concepts such as target cohorts, phenotypes, and datasets, as advocated in this article, can significantly enhance reproducibility in AI research. Encouraging researchers to share such definitions for their methodologies enables others to apply the same processes to different datasets, facilitating result comparison and broader applicability.

Our approach facilitates reproducibility across diverse data sources, which is essential for federated analysis of fragmented datasets. Achieving interoperability among datasets is a crucial requirement for federated machine learning applications, and our solution offers a transparent and traceable pipeline to accomplish this goal (31). Additionally, it enables validation for robustness, bias, and fairness across different sites, thereby enhancing the reliability and integrity of AI models deployed in healthcare settings. The framework and its implementation serve as an implementation guideline for EHDS vision, tackling how data sets across different sites in Europe can be harmonized and aggregated for secondary use purposes while also ensuring traceability and end-to-end transparency fulfilling the requirements of AI-Act.

4.2 Limitations and future work

Currently, the definition of Target Population, Feature Group, and Feature Set necessitates technical proficiency in crafting FHIR query and FHIRPath expressions. To enhance user accessibility and usability, we intend to augment the onfhir-feast implementation with a graphical user interface. This interface will empower users to define FHIR query and FHIRPath expressions through visual expression builders, streamlining the process and reducing the reliance on technical skills.

The scope of the pipeline and implementation is limited to tabular datasets production. As foundational models are trained on raw data for generic purposes, researchers may prefer to provide directly the FHIR formatted data rather than a dataset tailored for a specific AI use case. However, still there is an important use case for generative AI where this pipeline can be useful. Recently, synthetic data generation for privacy preserving data sharing is one of the hot topics in healthcare AI. Our pipeline can be part of such setups where a common dataset definition can be used in different healthcare settings to extract harmonized datasets locally and then apply generative AI to create synthetic datasets that maintains the statistical properties of original datasets. Then these synthetic datasets can be shared, combined and used in model training, and development without exposing sensitive patient information.

Furthermore, we aim to expand the capabilities of onfhir-feast by providing visual tools to data scientists. These tools will facilitate querying and exploration of source data during target population selection and feature set preparation. By offering visualizations, data scientists can better assess the adequacy of the datasets provided by data sources in addressing the research question at hand, enhancing overall data exploration and analysis. The existing implementation already includes basic statistics, such as the number of missing values, maximum, minimum, and average for numeric values, as well as cardinalities, within the feature set documentation. Our objective is to expand the underlying language and enhance the onfhir-feast implementation to enable querying additional statistics about datasets.

We plan to leverage this extension for two primary purposes. Firstly, we aim to utilize it for constructing a metadata catalog, which will serve as a comprehensive repository of dataset statistics and characteristics. Secondly, we intend to employ it for developing data set exploration user interfaces tailored for data analysts. These interfaces will facilitate the assessment of data quality across various dimensions, including conformance, completeness, and plausibility, enabling users to evaluate the quality of datasets effectively.

5 Conclusion

In summary, the proposed methodology and models offer significant contributions to the ML research community in healthcare by establishing standardized, transparent, and technology-agnostic dataset definitions. These definitions not only characterize the datasets themselves but also delineate the procedures for compiling them from Electronic Health Record (EHR) systems via standard FHIR interfaces. This innovative approach represents a crucial step towards establishing best practices for data harmonization. By creating reusable, transparent, and shareable dataset definitions, it addresses a critical need in setting up federated data sharing environments for the secondary use of EHR data, such as the European Health Data Spaces initiative. By promoting interoperability and standardization, these methodologies pave the way for more efficient and effective ML research in healthcare, ultimately leading to improved patient outcomes and advancements in medical knowledge.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/aiccelerate/common-data-model, https://github.com/aiccelerate/data-extraction-suite.

Author contributions

TN: Writing—original draft, Software, Methodology, Formal analysis, Conceptualization. AS: Writing—original draft, Software, Methodology, Conceptualization. SG: Writing—review and editing,

Software, Methodology, Conceptualization. CH: Writing—review and editing, Resources, Investigation. PG-C: Writing—review and editing, Resources, Investigation. AM: Writing—review and editing, Resources, Investigation. AE: Writing—review and editing, Resources, Investigation. GE: Writing—original draft, Methodology, Conceptualization.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of the article. The work presented in this manuscript was funded by the European Union's Horizon 2020 Research and Innovation Program under grant agreement no. 101016902 and the European Union's Horizon Europe Research and Innovation Program under grant agreements nos. 101057849 and 101080430.

Acknowledgments

We would like to acknowledge the support of the AICCELERATE, DataTools4Heart, and AI4HF consortiums.

Conflict of interest

TN, AS, SG, and GE were employed by the company Software Research and Development Consulting. The study presented in this manuscript is conducted in the scope of a research study.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2024.1393123/full#supplementary-material

References

1. European Commission. *Proposal for a regulation of the European Parliament and of the council laying down harmonised rules on artificial intelligence (Artificial intelligence act) SND smending certain union legislative acts COM/2021/206 final*. Brussels: European Commission (2024).

2. Mora-Cantallops M, Sánchez-Alonso S, García-Barriocanal E, Sicilia M. Traceability for trustworthy AI: A review of models and tools. *Big Data Cogn Comput.* (2021) 5:20. doi: 10.3390/bdcc5020020

3. Health Level 7 [HL7]. *Fat healthcare interoperability resources (FHIR)*. (2024). Available online at: https://www.hl7.org/fhir/ (accessed February 21, 2024).

4. Directorate-General for Health and Food Safety. *Proposal for a regulation - The European health data space COM(2022) 197/2*. Brussels: Directorate-General for Health and Food Safety (2022).

5. Williams E, Kienast M, Medawar E, Reinelt J, Merola A, Klopfenstein S, et al. A standardized clinical data harmonization pipeline for scalable AI application deployment (FHIR-DHP): Validation and usability study. *JMIR Med Inform.* (2023) 11:847. doi: 10.2196/43847

6. Tang S, Davarmanesh P, Song Y, Koutra D, Sjoding M, Wiens J. Democratizing EHR analyses with FIDDLE: A flexible data-driven preprocessing pipeline for structured clinical data. *J Am Med Inform Assoc.* (2020) 27:1921–34. doi: 10.1093/JAMIA/OCAA139

7. Xie F, Yuan H, Ning Y, Ong M, Feng M, Hsu W, et al. Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies. *J Biomed Inform.* (2022) 126:103980. doi: 10.1016/J.JBI.2021.103980

8. The Observational Health Data Sciences and Informatics [OHDSI]. *Program*. (2024). Available online at: https://www.ohdsi.org/ (accessed February 28, 2024).

9. Observational Health Data Sciences and Informatics [OHDSI]. *OMOP common data model*. (2024). Available online at: https://ohdsi.github.io/CommonDataModel/ (accessed February 22, 2024).

10. Belenkaya R, Gurley M, Golozar A, Dymshyts D, Miller R, Williams A, et al. Extending the OMOP common data model and standardized vocabularies to support observational cancer research. *JCO Clin Cancer Inform.* (2021) 5:12–20. doi: 10.1200/CCI.20.00079

11. Park W, Jeon K, Schmidt T, Kondylakis H, Alkasab T, Dewey B, et al. Development of medical imaging data standardization for imaging-based observational research: OMOP common data model extension. *J Imaging Inform Med.* (2024) 37:899–908. doi: 10.1007/s10278-024-00982-6

12. Liu D, Sahu R, Ignatov V, Gottlieb D, Mandl K. High performance computing on flat FHIR Files created with the new SMART/HL7 bulk data access standard. *AMIA Annu Symp Proc.* (2019) 2019:592–6.

13. Oehm J, Storck M, Fechner M, Brix T, Yildirim K, Dugas M. FhirExtinguisher: A FHIR resource flattening tool using FHIRPath. *Public Health Inform Proc MIE.* (2021) 2021:1112–3. doi: 10.3233/SHTI210369

14. Grimes J, Szul P, Metke-Jimenez A, Lawley M, Loi K. Pathling: Analytics on FHIR. *J Biomed Semant.* (2022) 13:1–19. doi: 10.1186/S13326-022-00277-1/FIGURES/7

15. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-scale machine learning on heterogeneous systems. *arXiv* [Preprint] (2015). arXiv:1603.04467.

16. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An imperative style, high-performance deep learning library. *Adv Neural Inform Process Syst.* (2019) 32:259.

17. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in python. *J Machine Learn Res.* (2011) 12:2825–30.

18. Health Level 7 [HL7]. *FHIRPath*. (2024). Available online at: https://hl7.org/fhirpath/ (accessed February 21, 2024).

19. Observational Health Data Sciences and Informatics [OHDSI]. *Chapter 10 defining cohorts - The book of OHDSI*. (2024). Available online at: https://ohdsi.github.io/TheBookOfOhdsi/Cohorts.html (accessed February 26, 2024).

20. Apache Software Foundation. *Apache SparkTM - unified engine for large-scale data analytics*. (2024). Available online at: https://spark.apache.org/ (accessed February 27, 2024).

21. Feature Store For Ml. *PRIYA*. (2024). Available online at: https://www.featurestore.org/ (accessed February 27, 2024)

22. Akka. *Self managed frameworks and runtimes for event-driven micro-services and APIs*. (2024). Available online at: https://akka.io/ (accessed February 28, 2024).

23. Apache Software Foundation. *Apache KafkaTM, open-source distributed event streaming platform*. (2024). Available online at: https://kafka.apache.org/ (accessed February 28, 2024).

24. Aiccelerate Project. *AI accelerator – a smart hospital care pathway engine (funded by the European Union's horizon 2020 framework under grant agreement no. 101016902)*. (2024). Available online at: https://aiccelerate.eu/ (accessed February 28, 2024).

25. DataTools4Heart. *A European health data toolbox for enhancing cardiology data interoperability, reusability and privacy (funded by the European union's horizon europe framework under grant agreement no. 101057849)*. (2024). Available online at: https://www.datatools4heart.eu/ (accessed February 28, 2024).

26. AI4HF. *Trustworthy artificial intelligence for personalised risk assessment in chronic heart failure (funded by the European Union's horizon europe framework under grant agreement no. 101080430)*. (2024). Available online at: https://www.ai4hf.com/ (accessed February 28, 2024).

27. toFHIR. *A high-performant and easy-to-use ETL (Extract, transform, load) tool to transform existing health datasets from various types of sources to HL7 FHIR*. (2024). Available online at: https://onfhir.io/tofhir/ (accessed February 28, 2024).

28. Sinaci A, Gencturk M, Teoman H, Erturkmen G, Alvarez-Romero C, Martinez-Garcia A, et al. A data transformation methodology to create findable, accessible, interoperable, and reusable health data: Software design, development, and evaluation study. *J Med Internet Res.* (2023) 25:822. doi: 10.2196/42822

29. onFHIR.io. *HL7 FHIR® based secure data repository*. (2024). Available online at: https://onfhir.io/ (accessed February 28, 2024).

30. Sohn E. The reproducibility issues that haunt health-care AI. *Nature*. (2023) 613:402–3. doi: 10.1038/D41586-023-00023-2

31. Sinaci A, Gencturk M, Alvarez-Romero C, Banu G, Erturkmen L, Martinez-Garcia A, et al. Privacy-preserving federated machine learning on FAIR health data: A real-world application. *Comput Struct Biotechnol J.* (2024) 24:136–45. doi: 10.1016/J.CSBJ.2024.02.014

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 27 August 2024
DOI 10.3389/fmed.2024.1396459

OPEN ACCESS

EDITED BY
Cecilia Ana Suarez,
National Scientific and Technical Research Council (CONICET), Argentina

REVIEWED BY
Marcelo Cardoso Dos Reis Melo,
Auburn University, United States
Tamás Micsik,
Semmelweis University, Hungary
Michael Liebman,
IPQ Analytics, United States

*CORRESPONDENCE
Julia Gehrmann
julia.gehrmann1@uk-koeln.de

RECEIVED 05 March 2024
ACCEPTED 06 August 2024
PUBLISHED 27 August 2024

CITATION
Gehrmann J, Soenarto DJ, Hidayat K, Beyer M, Quakulinski L, Alkarkoukly S, Berressem S, Gundert A, Butler M, Grönke A, Lennartz S, Persigehl T, Zander T and Beyan O (2024) Seeing the primary tumor because of all the trees: Cancer type prediction on low-dimensional data. *Front. Med.* 11:1396459. doi: 10.3389/fmed.2024.1396459

Seeing the primary tumor because of all the trees: Cancer type prediction on low-dimensional data

Julia Gehrmann[1]*, Devina Johanna Soenarto[1], Kevin Hidayat[1], Maria Beyer[1], Lars Quakulinski[1], Samer Alkarkoukly[1,2], Scarlett Berressem[3,4], Anna Gundert[3,4], Michael Butler[3,4], Ana Grönke[2], Simon Lennartz[5], Thorsten Persigehl[5], Thomas Zander[3,4] and Oya Beyan[1,2,6]

[1]Institute for Biomedical Informatics, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany, [2]Medical Data Integration Center (MeDIC), Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany, [3]Department of Internal Medicine, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany, [4]Center for Integrated Oncology Aachen Bonn Cologne Duesseldorf (CIO ABCD), Aachen, Germany, [5]Institute for Diagnostic and Interventional Radiology, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany, [6]Department of Data Science and Artificial Intelligence, Fraunhofer FIT, Sankt Augustin, Germany

The Cancer of Unknown Primary (CUP) syndrome is characterized by identifiable metastases while the primary tumor remains hidden. In recent years, various data-driven approaches have been suggested to predict the location of the primary tumor (LOP) in CUP patients promising improved diagnosis and outcome. These LOP prediction approaches use high-dimensional input data like images or genetic data. However, leveraging such data is challenging, resource-intensive and therefore a potential translational barrier. Instead of using high-dimensional data, we analyzed the LOP prediction performance of low-dimensional data from routine medical care. With our findings, we show that such low-dimensional routine clinical information suffices as input data for tree-based LOP prediction models. The best model reached a mean Accuracy of 94% and a mean Matthews correlation coefficient (MCC) score of 0.92 in 10-fold nested cross-validation (NCV) when distinguishing four types of cancer. When considering eight types of cancer, this model achieved a mean Accuracy of 85% and a mean MCC score of 0.81. This is comparable to the performance achieved by approaches using high-dimensional input data. Additionally, the distribution pattern of metastases appears to be important information in predicting the LOP.

KEYWORDS

oncology, Cancer of Unknown Primary, prediction, real-world data, classification

1 Introduction

The "Cancer of Unknown Primary" syndrome (CUP) is diagnosed if only metastases but no primary tumor can be found (1). Extensive examination and molecular analyzes without the support of AI currently enable predicting the location of the primary tumor (LOP) for 10–20% of CUP patients with an accuracy of 85–90% (2, 3). For these patients, an LOP-specific treatment can be chosen which significantly improves their prognosis.

Historically, about 3–5% of all cancer cases were diagnosed as CUP (4). Due to advances in diagnostics, this rate could be reduced to 1–2% in general, but it is still higher for patients living in areas with rudimentary clinical care (1, 5, 6). Additionally, the CUP syndrome is still among the 10 most common reasons for cancer-related deaths globally (1). Thus, further advances in LOP prediction are needed to improve the prognosis for CUP patients.

AI-driven data analysis can be a key component in achieving this and some promising approaches have already been developed (1, 7–14). They are described in Supplementary material. A major drawback of these related approaches is their dependency on high-dimensional input data measuring the transcriptome, the mutation pattern, or epigenetic features of the metastases. This data is not generated for cancer patients by default. Hence, the approaches introduce additional costs representing a potential translational barrier for clinical practice. In 2021, Lu et al. (15) have shown that the additional costs to generate transcriptomic, genetic, or epigenetic data might not be needed for most CUP cases. Although only using the sex of the patient and whole slide images (WSI) from pathological examinations as input data, they achieve comparably high classification performance in LOP prediction with a convolutional neural network (CNN) approach (15).

Motivated by the success of Lu et al. (15), we examined whether LOP prediction also works for even lower-dimensional data, i.e., to dispense with image files and instead only use a small number of structured clinical features as input data. Since such data is far less dimensional than genome data or images, the complexity of the task is reduced and the decision-making process becomes more comprehensible.

In 10-fold nested cross-validation (NCV) we examined the LOP prediction performance of a random forest (RF) classifier and a gradient boosted trees (GBT) classifier on three different input feature sets compiled from oncological real-world data (RWD) of non-CUP patients at University Hospital Cologne (UHC). An extensive extract transform load (ETL) process accompanied by interdisciplinary decisions ensured highest possible data quality. Comparing our results to the LOP prediction performance achieved by high-dimensional approaches, shows that our tree-based approach on input features such as the age, sex, histological specifications, lab results, and the distribution pattern of metastases can achieve classification performances as high as the complex approaches while being more transparent, accessible, affordable, and explainable. Especially, the distribution pattern of metastases proved to be a valuable source of information for well-performing classification.

2 Materials and methods

2.1 Data curation

In total, we compiled six datasets from clinical systems of UHC as shown in Figure 1. We included cancer cases of adult patients diagnosed with Lung, Pancreas, Kidney, Liver, Breast, Colorectal, Ears-Nose-Throat, or Upper GI cancer between 01.01.2000 and 30.06.2021. Patients having several cancer diagnoses within 5 years were excluded from the dataset. For each included cancer case, we compiled the age at diagnosis, the sex, histological specification, lab results, and the metastatic burden according to RECIST v1.1 (16). The histological specifications comprised the tumor grading as well as indicators for infestation of lymph nodes (N-value), lymph vessels (L-value) and veins (V-value). The lab results comprised the amount of leukocytes, C-reactive protein (CRP), Hemoglobin (HB), Carbohydrate Antigen 19–9 (CA 19–9), and Carcinoembryonic Antigen (CEA) in the blood. CA 19–9 and CEA are tumor markers (TM), i.e., proteins whose abundance can indicate certain types of tumors. The RECIST evaluations were translated to organ-specific Tumor Burden Scores (TBS) spanning from 0 (no infestation) to 4 (significant infestation). All TBS taken together represent the metastatic distribution pattern by indicating the tumor burden in individual organs. Based on the frequency of missing values for the individual features we created three feature sets:

1 "Core features" containing the age, the sex, histological specifications, leukocytes, CRP, and HB (frequency of missing values below 35%)
2 "Core features and TM" containing the core features and the TM CA 19–9 and CEA (frequency of missing values 77 and 69%, respectively).
3 "All features" containing the core features, the TM, and the organ-specific TBS, which indicate the distribution pattern of metastases (frequency of missing values 98%).

Due to the low availability of the TBS, we only included those cases in the "all features" dataset for which the TBS were available. As a result, four of the eight LOP classes were underrepresented, so we decided to create a four-class version of each dataset only containing the classes that were still well represented: Lung, Upper GI, Pancreas, and Liver. This resulted in a total of six datasets. Missing values were imputed in all six datasets using the R package "mice" in version 3.15.0 for Multiple Imputation by Chained Equations (MICE) (17–20). Eventually, the datasets were anonymized using the software tool ARX, which can anonymize structured data according to a variety of data privacy models (21). In particular, we deleted identifying features and established 5-anonymity with respect to the quasi-identifying features age and sex. This means that we generalized the age to age groups such that at least five patients share the same combination of age and sex. Additionally, ARX suppressed too specific cancer cases that would require huge age groups to achieve 5-anonymity. The sizes of the resulting datasets are depicted in Table 1. More details on the data curation process can be found in Supplementary material.

2.2 LOP prediction

We implemented LOP prediction by classifying the patients according to their type of cancer using a supervised ML approach. In particular, we applied a RF classifier and a GBT classifier on each of the six compiled datasets resulting in 12 classification runs in total. RF and GBT are tree-based ML methods, which have shown good performance in LOP prediction in related work (7–12). An additional advantage of these methods is their inherent explainability, which is a key requirement for AI-based decision support in medical contexts (22, 23). As supervised ML methods, both RF and GBT need class labels throughout model training. In our case, these class labels is the LOP. Therefore, we trained and evaluated the models on medical

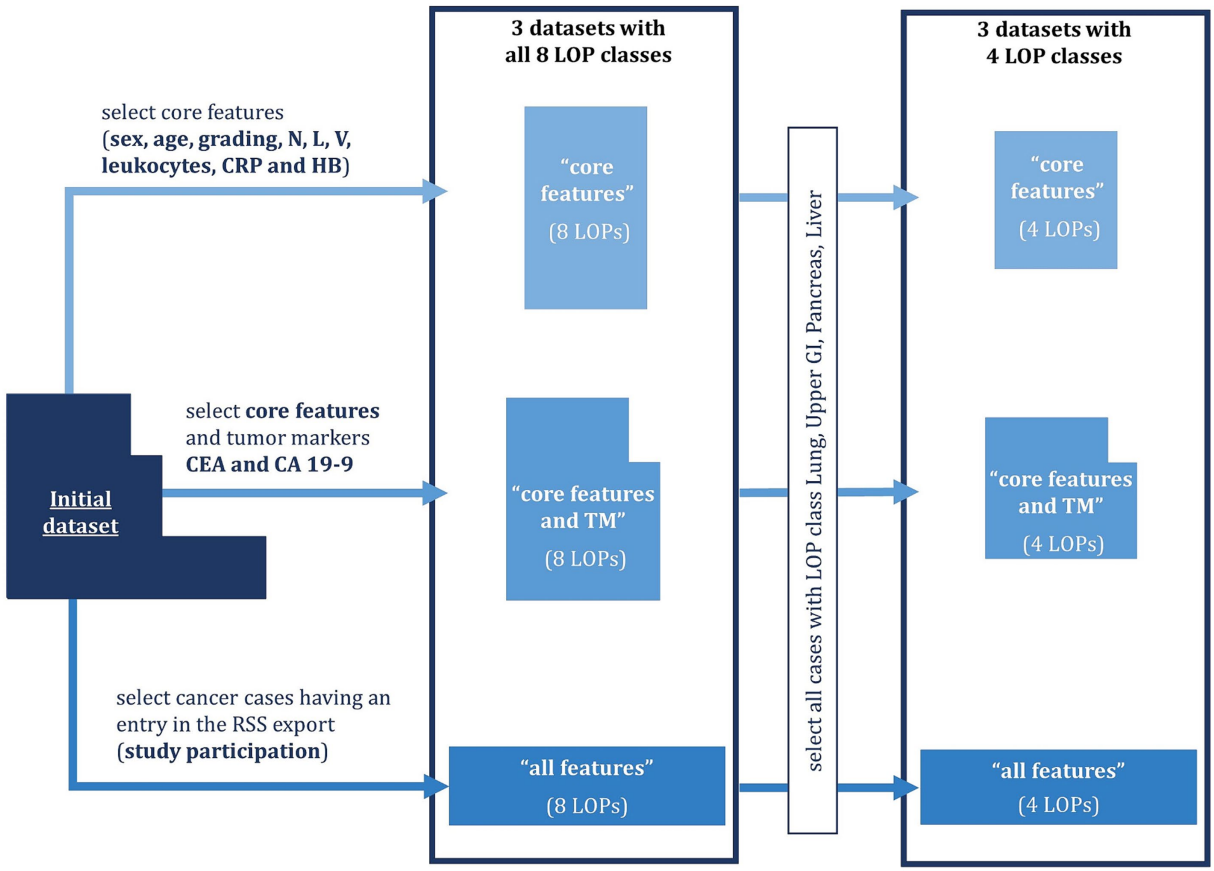

FIGURE 1
Dataset compilation process. Six datasets (nested boxes) were created which differ in the number of features (width of the boxes) and the number of cases (height of the boxes). The features always contain the core features listed above and optionally tumor markers (TM) as well as organ-specific Tumor Burden Scores (TBS). The number of cases is partly restricted by requiring the case to have an entry in the radiological study system (RSS) or the location of the primary tumor (LOP) to either be Lung, Upper GI, Pancreas, or Liver.

TABLE 1 Number of cancer cases in the three datasets "core features," "core features and TM," and "all features" before (blue) and after (green) anonymization when including all eight classes vs. only including four classes Lung, Upper GI, Pancreas, and Liver.

Number of classes	Cases in "core features" dataset (nine features)		Cases in "core features and TM" dataset (11 features)		Cases in "all features" dataset (30 features)	
	Before anonymization	After anonymization	Before anonymization	After anonymization	Before anonymization	After anonymization
8	13,861	13,764	13,861	13,712	336	328
4	4,295	4,271	4,295	4,271	299	297

RWD of cancer patients with known cancer types, i.e., on data of non-CUP patients.

We comprehensively evaluated the performance of the classifiers considering several performance metrics: accuracy, Precision, Recall, F1-score, and MCC score. This performance estimation was combined with 10-fold NCV to decrease the influence of randomness and to determine optimal hyperparameter values for the classifiers from a pre-defined parameter grid. For the RF, the parameter grid contained the values 5, 10, 20, 35, and 50 for the number of decision trees (DTs), the values 3, 5, 7, and 10 for the maximal depth of the DTs, the two entropy measures Gini-Index and Cross-Entropy, as well as training with and without bootstrapping. The parameter grid of the GBT contained the values 0.1, 0.2, and 0.5 for the learning rates and the values 3, 5, 7, and 10 for the maximal depth of the DTs in the GBT sequence. The optimal set of hyperparameter values was chosen by a grid search approach maximizing the MCC score of the classification. We have opted for an optimization according to MCC score due to the high class-imbalance in our datasets and the low sensitivity of the MCC score for such class-imbalances (24). The 10-fold NCV was stratified in order to maintain the class distribution in the test and training dataset. Eventually, we determined the importance of each input feature for LOP prediction based on the average decrease in class entropy over all splits in which the respective feature was the separating feature (25, 26). To enable a systematic comparison of individual features, we determined four groups of features according

to their feature importance (FI) for each classification setting, individually: low, medium low, medium high and high FI. The groups were defined based on the quartiles of the FI. More details on the methods and their implementation can be found in Supplementary material.

3 Results

3.1 LOP prediction performance

We applied 10-fold NCV to evaluate the classification performance of the tree-based ML algorithms on the six datasets. Figure 2 shows the mean performance values across the 10 NCV iterations for all examined classification settings, i.e., combinations of algorithm and dataset. In terms of average Accuracy, the performance spanned from 55.8 to 84.5% in the eight-class classification task and from 57.2 to 93.6% in the four-class classification task. The average MCC scores ranged from 0.42 to 0.81 when distinguishing eight LOP classes and from 0.34 to 0.92 when assigning the cancer cases to one of four LOP classes. The achieved performance values were stable across the 10 NCV iterations, which can be seen from the small standard deviations.

For both classification tasks (four and eight LOP classes), we observed that the values of all performance metrics increased with increasing numbers of features. The provision of the TBS ("all features") led to a particular increase in performance for both ML methods. Moreover, the GBT algorithm exhibited slightly higher performance scores than the RF in almost all combinations of metric and dataset. The only exceptions were the MCC score of the RF on the eight-class "all features" dataset and the Precision of the RF on the "core features" and "core features and TM" datasets. In these settings, the scores were slightly higher for the RF than for the GBT. Another striking observation was that Precision is usually higher than Recall in all classification runs. The only exceptions were the two classifiers trained to discriminate eight LOP classes based on "all features." These classifiers exhibited a slightly higher Recall than Precision. In general, including the TBS in the input dataset increased both Precision and Recall while decreasing their difference. Thus, including the TBS resulted in a more balanced decision making.

Considering individual combinations of datasets and ML algorithms, we observed that the Accuracy, Precision, Recall, and F1-score are higher in four-class classification than in the eight-class setting. The difference is particularly high on "all features," i.e., when the TBS are provided. In contrast to the simpler metrics, the MCC score is usually higher in the eight-class classification setting. Only the classification runs on "all features" achieve a higher MCC score when distinguishing between four instead of eight classes.

3.2 Feature importance

For each classification run, i.e., combination of feature set and ML algorithm, we determined the FI of individual features in every NCV iteration. The means of the FI values across NCV iterations are visualized in Figure 3 per feature and classification run.

Particularly striking is the overall decreased importance of the feature sex when not considering the LOP classes Breast, Colorectal, ENT, and Kidney. In this four-class setting, the FI is transferred from sex to all other features having a decent to high importance in the eight-class setting. The gain in FI is particularly high for the features CRP, leukocytes and the N-value. A medium gain can be observed for the other features contained in the feature set "core and TM." The highest increase in FI among the TBS, which indicate the distribution pattern of metastases, can be seen for the TBS of Pancreas, Lung, Esophagus, and Liver. These TBS features already had a rather high FI in the eight-class setting. The TBS for Brain, Stomach, Bones, and the group of Other Organs were subject to a medium increase in FI.

To enable a more systematic comparison of the FI in the different classification runs, we assigned the features to one of four groups: low, medium low, medium high and high FI. This grouping is based on the first, second, and third quartile of the mean FI value for each classification run and depicted in Table 2.

CRP, leukocytes, HB, the N-value and the age exhibited a high or medium high importance in the majority of classification runs. The feature sex was categorized diversely. When the TBS were not provided, the eight-class classification runs assigned a high importance to the sex while it was of low or medium low importance for almost all four-class classification runs. All approaches on the "all features" dataset categorized the sex to have a medium low importance. The grading had a medium high FI in the RF-based classification runs on the "all features" datasets. All other classification runs assigned a lower importance to it (medium low or low). The L- and the V-value both are categorized to have rather low FI. The TM CA 19–9 and CEA were assigned a rather high importance. Out of eight classification runs using the TM as input features, six categorized CEA to have a medium high and CA 19–9 to have high or medium high FI. The two eight-class runs on the "core and TM" dataset considered CA 19–9 to have a medium low importance and CEA to have a medium low or low importance. In general, CA 19–9 received higher FI scores than CEA.

The TBS were only provided as input features in four out of 10 classifications. In these four classifications, the group of highly important features mainly consists of TBS features. In particular, the TBS for Lymphnodes, Esophagus, Pancreas, Lung and Liver were assigned a high importance for LOP prediction. Only four non-TBS features were categorized as high importance features in a classification run on "all features": the N-value, CA 19–9, CRP, and the age.

A rather high importance was assigned to the TBS for Brain and Stomach while the TBS for Bladder received diverse categorizations. In the four-class classification runs, the TBS Bladder exhibited low importance for the LOP prediction while it had a medium high FI in the eight-class setting. The TBS for Kidney, Adrenal Gland, and Other Organs were assigned low or medium low FI in all four runs on "all features." The TBS for Intestine exhibited a medium low importance once. In all other classification runs, it had low importance. In two classifications it even achieved a mean FI of not more than 0. The TBS for Heart, Omentum, Skin, Spleen, Mamma, and Thyroid Gland belong to the features with low importance in all classification runs on "all features." It is noticeable that, with the exception of TBS Spleen, all these TBS have an average FI value of 0 in all four classifications. This means that the values of these TBS were not considered in any classification.

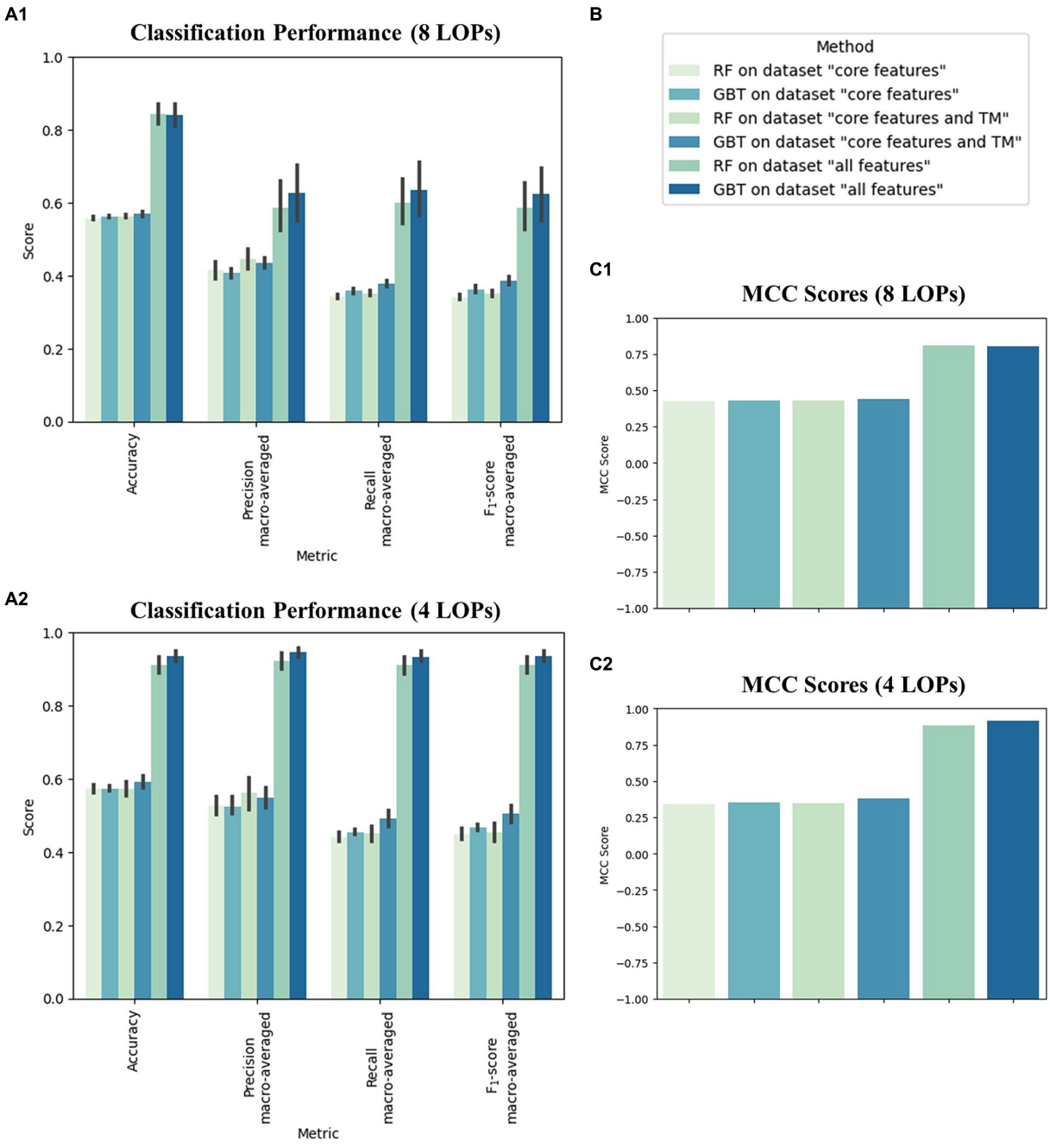

FIGURE 2
Performance of the two applied machine learning methods random forest (RF) and gradient boosted trees (GBT) on the three feature sets "core features," "core features and TM" and "all features" in predicting the location of the primary tumor (LOP). **(A)** Average classification performance of the six classifiers across the 10 iterations of the nested cross validation (NCV), measured by Accuracy, macro-averaged Precision, macro-averaged Recall, and macro-averaged F1-score all spanning from 0 to 1 **(B)** Legend displaying assignment of colors to classification settings. **(C)** Average classification performance across the 10 iterations of the NCV measured in terms of the Matthews correlation coefficient (MCC) spanning from –1 to 1. Sections **(A1)** and **(C1)** depict the performance in the eight-class classification task (Breast, Colorectal, ENT, Kidney, Liver, Lung, Pancreas, Upper GI). Sections **(A2)** and **(C2)** depict the performance in the four-class classification task (Lung, Upper GI, Pancreas, Liver).

4 Discussion

4.1 LOP prediction performance on low-dimensional data

We observed a generally higher LOP prediction performance when considering four instead of eight LOP classes. This was particularly true for the rather simple performance metrics Accuracy, Precision, Recall, and F1-Score. For these metrics, the baseline performance value of a predictor assigning classes randomly is higher with fewer classes. So, we explain the lower values of these metrics in the eight-class setting by the larger number of classes. The MCC score of the LOP prediction is slightly higher in the eight-class setting if the prediction model is provided with the "core features"

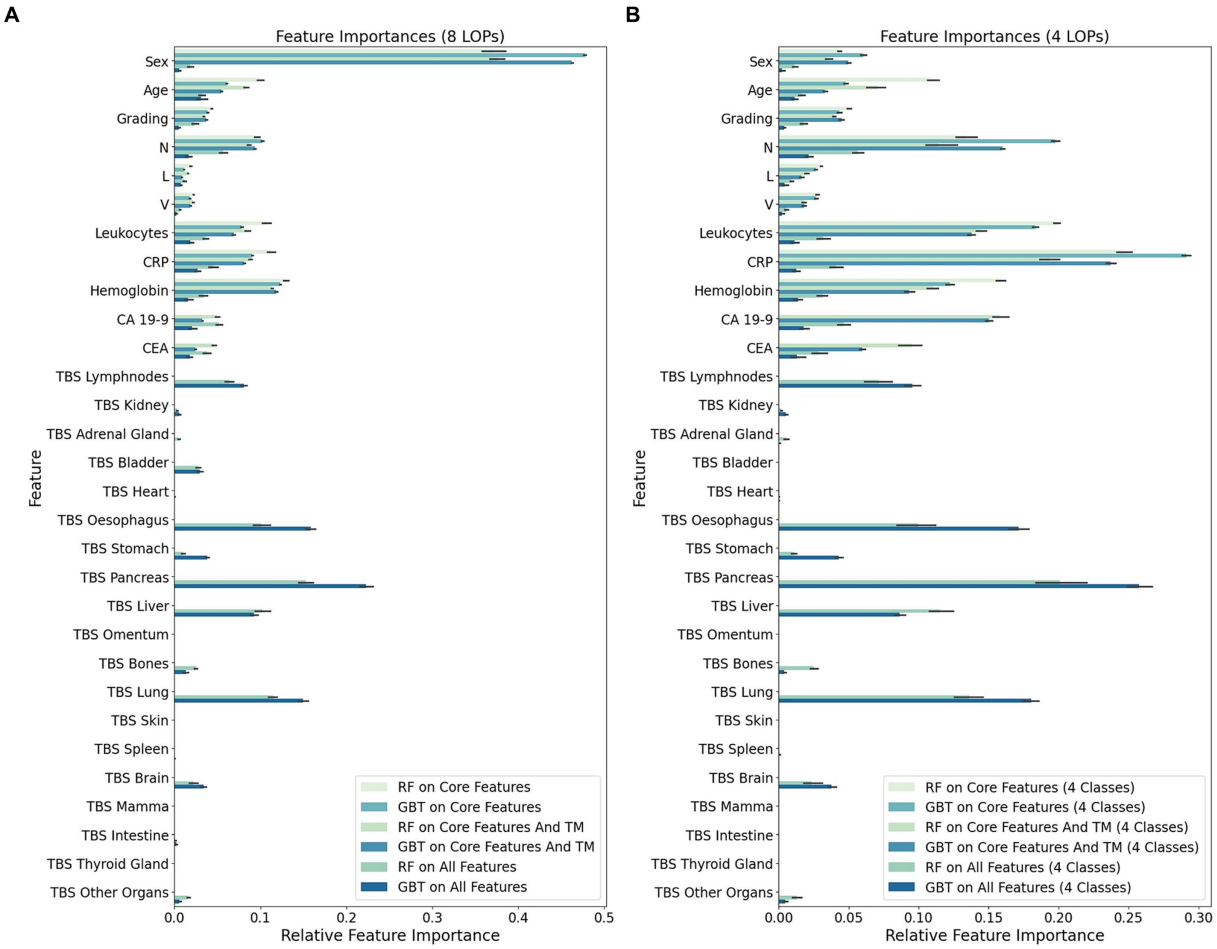

FIGURE 3
Mean feature importance (FI) for individual input features in 10-fold nested Cross-Validation (NCV). The FI values were determined in every NCV iteration for each combination of machine learning method [random forest (RF) or gradient boosted trees (GBT)] and feature set ["core features" (only first 9 features), "core features and TM" (only first 11 features), or "all features"]. This barplot visualizes the mean FI value of the individual features across 10 NCV iterations. **(A)** Importance of individual features in the eight-class classification task of assigning cancer cases to one of eight LOP classes (Breast, Colorectal, ENT, Kidney, Liver, Lung, Pancreas, Upper GI). **(B)** Importance of individual features in the four-class classification task of assigning cancer cases to one of four LOP classes (Lung, Upper GI, Pancreas, Liver).

or the "core features and TM." Since the difference in MCC scores in the two settings is very small, we consider this to be a random phenomenon. It is made possible by the restricted information content of the "core features" and the "core features and TM" feature sets. On the feature set "all features," the MCC score follows the same pattern as the other metrics, i.e., exhibits higher values in the four-class setting. Strikingly, the performance boost achieved by reducing to four classes was especially high on "all features." This can be explained by the fact that the eight-class version of the "all features" dataset contains four underrepresented classes that significantly degrade performance. This hypothesis is strengthened by the clinical observation, that the four cancer entities that are not included in the four-class setting (Kidney, Breast, Colorectal, ENT) do substantially differ from each other and the other four entities. This would mean that LOP prediction is clinically easier in our eight-class setting. Breast cancer is nearly exclusively seen in women and kidney cancer has a very different behavior. Therefore, the reduced performance in the eight-class setting will be mainly due to the mentioned class imbalance.

Regarding the ML methods used, we observed that the GBT method outperforms the RF. On one of the six datasets all measured performance values are higher for the GBT method (four-class "all features"). On the other five datasets, the majority of measured performance metrics is higher for the GBT method. This observation coincides with findings in ML research. These findings attribute a higher performance to the GBT method, in general, while the performance of the RF can be similarly high or even higher (27, 28).

Overall, we see that the LOP prediction performance on low-dimensional data is at the same level as the performance of related approaches using high-dimensional data (7–15). In our setting, this high performance (Accuracy: 93.6%, MCC: 0.917) was achieved with a GBT classifier on "all features," i.e., on the dataset containing the TBS. Including the TBS significantly increased the prediction performance although it has not reached the top performance of high-dimensional approaches (9, 11–14). Their LOP predictors achieved Accuracy values of 95–97%. We assume that the performance of our low-dimensional approach can be optimized further. This optimization could be achieved by including other or additional

TABLE 2 Features grouped by their importance for the LOP prediction.

Classification run	Low importance	1st quartile of FI	Medium low importance	2nd quartile of FI	Medium high importance	3rd quartile of FI	High importance
RF on core features (eight LOPs)	Grading (0.043), V (0.022), L (0.019)	0.043	N (0.096)	0.098	Leukocytes (0.107), Age (0.100)	0.110	Sex (0.371), HB (0.130), CRP (0.112)
GBT on core features (eight LOPs)	Grading (0.039), V (0.018), L (0.011)	0.039	Age (0.061)	0.070	CRP (0.091), Leukocytes (0.078)	0.096	Sex (0.477), HB (0.123), N (0.102)
RF on core features and TM (eight LOPs)	Grading (0.034), V (0.022), L (0.016)	0.040	CA 19–9 (0.049), CEA (0.046)	0.066	N (0.087), Leukocytes (0.085), Age (0.083)	0.087	Sex (0.377), HB (0.113), CRP (0.088)
GBT on core features and TM (eight LOPs)	CEA (0.024), V (0.019), L (0.009)	0.028	Grading (0.037), CA 19–9 (0.032)	0.046	CRP (0.081), Leukocytes (0.069), Age (0.055)	0.081	Sex (0.462), HB (0.119), N (0.094)
RF on all features (eight LOPs)	TBS Kidney (0.003), TBS Intestine (0.001), TBS Thyroid Gland (0.000), TBS Mamma (0.000), TBS Spleen (0.000), TBS Skin (0.000), TBS Omentum (0.000), TBS Heart (0.000)	0.004	TBS Brain (0.022), Sex (0.018), TBS Other Organs (0.016), L (0.012), TBS Stomach (0.010), V (0.006), TBS Adrenal Gland (0.005)	0.022	CEA (0.038), Leukocytes (0.036), HB (0.034), Age (0.032), TBS Bladder (0.028), TBS Bones (0.025), Grading (0.024)	0.040	TBS Pancreas (0.153), TBS Lung (0.115), TBS Liver (0.102), TBS Esophagus (0.101), TBS Lymphnodes (0.064), N (0.057), CA 19–9 (0.052), CRP (0.045)
GBT on all features (eight LOPs)	V (0.002), TBS Thyroid Gland (0.000), TBS Mamma (0.000), TBS Spleen (0.000), TBS Skin (0.000), TBS Omentum (0.000), TBS Heart (0.000), TBS Adrenal Gland (0.000)	0.002	TBS Bones (0.013), L (0.007), TBS Other Organs (0.006), Sex (0.006), TBS Kidney (0.005), Grading (0.005), TBS Intestine (0.003)	0.013	TBS Bladder (0.030), CRP (0.027), CA 19–9 (0.021), Leukocytes (0.019), CEA (0.018), HB (0.016), N (0.016)	0.030	TBS Pancreas (0.223), TBS Esophagus (0.158), TBS Lung (0.149), TBS Liver (0.092), TBS Lymphnodes (0.081), TBS Stomach (0.038), TBS Brain (0.034), Age (0.031)
RF on core features (four LOPs)	Sex (0.043), L (0.030), V (0.028)	0.043	Grading (0.050)	0.080	N (0.134), Age (0.110)	0.146	CRP (0.247), Leukocyte (0.199), HB (0.158)
GBT on core features (four LOPs)	Grading (0.043), V (0.027), L (0.027)	0.043	Age (0.048)	0.054	HB (0.122), Sex (0.060)	0.152	CRP (0.291), N (0.198), Leukocyte (0.183)
RF on core features and TM (four LOPs)	Sex (0.035), L (0.02), V (0.018)	0.038	Age (0.071), Grading (0.04)	0.083	N (0.114), HB (0.110), CEA (0.095)	0.114	CRP (0.193), CA 19–9 (0.158), Leukocytes (0.145)
GBT on core features and TM (four LOPs)	Age (0.033), V (0.018), L (0.016)	0.039	Sex (0.05), Grading (0.045)	0.055	Leukocytes (0.138), HB (0.093), CEA (0.06)	0.138	CRP (0.237), N (0.16), CA 19–9 (0.15)

(Continued)

TABLE 2 (Continued)

Classification run	Low importance	1st quartile of FI	Medium low importance	2nd quartile of FI	Medium high importance	3rd quartile of FI	High importance
RF on all features (four LOPs)	TBS Spleen (0.001), TBS Thyroid Gland (0.0), TBS Intestine (0.0), TBS Mamma (0.0), TBS Skin (0.0), TBS Omentum (0.0), TBS Heart (0.0), TBS Bladder (0.0)	0.001	TBS Other Organs (0.013), TBS Stomach (0.011), Sex (0.011), L (0.009), TBS Adrenal Gland (0.005), V (0.005), TBS Kidney (0.002)	0.013	Leukocytes (0.032), HB (0.031), CEA (0.029), TBS Bones (0.025), TBS Brain (0.023), Grading (0.018), Age (0.016)	0.034	TBS Pancreas (0.201), TBS Lung (0.136), TBS Liver (0.115), TBS Esophagus (0.1), TBS Lymphnodes (0.072), N (0.057), CA 19–9 (0.047), CRP (0.041)
GBT on all features (four LOPs)	TBS Thyroid Gland (0.0), TBS Intestine (0.0), TBS Mamma (0.0), TBS Spleen (0.0), TBS Skin (0.0), TBS Omentum (0.0), TBS Heart (0.0), TBS Bladder (0.0), TBS Adrenal Gland (0.0)	0.000	TBS Other Organs (0.005), TBS Kidney (0.005), TBS Bones (0.004), L (0.004), Grading (0.004), V (0.002), Sex (0.002)	0.005	CA 19–9 (0.018), HB (0.014), CEA (0.013), CRP (0.013), Leukocytes (0.011), Age (0.011)	0.019	TBS Pancreas (0.257), TBS Lung (0.18), TBS Esophagus (0.171), TBS Lymphnodes (0.096), TBS Liver (0.086), TBS Stomach (0.043), TBS Brain (0.038), N (0.022)

For each classification run, i.e., combination of dataset and ML algorithm, we created four groups of features according to their individual mean feature importance (FI) for the respective classification in 10-fold NCV. The groups were determined based on the first, second and third quartile of the FI among all features in the respective dataset for the respective classification setting: features with a FI at most 1st quartile (low importance), features with a FI above 1st quartile and at most 2nd quartile (medium low importance), features with a FI above 2nd quartile and at most 3rd quartile (medium high importance) and features with a FI above 3rd quartile (high importance). All values shown in this table are rounded to three decimal places.

routine clinical data. Moreover, other ML methods could be tested for their LOP prediction performance.

4.2 The predictive power of our feature sets

The LOP prediction performance on the feature sets "core features" and "core features and TM" was solid, but not remarkable. This is consistent with the conclusion from the previous paragraph that these feature sets are limited in their information content. This limitation reduces their predictive power in LOP prediction. By adding the TBS, i.e., the distribution pattern of metastases, to the dataset ("all features"), the LOP prediction performance increased significantly. Moreover, the TM and TBS received a large share of the overall FI when they were introduced to the dataset. As a consequence, the quartiles of the FI values decreased with increasing number of considered features. From these observations, we conclude that the TM and, in particular, the TBS add valuable information for LOP prediction to the dataset. This is a striking result considering that their limited availability makes the classification itself more difficult. Including the TM made the missing value imputation less stable due to the low availability of CEA and CA 19–9. Including the TBS reduced the dataset size significantly, because their extremely low availability required us to dispense with most cancer cases. Nevertheless, the TBS contributed to a remarkable increase in prediction performance. Furthermore, they led to a more balanced decision making which can be concluded from the reduced difference between Precision and Recall on "all features" compared to the other two feature sets. Reasoning on the predictive performance of individual features can be found in Supplementary material. A particularly striking observation was the decreased importance of the feature sex when not considering the LOP classes Breast, Colorectal, ENT, and Kidney anymore. This drop in FI for the sex could be due to the high number of female patients in the breast cancer group, while the ratio between men and women in the other entities is much more balanced.

4.3 The benefits of low-dimensional data for LOP prediction

When providing "all features" to the ML methods we achieved very high LOP prediction performance on low-dimensional data almost reaching the performance of high-dimensional approaches. Due to their slightly better performance, the high-dimensional approaches might appear more suitable for clinical LOP prediction. However, performance alone is not suitable for determining the quality of an LOP prediction system for clinical practice. This is because the performance only indicates how often the class predicted to be most probable was correct. Instead, it must be considered that the ML algorithm supports the oncologist in his decision; it does not

make the decision for him. It could therefore also output LOP probabilities instead of the most probable class alone. Based on such a probabilistic overview, the oncologist could make their decision including their own prior knowledge. Thus, eventually the output of the LOP prediction system would enhance the oncologist's knowledge in a data-driven manner instead of replacing it. The performance of the LOP prediction system alone cannot measure the quality of such a decision. We therefore believe that the small reduction in performance is justifiable; especially when one considers the clear advantages our low dimensional approach has in a practical setting. Our approach only needs routine clinical data, i.e., features readily available from a diverse patient population without specialized examinations. This restriction enables a cost-effective, user-friendly, and explainable LOP prediction for CUP patients which could be implemented by a clinical decision support system. The explainability is introduced by the chosen ML methods. While high-dimensional input data requires the application of artificial neural networks, which lack explainability, our low-dimensional approach allows the use of explainable tree-based methods like RF and GBT. Further decision support could be achieved by using probabilistic models such as Gaussian Process Models additionally to or instead of tree-based methods. Using such models would require some preprocessing of categorical variables but, on the other hand, add a statistically sound basis to the explainability of the LOP prediction. Moreover, as a future vision, our low-dimensional approach could enable a continuously learning LOP prediction system. Automated ETL processes could be used to update such a system with new patient data on an ongoing basis. These regular updates could improve the LOP prediction performance continuously. However, the data preparation process is currently still too complex and time-consuming for an automated ETL process (29). Overall, we consider the benefits of low-dimensional data for LOP prediction to outweigh the minor reductions in performance.

4.4 Limitations of our work

Our results show that low-dimensional data are well suited for LOP prediction, but our work has a few limitations beyond that. Firstly, our results do not reveal whether the performance improvements through adding the TM and TBS to the input data result specifically from these features. An alternative hypothesis is that the improved performance results merely from the ML method receiving more clinical information. Moreover, the significant performance gain through adding the TBS could be a result of a documentation bias. The radiologists knew the LOP when creating the RECIST evaluations of the cancer cases, based on which we created the TBS. The choice of documented target and non-target lesions might have been influenced by prior knowledge on the LOP. On the other hand, the RECIST guideline ensured the best possible objectivity. To improve the objectivity further, researchers could use different representations of the distribution of metastases. At UHC the documentation according to RECIST criteria was the only structured documentation representing the distribution of metastases.

Another limitation of our work is the restriction to eight rather broad LOP classes. Related works have considered more classes and sometimes even subclasses, which made their classification setting more difficult. Thus, for them, it was more difficult to achieve a high classification performance. We restricted to LOP classes that CUP patients have been assigned to post-mortem. So, we argue that many of the LOP classes considered by related work will not be relevant for deciding the treatment for CUP patients in practice. Additionally, some related works exceeded the capabilities of our approach by predicting the cancer subtype. Such an advanced prediction can further support treatment decisions. Moreover, some subtypes differ significantly in characteristics such as the distribution pattern of metastases. These significant differences may make differentiation of sybtypes easier than differentiation of higher-level cancer types. However, our results show that our RF- and GBT-based models can classify the different patterns that emerge in the subtypes into common cancer classes very well. Regarding the potential clinical disadvantage of not predicting the subtype, we argue that the subtype can be determined by entity-specific examinations once the LOP has been detected. What remains as a limitation is that we could not sufficiently test the feature set "all features" in the eight-class setting. When only considering four instead of eight classes, the FI of the feature sex dropped significantly. This clear reduction in FI shows that the eight-class classification task differs significantly from the four-class task. Due to the underrepresentation of the classes "Breast," "Colorectal," "ENT," and "Kidney" in the eight-class version of the "all features" dataset, we did not obtain a reliable performance measurement of the LOP prediction based on the TBS in the eight-class setting. This limitation could be mitigated by repeating the experiments on a more balanced dataset. The class balance could be increased by including data from further cancer centers also documenting their study progress according to RECIST v1.1. Another step remaining as future work is the clinical or external validation of our results. Such a validation should include examining the effects of our data compilation decisions on the LOP prediction.

4.5 Conclusion and future work

All in all, the robust classification performance on all datasets serves as a proof-of-concept that LOP prediction on low-dimensional clinical information works well. We achieved remarkable classification performance in particular when the prediction models were given the distribution pattern of metastases. The low dimensionality of our prediction approach increases its practical applicability in data-driven LOP prediction significantly. Future work could now aim for optimizing the classification results by using more or different clinical routine data as input values. Additional optimization is possible by increasing the number of cancer cases in the datasets through collaboration with further clinics. This would address the issues of the small dataset sizes and the biases possibly introduced by including the TBS. Moreover, other ML methods such as probabilistic models as well as ensembles of ML algorithms could be tested for their LOP prediction performance on low-dimensional clinical information. Above all, however, it is key to investigate whether our approach delivers reliable LOP predictions for CUP patients. Externally or clinically validating the reliability of our low-dimensional LOP prediction approach is crucial before deploying it in clinical practice. With its focus on practical applicability, our approach could optimize the prognosis of CUP patients effectively.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: the applied data anonymization strategy was agreed with the Ethics Committee and the Data Protection Department of University Hospital Cologne on the condition that the data is only made available to employees of the clinic. Reasonable requests to access the analyzed datasets should be directed to corresponding author upon which they will be discussed with the Ethics Committee and the Data Protection Department. Requests to access these datasets should be directed to JG, julia.gehrmann1@uk-koeln.de.

Ethics statement

The studies involving humans were approved by the Ethics Committee of the Medical Faculty of the University of Cologne. The studies were conducted in accordance with the local legislation and institutional requirements. The human samples used in this study were acquired from a by- product of routine care or industry. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

JG: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Visualization, Writing – original draft, Writing – review & editing. DS: Conceptualization, Data curation, Investigation, Writing – original draft, Writing – review & editing. KH: Data curation, Visualization, Writing – review & editing. MB: Investigation, Writing – review & editing. LQ: Investigation, Methodology, Writing – review & editing. SA: Data curation, Investigation, Writing – review & editing. SB: Data curation, Investigation, Writing – review & editing. AGu: Conceptualization, Data curation, Methodology, Project administration, Writing – review & editing. MB: Project administration, Writing – review & editing. AGr: Data curation, Project administration, Writing – review & editing. SL: Data curation, Methodology, Writing – review & editing. TP: Data curation, Methodology, Writing – review & editing. TZ: Data curation, Methodology, Supervision, Writing – review & editing. OB: Methodology, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. We acknowledge support for the Article Processing Charge from the DFG (German Research Foundation, 491454339). Data access and integration was enabled by BMBF (German Federal Ministry of Education and Research) through projects NUM-DIZ (FKZ: 01KX2121) and CORD (FKZ: 01ZZ1911D). Provided radiology data are partly funded by the EU-EFRE program (radCIO; project RA-1-1-014) and Jöster foundation (project AI4Onco).

Acknowledgments

We would like to thank the BMFB (German Federal Ministry of Education and Research) for its financial support of the data access and integration.

Conflict of interest

SL received author and speaker royalties from Amboss GmbH.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2024.1396459/full#supplementary-material

References

1. Laprovitera N, Riefolo M, Ambrosini E, Klec C, Pichler M, Ferracin M. Cancer of Unknown Primary: challenges and Progress in clinical management. *Cancers*. (2021) 13:451. doi: 10.3390/cancers13030451

2. Hainsworth JD, Fizazi K. Treatment for patients with unknown primary cancer and favorable prognostic factors. *Semin Oncol*. (2009) 36:44–51. doi: 10.1053/j.seminoncol.2008.10.006

3. Hübner G, Neben K, Stöger H. CUP syndrom–krebserkrankungen mit unbekanntem primärtumor Berlin, DGHO Leilinie (2018).

4. Pavlidis N, Pentheroudakis G. Cancer of Unknown Primary site. *Lancet*. (2012) 379:1428–35. doi: 10.1016/S0140-6736(11)61178-1

5. Rassy E, Pavlidis N. The currently declining incidence of Cancer of Unknown Primary. *Cancer Epidemiol*. (2019) 61:139–41. doi: 10.1016/j.canep.2019.06.006

6. Urban D, Rao A, Bressel M, Lawrence YR, Mileshkin L. Cancer of Unknown Primary: a population-based analysis of temporal change and socioeconomic disparities. *Br J Cancer*. (2013) 109:1318–24. doi: 10.1038/bjc.2013.386

7. Penson A, Camacho N, Zheng Y, Varghese AM, al-Ahmadie H, Razavi P, et al. Development of genome-derived tumor type prediction to inform clinical Cancer care. *JAMA Oncol*. (2020) 6:84–91. doi: 10.1001/jamaoncol.2019.3985

8. He B, Dai C, Lang J, et al. A machine learning framework to trace tumor tissue-of-origin of 13 types of cancer based on DNA somatic mutation. *Biochim Biophys Acta Mol basis Dis*. (2020) 11:165916. doi: 10.1016/j.bbadis.2020.165916

9. He B, Lang J, Wang B, Liu X, Lu Q, He J, et al. TOOme: a novel computational framework to infer Cancer tissue-of-origin by integrating both gene mutation and expression. *Front Bioeng Biotechnol*. (2020) 8:394. doi: 10.3389/fbioe.2020.00394

10. Nguyen L, Van Hoeck A, Cuppen E. Machine learning-based tissue of origin classification for Cancer of Unknown Primary diagnostics using genome-wide mutation features. *Nat Commun*. (2022) 13:4013. doi: 10.1038/s41467-022-31666-w

11. Liu H, Qiu C, Wang B, Bing P, Tian G, Zhang X, et al. Evaluating DNA methylation, gene expression, somatic mutation, and their combinations in inferring tumor tissue-of-origin. *Front Cell Dev Biol*. (2021) 9:619330. doi: 10.3389/fcell.2021.619330

12. Miao Y, Zhang X, Chen S, Zhou W, Xu D, Shi X, et al. Identifying cancer tissue-of-origin by a novel machine learning method based on expression quantitative trait loci. *Front Oncol*. (2022) 12:946552. doi: 10.3389/fonc.2022.946552

13. Zhao Y, Pan Z, Namburi S, Pattison A, Posner A, Balachander S, et al. CUP-AI-dx: a tool for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data and artificial intelligence. *EBioMedicine*. (2020) 61:103030. doi: 10.1016/j.ebiom.2020.103030

14. Vibert J, Pierron G, Benoist C, Gruel N, Guillemot D, Vincent-Salomon A, et al. Identification of tissue of origin and guided therapeutic applications in Cancers of Unknown Primary using deep learning and RNA sequencing (trans CUPtomics). *J Mol Diagn*. (2021) 23:1380–92. doi: 10.1016/j.jmoldx.2021.07.009

15. Lu MY, Chen TY, Williamson DFK, Zhao M, Shady M, Lipkova J, et al. AI-based pathology predicts origins for Cancers of Unknown Primary. *Nature*. (2021) 594:106–10. doi: 10.1038/s41586-021-03512-4

16. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer*. (2009) 45:228–47. doi: 10.1016/j.ejca.2008.10.026

17. Lin WC, Tsai CF. Missing value imputation: a review and analysis of the literature (2006–2017). *Artif Intell Rev*. (2020) 53:1487–509. doi: 10.1007/s10462-019-09709-4

18. Austin PC, White IR, Lee DS, van Buuren S. Missing data in clinical research: a tutorial on multiple imputation. *Can J Cardiol*. (2021) 37:1322–31. doi: 10.1016/j.cjca.2020.11.010

19. Kang H. The prevention and handling of the missing data. *Korean J Anesthesiol*. (2013) 64:402–6. doi: 10.4097/kjae.2013.64.5.402

20. Van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Softw*. (2011) 45:1–67. doi: 10.18637/jss.v045.i03

21. Prasser F, Kohlmayer F. Putting statistical disclosure control into practice: the ARX data anonymization tool. *Med Data Privacy Handb*. (2015) 27:111–48. doi: 10.1007/978-3-319-23633-9_6

22. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. *IEEE Trans Neural Netw Learn Syst*. (2021) 32:4793–813. doi: 10.1109/TNNLS.2020.3027314

23. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *In Fusion*. (2020) 58:82–115. doi: 10.1016/j.inffus.2019.12.012

24. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews correlation coefficient metric. *PLoS One*. (2017) 12:e0177678. doi: 10.1371/journal.pone.0177678

25. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. (2001) 29:1189–232. doi: 10.1214/aos/1013203451

26. Louppe G, Wehenkel L, Sutera A, Geurts P. Understanding variable importances in forests of randomized trees. In Advances in neural information processing systems. Curran Associates, Inc. (2013).

27. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on machine learning (2006), 161–168. Association for Computing Machinery.

28. Caruana R, Karampatziakis N, Yessenalina A. An empirical evaluation of supervised learning in high dimensions. In Proceedings of the 25th international conference on machine learning (2008), 96–103. Association for Computing Machinery.

29. Gehrmann J, Herczog E, Decker S, Beyan O. What prevents us from reusing medical real-world data in research. *Sci Data*. (2023) 10:459. doi: 10.1038/s41597-023-02361-2

frontiers | Frontiers in Medicine

TYPE Technology and Code
PUBLISHED 18 December 2024
DOI 10.3389/fmed.2024.1430676

OPEN ACCESS

EDITED BY
Gokce Banu Laleci Erturkmen,
Software Research and Development Consulting, Türkiye

REVIEWED BY
Andreas K. Triantafyllidis,
Centre for Research and Technology Hellas (CERTH), Greece
Ozgur Kilic,
Muğla University, Türkiye

*CORRESPONDENCE
Philipp Antczak
✉ philipp.antczak@uk-koeln.de

RECEIVED 10 May 2024
ACCEPTED 21 October 2024
PUBLISHED 18 December 2024

CITATION
Schmidt J, Arjune S, Boehm V, Grundmann F, Müller R-U and Antczak P (2024) Bridging health registry data acquisition and real-time data analytics.
Front. Med. 11:1430676.
doi: 10.3389/fmed.2024.1430676

Bridging health registry data acquisition and real-time data analytics

Johannes Schmidt[1], Sita Arjune[2,3,4], Volker Boehm[5], Franziska Grundmann[2], Roman-Ulrich Müller[2,3,4] and Philipp Antczak[2,4]*

[1]Bonacci GmbH, Cologne, Germany, [2]Department II of Internal Medicine and Center for Molecular Medicine Cologne, University of Cologne, Faculty of Medicine and University Hospital Cologne, Cologne, Germany, [3]Center for Rare Diseases Cologne, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany, [4]Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), Cologne, Germany, [5]Institute for Genetics, University of Cologne, Cologne, Germany

The number of clinical studies and associated research has increased significantly in the last few years. Particularly in rare diseases, an increased effort has been made to integrate, analyse, and develop new knowledge to improve patient stratification and wellbeing. Clinical databases, including digital medical records, hold significant amount of information that can help understand the impact and progression of diseases. Combining and integrating this data however, has provided a challenge for data scientists due to the complex structures of digital medical records and the lack of site wide standardization of data entry. To address these challenges we present a python backed tool, Meda, which aims to collect data from different sources and combines these in a unified database structure for near real-time monitoring of clinical data. Together with an R shiny interface we can provide a near complete platform for real-time analysis and visualization.

KEYWORDS

visualisation, shiny, database, healthcare, cohort

Introduction

The medical world has seen a paradigm shift in recent years, acknowledging that data collection and analysis is key to understand the most pressing challenges in human health. Particularly with rare diseases, where the low number of patients impact the statistical analysis of these, must ensure that high quality and systematic collection of data is optimized (1). Often retrospectively collected data is available within the hospitals medical record systems but are plagued by numerous free-text fields, simple collection of laboratory values where the measurement units are not standardized across the fields, and the sheer amount of variables that that have been accumulated into these databases over years of use (2, 3). Transitioning such database entities to more standardized and usable structures for clinical research or even simple oversight of departments within a healthcare organization can prove to be challenging and associated with a very high cost of implementation and transition.

Furthermore, quality control of such data is often performed only when data is extracted for clinical research and entry failures only noticed when compared to other individuals. This proves one of the major headaches for data scientists who aim to integrate and analyse such data in various contexts (4). Within the medical field, and especially for laboratory values, thresholds are known that describe compatibility with life, giving a first indication whether the values entered are reasonable. Given the broad spectrum of diseases and health states in humans it is not reasonable to assume that each medical professional knows and applies these thresholds, particularly when

they are early in their career. Written laboratory reports often include the range and thresholds to consider, but once provided within the database these are lost or stored in such a way that they are not directly accessible by the user (5). A more direct, and disease tailored, approach on the level of medical record oversight and data entry could lead to improved data quality and medical understanding.

In the last decade in Germany, there has been significant progress in the development of standardized interfaces to allow interoperability of data between health care institutions. The FHIR interface aims to provide a solution to transport data from one location to another and allow the sharing of patient data. While these developments are of great importance in the medical field, they do not fully address the internal and integrative use in clinical research. To this end, we have developed a small highly flexible and dynamic tool to collate, aggregate, integrate, and visualize clinical data. We opted to develop a centralized database structure, that pulls in data from multiple sources, formats, aligns, and tests them to ensure highest possible data quality. This database can then be connected to a visualization framework such as R shiny, Grafana, or Tableau to present the data in an aggregated fashion to healthcare professionals.

Implementation

Development of a universal translation service for medical data (MEDA)

Clinical registries are often based on data registration, management and storage designs which lack up-to-date database standards. These range from mere spreadsheets to specialized but non-standardized databases from various providers to collect and represent data (6). While these web-based tools often contain the ability to validate data entry, or limit the entry to specific datatypes, these features are often not used due to their complex configuration or lack of knowledge and experience by the initiating user. In addition, database structures are often inefficiently designed and variable names lack the descriptive nomenclature which allows other users to understand their values and implement these variables in their analyses. This then often requires the development of additional variable-dictionaries which provide extended definitions of the values. Particularly for the key aim of such datasets, i.e., downstream biostatistical analyses or real-time visualization, the initial data structure and simplicity of the database is an important aspect for implementation and use. Live data visualization for both data sharing and in-house observation of cohort development is hardly possible in this setting. This is especially important when the developer of the database has left the organization and the approaches and thoughts during the development process have not been documented accordingly. In most cases, the initial design allows questions posed by the developer and researchers associated with the project to be answered, however they can hinder the further use and analyses of these important data.

To address the challenges around clinical datasets described above, and to enable the utilization of existing resources, we have developed a Python and PostgreSQL application that is able to translate the existing information into a standardized database with a very well-defined data structure.[1] Specifically, we inherit the individual centric view fundamental to medical science and attach additional information as separated tables that can be brought together to analyse various questions. These tables separate cross-sectional and longitudinal data and are grouped based on their clinical relevance. The typical database structure we have utilized is provided in Figure 1 and highlights the components that are required to be configured within our tool.

To test this simple structure, we used a large patient cohort with chronic kidney disease available at the University Hospital Cologne and translated the currently utilized ClinicalSurveys.net (7) database using our tool. ClinicalSurveys is a web-based tool to design and collect patient relevant data through a simple survey based tool. It allows collaboration across multiple sites in a secure manner and enables a systematic data collection. This cohort data contains numerous, meticulously collected patient information ranging from different levels of laboratory values, questionnaires, family history, tomography, or historic clinical information. All in all, over 2000 variables were represented within this ClinicalSurveys database structure. The design of this database followed a fully patient centric approach where longitudinal data was encoded as repeated variables within its single database. While this can be a reasonable approach to collect prospective data on individuals over a longer period of time, it can be quite error prone as, particularly, longitudinal data may be entered in the wrong section of the database skewing downstream analysis. In addition, the long list of variables can lead to an increase in human errors during entry where misplaced punctuation marks or swapping of variables may occur. Downstream analyses and visualization may then be skewed by these data structures. Furthermore, quality assurance is more difficult to achieve since the large number of variables within a single database is challenging to evaluate for human individuals. Meda addresses some of these challenges in a semi-automated fashion. Most importantly, the tool automatically generates the database structure based on the configured data slots required. In essence, Meda follows a simple 5 steps approach:

Step 1: Reading Source Data:

> The pipeline begins by reading raw source data in a flat structure, where each value occupies its own column.

Step 2: Data Class Organization:

> The flat data is organized into nested data classes, which correspond to SQL-tables. When defining the data classes. At this point transformations or other computed variables can be generated through the provision of additional python functions.

Step 3: Data Class Factory:

> The data class factory populates the nested data classes from the flat data structure.

Step 4: DTO (Data-Transfer-Object) Factory:

> The DTO factory translates the nested data classes into DTOs that mirror the SQL structure.

Step 5: DTO Registry:

> The DTO registry manages the DTO factory and database connection. It generates a DTO from a data class and writes it to the database.

1 https://github.com/bonacci-johannes/meda

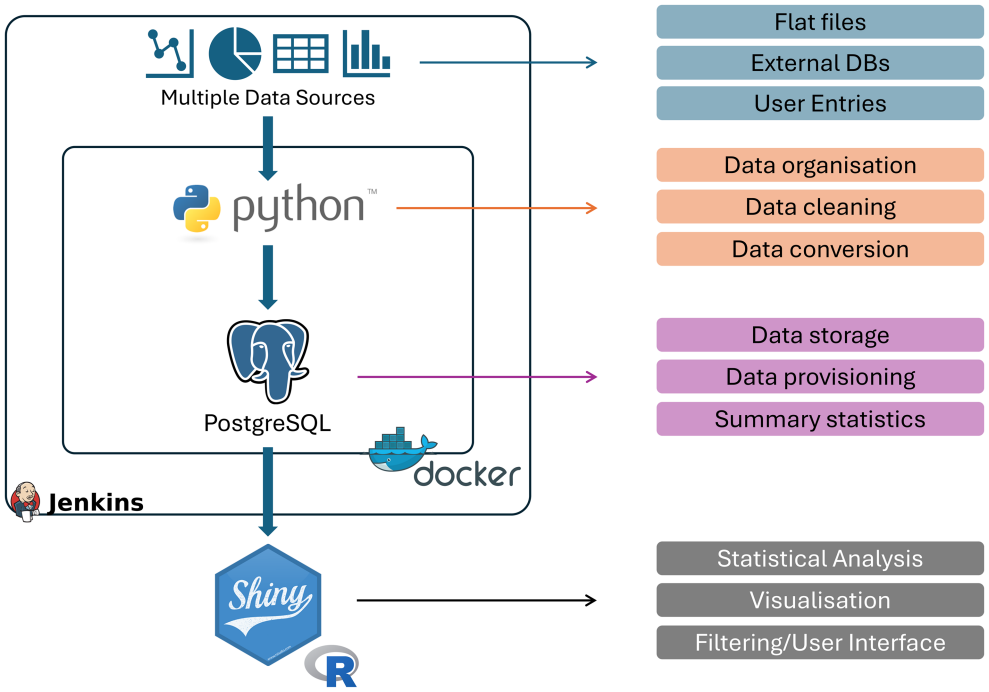

FIGURE 1
Schematic overview of Meda. The developed pipeline and containers used to transform clinical data to a real-time visualisation platform as well as their specific functions within the pipeline. Python and PostgreSQL are encased within a docker to ensure robust responses and compatibility as well as a Jenkins Butler to update, maintain and reset the database when necessary.

In addition, Meda aims to generate subsets of manageable chunks of data, following clinically relevant chunks of information. Import classes are defined that ensure errors can be caught and that data is imported in the right format. In prospective studies, where data is continuously collected over long periods of time, we are therefore able to import data on a regular basis. Next, based on the result of the class import an additional error table is generated that allows users to visualize these import errors and address them accordingly. We found that the visual representation had a significant influence on the motivation of our staff to fix and remedy the shown errors. Lastly, we implemented a threshold-based value verification system which aims to identify values that we deem to be "not compatible with life" and which are sent back to the users for verification.

Application

Setting up Meda for semi-automated data entry

As in our example, ClinicalSurveys was used to house and collect the data from various collaborators, we used Dockers surrounding our database and Meda tool to simplify the setting up and destruction of the database. Simply put, the PostgreSQL database is fully refreshed upon each update that is being made. This ensured that only one true data source was available and reduced the need for verification of data entries within our database. To manage the automated setting up and destruction of the database Jenkins was used. The Jenkins Butler (8) monitors changes in the source code, scripts, and classes that are required and updates the database as soon as changes are observed. The total workflow using this approach takes less than 3 min and can therefore be performed as often as daily if new data are expected on a regular basis. The aforementioned classes need to be implemented to ensure that the right data is entered into the database. A simple Patient centric import of individual characteristics is shown in Code Section 1. The utilized Feature keyword here is an included separate class which provides the information on how to construct a dataclass from a data dictionary and how to import it into the SQL table. It enables the use of transformers, specifications of the input key, specifications of target table type (error, crossectional, or longitudinal), and the potential defaults to consider.

```
class HeadADPKD(FeatureDataclass):
    patient_id: str = Feature(input_key='u_name', unique_index=True)

class Patient(HeadADPKD):
    # Columns
    clinical_survey_id: str = Feature(input_key='uid')
    clinical_survey_user: str = Feature(input_key='u_firstname')
    birthdate: Optional[date] = Feature(input_key='u_birth', null_defaults=NullDefaults.date)
    gender: str = Feature(input_key='u_gender')
    age_at_adpkd_diagnosis: Optional[int] = Feature(input_key='v_3726',
        null_defaults=NullDefaults.text, comment='years')

    # Sub tables
    race: Optional[Race] = Feature(input_key=('v_3381', 'v_3431'),
        transformer=transformer_race)
    health_status: Optional[HealthStatus]
    mutation: Optional[Mutation]
    meta: Optional[Meta]
    examinations: FrozenSet[MedicalExamination]
    family: Optional[Family]
    tolvaptan_dosing: Optional[TolvaptanDosing]
    tolvaptan_reaction: Optional[TolvaptanReaction]
    updosing: FrozenSet[Updosing]

    # errors
    import_errors: Optional[str] = Feature(is_error_field=True)
```

Code Section 1: Example Code for extracting data into the proposed database schema.

Automated evaluation and identification of missing and non-reliable data

During the data import, several additional steps are performed before adding the data to the database. First values are converted to a common reference unit. The unit conversion is a simple step but requires extensive configuration that covers all possible units. So far we have focused on the possible units within our CKD use case example and provide our conversion tables within the code. Code section 2 shows an example of such a configuration. This ensures that we do not need to store the unit information and that all data are converted to the relevant reference unit. Next, data are reviewed for known thresholds that are not compatible with life. Here a simple table (Table 1), which can be adjusted by user dynamically through a web-based interface, is evaluated and any values exceeding these thresholds collected within their own separate table. The results are presented to the user who can then adjust, if necessary, the value within the original table used as input. This also applies to any missing data encountered during the data import.

This workflow can easily be integrated into daily clinical routines and allows for direct evaluation and visualization of the data. In addition, the near instant visual response to the fixing of missing or non-reliable data results in a significantly increased data quality. Furthermore, enabling auditing within PostgreSQL can provide a continuous log of changes that have been performed and ensure that data consistency is preserved.

```
##############################
'density':
  ref_unit: 'g/L'
  conversion:
    'mg/L': 1000
    'mg/dL': 100
##############################
'particle_density':
  ref_unit: 'x 1E9/l'
  conversion:
    '/µl': 1000
    '/µL': 1000
    'x 1E3/µl': 1
    'x 1E12/l': 0.001
    'x 1E6/µl': 0.001
##############################
'molar_density':
  ref_unit: 'mol/L'
  conversion:
    'mol/l': 1
    'mmol/L': 1.e+3
    'µmol/L': 1.e+6
    'pmol/L': 1.e+12
    'pmol/l': 1.e+12
```

Code Section 2: Automated conversion of units during import and plausibility check.

Visualization and continuous evaluation of data provides new insights into patient health

The last step in our pipeline is the development of a visual representation of the data imported by Meda. Here we decided to develop an R (9) shiny application. While other types of frameworks exist to provide real-time views of such data, they are limited in their

TABLE 1 Example threshold table used during data import.

Column	Review_high	Invalid_low	Invalid_high
Natrium	160	115	160
Kalium	7	2	7
Lipase	1000	0	3000
Osmolarity	330	240	350
Hematocrit	50	20	70
Mcv	105	50	120
Calcium	3	1	4
Phosphat	2.5	0.2	6
Creatinine	3	0.2	20
Urea	200	10	500
Uric_acid	12	0.2	25
Albumin	60	5	100

statistical analyses that can provide useful information in a clinical setting (Figure 2).

Integration and functionality of shiny application in R

The development of the Shiny application (10) represents a significant progression in the implementation of the ClinicalSurveys database. The Shiny framework in R facilitates the creation of dynamic web applications that offer the ability to visualize and analyze data in real time. Through the utilization of this technology, the application converts unprocessed clinical data into user-friendly, interactive dashboards and reports. As a result, in the future healthcare providers are provided with instantaneous access to patient information and trends. The Shiny application has been carefully designed to accommodate the particular requirements of healthcare professionals. The platform provides a collection of interactive tools that enable users to analyze demographic patient data along multiple axes, including time, disease advancement, and treatment results. At this degree of engagement, a more profound comprehension of patient health patterns is fostered, which empowers the development of individualized patient care plans and the discovery of effective treatment protocols.

Continuous evaluation for proactive healthcare

One of the most significant features of the Shiny application is its capability for continuous data evaluation. As the PostGreSQL database is refreshed with each update (once daily), the application automatically incorporates the latest data, ensuring that healthcare providers have access to the most current patient information. The insights garnered from the continuous evaluation of patient data have profound implications for both patient care and clinical research. For patient care, it enables a shift toward more proactive and personalized

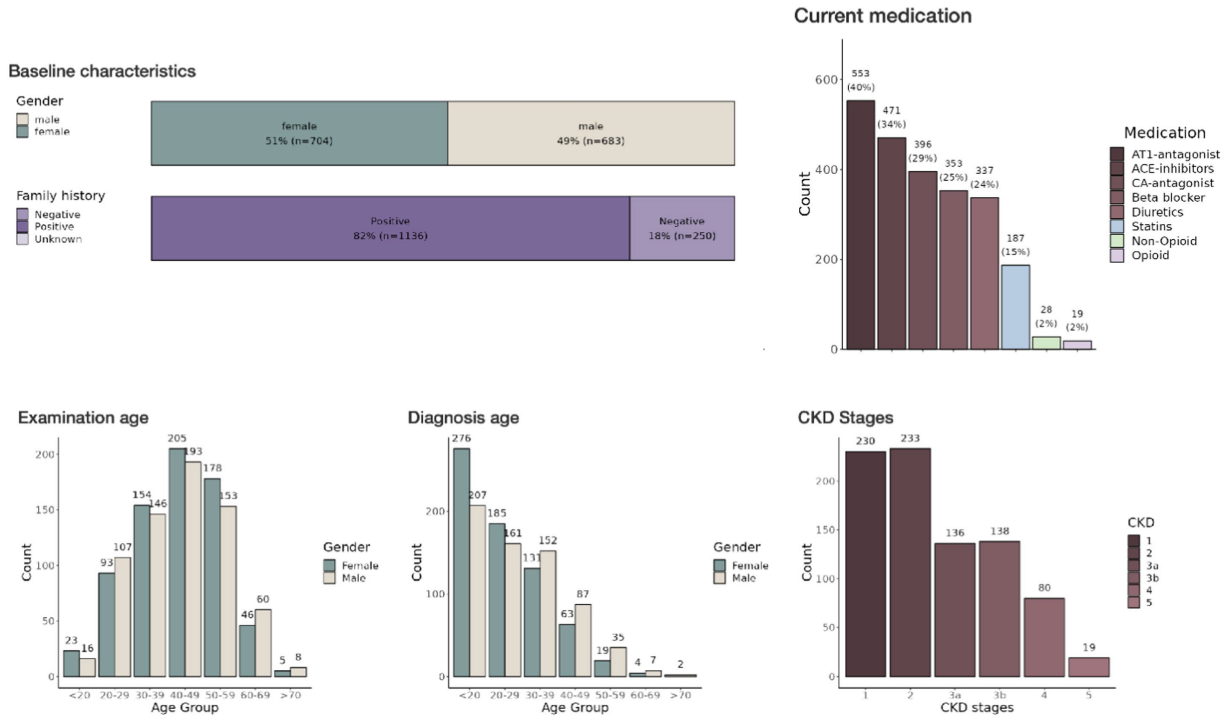

FIGURE 2
Interactive dashboard of the ADPKD registry in a shiny application—This image displays part of the interactive dashboard from the shiny application used to visualize data from the large patient cohort with chronic kidney disease. Several pages and filter options are available. The dashboard itself is segmented into tabs including 'Patient Characteristics', 'Disease Stage', 'Medication', and others, allowing for a comprehensive overview of the data categories. The 'Baseline characteristics' section provides histograms detailing demographic information.

healthcare strategies, significantly improving patient outcomes. In the realm of research, the application provides a rich dataset for analyzing treatment efficacy, patient responses, and disease patterns, thereby contributing to the advancement of medical knowledge and the development of new treatment modalities.

Discussion

The Meda pipeline was developed to bridge health registry data and real-time analysis of the available data. Our key approach was to develop a system where any type of clinical information could be imported, through the provision of simple configuration files, and where data could be displayed in near real-time to the user. Meda restructures and standardizes such information and provides programmable access to this data. While we developed this in the context of clinical registries, its approach can be used for whole clinical databases that over the years have increased in complexity.

The choice of webfront was driven by the requirements within our statistical analyses. While there are a number of real-time visualization frameworks available, such as Grafana (11), Metabase, or Tableau (12), they are not designed to handle clinical information and the underlying statistics within the biomedical domain. The shiny front, in combination with the many R packages available, allows us to generate and display any type of statistical analysis based on the data. These have been widely used in clinical data visualization and several packages have been generated to fulfil the requirements by the relevant health professionals (13–15). Shiny, and therefore R, bring additional obstacles into this development as R is generally slow in utilizing database queries, has a complex memory management, and can be inefficient in the use of data structures. To address these shortcomings we have opted to preprocess the database data every morning, and on demand, which generates the objects required for visualization and statistical analysis and are loaded through serialized R object storage. This results in a much faster visualization but limits the real-time application of our approach. Given that our registry data does not change on a daily basis and that data entry can be delayed based on clinical workload we struck a balance between functionality and overall speed in our approach. Further development of existing, faster, frameworks for visualization would remedy this.

While our tool is not the first visualization platform available (16, 17), our tool expands on the purely visual aspects of healthcare data. As databases across the healthcare sector are growing and are often based on grandfathered implementations developed in the last decades, access to this data is often complex and convoluted. In addition, the interpretation of this huge amount of data is challenging and requires a more computational and visual approach (18). Particularly, the growing number of complex cohorts, with both retrospective and prospective data collection, has proved to be challenging due to the heterogeneity in collection systems, the lack of standardization across healthcare institutions, and differences in ethical considerations. Our tool aims to address a number of these issues by enabling the integration and near real-time representation of data. By interfacing directly with a hospitals clinical data

repository our tool could show important statistics and analyses in near real-time to clinical staff, ensuring an efficient and effective oversight of data entry in various settings as well as allow for AI based decision support systems to be made available (19). While raw data is the preferred data-type, the tool would also be able to collect already computed statistics and integrate data from multiple institutions to visualize the state of healthcare institutions over a larger geographical area while not exceeding the ethical considerations of each institution.

New approaches to sharing data between institutions using the FHIR (Fast Healthcare Interoperability Resource), provides means of interfacing and exchanging data in a save and standardized environment. While our tool does not currently contain a plugin for including FHIR resources, these are often best placed at the database to database interface (20) where our tool performs best. FHIR has been used extensively for data capture, standardization, recruitment, and consent management (20). Our tool can utilize such information directly from the associated database and provide a suitable visualization and update for healthcare professionals. However, direct implementation of such plugins is possible within the framework of MEDA. Our open approach via data class factories and classes can enable any type of direct interoperability with the standards utilized at any given institution.

Overall, we have established a tool that addresses the current scientific and clinical challenges in working with larger cohorts and provides a standardized structure for use within data science groups. We hope to enable a faster and simpler pipeline for clinical questions from data to results and drive the knowledge generation within medicine.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: patient data can only be published using summary statistics as described by the ethical agreements of this cohort. As much of the publication is code to use such data, we do not intend to publish the additional data. Requests to access these datasets should be directed to roman-ulrich.mueller@uk-koeln.de.

Author contributions

JS: Conceptualization, Data curation, Methodology, Software, Writing – review & editing. SA: Data curation, Investigation, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing. VB: Methodology, Software, Visualization, Writing – original draft. FG: Data curation, Project administration, Writing – review & editing. R-UM: Conceptualization, Data curation, Funding acquisition, Investigation, Supervision, Writing – review & editing. PA: Conceptualization, Methodology, Project administration, Resources, Software, Supervision, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The research at hand was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – CECAD, EXC 2030 – 390661388. R-UM was supported by Marga and Walter Boll-Stiftung.

Conflict of interest

JS was employed by Bonacci GmbH.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Paganelli AI, Mondéjar AG, da Silva AC, Silva-Calpa G, Teixeira MF, Carvalho F, et al. Real-time data analysis in health monitoring systems: a comprehensive systematic literature review. *J Biomed Inform*. (2022) 127:104009. doi: 10.1016/j.jbi.2022.104009

2. Kannampallil TG, Schauer GF, Cohen T, Patel VL. Considering complexity in healthcare systems. *J Biomed Inform*. (2011) 44:943–7. doi: 10.1016/j.jbi.2011.06.006

3. Lehmann CU, Kim GR, Johnson KB. Pediatric Informatics: Computer Applications in Child Health Springer (2009). doi: 10.1007/978-0-387-76446-7

4. Galitsky B, Goldberg S. Artificial intelligence for healthcare applications and management. Cambridge: Academic Press (2022).

5. Clot-Silla E, Argudo-Ramirez A, Fuentes-Arderiu X. Letter to the editor: measured values incompatible with human life. *EJIFCC*. (2011) 22:52–4.

6. Schweinar A, Wagner F, Klingner C, Festag S, Spreckelsen C, Brodoehl S. Simplifying multimodal clinical research data management: introducing an integrated and user-friendly database concept. *Appl Clin Inform*. (2024) 15:234–49. doi: 10.1055/a-2259-0008

7. Vehreschild JJ, Rüping MJ, Cornely OA. A web-based research portal for rare infectious diseases [text/html]. 10. Kongress Für Infektionskrankheiten und Tropenmedizin (KIT 2010) (2010) 10. doi: 10.3205/10KIT127,

8. Jenkins. (2011). Jenkins. Available at:https://www.jenkins.io/

9. R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Available at:https://www.R-project.org/

10. Chang W., Cheng J., Allaire J. J., Sievert C., Schloerke B., Xie Y., et al. (2024). Shiny: web application framework for R. Available at:https://shiny.posit.co/

11. Abbasian M, Khatibi E, Azimi I, Rahmani AM. PHAS: an end-to-end, open-source, and portable healthcare analytics stack. *Procedia Comp Sci*. (2023) 220:511–8. doi: 10.1016/j.procs.2023.03.065

12. Ko I, Chang H. Interactive visualization of healthcare data using tableau. *Healthcare Infor Res*. (2017) 23:349–54. doi: 10.4258/hir.2017.23.4.349

13. Heinsberg LW, Koleck TA, Ray M, Weeks DE, Conley YP. Advancing nursing research through interactive data visualization with R shiny. *Biol Res Nurs*. (2023) 25:107–16. doi: 10.1177/10998004221121109

14. Miller DM, Shalhout SZ. StoryboardR: an R package and shiny application designed to visualize real-world data from clinical patient registries. *JAMIA Open*. (2023) 6:ooac109. doi: 10.1093/jamiaopen/ooac109

15. Owen RK, Bradbury N, Xin Y, Cooper N, Sutton A. MetaInsight: an interactive web-based tool for analyzing, interrogating, and visualizing network meta-analyses using R-shiny and netmeta. *Res Synth Methods*. (2019) 10:569–81. doi: 10.1002/jrsm.1373

16. Abudiyab NA, Alanazi AT. Visualization techniques in healthcare applications: a narrative review. *Cureus*. (2022) 14:e31355. doi: 10.7759/cureus.31355

17. Elshehaly M, Randell R, Brehmer M, McVey L, Alvarado N, Gale CP, et al. QualDash: adaptable generation of visualisation dashboards for healthcare quality improvement. *IEEE Trans Vis Comput Graph*. (2021) 27:689–99. doi: 10.1109/TVCG.2020.3030424

18. Menon A., Aishwarya M. S, Joykutty A. Maria, Av A. Y., Av A. Y.,. (2021). Data visualization and predictive analysis for smart healthcare: tool for a hospital. 2021 IEEE Region 10 Symposium (TENSYMP), 1–8

19. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ*. (2023) 23:689. doi: 10.1186/s12909-023-04698-z

20. Vorisek CN, Lehne M, Klopfenstein SAI, Mayer PJ, Bartschke A, Haese T, et al. Fast healthcare interoperability resources (FHIR) for interoperability in Health Research: systematic review. *JMIR Med Inform*. (2022) 10:e35724. doi: 10.2196/35724

Frontiers in Medicine

Translating medical research and innovation into improved patient care

A multidisciplinary journal which advances our medical knowledge. It supports the translation of scientific advances into new therapies and diagnostic tools that will improve patient care.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact