Data interoperability in context: the importance of open-source implementations when choosing open standards

Daniel Kapitan   Femke Heddema   Andre Dekker

Melle Sieswerda   Bart-Jan Verhoeff   Matt Berg

In response to the proposal of Tsafnat et al. to converge towards three open health data standards, this viewpoint provides a critical reflection on the proposed alignment of using openEHR, FHIR and OMOP as the default standards for clinical care and administration, data exchange and longitudinal analysis, respectively. We argue that open standards are a necessary but not sufficient condition to achieve health data interoperability. The ecosystem of open-source software needs to be considered when choosing an appropriate standard for a given context. We discuss two specific contexts, namely standardization of i) health data for federated learning, and ii) health data sharing in low and middle income countries (LMICs). Specific design principles, practical considerations and implementation choices for these two contexts are described, based on ongoing work in both areas. In the case of federated learning, we observe convergence towards OMOP and FHIR, where the two standards can effectively be used side-by-side given the availability of mediators between the two. In the case of health information exchanges in LMICs, we see a strong convergence towards FHIR as the primary standard. We propose practical guidelines for context-specific adaptation of open health data standards.

Open standards are a necessary but not sufficient condition for interoperability

"A paradox of health care interoperability is the existence of a large number of standards with significant overlap among them," say Tsafnat et al., followed by a call to action towards the health informatics community to put effort into establishing convergence and preventing collision [1]. To do so, they propose to converge on three open standards, namely i) openEHR for clinical care and administration; ii) Fast Health Interoperability Resources (FHIR) for data exchange and iii) Observational Medical Outcomes Partnership Common Data Model (OMOP) for longitudinal analysis. They argue that open data standards, backed by engaged

communities, hold an advantage over proprietary ones and therefore should be chosen as the steppingstones towards achieving true interoperability.

While we support their high-level rationale and intention, we feel their proposed trichotomy does not do justice to details that are crucial in real-world implementations. This viewpoint provides a critical reflection on their proposed framework in three parts. First, we reflect on salient differences between the three open standards from the perspective of the notion of openness of digital platforms [2], the paradox of open [3] and the hourglass model of open architectures [4,5]. Subsequently, we outline the importance of the open-source software (OSS) by reflecting on our considerations in designing and implementing health data platforms in two specific contexts, namely i) platforms for federated learning on shared health data in high income countries; and ii) health data platforms for low and middle income countries (LMICs). These case studies illustrate the limitations of the trichotomy proposed by Tsafnat et. Particularly, we argue that of the three standards, FHIR stands out as being the most practical and adaptable which allows it to be used for longitudinal analysis and routine collection of clinical data as well, besides its original purpose as a health data exchange standard. We conclude this viewpoint with practical implications of these findings and directions for future research of open health data standards.

Digital platforms require extensibility, availability of complementary components and availability of executable pieces of software

In their editorial, Tsafnat et al. argue that i) the paradox of interoperability of having overlapping standards can be addressed by converging on just three standards; ii) practical and socio-technical considerations are as important as, if not more important than, technical superiority and therefore balancing of customizability and rigidity is of the essence; and iii) open standards, backed by engaged communities, hold an advantage over proprietary ones. While we concur with these points, we argue that these are necessary, but not sufficient conditions for convergence of health data standards. Existing research on digital platforms underlines the importance of the platform's openness, not only in terms of open standards, but also in terms of availability of executable pieces of software, extensibility of the code base and availability of complements to the core technical platform (in this case the health data standard is a critical, defining component of the core technical platform) [2]. Openness in this context pertains to the software modules that constitute the digital platform. Realizing openness can be achieved through open sourcing the core components of the platform or defining standardized interfaces through which components can interact [6]. Only when the majority of these aspects of digital platforms are met can we reasonably expect that the digital platform will indeed flourish and be long lived.

> **i** Textbox 1: Conceptual background of the digital platform.
>
> **Digital platforms** are software-based online infrastructures that facilitate interactions and transactions between users. In the context of this paper, digital platforms serve as an interface used to interact with data systems. **Data systems** describe a set of technologies, tools and processes that extract, manage and deliver data. Where the **data system** describes the functional implementation, the **data architecture** specifies the design framework, outlining how the data flows in its collection, storage, processing and governance. Its key components are **data sources** (original 'raw' data that is collected before any processing), data repositories like **databases**, **data warehouses** or **lakes** and data processing engines and pipelines that transform raw data into a usable format for analysis.
> All architectures include a **core technical platform** (the foundational infrastructure) that can be extended to facilitate the necessary digital services. **Data architectures** contain different levels of specifications for the technical components entailed in the system. These levels include a systems **code base** (machine-readable text describing how to extract and process certain data), **software tools** (programs and applications enabling digital operations) and **stacks** (layers of software systems working together).

If open digital platforms are what we want, the question is how to achieve that. In what they frame as 'the paradox of open', Keller and Tarkowski argue that open platforms and their associated ecosystems can only flourish if two types of conditions are met [3]. The first condition states that many people need to contribute to the creation of a common resource. "This is the story of Wikipedia, OpenStreetMap, Blender.org, and the countless free software projects that provide much of the internet's infrastructure." [3] Indeed, Tsafnat et al. have explicitly taken into account that "an engaged and vibrant community is a major advantage for the longevity of the data standards it uses," which has informed their proposal to converge towards OMOP, FHIR and openEHR over other existing health data standards. However, the importance of OSS is somewhat overlooked. This point is only mentioned in passing when Tsafnat et al. reference work done by Reynolds and Wyatt who already argued in 2011 "… for the superiority of open-source licensing to promote safer, more effective healthcare information systems. We claim that open-source licensing in health care information systems is essential to rational procurement strategy" [7]. Hence, we extend the line of reasoning of Tsafnat et al. by emphasizing that the availability of executable OSS components, which inherently make it easier to extend the code base of the health data standard and thereby driving greater availability of complementary components, is an important criterion which needs to be explicitly taken into account when choosing which standard to adopt.

The second condition put forward by Keller and Tarkowski is that open ecosystems have proven fruitful when "opening up" is the result of external incentives or requirements, rather than voluntary actions. Examples of such external incentives are "… publicly funded knowledge production like Open Access academic publications, cultural heritage collections in the Public Domain, Open Educational Resources, and Open Government data." [3] Another canonical

example is the birth of the GSM standard, which was mandated by European legislation [8]. Reflecting on this condition in the context of open health data ecosystems, we observe a salient difference between FHIR versus openEHR and OMOP, namely that the former is the only one that has been mandated - or at least strongly recommended - in some jurisdictions. Survey results on the state of FHIR show that the FHIR standard has been mandated or advised in 20 countries [9]. Notably, the European Electronic Health Record Exchange Format (EHRxF), introduced by the European Commission in 2019 with the aim to ensure secure, interoperable, cross-border access to electronic health data across the EU, decided in 2022 to adopt HL7 FHIR as the exchange format for future priority data categories [10]. In the US, the Office of the National Coordinator for Health Information Technology and the Centers for Medicare and Medicaid Services have introduced a steady stream of new regulations, criteria, and deadlines in Health IT that has resulted in significant adoption of FHIR [11]. In India, the open Health Claims Exchange protocol specification - which is based on FHIR - has been mandated by the Indian government as the standard for e-claims handling [12,13]. The African Union recommends all new implementations and digital health system improvements use FHIR as the primary mechanism for data exchange [14], but doesn't say anything about the use of, for example, openEHR for clinical point-of-service systems.

Our third critical reflection on choosing health data standards pertains to the notion of the hourglass model [4,5] and the concept of open architectures [15]. The hourglass model is "… an approach to design that seeks to support a great diversity of applications (at the top of the hourglass) and allow implementation using a great diversity of supporting services (at the bottom)." [5] The center of the hourglass - the waist or also called the spanning layer in the information systems parlance - is defined by a set of minimal standards which mediates all interactions between the higher and lower layers. In the case of the Internet, the spanning layer is defined by the TCP/IP protocol, which is supported by a variety of underlying connectivity services (many different physical networks) on top of which many different applications can be built (email, videoconferencing etc.). We argue that FHIR has an added benefit over openEHR and OMOP because it can act as the spanning layer within an open health data platform. Because FHIR is inherently designed to function as a data exchange standard, it can function as a mediator between different components of the health data platform. The modularity of the various components that are part of the FHIR ecosystem allow it to be used effectively to implement subsystems.

We argue that i) the external incentives that have mandated FHIR in certain jurisdictions, and ii) the inherent modularity of the FHIR standard have resulted in a large boost in both commercial and OSS development activities in the FHIR ecosystem. Illustrative of this is the speed with which the Bulk FHIR API has been defined and implemented in almost all major implementations [16,17], and the SQL-on-FHIR specification to make large-scale analysis of FHIR data both accessible to a larger audience as well as portable between systems [18].

Table 1: Number of contributors and number of repositories on GitHub for the three healthcare data standards as per 28-01-2025.

(a) Last three months

search term	# contrib.	# repos.
"openEHR"	82	49
"OMOP" or "OHDSI"	446	221
"FHIR	1,648	756

(b) All time

# contrib.	# repos.
429	450
1,019	113
8,497	8,617

> ℹ **Textbox 2:** Conceptual background of data processing pipelines for analytics.
>
> **Data pipelines** define a sequence or workflow of processes for data. **Data processing engines** are tools that process, transform and analyze large-scale data and as such provide the foundational infrastructure to implement data pipelines. **Computing workloads** are specific tasks executed across data systems, like data processing and analytics. **Data transformation** entails all the processing pipelines that convert data into usable insights. **Mappings** are specific data transformations that aim to align data from different sources with a unified structure. **Granular mappings** transform data at the most detailed level, translating data elements across different schemas. **Queries** are built on top of transformed data, and retrieve data for insights generation, sometimes requiring further data processing.

The external incentives have also led to more people voluntarily contributing to FHIR-related OSS projects, which has resulted in a wide offering of FHIR components across major technology stacks (Java, Python, .NET), thereby strengthening the first condition for establishing openness. By comparison, OMOP and openEHR have profited less from external incentives to spur the adoption and thereby growing the ecosystem beyond a certain critical mass. To illustrate this, a quick-scan of the available OSS components listed on the website of the three governing bodies HL7 [19], OHDSI [20] and openEHR [21], indicates that the ecosystem of FHIR and OMOP have a significantly larger offering of extensible and complementary OSS components than openEHR, although for the latter notable a mature OSS implementation is available with EHRbase [22]. Taking GitHub as an proxy of worldwide development activities, Table 1 shows the number of contributors and repositories for three different search terms. Note that these numbers should be taken as rough indicators. Given that the FHIR standard has more application areas, one would expect more GitHub projects than for openEHR.

In summary, we stress that beyond evaluating the intrinsic structure of an open standard and the community that supports the standard, we need to consider the wider ecosystem of OSS implementations and availability of complementary components. From this wider perspective of the whole ecosystem surrounding the three standards, FHIR stands out as having the most diverse and rich ecosystem because it has been mandated in certain jurisdictions and

because its technical foundations are inherently more broad and modular. This is relevant when comparing these standards in real-world implementations. We now turn to two specific use cases where these considerations are at play.

Standardization of health data for federated learning

The current fragmentation in health data is one of the major barriers towards leveraging the potential medical data for machine learning (ML). Without access to sufficient data, ML will be limited in its application to health improvement efforts and, ultimately, from making the transition from research to clinical practice. High quality health data, obtained from a research setting or a real-world clinical practice setting, is hard to obtain, because health data is tightly regulated.

> **ℹ Textbox 3: Conceptual background of distributed data systems.**
>
> **Data systems** often have a centralized architecture, where data is collected in a single repository or location. However, data systems can also **distribute** the storage and processing of data across different nodes or locations such as servers and edge devices. **Servers** act as the central processing units in data architecture, supporting computing workloads in data extraction, storage and transformation of data. **Edge devices** mainly provide support to the data extraction and preprocessing, generally located near the source of the data.
>
> **Federated learning** is an approach where machine learning models are trained across a distributed data system. Data transformations and analyses are performed on locally held data across multiple nodes, typically using edge devices or local servers. In this setup, the server that hosts the machine learning model does not need direct access to the source data. Instead, it aggregates the outputs of the local nodes (the updated model parameters) to train a global model. This method ensures that sensitive data remains local, preserving privacy while still enabling collaborative model training across distributed systems.

Federated learning (FL) is a learning paradigm that aims to address these issues of data governance and privacy by training algorithms collaboratively without moving (copying) the data itself [23,24]. Based on ongoing work with the PLUGIN healthcare consortium [25], we have detailed an architecture for FL for secondary use of health data for hospitals in the Netherlands. The starting point for this implementation is the National Health Data Infrastructure agreements for research, policy and innovation for the Dutch healthcare sector, which have been adopted at the beginning of 2024 [26]. Figure 1 shows a high level reference architecture of the infrastructure to be, comprising three areas (multiple use, applications and generic features) and a total of 26 functional components [26]. One of the prerequisites of this architecture is that organizations that participate in a federation of 'data stations' use the same common data model to make the data Findable, Accessible, Interoperable and Resusable

(FAIR). These FAIR data stations comprise components 7, 8 and 9 in Figure 1, i.e. the data, metadata and APIs, respectively, through which the the data station can be accessed and used.
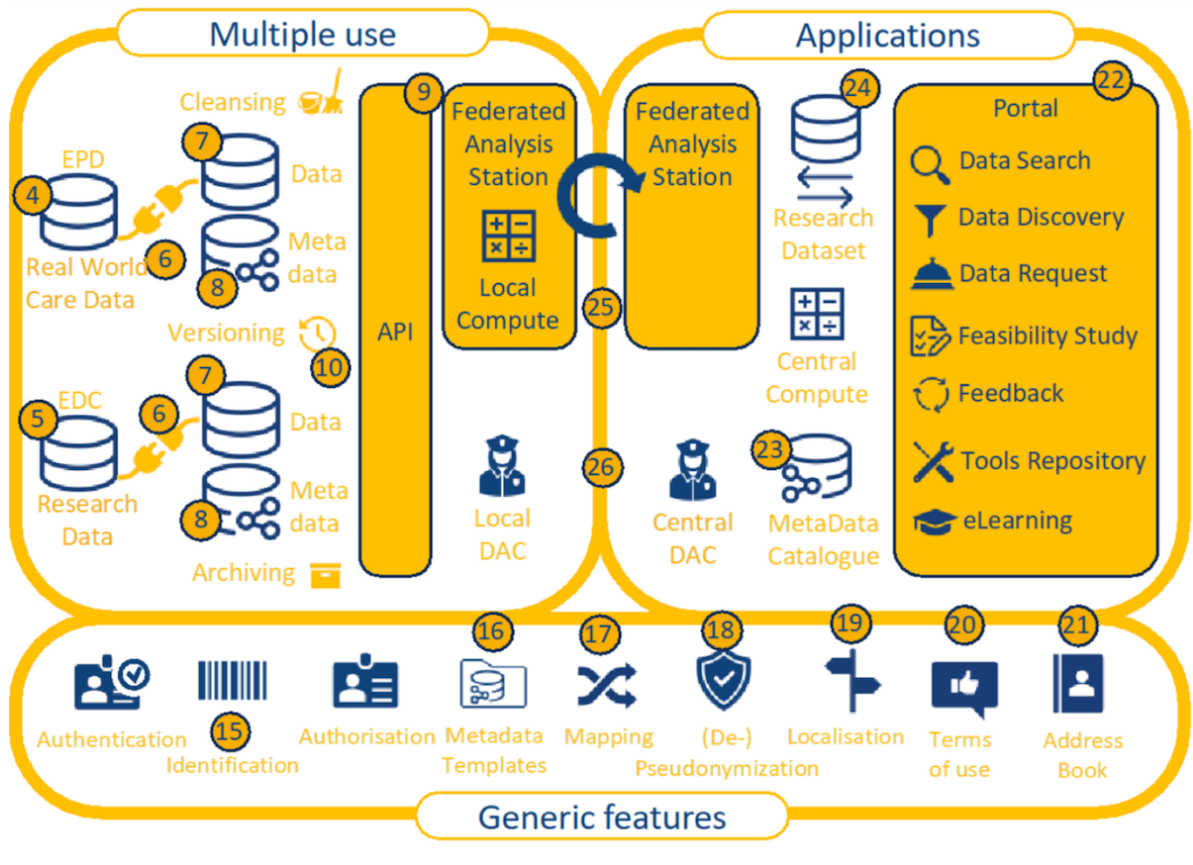

Figure 1: Reference architecture for the Dutch health data infrastructure for research and innovation [26]

Following the line of reasoning of Tsafnat et al., OMOP would be the go-to standard for storing the longitudinal data in each of the data stations, where data is transformed from the original source (component 6), stored using a common data model (component 7) and properly annotated with metadata (component 8). Indeed, by now there are quite a real-world implementations of federated learning networks based on the OHDSI-OMOP stack, including a global infrastructure with 22 centers for COVID19 prediction models [27], FeederNet in South Korea with 57 participating hospitals [28], Dutch multi-cohort dementia research with 9 centers [29], the European severe heterogeneous asthma research collaboration [30] and the recently initiated Belgian Federated Health Innovation Network (FHIN) [31].

For the PLUGIN project, however, we choose to adopt FHIR as a data model because it is more compatible with the data model of the clinical administration systems. As PLUGIN

focuses on secondary use of routine health data, we feel it is more suitable than OMOP, the latter being more suitable for clinical research data. OpenEHR might have been an option, too, if more implementations and complementary components had been available. Another reason for choosing FHIR is its practicality and extensibility to be used in a Python-based data science stack, provenance of RESTful APIs out-of-the-box to facilitate easy integration with the container-based vantage6 FL framework, and the support of many healthcare terminologies and flexibility through its profiling mechanism [32–34]. Increasingly, other projects have reported the use of FHIR for persistent, longitudinal storage for FL. A scoping review on the use of FHIR for clinical research shows that it is increasingly being used for data preparation, cohort selection and secondary data sharing [35]. The CODA platform, which aims to implement a FL infrastructure in Canada similar to the PLUGIN project, compared OMOP and FHIR and chose the latter as it has been found to support more granular mappings required for analytics [36]. The fair4health project used FHIR as part of a FAIRification workflow to simplify the process of data extraction and preparation for clinical study analyses [37].

Given that OMOP can, conceptually, be viewed as a strict subset of FHIR, hybrid solutions using a combination of OMOP and FHIR have also been reported, such as the German KETOS platform [38], and the preliminary findings from the European GenoMed4All project which aims to connect clinical and -omics data [39]. A collaboration of 10 university hospitals in Germany have shown that standardized ETL-processing from FHIR into OMOP can achieve 99% conformance [40], which confirms the feasibility of the solution pattern where FHIR acts as an intermediate sharing standard through which data from (legacy) systems are extracted and made available for reuse in a common data model. One could argue that the distinction between FHIR and OMOP becomes less relevant if data can be effectively stored in either standard. We are hopeful that initiatives like OMOP-on-FHIR indeed will foster convergence rather than collision between these two standards [41].

In the case of PLUGIN, another important consideration for choosing FHIR over OMOP is, that from a data architecture perspective, the mechanism of FHIR Profiles can be tied to principle of late binding commonly applied in data lake/warehouse architectures (Figure 2): allow ingest of widely different sources, and gradually add more constraints and validations as you move closer to a specific use case. If machine learning is the primary objective for secondary use, one wants to be able to cast a wider net of relevant data, rather than being too restrictive when ingesting the data at the start of the processing pipeline. Late binding in data warehousing is a design philosophy where data transformation and schema enforcement are deferred as late as possible in the data processing pipeline, sometimes even until query time. This approach contrasts with early binding, where data is transformed and structured as it is ingested into the data warehouse. The advantage of this design is that it allows for greater flexibility and allows us to leverage new standards and technologies using the lakehouse architecture and the composable data stack for the implmentation of the data stations (see info box 4). During the initial ingestion of the data, we only require the data to conform to the minimal syntactic standard defined by the base FHIR version (R4 in the diagram). As the data is processed, more strict checks and constraints are applied, whereby ultimately different profiles can co-exists next to one another (the two most inner rectangles), within a larger

rectangle with fewer restrictions. Note that if any of the profiles includes a FHIR extension, such as adding a field to include a birth name, the profiles are no longer strictly concentric. Hence extra care needs to be taken when dealing with extensions when applying the principle of late binding.
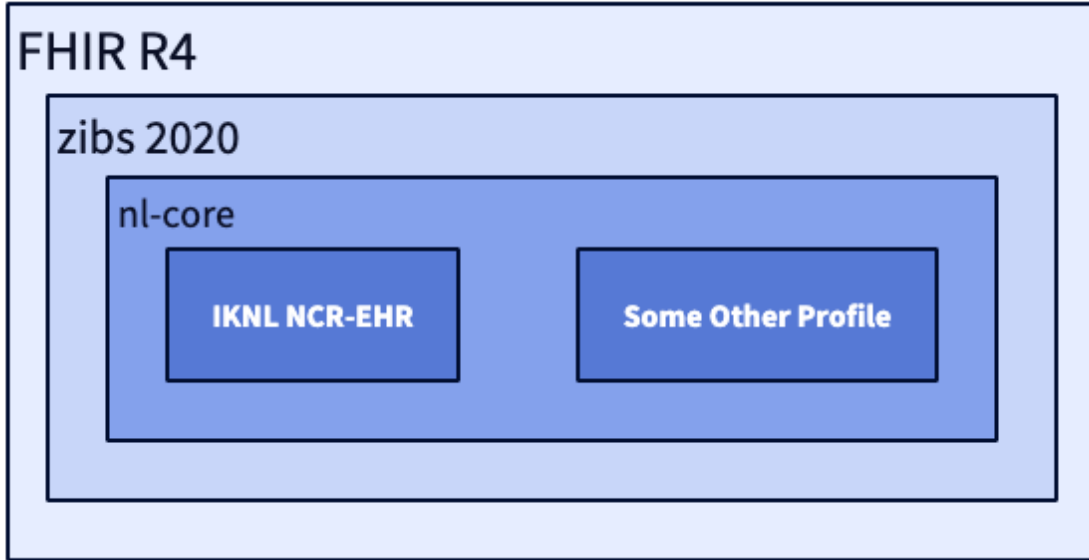

Figure 2: Principle of late binding with FHIR profiling mechanism, illustrated with FHIR Profiles that are currently in use in the Netherlands.

ℹ Textbox 4: Lakehouse architecture and the composable data stack

Data lakehouses typically have a zonal architecture that follows the Extract-Load-Transform (ELT) parttern where data is ingested from the source systems in bulk (E), delivered to storage with aligned schemas (L) and transformed into a format ready for analysis and re-use (T) [42–44]. The discerning characteristic of the lakehouse architecture is its foundation on low-cost and directly-accessible storage that also provides traditional database management and performance features such as ACID transactions, data versioning, auditing, indexing, caching, and query optimization [45].

The **composable data stack** [46] is a new set of technologies and open standards for fast processing of data using columnar data formats, including Apache Arrow as the standard columnar in-memory format with RPC-based data movement [47]; Apache Parquet as the standard columnar on-disk format [48]; and Apache Iceberg as the open table format [49,50]. This design also enables the use of new embedded, in-memory data processing engines. In turn, this opens up possibilities to bring computing workloads to edge devices,

> such as running DuckDB in the browser on top of WebAssembly [51].
> Using these technologies, full separation of storage and compute can be achieved which allows for cost-effective implementation of data stations.

One of the key challenges in using FHIR in this way pertains to the need for upgrading the whole ELT pipeline when upgrading to a new primary FHIR version, for example R6. The potential technical debt of version upgrades in the future is not specific to FHIR, but being a younger standard changes are more frequent compared to OMOP and openEHR. However, we expect that the development time required to upgrade FHIR versions is significantly less than the initial migration to FHIR.

The considerations above also show the conceptual difference of FHIR as a health data exchange standard versus openEHR as a persistent storage of routine healthcare data and OMOP as a persistent storage of health research data. For health data exchange and federated learning, the recipient of the data determines, to a large extent, what subset of data in the source needs to be made available – i.e. the target data model is known late and this favors late binding. In the case of routine collection of data, the holder of the source data determines what data needs to be stored – and typically everything – which favors early binding.

Health data standards in LMICs

It is a widely held belief that digital technologies have an important role to play in strengthening health systems in LMICs. Yet, also here the current fragmentation of health data stands in the way of scaling up digital health programs beyond project-centric, vertical solutions into sustainable health information exchanges [52]. In the context of global digital health developments, Mehl et al. have also called for convergence to open standards, similar to Tsafnat et al., but additionally stress the need for OSS (also our main argument in this paper), open content (representations of public health, health system or clinical knowledge to guide implementations) and open architectures (reusable enterprise architecture patterns for health systems) [15]. As for the open architecture, we see a convergence towards the OpenHIE framework [53], which has been adopted by many African countries as the architectural blueprint for implementing nation-wide health information exchanges (HIE) [54], including Nigeria [55], Kenya [56] and Tanzania [57]. Figure 3 shows an overview of the OpenHIE architecture.

While the OpenHIE specification is agnostic to which data standards should be used, in practice the digital health community in LMICs have converged towards FHIR as the primary standard for health information exchange, in line with the proposal by Tsafnat et al. To illustrate this point, consider the OpenHIM Platform architecture (Figure 4), which is currently the largest OSS implementation of the OpenHIE specification. In OpenHIM, clients (Point-of-Service systems) can initiate various workflows to submit or query patient data. The Shared Health Record (SHR) acts as the core transactional system for the health information exchange, which
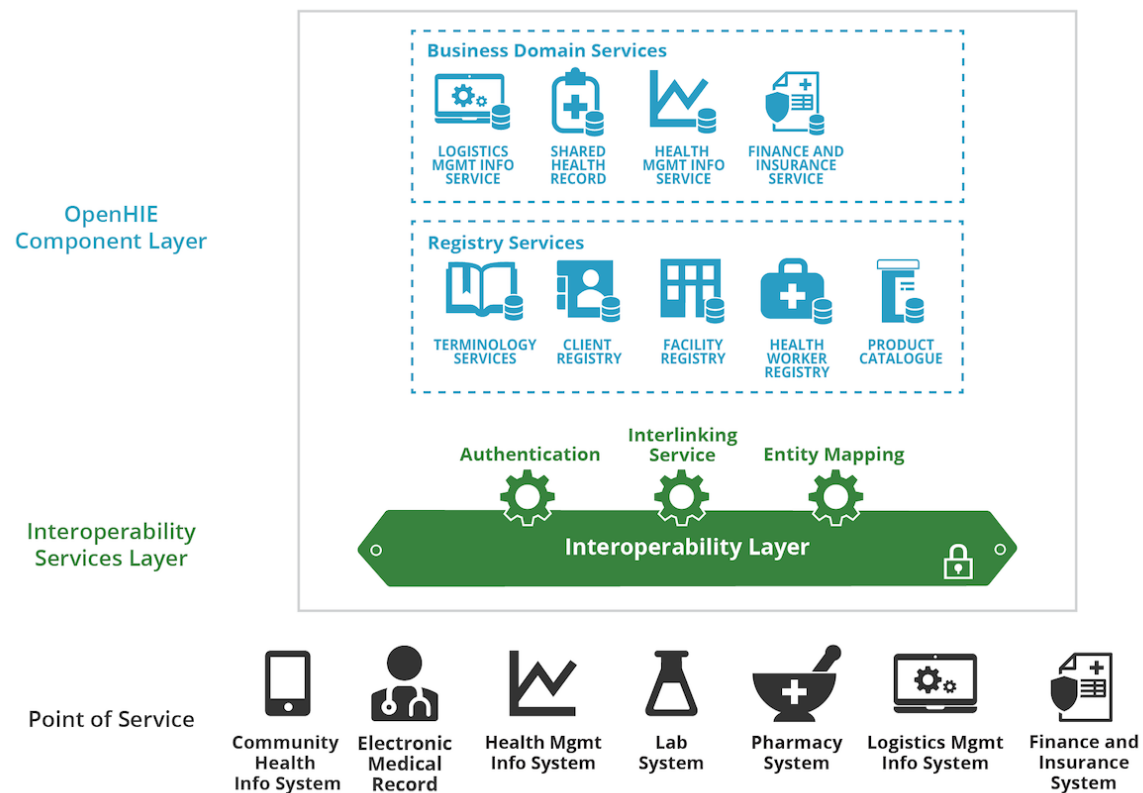
Figure 3: OpenHIE architecture showing the Point of Service systems (black), the Interoperability Layer (green) and the Component Layer (blue).

in this case is realized with the HAPI FHIR server, being one of the most widely used open-source FHIR server implementations [58].
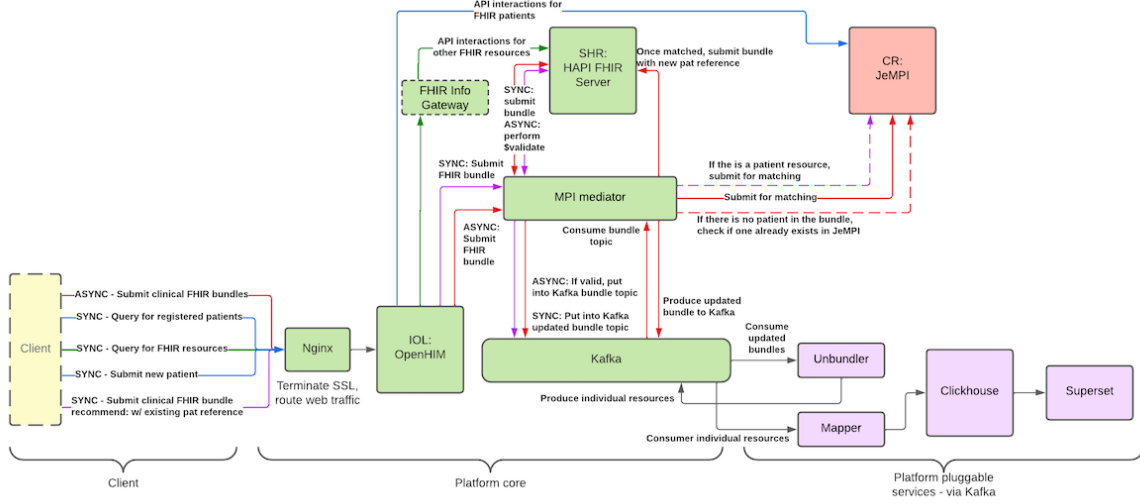

Figure 4: OpenHIM Platform Architecture, illustrating the use of FHIR-based workflows between the components as specified in OpenHIE. CR: Client Registry. IOL: Interoperability Layer. MPI: Master Patient Index. SHR: Shared Health Record. Image source: https://jembi.gitbook.io/.

Looking at the Point-of-Service systems, we see that as of today openEHR is rarely used as the standard for routine collection of clinical data in LMICs. The largest OSS electronic health record (EHR) implementations for low-resource settings are based on non-standardized data models, and it is unlikely this will change any time soon [59]. Instead, we see that FHIR-native software development frameworks such as OpenSRP [60] and the Open Health Stack [61] are being used more and more. In this approach, health professionals use Android apps to register and collect routine health data (Figure 5). As an example, OpenSRP has been deployed in 14 countries targeting various patient populations, amongst which is the implementation of the WHO antenatal and neonatal care guidelines for midwives in Lombok, Indonesia [62,63]. Beda EMR takes a similar approach and provides a FHIR native front-end that can be used in combination with any FHIR server as a backend [64]. Such a solution design is particularly useful for mid-size and smaller healthcare facilities, which are often resource constrained, lacking basic IT infrastructure to deploy a full-blown electronic medical record system. Hence the FHIR-based SHR functions as both, the administrative system-of-record and as the hub for information exchange at the same time.

Finally, regarding longitudinal data analysis, we also see a convergence towards FHIR as the primary standard in LMICs. As in the case of federated learning, the choice for FHIR to implement data warehouse and analytic platforms is the preferred method due to the widespread availability of complementary OSS components. FHIR-specific technologies such as Bulk FHIR
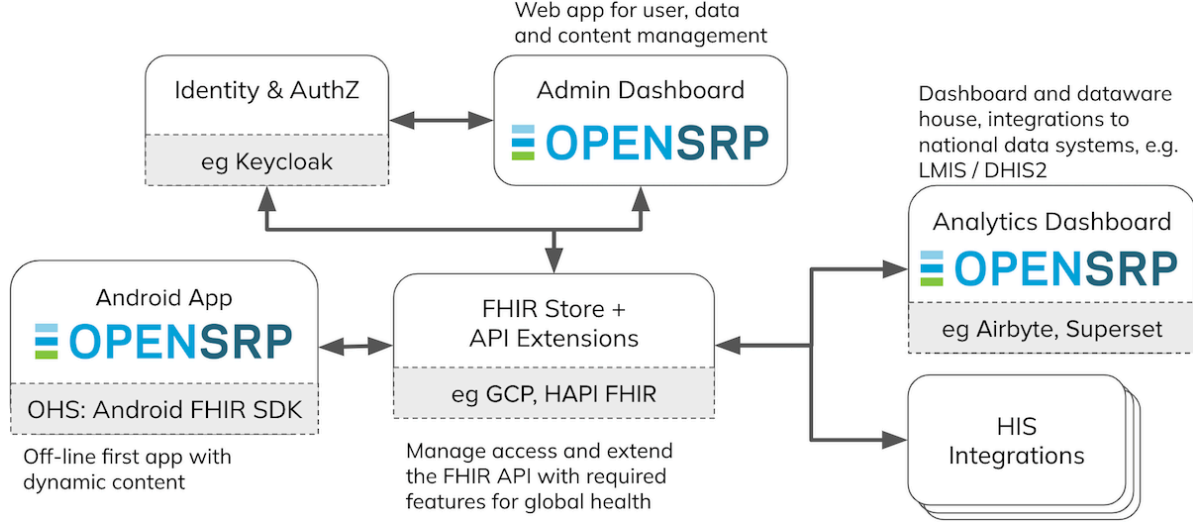
Figure 5: Overview of OpenSRP2 OSS framework for building clinical administration apps. HIS: health information systems. Image source: https://docs.opensrp.io/.

data access and SQL-on-FHIR mentioned earlier, allow the FHIR ecosystem to be used, complemented and integrated with generic OSS data warehousing components such as Clickhouse [65] and dbt [66]. Recently, more studies have pointed to the potential that FHIR brings when it is used in conjunction with machine learning and AI [67]. FHIR-based shared health records can act as systems of records for countries, thereby enabling reuse by health researchers, foundations, etc. to create public value with this data.

All in all, we see that in the context of LMICs, the standardization of the three domains put forward by Tsafnat merge into one. The SHR, as the key component within the OpenHIE specification, serves as the back-end of the system-of-record and provides a transactional, persistent storage engine for information exchange. Downstream longitudinal data stores continue to use FHIR as the common data model for analytical purposes. One could argue that it is in fact advantageous to converge to just one standard, thereby reducing complexity and cost of the total system. Such a perspective ties in with the notion of the hourglass model and open architectures: because FHIR is inherently designed to make optimal use of internet standards, such as the json file format and REST APIs, it is very modular and developer friendly. The many components that make up the FHIR allows the standard to be used effectively to implement subsystems, such as a facility registry or a health worker registry. By comparison, OMOP and openEHR are designed with a smaller scope with fewer application areas and are thereby less suitable as a standard to implement the subsystems defined in the OpenHIE specification.

Discussion, conclusion and future research

We agree with Tsafnat et al. that there is a dire need to converge to open data standards in healthcare, and support their proposal to focus on openEHR, FHIR and OMOP in healthcare informatics going forward. However, open standards are a necessary but not sufficient condition for the convergence of health data standardization. The availability of OSS implementations and complementary technologies is important when choosing which open standard to use. We find that the proposed trichotomy is too restrictive and therefore of limited use in guiding design choices to be made in real-world scenarios. Instead, we think that the full-STAC approach described by Mehl et al. is more comprehensive [15]. Furthermore, we argue that FHIR has the potential of acting as the spanning layer for health data interoperability, thereby enabling much wider standardization and adoption within the health data ecosystem at large. This is illustrated by the two cases considered in this paper, where FHIR is used beyond its original scope as a health data exchange standard.

In the case of FL, FHIR can be used interchangeably with OMOP for longitudinal analysis. Also, due to its inherently modular design, FHIR can be used in conjunction with the principle of late binding, as opposed to early binding for OMOP and openEHR, which is a relevant design criterion for implementing federated data platforms for secondary use. In the case of LMICs, we see that FHIR is emerging as the standard for all three domains of routine health data collection at the clinical point-of-service, data exchange and longitudinal analysis. We believe this is driven by the resource constrained setting in LMICs, the modularity of FHIR and the lower complexity and shallower learning curve of FHIR compared to openEHR. We expect that FHIR will play a major role in driving health data convergence in LMICs, because the availability of OSS implementations and complementary components are important enablers in these resource-constrained environments. We strongly support ongoing developments to increase the availability of OSS implementations as digital public goods [68] and integration projects such as Instant OpenHIE [69], which will improve health data interoperability in LMICs.

Although openEHR has not been chosen as the standard for the two use cases presented here, we want to stress that it is not our intention to argue for or against any of the three standards *a priori* nor do we intend to dismiss openEHR outright. Instead, our aim is to illustrate the kind of design choices and trade-offs that need to be made, particularly those related to the availability and complementary of OSS components. Significant developments and uptake of openEHR as a clinical data respository have been reported, with currently 17 openEHR solution providers that have been implemented in thousands clinics and research organizations worldwide [70]. Additionally work is underway to integrate openEHR with FHIR for data exchange [71,72]. Some experts agree that openEHR is the only specification that provides a comprehensive solution for building a standardized EHR [73]. Furthermore, openEHR is not only being deployed as a clinical system-of-record, but also as persistent clinical data repository for implementing national health information exchanges in European Nordic countries [74] and Slovenia [75], which is very similar to the solution design of the SHR within

the OpenHIE architecture presented here. An ongoing program in the south of the Netherlands has demonstrated a decentralized data sharing ecosystem using separate openEHR data stores, where federated queries are supported by the openEHR Archetype Query Language [76].

The two cases allow us to reflect and revisit the key arguments of this paper, namely, the importance of OSS implementations and availability of complementary components for wide-scale adoption of health data standards. There is an important and equally complex interplay between OSS development and standardization, where OSS implementation can occur before, after or in parallel to standardization efforts [77,78]. Various studies have provided increasing evidence that OSS is a key success factor in driving software-related standardization [78], and by extension we think it is critical when aiming to achieve data standardization. The history of how the DICOM imaging standard came to be is a good example how OSS development was pivotal in achieving wide-scale adoption of this standard [79,80].

In contrast, the phenomena of forking, fragmentation and splintering are known to hinder an industry to consolidate towards a set of open standards [81]. Given the specific characteristics of data as an artefact, fragmentation is arguably the most relevant of these phenomena. De Reuver et al. expect fragmentation to persist for some time in the evolution of data platforms and associated ecosystems [6]. The case of the Unix operating system is an interesting example where fragmentation hampered standardization, next to market dynamics and issues related to intellectual property rights [81].

But even when OSS has successfully contributed to 'tip' the healthcare industry to a set of health data standards, issues remain regarding sustainability of the OSS ecosystem itself. The market dynamics and economics of OSS ecosystems differ considerably from industry to industry: sustainability of OSS in the context of, say, the cultural and scientific heritage sector will be different to the challenges of OSS projects that are used as mission-critical components of open digital infrastructures worldwide. In case of the latter, underfunding is a critical issue and initiatives such as the German Sovereign Tech Agency have been launched to alleviate this [82]. In the context of open health data standards, we believe that risks related to underfunding are lower and more manageable. Within the digital health community there is a range of commercial companies supporting the OSS projects and creating sustainable business from it using various business models like offering support contracts, split licensing and complementary closed source products [83]. Regarding the dynamics of forking and fragmentation mentioned earlier, we feel that code forking on balance has a net positive effect on long-term sustainability of OSS at the level of the software itself, the community and the ecosystem [84].

Going forward, we suggest the following directions for future research. Given that health data standardization will continue to require mappings, we propose to explore the use of machine learning, and particularly large-language models, as a means to reduce the development effort required to create transformations between various health data formats. New machine learning methods can also be developed to assess and improve data quality across the various stages of the data processing pipelines. In terms of data integration, we expect that health data will increasingly be used in conjunction with data from social services and the welfare domain, which requires new techniques to integrate different data domains, for example using

knowledge graphs and ontologies. Last, but certainly not least, future research should not only explore the technical but also the social implications of implementing OSS components for data standardization across the healthcare system, specifically in settings where governance or ethical considerations of data interoperability have not specifically been addressed at a regulatory level. In line with the embedding of open standards in the open-source ecosystem, we assert that the benefits of health data standardization will only be realized if they are coupled with collaborative, community-driven governance models. It remains essential to ensure that the development, adoption, and evolution of standards remain inclusive, transparent, and responsive to the diverse needs within the health system.

Authors' Contributions

DK contributed to the concept and design of the manuscript and prepared the first draft. AD, MS and BJV contributed to the section on federated learning. FH and MB contributed to the section on LMICs. DK and FH revised the manuscript based on the feedback from the peer reviewers. All authors contributed to the final revision and approved the final manuscript.

Acknowledgements

Conflicts of interests

DK received funding from PharmAccess as a contractor to conduct the work on LMICs reported here. MB/Ona is the core developer of the open-source OpenSRP 2 framework.

Abbreviations

ACID: Atomicity, Consistency, Isolation, Durability API: Application Programming Interface EHDS: European Health Data Space EHR: Electronic Health Record ELT: Extract, Load, Transform FAIR: Findability, Accessibility, Interoperability, and Reusability FHIR: Fast Healthcare Interoperability Resources FL: Federated Learning Full-STAC: Concept that advocates for open standards, open technology, open architecture and open content GSM: Global System for Mobile Communications HIE: Health Information Exchange HL7: Health Level 7 LMIC: Low and middle-income countries ML: Machine Learning OHDSI: Observational Health Data Sciences and Informatics OMOP: Observational Medical Outcomes Partnership OSS: Open-Source Software SHR: Shared Health Record REST: representational state transfer TCP/IP: Transmission Control Protocol/Internet Protocol

References

1. Tsafnat G, Dunscombe R, Gabriel D, Grieve G, Reich C. Converge or Collide? Making Sense of a Plethora of Open Data Standards in Health Care. *Journal of Medical Internet Research.* 2024;26(1):e55779. doi:10.2196/55779

2. de Reuver M, Sørensen C, Basole RC. The Digital Platform: A Research Agenda. *Journal of Information Technology.* 2018;33(2):124-135. doi:10.1057/s41265-016-0033-3

3. Keller P, Tarkowski A. The Paradox of Open. *Open Future.* Published online March 5, 2021. Accessed March 25, 2024. https://openfuture.pubpub.org/pub/paradox-of-open/release/1

4. Estrin D, Sim I. Health care delivery. Open mHealth architecture: An engine for health care innovation. *Science.* 2010;330(6005):759-760. doi:10.1126/science.1196187

5. Beck M. On the hourglass model. *Communications of the ACM.* 2019;62(7):48-57. doi:10.1145/3274770

6. de Reuver M, Ofe H, Agahari W, Abbas AE, Zuiderwijk A. The openness of data platforms: A research agenda. In: *Proceedings of the 1st International Workshop on Data Economy.* DE '22. Association for Computing Machinery; 2022:34-41. doi:10.1145/3565011.3569056

7. Reynolds CJ, Wyatt JC. Open Source, Open Standards, and Health Care Information Systems. *Journal of Medical Internet Research.* 2011;13(1):e1521. doi:10.2196/jmir.1521

8. GSM. In: *Wikipedia.*; 2024. Accessed September 20, 2024. https://en.wikipedia.org/w/index.php?title=GSM&oldid=1245675274

9. *The State of FHIR 2024 Survey Results.* HL7; 2024. Accessed April 4, 2024. https://www.hl7.org/documentcenter/public/white-papers/2024%20StateofFHIRSurveyResults_final.pdf

10. Carmo A, Martins H. *D2.2 - EHRxF in a Nutshell-WP2-ISCTE.*; 2024. https://ehr-exchange-format.eu/wp-content/uploads/2024/10/D2.2-v20240704-EHRxF-in-a-Nutshell-WP2-ISCTE.pdf

11. Firely. *FHIR in US Healthcare Regulations.*; 2023. Accessed May 30, 2024. https://simplifier.net/organization/firely/news/153

12. *National Digital Health Mission.* India National Health Authority; 2020. https://www.niti.gov.in/sites/default/files/2023-02/ndhm_strategy_overview.pdf

13. *HCX Protocol V0.9.*; 2023. Accessed September 18, 2024. http://hcxprotocol.io/

14. Tilahun B, Mamuye A, Yilma T, Shehata Y. *African Union Health Information Exchange Guidelines and Standards.*; 2023. https://africacdc.org/download/african-union-health-information-exchange-guidelines-and-standards/

15. Mehl GL, Seneviratne MG, Berg ML, et al. A full-STAC remedy for global digital health transformation: Open standards, technologies, architectures and content. *Oxford Open Digital Health.* 2023;1:oqad018. doi:10.1093/oodh/oqad018

16. Mandl KD, Gottlieb D, Mandel JC, et al. Push Button Population Health: The SMART/HL7 FHIR Bulk Data Access Application Programming Interface. *npj Digital Medicine.* 2020;3(1):1-9. doi:10.1038/s41746-020-00358-4

17. Jones J, Gottlieb D, Mandel JC, et al. A landscape survey of planned SMART/HL7 bulk FHIR data access API implementations and tools. *Journal of the American Medical Informatics Association.* 2021;28(6):1284-1287. doi:10.1093/jamia/ocab028

18. *SQL on FHIR V2.0.0-Pre.* Accessed September 20, 2024. https://build.fhir.org/ig/FHIR/sql-on-fhir-v2/

19. FHIR Open Source Implementations. September 20, 2024. Accessed September 20, 2024. https://confluence.hl7.org/display/FHIR/Open+Source+Implementations

20. Software Tools – OHDSI. Accessed September 20, 2024. https://www.ohdsi.org/software-tools/

21. Beale SH Thomas. openEHR Platform. Accessed September 20, 2024. https://openehr.org/products_tools/platform/

22. EHRbase 2.0 website. Published online March 19, 2024. Accessed September 20, 2024. https://www.ehrbase.org/

23. Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. *npj Digit Med.* 2020;3(1, 1):1-7. doi:10.1038/s41746-020-00323-1

24. Teo ZL, Jin L, Liu N, et al. Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture. *Cell Reports Medicine.* 2024;5(2):101419. doi:10.1016/j.xcrm.2024.101419

25.   PLUGIN – Platform voor Uitwisseling en Hergebruik van Klinische Data Nederland. Accessed March 14, 2025. https://plugin.healthcare/

26.   Health-RI. Agreements on the National Health Data Infrastructure for Research, Policy and Innovation - Health-RI Nationale Gezondheidsdata-infrastructuur - Confluence. January 29, 2024. Accessed June 3, 2024. https://health-ri.atlassian.net/wiki/spaces/HNG/pages/249073646/Agreements+on+the+National+Health+Data+Infrastructure+for+Research+Policy+and+Innovation

27.   Khalid S, Yang C, Blacketer C, et al. A standardized analytics pipeline for reliable and rapid development and validation of prediction models using observational health data. *Computer Methods and Programs in Biomedicine.* 2021;211:106394. doi:10.1016/j.cmpb.2021.106394

28.   Lee S, Kim C, Chang J, Park RW. FeederNet (Federated E-Health Big Data for Evidence Renovation Network) platform in Korea – OHDSI. 2022. Accessed June 4, 2024. https://www.ohdsi.org/2022showcase-33/

29.   Mateus P, Moonen J, Beran M, et al. Data harmonization and federated learning for multi-cohort dementia research using the OMOP common data model: A Netherlands consortium of dementia cohorts case study. *Journal of Biomedical Informatics.* 2024;155:104661. doi:10.1016/j.jbi.2024.104661

30.   Kroes JA, Bansal AT, Berret E, et al. Blueprint for harmonising unstandardised disease registries to allow federated data analysis: Prepare for the future. *ERJ Open Research.* 2022;8(4). doi:10.1183/23120541.00168-2022

31.   Deltomme C, Denturck K, De Jaeger P, et al. Federated Health Innovation Network (FHIN). Published online September 20, 2024. https://www.ohdsi-europe.org/images/symposium-2024/Posters/poster%20OHDSI%20FHIN%20Camille%20Deltomme%20-%20Camille%20Deltomme.pdf

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;2020:870-877. Accessed September 21, 2024. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8075508/

33.   Choudhury A, van Soest J, Nayak S, Dekker A. Personal Health Train on FHIR: A Privacy Preserving Federated Approach for Analyzing FAIR Data in Healthcare. In: Bhattacharjee A, Borgohain SKr, Soni B, Verma G, Gao XZ, eds. *Machine Learning, Image Processing, Network Security and Data Sciences.* Communications in Computer and Information Science. Springer; 2020:85-95. doi:10.1007/978-981-15-6315-7_7

34. Smits D, Van Beusekom B, Martin F, Veen L, Geleijnse G, Moncada-Torres A. An Improved Infrastructure for Privacy-Preserving Analysis of Patient Data. In: Mantas J, Gallos P, Zoulias E, et al., eds. *Studies in Health Technology and Informatics.* IOS Press; 2022. doi:10.3233/SHTI220682

35. Duda SN, Kennedy N, Conway D, et al. HL7 FHIR-based tools and initiatives to support clinical research: A scoping review. *Journal of the American Medical Informatics Association.* 2022;29(9):1642-1653. doi:10.1093/jamia/ocac105

36. Mullie L, Afilalo J, Archambault P, et al. CODA: An open-source platform for federated analysis and machine learning on distributed healthcare data. *Journal of the American Medical Informatics Association.* Published online December 21, 2023:ocad235. doi:10.1093/jamia/ocad235

37. Sinaci AA, Gencturk M, Alvarez-Romero C, et al. Privacy-preserving federated machine learning on FAIR health data: A real-world application. *Computational and Structural Biotechnology Journal.* 2024;24:136-145. doi:10.1016/j.csbj.2024.02.014

38. Gruendner J, Schwachhofer T, Sippl P, 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(10):e0223010. doi:10.1371/journal.pone.0223010

39. Cremonesi F, Planat V, Kalokyri V, et al. The need for multimodal health data modeling: A practical approach for a federated-learning healthcare platform. *Journal of Biomedical Informatics.* 2023;141:104338. doi:10.1016/j.jbi.2023.104338

40. 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. *International Journal of Medical Informatics.* 2023;169:104925. doi:10.1016/j.ijmedinf.2022.104925

41. OMOPonFHIR. Accessed September 20, 2024. https://omoponfhir.org/

42. Hai R, Koutras C, Quix C, Jarke M. Data Lakes: A Survey of Functions and Systems. *IEEE Transactions on Knowledge and Data Engineering.* 2023;35(12):12571-12590. doi:10.1109/TKDE.2023.3270101

43. Harby AA, Zulkernine F. From Data Warehouse to Lakehouse: A Comparative Review. In: *2022 IEEE International Conference on Big Data (Big Data).* IEEE; 2022:389-395. doi:10.1109/BigData55660.2022.10020719

44. Harby AA, Zulkernine F. Data Lakehouse: A Survey and Experimental Study. doi:10.2139/ssrn.4765588

45. Armbrust M, Ghodsi A, Xin R, Zaharia M. Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. In:; 2021:8. https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf

46. Pedreira P, Erling O, Karanasos K, et al. The Composable Data Management System Manifesto. *Proc VLDB Endow.* 2023;16(10):2679-2685. doi:10.14778/3603581.3603604

47. *Apache Arrow.*; 2024. Accessed September 20, 2024. https://arrow.apache.org/

48. *Apache Parquet.*; 2024. Accessed September 20, 2024. https://parquet.apache.org/

49. Jain P, Kraft P, Power C, Das T, Stoica I, Zaharia M. Analyzing and Comparing Lakehouse Storage Systems. In:; 2023. https://www.cidrdb.org/cidr2023/papers/p92-jain.pdf

50. *Apache Iceberg.* Accessed September 20, 2024. https://iceberg.apache.org/

51. An in-process SQL OLAP database management system. Accessed October 10, 2024. https://duckdb.org/

52. Karamagi HC, Muneene D, Droti B, et al. eHealth or e-Chaos: The use of Digital Health Interventions for Health Systems Strengthening in sub-Saharan Africa over the last 10 years: A scoping review. *J Glob Health.* 2022;12:04090. doi:10.7189/jogh.12.04090

53. *OpenHIE Framework V5.2-En.*; 2024. Accessed August 27, 2024. https://ohie.org/

54. Mamuye AL, Yilma TM, Abdulwahab A, et al. Health information exchange policy and standards for digital health systems in africa: A systematic review. *PLOS Digital Health.* 2022;1(10):e0000118. doi:10.1371/journal.pdig.0000118

55. Dalhatu I, Aniekwe C, Bashorun A, et al. From Paper Files to Web-Based Application for Data-Driven Monitoring of HIV Programs: Nigeria's Journey to a National Data Repository for Decision-Making and Patient Care. *Methods Inf Med.* 2023;62(03/04):130-139. doi:10.1055/s-0043-1768711

56. Thaiya MS, Julia K, Joram M, Benard M, Nambiro DA. Adoption of ICT to Enhance Access to Healthcare in Kenya. *IOSR-JCE.* 2021;23(2):45-50.

57. Nsaghurwe A, Dwivedi V, Ndesanjo W, et al. One country's journey to interoperability: Tanzania's experience developing and implementing a national health information exchange. *BMC Medical Informatics and Decision Making.* 2021;21(1):139. doi:10.1186/s12911-021-01499-6

58. HAPI FHIR - The Open Source FHIR API for Java. Accessed September 20, 2024. https://hapifhir.io/

59. Syzdykova A, Malta A, Zolfo M, Diro E, Oliveira JL. Open-Source Electronic Health Record Systems for Low-Resource Settings: Systematic Review. *JMIR Medical Informatics.* 2017;5(4):e44. doi:10.2196/medinform.8131

60. Mehl G. Open Smart Register Platform (OpenSRP). *OpenSRP.* 2020;5:42-43. Accessed January 21, 2023. https://lib.digitalsquare.io/handle/123456789/77592

61. Open Health Stack. Accessed September 20, 2024. https://developers.google.com/open-health-stack

62. Development SI for. BUNDA App. May 9, 2023. Accessed January 18, 2024. https://www.sid-indonesia.org/post/bunda-app

63. Kurniawan K, FitriaSyah I, Jayakusuma AR, et al. Midwife service coverage, quality of work, and client health improved after deployment of an OpenSRP-driven client management application in Indonesia. In: Atlantis Press; 2019:155-162. doi:10.2991/ichs-18.2019.21

64. Beda EMR. Accessed December 30, 2024. https://beda.software/emr

65. ClickHouse. Clickhouse: Fast Open-Source OLAP DBMS. Accessed September 20, 2024. https://clickhouse.com

66. Dbt. Accessed September 20, 2024. https://www.getdbt.com/index

67. Balch JA, Ruppert MM, Loftus TJ, et al. Machine Learning–Enabled Clinical Information Systems Using Fast Healthcare Interoperability Resources Data Standards: Scoping Review. *JMIR Medical Informatics.* 2023;11(1):e48297. doi:10.2196/48297

68. Digital Public Goods Alliance. 2024. Accessed February 5, 2024. https://digitalpublicgoods.net/

69. Instant OpenHIE V2. Published online July 3, 2024. Accessed September 20, 2024. https://jembi.gitbook.io/instant-v2/

70. Delussu G, Frexia F, Mascia C, et al. A survey of openEHR Clinical Data Repositories. *International Journal of Medical Informatics.* 2024;191:105591. doi:10.1016/j.ijmedinf.2024.105591

71. FHIR Connect specfication. Published online October 10, 2024. Accessed February 4, 2025. https://github.com/better-care/fhir-connect-mapping-spec

72. Welcome to openFHIR's documentation! — openFHIR 0.9.3 documentation. Accessed February 4, 2025. https://open-fhir.com/documentation/index.html

73. Pedrera-Jiménez M, García-Barrio N, Frid S, et al. Can OpenEHR, ISO 13606, and HL7 FHIR Work Together? An Agnostic Approach for the Selection and Application of Electronic Health Record Standards to the Next-Generation Health Data Spaces. *Journal of Medical Internet Research.* 2023;25(1):e48702. doi:10.2196/48702

74. Pohjonen H. Norway, Sweden, and Finland as forerunners in open ecosystems and openEHR. In: Hovenga E, Grain H, eds. *Roadmap to Successful Digital Health Ecosystems.* Academic Press; 2022:457-471. doi:10.1016/B978-0-12-823413-6.00011-2

75. Bajrić S. Building a Sustainable Ecosystem for eHealth in Slovenia: Opportunities, Challenges, and Strategies. *DIGITAL HEALTH.* 2023;9:20552076231205743. doi:10.1177/20552076231205743

76. *Demonstratie Proof of Concept Regionaal Data-Ecosysteem Zuid-Limburg.*; 2025. Accessed February 26, 2025. https://www.youtube.com/watch?v=jT5UTLRX5VQ

77. Wright SA, Druta D. Open source and standards: The role of open source in the dialogue between research and standardization. In: *2014 IEEE Globecom Workshops (GC Wkshps).* IEEE; 2014:650-655. doi:10.1109/GLOCOMW.2014.7063506

78. Blind K, Böhm M. *The Relationship Between Open Source Software and Standard Setting.* Publications Office of the European Union; 2019. doi:10.2760/163594

79. Erickson BJ, Langer S, Nagy P. The Role of Open-Source Software in Innovation and Standardization in Radiology. *Journal of the American College of Radiology.* 2005;2(11):927-931. doi:10.1016/j.jacr.2005.05.004

80. Nagy P. Open Source in Imaging Informatics. *J Digit Imaging.* 2007;20(S1):1-10. doi:10.1007/s10278-007-9056-1

81. Simcoe T, Watson J. Forking, Fragmentation, and Splintering. *Strategy Science.* 2019;4(4):283-297. doi:10.1287/stsc.2019.0094

82. Sovereign Tech Agency. February 27, 2025. Accessed February 28, 2025. https://www.sovereign.tech/

83. Chang V, Mills H, Newhouse S. From Open Source to long-term sustainability: Review of Business Models and Case studies. In:; 2007. Accessed February 28, 2025. https://eprints.soton.ac.uk/263925/

84. Nyman L, Lindman J. Code Forking, Governance, and Sustainability in Open Source Software. *Technology Innovation Management Review.* 2013;January. http://timreview.ca/article/644