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 implementations 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, with as yet limited adoption of OMOP and openEHR. 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 ecosystem 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 clinical administration as well. 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 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 emphasis on open-source implementations 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 open-source pieces of software, 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. 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 [9]. 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 [10,11]. The African Union recommends all new implementations and digital health system improvements use FHIR as the primary mechanism for data exchange [12], but doesn't say anything about the use of, for example, openEHR for administrative point-of-service systems. The upcoming legislation on the European Health Data Space (EHDS) mandates interoperability between electronic health record systems but has not specified which standard is to be used, although FHIR and openEHR have both been mentioned in the legislative discussion. At the time of writing, the results from the HealthData@ EU Pilot regarding interoperability standards was still unavailable [13].

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 [14]. 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 open-source 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 [15,16], and the SQL-on-FHIR specification to make large-scale analysis of FHIR data accessible to a larger audience and portable between systems [17].

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 work-**

loads 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.

It has also led to more people voluntarily contributing to FHIR-related open-source 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 not yet profited 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 open-source components listed on the website of the three governing bodies HL7 [18], OHDSI [19] and openEHR [20], indicates that the ecosystem of FHIR and OMOP have a significantly larger offering of extensible and complementary open-source components than openEHR, although for the latter notable mature open-source implementation are also emerging such as EHRbase [21]. Taking GitHub as an indicator of worldwide development activities, we find that a search on "FHIR" yields 8.2 thousand results, "OMOP or OHDSI" has one thousand results, and "openEHR" returns 400 results. Note that these numbers should be taken as rough indicators. Indeed, given that the FHIR standard is more modular, we would expect more GitHub projects than for, say, openEHR.

In sum, we stress that beyond evaluating the intrinsic structure of an open standard and the community that supports the standard, we need to take into account the wider ecosystem of open-source 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 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 highly sensitive and its usage is tightly regulated.

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 serve 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 analysis occur 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 [22,23]. Based on ongoing work with the PLUGIN healthcare consortium [24], we have detailed an architecture for FL for secondary use of health data for hospitals in the Netherlands. The starting point for this implementation are 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 [25]. 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 (for details please refer to [25]). 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.

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 in common data model (component 7) and properly annotated with metadata (component 8). Indeed, by now there are quite a few reports of real-world implementations of federated learning networks based on the OHDSI-OMOP stack, including a global infrastructure with 22 centres for COVID19 prediction models [26], FeederNet in South Korea with 57 participating hospitals [27], Dutch multi-cohort dementia research with 9 centres [28], the European severe heterogeneous asthma research collaboration [29] and the recently initiated Belgian Federated Health Innovation Network (FHIN) [30].

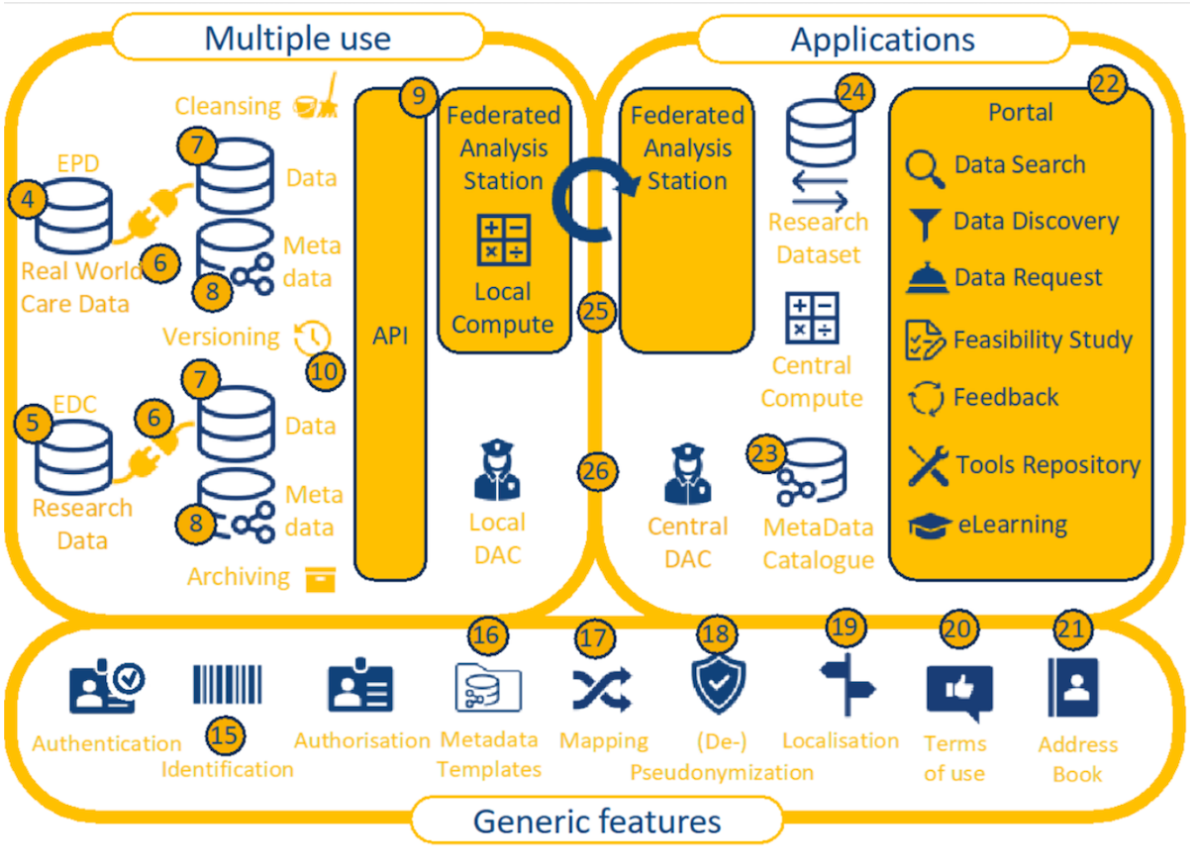

Figure 1: Reference architecture for the Dutch health data infrastructure for research and innovation [25]

For the PLUGIN project, however, we choose to adopt FHIR because the data model 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 the profiling mechanism [31–33]. Increasingly, other projects have reported the use of FHIR for persistent, longitudinal storage for FL. 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 [34]. The fair4health project used FHIR as part of a FAIRification workflow to simplify the process of data extraction and preparation for clinical study analyses [35].

Given that OMOP can be conceptually 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 [36], and the preliminary findings from the European GenoMed4All project which aims to connect clinical and -omics data [37]. A collaboration of 10 university hospitals in Germany have shown that standardized ETL-processing from FHIR into OMOP can achieve 99% conformance [38], 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 [39].

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, we want 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 advantages of this design is that it allows for greater flexibility. 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 circles), within a larger circle with fewer restrictions. Note that if any of the profiles includes a FHIR extension, such as adding a field to include a female's maiden 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.

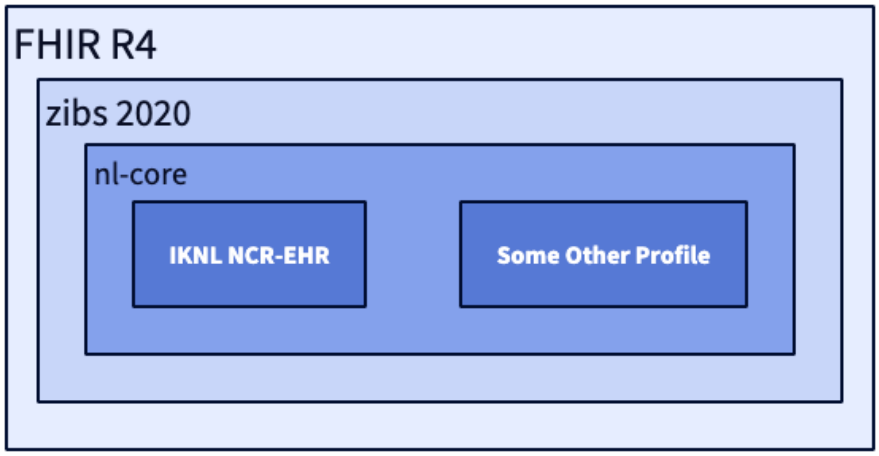

Figure 2: Principle of late binding with FHIR profiling mechanism, illustrated with FHIR Profiles that are currently in use in the Netherlands.

We found that this principle of late binding also allows flexible and efficient implementations of the data stations that make use of the current best practices of a lakehouse architecture of [40–42] and the composable data stack [43]. Lakehouses typically have a zonal architecture that follow the Extract-Load-Transform pattern (ELT) 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 (T) [40]. 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 [44]. Lakehouses thus combine the key benefits of data lakes and data warehouses: low-cost storage in an open format accessible by a variety of systems from the former, and powerful management and optimization features from the latter. By explicitly aligning the mechanism of FHIR Profiles with this design pattern of a data lakehouse enables us to use complementary standards and open-source components, most notably Apache Arrow as the standard columnar in-memory format with RPC-based data movement [45]; Apache Parquet as the standard columnar on-disk format [46]; and Apache

Iceberg as the open table format [47,48]. This design also enables the use of new embedded, in-process data processing engines, which in turn opens up possibilities to bring computing workloads to edge devices, such as running DuckDB in the browser on top of WebAssembly [49].

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 above considerations 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 available in the source needs to be made available – i.e. the target data model is known late and this favors late binding. In a persistent storage setting, 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 programmes beyond project-centric, vertical solutions into sustainable health information exchanges [50]. 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 open-source technologies (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) [14]. As for the open architecture, we see a convergence towards the OpenHIE framework [51], which has been adopted by many sub-Saharan African countries as the architectural blueprint for implementing nation-wide health information exchanges (HIE) [52], including Nigeria [53], Kenya [54] and Tanzania [55]. Figure 3 shows an overview of the OpenHIE architecture.

While the OpenHIE specification is agnostic to which data standard should be used, in practice the digital health community in LMICs have *de facto* 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 open-source implementation of the OpenHIE specification. 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

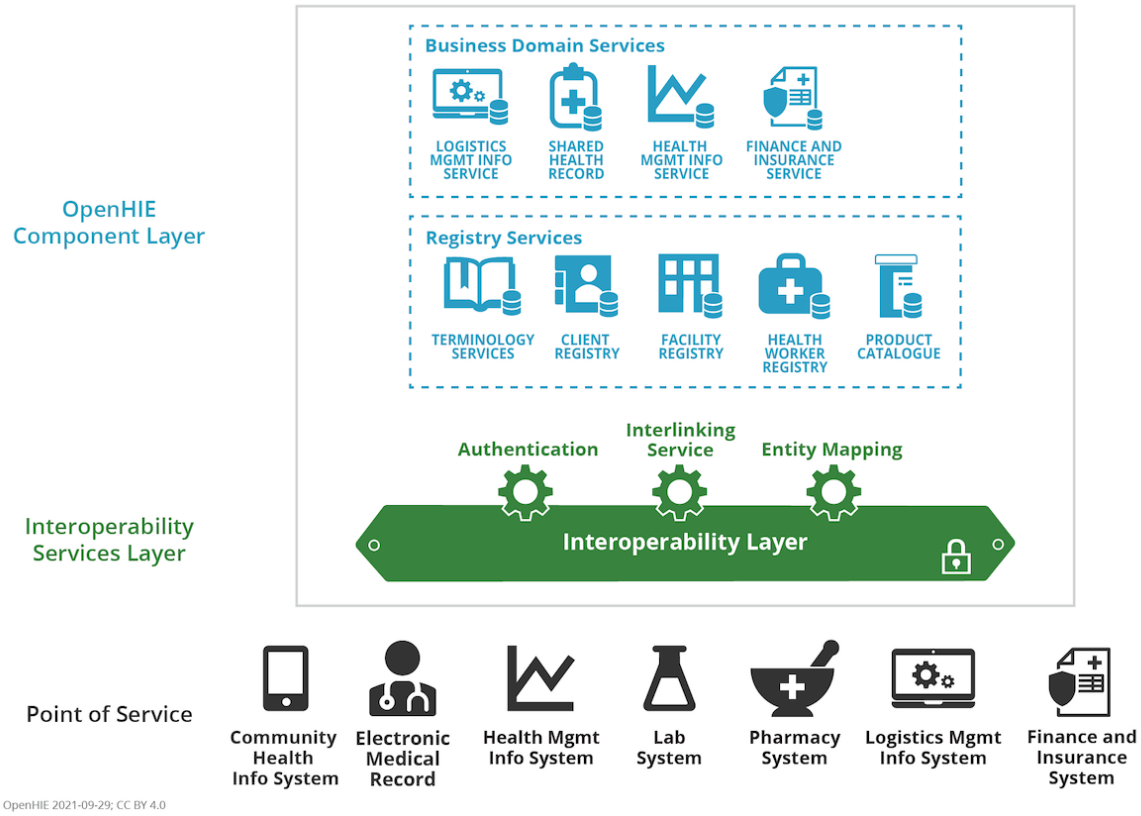

Figure 3: OpenHIE architecture showing the Point of Service systems (black), the Interoperability Layer (green) and the Component Layer (blue).

in this case is realized with the HAPI FHIR server, being one of the most widely used open-source FHIR server implementations [56].

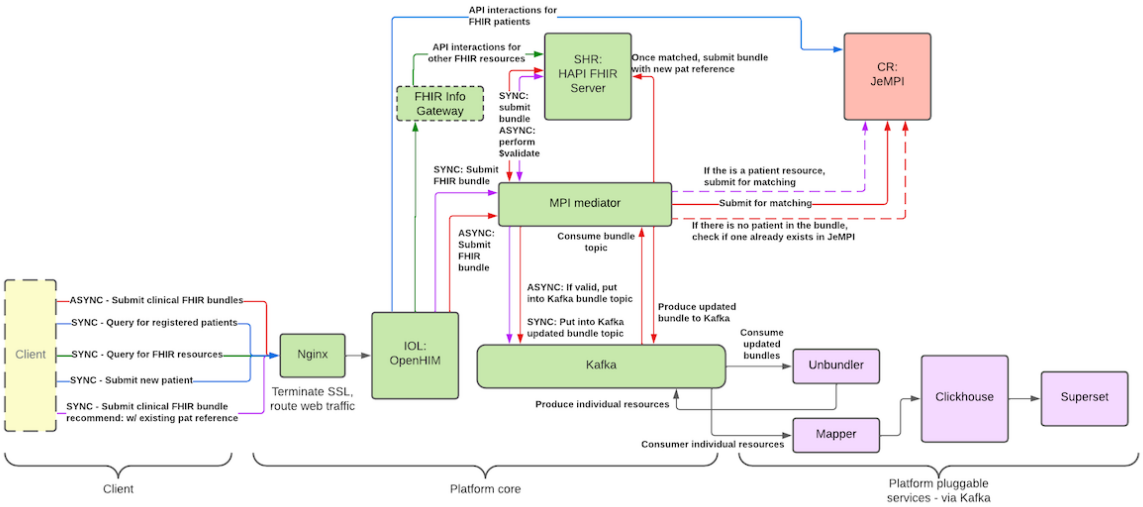

Figure 4: OpenHIM Platform Architecture, illustrating the use of FHIR-based workflows between the components as specified in OpenHIE. CR: Client Registry. IOL: Interoperability Layer. MPI: Master Patient Index. SHR: Shared Health Record. Image taken from https://jembi.gitbook.io/.

Looking at the Point-of-Service systems, we see that as of today openEHR is rarely used as the standard for clinical administration in LMICs. The largest open-source electronic health record (EHR) implementations are based on proprietary data models, and it is unlikely this will change any time soon [57]. Instead, we see that FHIR-native software development frameworks such as OpenSRP [58] and the Open Health Stack [59] 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 a reference implementation of the WHO antenatal and neonatal care guidelines for midwives in Lombok, Indonesia [60,61]. 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 [62]. Such asolution 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, by necessity, the FHIR-based SHR functions as 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 open-source technologies. FHIR-specific technologies such as Bulk FHIR data access and SQL-on-FHIR mentioned earlier, allow the FHIR ecosystem to be

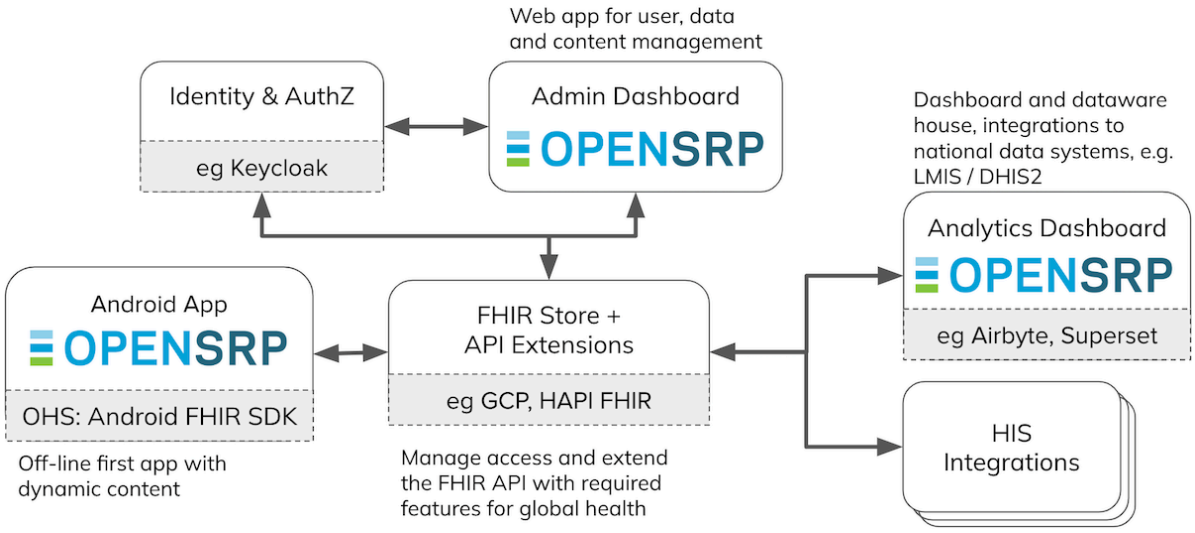

Figure 5: Overview of OpenSRP2 open-source framework for building clinical administration apps. HIS: health information systems. Image source: https://docs.opensrp.io/.

used, complemented and integrated with generic open-source data warehousing technologies such as Clickhouse [63] and dbt [64]. Increasingly more studies have pointed to the potential that FHIR can bring when it is used in conjunction with machine learning and AI [65]. 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 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 less modular in their design and are thereby less suitable as a standard to implement the subsystems defined in the OpenHIE specification.

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 health-

care informatics going forward. However, open standards are a necessary but not sufficient condition for the convergence of health data standardization. The availability of open-source implementations and complementary technologies are as 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 [14]. Furthermore, we argue that FHIR has the potential of acting as the spanning layer within the open health data system at large, thereby enabling much wider standardization and adoption within the health data ecosystem at large. This is illustrated by the two cases considered here, 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 clinical administration, data exchange and longitudinal analysis. We expect that FHIR will play a major role in driving health data convergence in LMICs, because the availability of open-source implementations and complementary components are important enablers in these resource-constrained environments. We strongly support ongoing developments to increase the availability of open-source implementations as digital public goods [66] and integration projects such as Instant OpenHIE [67], through which we have a fighting chance to move the needle in health data standardization for LMICs.

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 open-source 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. All authors contributed to the final revision and approved the final manuscript.

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

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

10. *National Digital Health Mission*. India National Health Authority; 2020.

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

12. Tilahun B, Mamuye A, Yilma T, Shehata Y. *African Union Health Information Exchange Guidelines and Standards.*; 2023.

13. Recommendations of standards for data interoperability, querying and exchange and on QC/QA & provenance (WP8) - EHDS2 Pilot - Official website. December 17, 2024. Accessed December 30, 2024. https://ehds2pilot.eu/upcoming_results/recommendations-of-standards-for-data-interoperability-querying-and-exchange-2/

14. 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

15. 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

16. 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

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

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

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

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

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

22. 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

23. 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

24. PLUGIN – Platform voor Uitwisseling en Hergebruik van Klinische Data Nederland. Accessed September 21, 2024. https://plugin.healthcare/

25. 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

26. 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

27. 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/

28. 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

29. 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

30. 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

31. 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/

32. 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

33. 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

34. 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

35. 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

36. 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

37. 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

38. 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

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

40. 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

41. 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

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

43. 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

44. 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.

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

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

47. Jain P, Kraft P, Power C, Das T, Stoica I, Zaharia M. Analyzing and Comparing Lakehouse Storage Systems. Published online 2023.

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

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

50. 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

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

52. 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

53. 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

54. 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.

55. 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

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

57. 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

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

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

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

61. 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

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

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

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

65. 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

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

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