Design and implemention of open health data systems with standardized data and analytics capabilities for low- and middle income countries

Daniel Kapitan Femke Heddema Julie Fleischer
Irni Gemzon Chris Ihure Steven Wanyee
Alessandro Pietrobon Ryan Chrichton John Grimes

TO DO: add abstract.

Introduction

The need for standardized open health data systems

It is a widely held belief in the global health community that digital technologies have an important role to play in strengthening healthcare in low- and middle income countries (LMICs). Digital technologies have the potential to increase the availability, accessibility, acceptability, and quality of health services; make healthcare more preventive, personalised, and mobile; and enfranchise patients and communities, particularly those who are most vulnerable (Kickbusch et al. 2021).

Yet, to date the global digital health ecosystem is still project-centric, resulting in data fragmentation and technology lock-in, compromising health care delivery (Mehl et al. 2023). There is no dearth of the number of digital health interventions: sub-Saharan Africa has seen over 700 such projects in the past 10 years. There is, however, a fundamental lack in coordination, integration, scalability, sustainability, and equitable distribution of investments in digital health. Health policymakers and the global health community at large need to urgently institute coordination mechanisms to terminate unending duplication and disjointed vertical implementations and manage solutions for scale (Karamagi et al. 2022).

To achieve scaleability of digital health interventions, we need to design and implement standardized open health data systems (OHDS) and their associated ecosystems that can support improvements in the wider health sector along five dimensions, namely (Kelley et al. 2020):

1. overall quality and continuity of care;
2. adherence to clinical guidelines and best practices;
3. efficiency and affordability of services and health commodities, by reducing duplication of effort and ensuring effective use of time and resources;
4. health-financing models and processes, regulation, oversight, and patient safety resulting from increased availability of performance data and reductions in errors; and
5. health policy-making and resource allocation based on better quality data.

Data & analytics functions are essential in open health data systems

Learnings from digital health interventions have shown that improvements along the five aforementioned dimensions are within reach. As an example, in our experience with the MomCare programme, we have demonstrated that routinely collected health data, combined with financial data, can be effectively used to gain insight into the continuity of care, improve clinical adherence whilst maintaining efficiency and affordability of health services in LMICs. By actively coaxing pregnant mothers to undergo 4 antenatal check-ups, outcomes were improved whilst maintaining the average cost of maternal, newborn and child health (MNCH) services (Huisman et al. 2022; Sanctis et al. 2022; Izudi et al. 2023).

MomCare is critically dependent on the availability of data and analytics functions within its underlying supportive OHDS. For example, to implement the value-based healthcare business logic of MomCare, detailed analysis of patient journeys is required. This functionality is currently not available as a standard; standardized reports that are available, such as DHIS2, contained insufficient information and analytical functions to support intervention that aim to improve continuity of care and adherence to clinical guidelines. As such, a large part of the day-to-day operation of MomCare revolved performaing data and analytics: data acquisition, data integration, analysis etc.

This situation is not unique to MomCare. In fact, many in the global digital health community think that data & analytics services should be an essential capibility of OHDSs going forward. This is exemplified how the OpenHIE reference architecture ("OpenHIE Framework V5.2-En" 2024) has been adopted by many sub-Saharan African countries as the blueprint for implementing nation-wide health information exchanges (HIE) (Mamuye et al. 2022), including Nigeria (Dalhatu et al. 2023), Kenya (Mbugua et al. 2021) and Tanzania (Nsaghurwe et al. 2021). These countries have, as a matter of course, extended the framework to include "data & analytics services" as an additional domain (Mbugua et al. 2021; Dalhatu et al. 2023). In terms of the often-used distinction between primary and secondary health data use (Cascini et al. 2024), these countries aim to extend OpenHIE beyond it original scope of primary data sharing to also include secondary use of health data for academic research, real-world evidence studies etc. If we are to use the OpenHIE framework for secondary use, supported by data & analytics as well, we need to extend the standards, technologies and architecture to include functionality to do so. The lack of detailed specifications and consensus of this addition to

OpenHIE currently stands in the way of development projects that aim to establish more comprehensive platforms to support primary and secondary health data sharing in LMICs.

Objects of openness in secondary data sharing

To set the scene of for the main contributions of this viewpoint paper, consider four types of secondary data sharing as shown in Table 1, where we follow the research agenda proposed by de Reuver et al. to scrutinize how openness of data platforms can be achieved (de Reuver et al. 2022).

Table 1: Types of secondary data sharing, and there relevance to the proposed extension of the OpenHIE specification.

	Type of data sharing	**Relevance to extension of OpenHIE specification**
1	Data at the most granular level with which the patient journey (timeline) can be reconstructed and used for various analytic tasks.	The Shared Health Record (SHR) is specified as an operational, real-time transactional data source, distinct from a data warehouse. A seperate specification of data and analytics functions, typically provided by a datawarehouse, is required.
2	Aggregated data, typically used for routine reports and benchmarking.	The Health Management Information System (HMIS) should support Aggregate Data Exchange (ADX) workflow standard. More flexible and extensive workflows are emerging based on FHIR.
3	Data analytics modules, that provide secure and privacy-preserving computational environments to work with the data.	Federated learning (FL) (Rieke et al. 2020) and privacy- enhancing technologies (PETs) (Scheibner et al. 2021; Jordan, Fontaine, and Hendricks-Sturrup 2022) provide new paradigms that address the problem of data governance and privacy by training algorithms collaboratively without exchanging the data itself. Requires use of a common data model, such as FHIR, to analyze the data in a collaborative, decentralized fashion.
4	Trained models that have been derived from the data and can be used stand-alone for decision support.	Increasing need to open source trained AI models (“The Open Source AI Definition V0.0.9” 2024), enabled by technologies such as ONNX (“ONNX V1.15.0” 2023).

[TO DO: add more text here to clarify our position and intention with this paper.]

Outline

The main contribution of this paper is to propose how recent standards and open source implementations from the data engineering community can be integrated into the OpenHIE framework. In the following, we first describe how the lakehouse design pattern, being the most widely used data & analytics solution architecture, can be integrated in OpenHIE. To demonstrate the feasibility of this design, we present a proof-of-concept impelementation using open source technologies within the context of the MomCare programme. Code and digital artifacts of this demonstrator is available as supplementary material [TO DO: include links to support GitHub repositories]. Subquently, we compare this solution design with two widely used and operational OpenHIE-compliant open source frameworks, namely the OpenHIM platform (https://jembi.gitbook.io/openhim-platform/) and the OnaData platform https://ona.io/home/products/ona-data/features/. Finally, we discuss our findings and propose routes for future development.

We take a narrative approach in presenting our design and the case studies, surveying existing scientific studies on OHDSs, focusing on the seminal reports and subsequently searching forward citations. In addition, we have searched the open source repositories (most notably GitHub) and the online communities (OpenHIE community, FHIR community) to search for relevant open standards, technologies and architectures. This paper should not be considered as a proper systematic review. The main contributions of this paper are i) description of a framework for the components of the Data & Analysis Services that builds on current best practices from the data engineering community into the OpenHIE framework; and ii) evaluation of different implementations and design options for various data sharing scenarios within an extended OpenHIE architecture.

Extending OpenHIE to include modern data and analytics standards

High-level solution design

The original OpenHIE specification discerns four domains, namely Point-of-Service systems, the Interoperability Layer, Common Services and Business Services (Figure 1). We propose to extend the OpenHIE architecture with a "Data and Analytics Services" domain with different zones taken from the data lakehouse architecture, which currently is the most commonly used design pattern in this domain (Armbrust et al. 2021; Hai et al. 2023; Harby and Zulkernine 2022, 2024). 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) (Hai et al. 2023). 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 (Armbrust et al. 2021). 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.

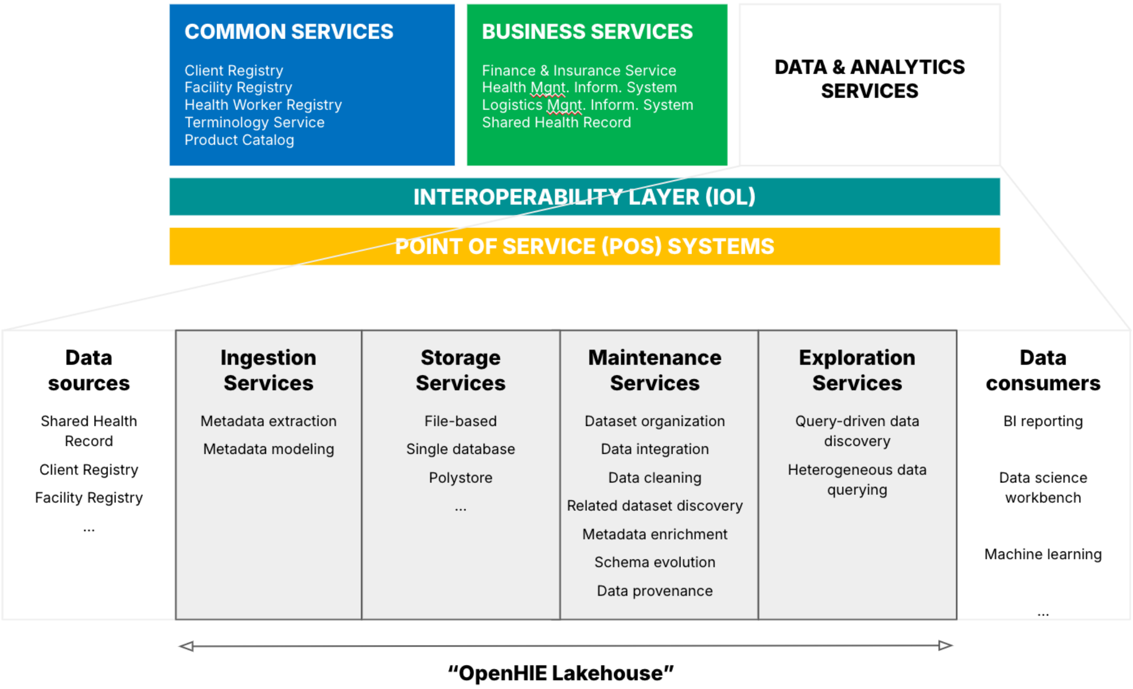

Figure 1: Proposed extension of the OpenHIE architecture that includes "Data and Analytics Services" as an additional service domain.

Following the terminology proposed by Hai et al. (Hai et al. 2023), summarizes how the different zones of the lakehouse architecture can be adapted for healthcare and integrated into the OpeHIE specification. Taking Fast Healthcare Interoperability Resources (FHIR) as the open data standard[1], we envisage the extended OpenHIE architecture to include a 'OpenHIE Lakehouse' with the following healthcare specific adaptations of the various data & analytics services:

- **Ingestion Services:** use of FHIR standard to harmonize all incoming healthcare data to a common data model, including metadata extraction and metadata modeling. Should support both single-records streaming ingest as well as bulk data ingestion in batches using the Bulk FHIR API as interface (Mandl et al. 2020; Jones et al. 2021).
- **Storage Service:** should support columnar storage engines optimized for analytical workloads. In case of file-based storage, use columnar file formats such as Apache Parquet.

[1] We have argued the choice of FHIR as the common data model elsewhere, working paper to be submitted (link).

In case of databases, prefer open source engines such as Clickhouse or PostgreSQL that also support external, file-based tabels (Pedreira et al. 2023).

- **Maintenance:** use one of the open table formats such as Apache Iceberg, Apache Hudi and/or Delta Lake (Jain et al. 2023) to realize dataset organization, data integration, schema evolution and data provenance. Use of SQL-on-FHIR v2 View definitions to facilitate access to FHIR resources in flattened, tabular format ("SQL on FHIR Speciification V0.0.1-Pre" 2024)
- **Exploration:** use on-demand, read-only analytical processing engines to provide a unified querying interface to access the heterogeneously structured data using new dataprocessing technologies such as DuckDB, polars etc. Should support SQL-on-FHIR Runners, such as Pathling, to generate the standardized views on demand.

Strictly speaking, data consumer services are not part of the lakehouse solution design. In practice, these services are implemented using a combination of business intelligence (BI) reporting tools, and interactive development environment (IDE) to perform SQL queries and/or an interactive notebook computing environment (Granger and Perez 2021). In the discussion we will address future support of federated learning and/or secure multiparty computation network.

Parking lot: following snippets need reviewing (integrate or discard)

SQL-on-FHIR v2 as an intermediate representation for FHIR data in tabular format

The premise of separating the user interface from the execution engine is directly related to the key objective of the SQL-on-FHIR project (https://build.fhir.org/ig/FHIR/sql-on-fhir-v2/), namely to make large-scale analysis of FHIR data accessible to a larger audience, portable between systems and to make FHIR data work well with the best available analytic tools, regardless of the technology stack. However, to use FHIR effectively analysts require a thorough understanding of the specification as FHIR is represented as a graph of resources, with detailed semantics defined for references between resources, data types, terminology, extensions, and many other aspects of the specification. Most analytic and machine learning use cases require the preparation of FHIR data using transformations and tabular projections from its original form. The task of authoring these transformations and projections is not trivial and there is currently no standard mechanisms to support reuse.

The solution of the SQL-on-FHIR project is to provide a specification for defining tabular, use case-specific views of FHIR data. The view definition and the execution of the view are separated, in such a way that the definition is portable across systems while the execution engine (called runners) are system-specific tools or libraries that apply view definitions to the underlying data layer, optionally making use of annotations to optimize performance.

Ingestion

- Default workflow is extraction of data from SHR using Bulk FHIR API. Data contains metadata (incl. FHIR versions) and fully qualified semantics, for example, coding systems. Despite this, metadata extraction and metadata modeling is still required to meet the FAIR requirements. Issues that need to be solved by these services:
- To prepare for future updates of FHIR versions
- Implement late-binding principle of having increasingly more specific FHIR profiles as bulk FHIR data propagates through lakehouse
- FHIR vs. FAIR
    - How does FHIR relate to approaches taken by the FAIR community, which tend to take more an approach of using knowledge graphs. For example, VODAN Africa (Gebreslassie et al. 2023; Purnama Jati et al. 2022).
    - FAIR principles vs FHIR graph: is FHIR a FAIR Data Object
- Since we use FHIR, we don't need a semantic layer because that is already provided
- We do need different semantic layer, namely with metrics. Explain different types of semantics.
    - The metrics layer same function as CQL. Discuss CQL vs generic metrics layer.

Storage

- File-based:
    - from ndjson to parquet
    - possibly used delta lake for time versioning
    - separation of storage from compute not only for benefits of lower TCO, but also be ready for federated learning and MPC in future

Maintenance

- SQL-on-FHIR Views provide new standard to support mADX aggregate reporting !! We need to stress this, because this is an existing OpenHIE workflow
- Maintenance-related functions remain the same
- NB: orchestration falls under data provenance
- NB: make comparison with HMIS component
    - workflow requirements: Report aggregate data (link): receiver is HMIS, mADX; this is not analytics!
    - Functional requirements: https://guides.ohie.org/arch-spec/openhie-component-specifications-1/openhie-health-management-information-system-hmis
    - Requirements are similar, but implementation differs: Datamodel is non-FHIR, focused on DataValue, which conceptually equates to FHIR Measure

Case study of implementations

PharmAccess demonstrator Momcare programme

MomCare was launched in Kenya (Huisman et al. 2022; Sanctis et al. 2022) and Tanzania (Shija et al. 2021; Mrema 2021) in 2017 and 2019 respectively, with the objective to improve health outcomes for maternal and antenatal care. MomCare distinguishes two user groups: mothers are supported during their pregnancy through reminders and surveys, using SMS as the digital mode of engagement. Health workers are equipped with an Android-based application, in which visits, care activities and clinical observations are recorded. Reimbursements of the maternal clinic are based on the data captured with SMS and the app, thereby creating a conditional payment scheme, where providers are partially reimbursed up-front for a fixed bundle of activities, supplemented by bonus payments based on a predefined set of care activities.

In its original form, the MomCare programme used closed digital platforms. In Kenya, M-TIBA is the primary digital platform, on top of which a relatively lightweight custom app has been built as the engagement layer for the health workers (Huisman et al. 2022). M-TIBA provides data access through its data warehouse platform for the MomCare programme, however, this is not a standardized, general purpose API. In the case of Tanzania, a standalone custom app is used which does not provide an interface of any kind for interacting with the platform (Mrema 2021). Given these constraints, the first iteration of the MomCare programme used a custom-built data warehouse environment as its main data platform, on which data extractions, transformations and analysis are performed to generate the operational reports. Feedback reports for the health workers, in the form of operational dashboards, are made accessible through the app. Similar reports are provided to the back-office for the periodic reimbursement to the clinics.

Clearly, a more open and scaleable platform was required if MomCare was to be implemented in more regions. This need led to a redesign of the underlying technical infrastructure of the MomCare project. The objectives of this work were in fact to demonstrate a solution design that could support the first three types of data sharing. First, to investigate the viability of using FHIR for bulk data sharing, MomCare Tanzania was used a testbed to assess the complexity and effort required to implement the facade pattern to integrate the legacy system into the FHIR data standard. Using the longitudinal dataset from approximately 28 thousand patient records, FHIR transformations script were developed and deployed using the mediator function of the IOL. The data was transformed into 10 FHIR v4 resources and the conceptual data model of the existing MomCare app could readily be transformed into the FHIR standard using SQL and validated with a Python library (Islam 2023). The largest challenge during the transformation process pertained to the absence of unique business identifiers for patients and healthcare organizations. For patients, either the mobile phone number or the healthcare insurance number was taken, depending on availability. A combination of name, address and latitude/longitude coordinates were used to uniquely identify organizations and locations, as Tanzania does not have a system in place for this purpose.

The second objective was to reproduce existing analytic reports, using the bulk FHIR data format as input. Here, the focus was to standardize the logic required for producing metrics and reports. The transformed and validated data is uploaded into the FHIR server on a daily basis using an automated cloud function. Analysis of bulk data was done by directly reading the standard newline delimited JSON into the Python pandas data analysis library. Cross checking the output with queries on the original data confirmed that the whole data pipeline produced consistent results. For example, the report of the antenatal coverage metric (number of pregnancies with four or more visits) could be reproduced per patient journey and aggregated (per year, per organization etc.) as required for the MomCare reports.

TO DO: explain logic of patient-timeline table. Write standard transformation to go from FHIR resources to this standard table. On top of that the actual metrics and reporting. Explain serverless: we wanted to get rid of resource-heavy data visualization tools. This led to the idea of serverless: using duckdb-wasm and pipelines of cloud functions.

The third objective was to run a technical feasibility test for federated analytics. Using the MPC platform of Roseman Labs, we managed to do aggregations in the blind … TO DO: explain that we managed to reproduce the reports we generated in the clear, but then in the blind. Note, however, that in the remainder we will focus on first two types of data sharing.

Based on these experiments, we arrived at the following design for the data & analytics services

- Use ‘serverless’ file-based storage: bulk copy of data as-is in parquet
  - Tension: how to manage change data capture
  - Tension: how to manage access rights
- Use SQL-on-FHIR-v2 to create tabular views.
  - Example: patient timeline
  - TO DO: rewrite patient timeline queries with SQL-on-FHIR-v2 and run it with Pathling
- Use semantic modeling layer to define metrics
  - There are many options: dbt, cube.dev
  - Fulfills same function as ADX/mADX IHE profile in OpenHIE specification
  - Tension: going from patient-timeline to reported metrics still isn’t standardized. This is where Ibis/Substrait comes in. Substrait as IR for cross-language serialization for relational algebra. Can be executed on different backends. Write once, run on different engines.
- Distribute and publish reports on resource-constrained devices

  - duckdb
  - sveltekit

TO DO: Add diagram

Jembi OpenHIM platform

To evaluate the extended OpenHIE architecture described above, we first consider the OpenHIM Platform. The Open Health Information Mediator (OpenHIM, http://openhim.org/)) component is the reference implementation of the Interoperability Layer (IOL) as defined in the OpenHIE specification. The most current version (8.4.2 at the time of writing) provides all the core functions including central point of access for the services of the HIE; routing functions; central logging for auditing and debugging purposes; and orchestration/mediation mechanisms to co-ordinate requests. By extension, the OpenHIM Platform (https://jembi.gitbook.io/openhim-platform) is a reference implementation of a set of Instant OpenHIE configurations, refered to as 'recipes' in the documentation. In the following we will evaluate the recipe for "a central data repository with a data warehouse" that provides "A FHIR-based Shared Health record linked to a Master Patient Index (MPI) for linking and mathing patient demographics and a default reporting pipeline to transform and visualise FHIR data" (https://jembi.gitbook.io/openhim-platform/recipes/central-data-repository-with-data-warehousing).

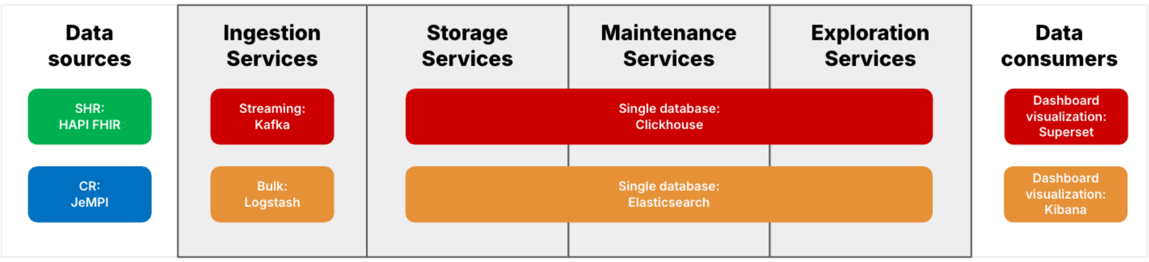

Figure 2: Overview of the default data stack of the OpenHIM Platform. The default stack (top, red) consists of Kafka, Clickhouse and Superset. An alternative solution based on the ELK stack is also supported (bottom, orange), consisting of Elasticsearch, Logstash and Kibana.

Figure 2 shows a schematic overview of two data stacks that are supported in the OpenHIM platform. The Shared Health Record (SHR, implemented with HAPI FHIR server) and the Client Registry (CR, implemented with JeMPI server) are the sources that store clinical FHIR data and patient demographic data, respectively. The default data stack is based on streaming ingestion using Kafka into a Clickhouse database. As part of the ingestion, incoming FHIR bundles that contain multiple FHIR resources are unbundled in separate topics using a generic Kafka utility component. Subsequently, each FHIR resource topic is flatted with

Kafka mappers that use FHIRPath. Superset is used as the tool for consuming the data to create dashboard visualizations.

The OpenHIM platform also support data and analytics based on the ELK stack, where data is ingested in bulk using Logstash, stored in Elasticsearch and made available for consumption in Kibana. Also here, the incoming FHIR bundles are unbundled in Logstash into separate FHIR resources. However, given that Elasticsearch is a document-based search engine, the FHIR resources are stored as-is with no flattening. Exploring and analysing the data requires writing queries in Elasticsearch Query Language (ES|QL), either through the query interface of Elasticsearch or using Kibana.

Evaluating these two data stacks, we see the following:

- Pattern of flattening FHIR resources with FHIRPath expressions is very close to the idea of SQL-on-FHIR. Although it doesn't adhere to this new standard in the strict sense, the philosophy of generating tabular views is the same
- When using the ELK stack, flattening is done at the end. Implementations of FHIRPath support Elasticsearch as an execution engine, also here
- Main limitations: both Clickhouse en Elasticsearch don't follow decomposition of storage, compute and UI. Therefore, downward scaleability is limited.

ONA OpenSRP 2

Continuing our evaluation of the extended OpenHIE architecture, we can see a different flavor in the implementation driven by Ona. Ona is a social enterprise that has pioneered the adoption of FHIR data standard via the development of OpenSRP2, a FHIR-based data collection app built using Google's FHIR SDK and focused on enabling offline-first workflows for community-based care. OpenSRP 2 is a complete rewrite of the original OpenSRP application, a global public good maintained by Ona and deployed in XX countries worldwide.

OpenSRP2 applications are currently implemented in the field in three countries (Uganda, Liberia, and Madagascar) in collaboration with local Ministries of Health and with international donors such as UNICEF, supporting a variety of different workflows including antenatal care (ANC), postnatal Care (PNC), immunization, and last-mile logistics. Besides the OpenSRP Android application and HAPI-FHIR backend, in each of its projects Ona also implements a companion set of tools that support analytics and various reporting needs.

Requirements for data sharing

Based on years of work in global health, Ona has learned that the data stack implemented to support a national-scale implementation of its FHIR-based application responds to the following requirements.

Table 2: Requirements and rationale for open health data platform developed and used by ONA, based on OpenSRP 2 (https://opensrp.io/).

Requirement	Rationale
Ingest data from multiple sources, both FHIR and non-FHIR based.	While most health record data can be collected and aggregated in FHIR, Ministries of Health rely on other data sources to govern their operations. For example, operationalizing an immunization campaign usually includes tracking against specific targets for locations to be visited on specific days and number of children to immunize per day. Such targets are often stored in spreadsheets or other applications where the data is not FHIR.
Ingest data in batches.	Most data ingestion can happen in batches, since Ona's applications are deployed in hard to reach areas where connectivity is an issue. Data ingestion closer to real-time can be relevant for disaster-response and other time-sensitive applications, but this is not a priority.
Support national-scale data volumes.	A data store that can grow from dozens to thousands of devices and where data can be aggregated up to the national level, matching the scale of implementation of data collection applications in the field.
Pre-compute complex business metrics.	Reporting on health systems requires pre-computing complex metrics and often performing cohort analyses to map trends in service provision. For example, understanding quality of care for children requires computing metrics such as the percentage of children fully immunized on schedule (i.e. children 6-59 months that have received the set of vaccines required by the Ministry of Health, and have received each of those vaccines within the expected age-window). For Business Intelligence applications, calculating such a vital metric cannot be performed at run time, to avoid long and expensive queries.
Outbound integrations.	While aggregated data and reports should be accessible by other applications such as BI platforms via pulls, there should be an easy integration framework to push data to other applications used by the Ministry of Health for other purposes, such as DHIS2 for health systems management or RapidPro for communications with program beneficiaries.
Open source and easily deployable in-country.	Given the extremely sensitive nature of health data, it is paramount for governments to have the flexibility to deploy the stack in various different environments, both on premise and in private clouds.

The architecture Learning from experience in the field and internal research and development, Ona has developed preferences for a specific data stack responding to the aforementioned requirements.

[[graphic]]

Core toolings in the stack include: Data ingestion with Airbyte. Ona uses Airbyte as the primary data ingestion tool, leveraging the wide array of connectors that come standard with the application as well as a dedicated suite of connectors developed internally by Ona, including HAPI FHIR, RapidPro, Ona Data, Kobo Toolbox and others.
Data storage with Clickhouse. While different health projects have varying requirements, Ona has found success in using Clickhouse as the main analytics data store in its most recent implementations. Clickhouse supports the scale required for analytics at a national level, as well as the speed that enables cross-application integrations and more real-time analytics. For example, in Madagascar Ona uses its reporting suite to identify facilities with stock in need of maintenance and can trigger the scheduling of a maintenance visit ad hoc. Data transformation with dbt. Following global best practice, Ona leverages dbt to segregate the data warehouse in different levels (staging, marts, metrics), as well as pre-computing complex indicators for ease of reporting and for transmissions into other systems. For example, in Liberia Ona implements OpenSRP at community health worker level, but can aggregate immunization data at facility level in the data warehouse and then push quarterly summary metrics to DHIS2. No recommendation on reporting / BI tooling. Ona recognizes that business users have their own strong preferences for BI tooling, and some already have licenses for specific software, so the architecture is flexible to provide easy connections to different BI tools.

Evaluation

Evaluating the data stack, we see the following: Use of generic best-of-breed tooling. Ona focused on utilizing Open HIE tools that are widely adopted outside of the global health and development sectors. This approach aims to provide assurance on two main fronts, the ability to handle performance at scale and the long term dependability of the tools, rather than relying on smaller projects with uncertain long term funding or unproven implementations. Columnar data warehouse for analytics. The scale of Ona's project requires the implementation of a dedicated database for analytics. While original data can still be stored as parquet or other file system, being able to ingest it into a relational data store allows to create well defined indicators. Using clickhouse as a tool helps and combine the need accuracy with the speed of reporting as new data is ingested.
Strong emphasis on SQL. While Ona has tested and experimented with FHIR-specific tooling, such as the definition of data projections using sql-on-fhir, Ona found that relying on sql for coding business logic remained the faster and most scalable approach.

In summary, for Ona building analytics with FHIR data looks similar to building analytics with any other type of data. While FHIR provides a clear and standard data model, managing information for most health systems requires custom integration of data between different sources, as well as computing indicators using business logic specific to the needs of the local users. Building upon well established best-of-breed tools allows Ona to implement FHIR applications at scale and provide trusted analytics on top.

Discussion

Given the solution design, and evaluation of the three case studies, we discuss the pros and cons from different perspectives that we believe are essential design principles to realise solidarity-based OHDSs, namely [TO DO: this is just first shot at formulating our design principles; needs refinement]:

- inclusive-by-design, based on the notions of datasolidarity and maximising autonomy of all future participants in the ecosystem;
- scaleable-by-design, particularly focusing on downward scaleability to support a decentralized platform topology to allows for bottom-up deployment scenarios (from local care networks –> county-level networks –> national networks) instead of top-down national roll-out;
- open-by-design, whereby a balance is found to resort to minimal standards and allow for a large diversity of partners and technologies to be used;

Inclusive-by-design: datasolidarity, FAIRness and autonomy

- which can be framed within the context of ongoing efforts towards Findable, Accessible, Interoperable and Reusable (FAIR) sharing of health data (Guillot, Bøgsted, and Vesteghem 2023).
- equitable data sharing requires more than just FAIRification (**evertz2023what?**)

Scaleable-by-design

Today, many components of the OpenHIE specification are now available as a digital public goods. Typically, these open source components are intended to support deployments in small countries (population up to 10 million) or large NGOs out of the box, and should provide a stepping stone for customized deployments in medium-sized countries (population around 40 million).[2] To further ease the development, configuration and deployment of health information exchanges, the concept of 'Instant OpenHIE' has been championed to (i) allow implementers to engage with a preconfigured health information exchange solution and running tools (based on the architecture) and test their applicability and functionality with a real health context problem; and (ii) have a packaged reference version of the OpenHIE architecture that is comprised of a set of reference technologies and other appropriate tools that form the building blocks of the health information exchange that can be configured and extended to

[2] Although the OpenHIE specification does not include details on dimensioning, these are typically the requirements that are used within the community. See OpenHIE Community Wiki.

support particular use cases ("Instant OpenHIE V2.2.0" 2024). Besides the core functional components of the OpenHIE architecture, the Instant OpenHIE toolkit allows packaging and integration of generic components such as Identity and Access Managment (IAM) and a reverse proxy gateway. In the following, we will evaluatie three of such configurations, with the aim to conceptualize and evaluate the proposed Data and Analytics Services domain of of the OpenHIE architecture.

We also posit that a decentralized platform is more conducive to realize a solidarity-based approach to health data sharing that i) gives people a greater control over their data as active decision makers; ii) ensures that the value of data is harnessed for public good; and iii) moves society towards equity and justice by counteracting dynamics of data extraction (Prainsack et al. 2022). With this approach, we purposefully challenge the dominant paradigm of designing and implementing centralized platforms to support the digital transformation of healthcare in LMICs (Ogundaini and Achieng 2022) with the aim to make digital platforms work *for* development (Hermes et al. 2020).

[TO DO: elaborate on how we see this solution design can be implemented from the bottom-up, typically in a primary care network serving a population of around 80,000 people with level 3 facilities that have limited resources]

Table 3: Number of healthcare facilities in Kenya. Source: Kenya Health Facility Census, Ministry of Health, September 2023.

Level	Description	Number of facilities
2	Dispensaries and private clincs, typically located in a school, industrial plant or other organization that dispenses medication and sometimes basic medical and dental treatment	8,806
3	Health centres, medium-sized units which cater for a population of about 80,000 people	2,559
4	Sub-county hospital, similar to health centres with additional facilities for more complex procedures	971
5	County referral hospital, regional centres which provide specialised care	34
6	National referral hospital	5

Open-by-design: mitigating risk and rebalancing asymmetries

The shift in perspective from digital platforms to data platforms coincides with the paradox of open (Keller and Tarkowski 2021). Originally, openness of digital platforms focused on open source and open standards (as shown above for OpenHIE) which by has been superseded by "… conflicts about privacy, economic value extraction, the emergence of artificial intelligence, and the destabilizing effects of dominant platforms on (democratic) societies. Instead of access to information, the control of personal data has emerged in the age of platforms as the critical contention." (Keller and Tarkowski 2021). These conflicts are particularly salient in the healthcare domain, where people are generally willing to share their health data to receive the best care (primary use, which is aligned with the concept of digital platforms), while the attitude towards secondary use of health data (conceptually aligned with a data platform) varies greatly depending on the type and context (Cascini et al. 2024). The shift in perspective from digital platforms supporting primary data sharing toward data platforms supporting secondary data sharing is one of the key issues surrounding the polemic of data spaces (Otto, Ten Hompel, and Wrobel 2022) and data solidarity (Kickbusch et al. 2021; Prainsack et al. 2022; Prainsack and El-Sayed 2023; Purtova and van Maanen 2023).

- Risk of openness: What are the novel (negative) implications of opening up data platforms? How can reflexivity in design help providers to resolve the negative implications of openness?
- Answers/insights to above:
  - Openness of standardized view on FHIR data and cross-language serialization of relational algebra makes it possible to fully standardize the workflow from start to finish
  - Platform-to-platform: MPC
  - Risk of openness: difficult to answer …
- Paradox of open in disucssion: we started with hypotheses that a decentralized approach will lead to distribution of power, hence … But is this really the case? Will open source not backfire and strengthen their position?

Limitation and future work

- Access control is still a pain-point, can we move to Attribute-based access control?
  - TO DO: if you have generated flattened SQL tables, how are you going to manage security?
  - Cerbos, attribute based on lineage or anonymized tables
  - Catalogs solve this: Tabular.io, Google BigLake. What is open source option?
- Federated learning and multiparty computation
  - Lakehouse serves as datastations
  - Explain first results Roseman Labs
- … [more future work items here]

Conclusion

Acknowledgements

Conflicts of interest

DK received funding from PharmAccess to conduct this work as a contractor. SW/IntelliSOFT received funding from PharmAccess as an implementation partner in various projects. AP/ONA received funding from PharmAccess to develop an improved version of one of the components of the open source OpenSRP 2 framework.

Abbreviations

ACID	Atomicity, Consistency, Isolation, and Durability
CLI	Command-line Interface
CR	Client Registry
DHI	Digital Health Intervention
ELK	Elasticsearch, Logstach and Kibana stack
ELT	Extract, Load and Transform
FAIR	Findable, Accessible, Interoperable and Reusable
FHIR	Fast Healthcare Interoperability Resources
FL	Federated learning
HIE	Health Information Exchange
LMIC	Low- and middle income countries
MPC	Multiparty Computation
PET	Privacy-enhancing technologies
OHDS	Open health data system
SHR	Shared Health Record

References

Armbrust, Michael, Ali Ghodsi, Reynold Xin, and Matei Zaharia. 2021. "Lakehouse: A New Generation of Open Platforms That Unify Data Warehousing and Advanced Analytics." In *11th Annual Conference on Innovative Data Systems Research (CIDR '21)*, 8.

Cascini, Fidelia, Ana Pantovic, Yazan A. Al-Ajlouni, Valeria Puleo, Lucia De Maio, and Walter Ricciardi. 2024. "Health Data Sharing Attitudes Towards Primary and Secondary Use of Data: A Systematic Review." *eClinicalMedicine* 71 (May): 102551. https://doi.org/10.1016/j.eclinm.2024.102551.

Dalhatu, Ibrahim, Chinedu Aniekwe, Adebobola Bashorun, Alhassan Abdulkadir, Emilio Dirlikov, Stephen Ohakanu, Oluwasanmi Adedokun, et al. 2023. "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 of Information in Medicine* 62 (03/04): 130–39. https://doi.org/10.1055/s-0043-1768711.

de Reuver, Mark, Hosea Ofe, Wirawan Agahari, Antragama Ewa Abbas, and Anneke Zuiderwijk. 2022. "The Openness of Data Platforms: A Research Agenda." In *Proceedings of the 1st International Workshop on Data Economy*, 34–41. DE '22. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3565011.3569056.

Gebreslassie, Tesfit Gebremeskel, Mirjam van Reisen, Samson Yohannes Amare, Getu Tadele Taye, and Ruduan Plug. 2023. "FHIR4FAIR: Leveraging FHIR in Health Data FAIRfication Process: In the Case of VODAN-A." *FAIR Connect* 1 (1): 49–54. https://doi.org/10.3233/FC-230504.

Granger, Brian E., and Fernando Perez. 2021. "Jupyter: Thinking and Storytelling With Code and Data." *Computing in Science & Engineering* 23 (2): 7–14. https://doi.org/10.1109/MCSE.2021.3059263.

Guillot, Paul, Martin Bøgsted, and Charles Vesteghem. 2023. "FAIR Sharing of Health Data: A Systematic Review of Applicable Solutions." *Health and Technology* 13 (6): 869–82. https://doi.org/10.1007/s12553-023-00789-5.

Hai, Rihan, Christos Koutras, Christoph Quix, and Matthias Jarke. 2023. "Data Lakes: A Survey of Functions and Systems." *IEEE Transactions on Knowledge and Data Engineering* 35 (12): 12571–90. https://doi.org/10.1109/TKDE.2023.3270101.

Harby, Ahmed A., and Farhana Zulkernine. 2022. "From Data Warehouse to Lakehouse: A Comparative Review." In *2022 IEEE International Conference on Big Data (Big Data)*, 389–95. Osaka, Japan: IEEE. https://doi.org/10.1109/BigData55660.2022.10020719.

———. 2024. "Data Lakehouse: A Survey and Experimental Study." {{SSRN Scholarly Paper}}. Rochester, NY. https://doi.org/10.2139/ssrn.4765588.

Hermes, Sebastian, Tobias Riasanow, Eric K. Clemons, Markus Böhm, and Helmut Krcmar. 2020. "The Digital Transformation of the Healthcare Industry: Exploring the Rise of Emerging Platform Ecosystems and Their Influence on the Role of Patients." *Business Research* 13 (3): 1033–69. https://doi.org/10.1007/s40685-020-00125-x.

Huisman, Liesbeth, Shannen Mc van Duijn, Nuno Silva, Rianne van Doeveren, Jacinta Michuki, Moses Kuria, David Otieno Okeyo, et al. 2022. "A Digital Mobile Health Platform Increasing Efficiency and Transparency Towards Universal Health Coverage in Low- and Middle-Income Countries." *Digital Health* 8: 20552076221092213. https://doi.org/10.1177/20552076221092213.

"Instant OpenHIE V2.2.0." 2024. https://jembi.gitbook.io/instant-v2/.

Islam, Md Nazrul. 2023. "Fhir.resources: FHIR Resources as Model Class."

Izudi, Jonathan, Henry Odero Owoko, Moussa Bagayoko, and Damazo Kadengye. 2023. "Experiences of Mothers and Health Workers with MomCare and SafeCare Bundles in Kenya and Tanzania: A Qualitative Evaluation." *PloS One* 18 (11): e0294536. https://doi.org/10.1371/journal.pone.0294536.

Jain, Paras, Peter Kraft, Conor Power, Tathagata Das, Ion Stoica, and Matei Zaharia. 2023.

"Analyzing and Comparing Lakehouse Storage Systems."

Jones, James, Daniel Gottlieb, Joshua C Mandel, Vladimir Ignatov, Alyssa Ellis, Wayne Kubick, and Kenneth D Mandl. 2021. "A Landscape Survey of Planned SMART/HL7 Bulk FHIR Data Access API Implementations and Tools." *Journal of the American Medical Informatics Association* 28 (6): 1284–87. https://doi.org/10.1093/jamia/ocab028.

Jordan, Sara, Clara Fontaine, and Rachele Hendricks-Sturrup. 2022. "Selecting Privacy-Enhancing Technologies for Managing Health Data Use." *Frontiers in Public Health* 10 (March): 814163. https://doi.org/10.3389/fpubh.2022.814163.

Karamagi, Humphrey C, Derrick Muneene, Benson Droti, Violet Jepchumba, Joseph C Okeibunor, Juliet Nabyonga, James Avoka Asamani, Moussa Traore, and Hillary Kipruto. 2022. "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." *Journal of Global Health* 12 (December): 04090. https://doi.org/10.7189/jogh.12.04090.

Keller, Paul, and Alek Tarkowski. 2021. "The Paradox of Open." *Open Future*, March.

Kelley, Edward, Diana Zandi, Ramesh Krishnamurthy, Garrett Mehl, David Novillo, Derrick Muneene, Mohamed Nour, et al. 2020. "Digital Health Platform Handbook: Building a Digital Information Infrastructure (Infostructure) for Health."

Kickbusch, Ilona, Dario Piselli, Anurag Agrawal, Ran Balicer, Olivia Banner, Michael Adelhardt, Emanuele Capobianco, et al. 2021. "The Lancet and Financial Times Commission on Governing Health Futures 2030: Growing up in a Digital World." *The Lancet* 398 (10312): 1727–76. https://doi.org/10.1016/S0140-6736(21)01824-9.

Mamuye, Adane L., Tesfahun M. Yilma, Ahmad Abdulwahab, Sean Broomhead, Phumzule Zondo, Mercy Kyeng, Justin Maeda, Mohammed Abdulaziz, Tadesse Wuhib, and Binyam C. Tilahun. 2022. "Health Information Exchange Policy and Standards for Digital Health Systems in Africa: A Systematic Review." *PLOS Digital Health* 1 (10): e0000118. https://doi.org/10.1371/journal.pdig.0000118.

Mandl, Kenneth D., Daniel Gottlieb, Joshua C. Mandel, Vladimir Ignatov, Raheel Sayeed, Grahame Grieve, James Jones, Alyssa Ellis, and Adam Culbertson. 2020. "Push Button Population Health: The SMART/HL7 FHIR Bulk Data Access Application Programming Interface." *Npj Digital Medicine* 3 (1): 1–9. https://doi.org/10.1038/s41746-020-00358-4.

Mbugua, Samuel, Julia Korongo, Mutai Joram, Masese Chuma Benard, and Alice Nambiro. 2021. "Adoption of ICT to Enhance Access to Healthcare in Kenya." *IOSR Journal of Computer Engineering* 23 (April): 45–50. https://doi.org/10.9790/0661-2302024550.

Mehl, Garrett L, Martin G Seneviratne, Matt L Berg, Suhel Bidani, Rebecca L Distler, Marelize Gorgens, Karin E Kallander, et al. 2023. "A Full-STAC Remedy for Global Digital Health Transformation: Open Standards, Technologies, Architectures and Content." *Oxford Open Digital Health* 1 (January): oqad018. https://doi.org/10.1093/oodh/oqad018.

Mrema, Anunsiatha. 2021. "Application of Digital Platform to Enhance Quality Improvement in Momcare Facilities in Manyara." *AIJR Abstracts*, July.

Nsaghurwe, Alpha, Vikas Dwivedi, Walter Ndesanjo, Haji Bamsi, Moses Busiga, Edwin Nyella, Japhet Victor Massawe, et al. 2021. "One Country's Journey to Interoperability: Tanzania's Experience Developing and Implementing a National Health Information Exchange." *BMC Medical Informatics and Decision Making* 21 (1): 139. https://doi.org/10.1186/

s12911-021-01499-6.

Ogundaini, Oluwamayowa O., and Mourine S. Achieng. 2022. "Systematic Review: Decentralised Health Information Systems Implementation in Sub-Saharan Africa." *Journal for Transdisciplinary Research in Southern Africa* 18 (1): 1–10. https://doi.org/10.4102/td.v18i1.1216.

"ONNX V1.15.0." 2023. https://onnx.ai/.

"OpenHIE Framework V5.2-En." 2024. *OpenHIE*. https://ohie.org/.

Otto, Boris, Michael Ten Hompel, and Stefan Wrobel, eds. 2022. *Designing Data Spaces: The Ecosystem Approach to Competitive Advantage*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-93975-5.

Pedreira, Pedro, Orri Erling, Konstantinos Karanasos, Scott Schneider, Wes McKinney, Satya R Valluri, Mohamed Zait, and Jacques Nadeau. 2023. "The Composable Data Management System Manifesto." *Proceedings of the VLDB Endowment* 16 (10): 2679–85. https://doi.org/10.14778/3603581.3603604.

Prainsack, Barbara, and Seliem El-Sayed. 2023. "Beyond Individual Rights: How Data Solidarity Gives People Meaningful Control over Data." *The American Journal of Bioethics* 23 (11): 36–39. https://doi.org/10.1080/15265161.2023.2256267.

Prainsack, Barbara, Seliem El-Sayed, Nikolaus Forgó, Łukasz Szoszkiewicz, and Philipp Baumer. 2022. "Data Solidarity: A Blueprint for Governing Health Futures." *The Lancet Digital Health* 4 (11): e773–74. https://doi.org/10.1016/S2589-7500(22)00189-3.

Purnama Jati, Putu Hadi, Mirjam van Reisen, Erik Flikkenschild, Fransisca Oladipo, Bert Meerman, Ruduan Plug, and Sara Nodehi. 2022. "Data Access, Control, and Privacy Protection in the VODAN-Africa Architecture." *Data Intelligence* 4 (4): 938–54. https://doi.org/10.1162/dint_a_00180.

Purtova, Nadya, and Gijs van Maanen. 2023. "Data as an Economic Good, Data as a Commons, and Data Governance." *Law, Innovation and Technology* 0 (0): 1–42. https://doi.org/10.1080/17579961.2023.2265270.

Rieke, Nicola, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, et al. 2020. "The Future of Digital Health with Federated Learning." *Npj Digital Medicine* 3 (1): 1–7. https://doi.org/10.1038/s41746-020-00323-1.

Sanctis, Teresa De, Mary-Ann Etiebet, Wendy Janssens, Mark H. van der Graaf, Colette van Montfort, Emma Waiyaiya, and Nicole Spieker. 2022. "Maintaining Continuity of Care for Expectant Mothers in Kenya During the COVID-19 Pandemic: A Study of MomCare." *Global Health: Science and Practice* 10 (4). https://doi.org/10.9745/GHSP-D-21-00665.

Scheibner, James, Jean Louis Raisaro, Juan Ramón Troncoso-Pastoriza, Marcello Ienca, Jacques Fellay, Effy Vayena, and Jean-Pierre Hubaux. 2021. "Revolutionizing Medical Data Sharing Using Advanced Privacy-Enhancing Technologies: Technical, Legal, and Ethical Synthesis." *Journal of Medical Internet Research* 23 (2): e25120. https://doi.org/10.2196/25120.

Shija, Liberatha, Johnson Yokoyana, Anunsiatha Mrema, Jonia Bwakea, Theodora Kiwale, and Neema Massawe. 2021. "Access to Essential Maternal Health Commodities Key to Improving Quality and Adherence to Maternal Healthcare Regimes: Experience from a MomCare Project in Northern Tanzania." *AIJR Abstracts*, July.

"SQL on FHIR Speciification V0.0.1-Pre." 2024. https://build.fhir.org/ig/FHIR/sql-on-fhir-v2/.

"The Open Source AI Definition V0.0.9." 2024. *HackMD*. https://hackmd.io/@opensourceinitiative/osaid-0-0-9.