Deep Science Publishing, 2026
https://doi.org/10.70593/978-93-7185-978-3_2

DeepScience
Open Access Books

Chapter 2: Architecture of Interconnected Clinical Data Platforms

2.1. Introduction

The interaction of digital clinical data platforms, as formally defined, underlies the paradigm of clinical data interconnectivity. Being a specialisation of the wider area of digital data interoperability, the clinical data interconnection—besides the artificial intelligence technology—has been recognised as critical for the development of a sustainable global healthcare space based on cloud and service-oriented architectures. Clinical data interconnectivity is also a pre-requisite for other forms of data exchange in the healthcare realm, such as data sharing and data publication, which are supported by additional processes, services, solutions, and tools for privacy protection and risk management.

The need for different types of interconnected clinical data platforms originates from the foremost requirement of clinical data provenance in any healthcare applications and its fulfilment by common data models developed in parallel several decades ago and meant to serve as a bridge between different clinical databases for possibly different populations, disease entities, and even data-acquisition modalities. The frameworks and guidelines for applying risk assessment and mitigation are indicative of the emergent introduction of techniques for machine learning and natural language processing.

2.1.1. Overview of Clinical Data Interconnectivity

Accessibility, interconnectedness, and use of clinical data are often regarded as vital factors for the advancement of precision medicine and health research. Interconnectivity facilitates knowledge transfer among research and clinical institutions that gather, curate, and maintain different types and components of collected clinical data. Data processed and made available on the basis of a common framework or basic structure can enable easier queries and reuse for the benefit of specific user groups. Scientific findings can

thus be verified and replicated by independent user groups, such as through the testing of biomedical hypotheses or validation of blood products for transfusion.

Interconnection allows data gathered for a primary purpose to be reused repeatedly for secondary purposes, for example, enabling disease prediction and detection based on cardiotoxicity biomarkers. Increasing numbers of minor health care institutions are digitizing their paper records and linking them to resource libraries and specialized databases through basic data elements and metadata to facilitate information reuse and integration. Ominibus studies that combine a large variety of information types tie together multiple aspects of health and disease across both longitudinal and transversal approaches.

2.2. Rationale for Interconnection in Clinical Data

Interconnection of clinical data from multiple data platforms is increasingly being recognized as important. The availability of interconnected clinical data can facilitate novel research endeavours that were not previously feasible, and at the same time accelerate a plethora of use cases over existing data that require the ability to query multiple studies or institutions collectively. Examples of such use cases include population-size constrained validation, model learnings with scarce events, uncertainty quantification for rare diseases and biomarker discovery for multicentre trials or trials in low-prevalence diseases. The capability to analyse clinical research data that is derived from multiple studies or institutions can improve robustness and provide new insights, especially when external validation cohorts are difficult to establish due to scarcity of events.

Another prominent example is the integration of paediatric data for risk prediction and prognostication in acute respiratory illness. Here, two paediatric risk scores from separate translational cohorts are directly compared, and transferability to a third cohort is assessed. In another example, the risk of disease progression in patients with advanced Parkinson's disease is predicted using clinical and laboratory data collected at a single time point. Here, an important caveat is that these predictions are evaluated on the same dataset to which the model was fitted. These types of analyses necessitate the integration of data from multiple disease cohorts to provide a robust evaluation of the models, as ideally, the predictions should be tested on data not used to fit the model.

2.2.1. Justification for Clinical Data Interconnectivity

In recent years, the creation of a myriad of interconnected clinical data repositories has been seen as important for various applications, including data-driven personalized and

precision medicine and clinical research. The need for interconnectivity has been emphasized by numerous community initiatives such as the Broad Institute's Genomics Platform, the National Institute of Health-funded Accelerating Medicines Partnership in Osteoarthritis, the UK Biobank, etc. When data interconnectivity is achieved, clinical data repositories, typically developed for a single purpose, can serve additional purposes without incurring the cost of duplicating data generation efforts. Clinical data reuse across different domains of research often reduces the time required to conduct the research, improve the results, and in so doing potentially improve the quality of patient care.

As a showcase example, The Cancer Genome Atlas (TCGA) focuses exclusively on identifying the genetic, epigenetic and transcriptomic changes involved in 33 cancer types using genomic sequencing and genotyping data from over 10,000 patients. Nevertheless, it has been widely used for other applications, including the determination of human leukocyte antigen allelic expression, the identification of CtIP as a haploinsufficient breast cancer gene, and the characterization of molecular subtypes of head and neck squamous cell carcinoma among others. The data from TCGA, as well as other National Cancer Institute initiatives, have been hosted within the Genomic Data Commons to simplify access and support greater collaborative efforts.

2.3. Architectural Paradigms and Reference Models

Architectural patterns provide a conceptual structure for systems, while architectural models represent the ideas in a known format to facilitate discussion and communication. A layered architecture consists of vertical stacks representing different concerns and often employs a service-oriented approach. However, it scales poorly, and increasingly complex distributed applications are better visualized in a microservices pattern. Two well-known microservices reference models are the Twelve-Factor App methodology for application development and the microservices architecture for application deployment of the Martin Fowler and James Lewis.

Modern applications are hosted as a dynamic composition of independent, loosely coupled services structured around business capabilities. Each service is designed to fulfill a specific business task, supports a standard protocol with a well-defined interface, runs in its own process, is maintained by a small agile team, and can be supported by any language or technology. Applications interconnect services via HTTP, REST, messaging queues, and user-facing channels and feature crosscutting support for authentication, logging, and security. Optimal service autonomy is achieved via the data management principle, which specifies that each service should access its private data store. A robust recommendation set for microservices design incorporates the Twelve-Factor App methodology.

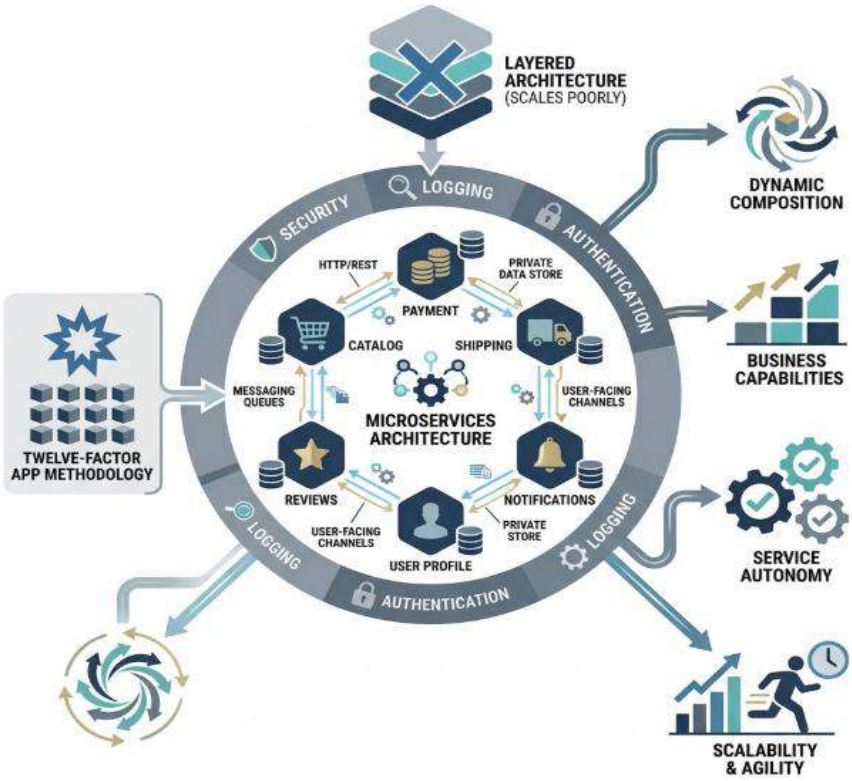

Fig 2.1: Architectural Evolution: Transitioning from Layered Structures to Autonomous Microservices and Twelve-Factor App Methodologies

2.3.1. Layered Architectures

A layered architecture of clinical data platforms typically consists of a User Layer, Application Layer, Data Layer, Infrastructure Layer, and Network Layer. The User Layer consists of users of the system, with privileges determining what data and functionality they can access. The Application Layer contains all platform logic, including functions, services, and maintenance utilities. The Data Layer incorporates a unified clinical database, data pipelines to ingest clinical data with transformations and qualities applied as necessary, orchestrators to manage the pipelines, and any other functions and tools necessary for properly managing the clinical data. The Infrastructure Layer consists of the physical resources needed to run the software and hardware in the User, Application, and Data Layers. The Network Layer formalises how the User Layer connects to the Application Layer, but does not specify the network infrastructure the layers use.

A general-purpose clinical data platform may also adopt a layered architecture, although the layers may span widespread teams and technologies. Such a platform enables development teams to build services that can be reused by other teams, so that secondary specialisms can still be incorporated into the product and services hosted within the platform. The platform itself may contain normal services, although ideally only for activities seldom used by other teams in the enterprise. Using the platform makes sense when interest crosses discipline boundaries, and a wide-scale or unique common interest has been recognised.

2.3.2. Service-Oriented and Microservices Approaches

Interconnected clinical data platforms can be built using service-oriented and microservices architectural styles. Service-oriented architectures (SOAs) are designed to provide decomposed functionalities packaged as services accessible via standardized interfaces. An SOA platform typically contains myriad services hosted in service containers, a service registry enabling the discovery and access of services, and a service orchestration component responsible for coordinating the execution of multiple services. SOAs have the advantages of being reusable, allowing services to be combined dynamically and/or simultaneously, and supporting interoperability across heterogeneous information technology infrastructures.

Microservices, or microservices architectures, are an architectural variant of SOAs that decompose an application into numerous loosely coupled and independently deployable services. In-microservices frameworks, each service performs its functionality in a self-contained manner and can be developed in any programming language. The typical communications mechanism across microservices architectures is lightweight Representational State Transfer APIs using Hypertext Transfer Protocol. Microservices architectures have the merits of shortening the development and deployment life cycle, providing flexibility and scale in the testing and refinement of services, and being able to use independent technology stacks.

2.4. Data Models, Standards, and Interoperability

Data models provide a vocabulary for defining the structure of data and its content, supported by a set of rules governing operations applicable to the data represented in that model, as well as the relationships that such data can maintain with other data objects. Common Data Models (CDMs), namely structures defined for a particular domain, or ontology models, such as formal specifications of a domain that usually describe concepts, relationships, and any associated rules therein, often serve distinct objectives. Nonetheless, ontological structures need to be operationalized for systems to execute

inference. Thus, the term Common Data Model is often broadened to incorporate an ontology-based operationalization that adheres to the original, or a close, intent when it comes to delivering the underlying data infrastructure for a particular domain. The pharmaceutical industry, for instance, has specified the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM) and its graph-enabled ODM-XML. Scientific facilities working on rare diseases and data for multiple nations have specified the Global Alliance for Genomics and Health (GA4GH) Genomic Knowledge Network (GDKN) as a Common Data Model for a myriad of data exchanges.

Interoperability denotes the capacity of a system (e.g. services, systems, products, devices or applications) to interact and exchange data with other systems and to utilize the exchanged data. Interoperability is needed at all levels of the architecture, namely messaging, semantic, de facto functional and operational. The need for standardized messages at the messaging level supports the adoption of a representative website thereof. Configuration guides for vendors and implementers at the de facto functional level increase the chance of services working together during an integration process. Reference architectures, such as the one produced by the Service Availability Forum (SA Forum) or the ISO Open Distributed Processing (ODP) Reference Model, are good basis for a study into the consistency of concepts present in the architecture of the interconnected clinical data platforms.

2.4.1. Common Data Models and Ontologies

Different types of clinical data platforms rely on a variety of data models. Data integration is made less complex when an organisation uses a standard logical data model, a common data model (CDM). When it encompasses a shared ontology it becomes more sophisticated. An ontology clarifies the semantics of data elements: their set of possible values along with relevant relationships with other data variables. Ontologies also aid data quality efforts by facilitating the detection, monitoring, and remediation of data quality issues.

For drug development, the Clinical Data Interchange Standards Consortium (CDISC) has developed a suite of SDTM and ADaM data models. The Observational Health Data Sciences and Informatics (OHDSI) community has developed a CDM to support observational health research. Within the paediatric domain, the Paediatric Translational Medicine Consortium is establishing a Paediatric Data Warehouse that merges genetic, genomic, and clinical data from multiple sources residing at multiple institutions. Its overarching goal is to model and understand the development of cancers in paediatric populations. The Paediatric Data Warehouse uses a data model built on the CDISC standardised ontology but is extending it to support integration efforts with the TGen Paediatric Cancer Genome Project.

2.4.2. Interoperability Standards and Messaging Protocols

OSLC and HL7 FHIR are two standards that address different levels of the interoperability spectrum: ontology and messaging, respectively. The Open Services for Lifecycle Collaboration (OSLC) is an open initiative to define key concepts and elements that appear across domains in the development and management of software. OSLC complements application programming interfaces (APIs) by enabling semantic interoperability across different tools from different vendors and thus facilitating service-oriented and plug-and-play capabilities. OSLC enables integration at a higher semantic level than pure REST APIs, capturing concepts like change sets, resources, important and less important change requests, launchable resources, and resource type relationship hierarchies. It specifies things such as resource types, operations, and the content of important resources, and it provides ontology for request types and resource types so third parties can create OSLC service providers and consumers. An OSLC link enables two applications to connect their users' contexts and share a resource relevant to that context. OSLC specifications define Ontology Resource Definitions (ORDs) as the basis for OSLC link choices.

HL7 FHIR is a standard to enable the secure exchange of health care data. It combines and leverages the best aspects of HL7's previous and current standards. FHIR consists of a set of logical and physical data models and a RESTful API and supports multiple transport protocols, including HTTP, MQTT, and WebSockets. The FHIR specification supplements RESTful API implementation guidance with best practices for establishing FHIR-based systems that communicate well through custom solutions with unparalleled ease of use. Each FHIR resource is a defined set of data that describes an aspect of health care, such as patients, medications, observations, schedules, and so on. FHIR resources are designed to be easily understandable by both human and machine, and they are represented and can be exchanged in JSON or XML.

2.5. Data Provenance, Quality, and Governance

Quality of data used in medical contexts is a growing concern since its origin (provenance) and any possible modifications are often unknown. This is a major drawback in interoperability solutions for medical data in general, and for the integration of data platforms of clinical and diagnostic laboratories in particular. In the absence of evidence-based approaches that connect the collected data with a source of clinical truth for quality evaluation, provenance tracking is key to the assessment of its fitness for purpose. Provenance-tracking mechanisms need to be complemented by the definition of quality dimensions, a framework for data quality assessment, and domains of clinical data quality, as well as the design of quality-flagging tools. Major areas of development encompass risk assessment, quality evaluation during data integration and exchange, the

generation of Medical Data Quality Reports, and the identification of quality-related research data management tasks.

Data quality is crucial in medical contexts, as errors may lead to wrong decisions with a high toll in personal suffering and public spending. The need for provenance tracking is especially relevant in clinical data interchanges and integrations, particularly when quality assessment against clinical truth is impossible. Provenance tracking alone cannot guarantee fitness for purpose, however, since not all medical databases are equally reliable and some do require quality evaluations. for medical data quality dimensions, a general-purpose data quality framework relevant in clinical settings in particular, and data quality domains specific to clinical data.

2.5.1. Provenance Tracking Mechanisms

A key requirement for research-relevant clinical data is clear documentation of data provenance, encompassing data sources, origins, transformations, quality, and accountability. Well-defined provenance information is important, not only for trustworthy information retrieval but also for reliable downstream analytical and machine-learning workflows. Indeed, machine-learning analysis can inadvertently amplify biases present in even small training datasets, resulting in unjust outcomes in practice.

Several platforms implement provenance-tracking mechanisms capable of capturing these provenance dimensions, including data origination, preparation, curating party, and quality. The administrative and computational complexities associated with retaining exhaustive provenance information, however, can limit provenance tracking adopting standard methods. Consequently, some data-science pipelines follow less rigorous approaches, focusing predominantly on data origination and preparation. Nevertheless, systems without comprehensive provenance tracking remain susceptible to detectability concerns and gaps associated with accountability and data quality. Data use–based governance models that efficiently whitelist permitted uses can help mitigate concerns related to data origination falsification, such as unauthorized sales of information.

2.5.2. Data Quality Frameworks

Reliable data depend not only on being accurate and timely, but also on conforming to community standards across disparate sources. The trusted vendor framework from Snowflake, designed for use in data warehousing, uses readiness, application, and science measures to determine data fitness. Readiness relies on logical quality

dimensions like completeness, consistency, and uniqueness. Snowflake incorporates the Pass protocol for sharing assessments of data-source readiness and can receive external-authenticated production code or backfill data from third parties such as Micron and Western Digital (for memory chips) and Goldman Sachs (market data). Application dimensions cover considerations such as likeness to the user, operational speed, and predictive power, while science measures incorporate covariates and assumptions about missing hazard values.

Assessments of communication, transmission, consumption, and request quality can be applied to exchanges and can also be integrated with risk and compliance assessments. A logical-data-quality-management framework maps quality dimensions, roles, and techniques for each step of the data life cycle. Tracing protocols, such as Provenance and W3C's PROV Ontology, cross-sectionally encode metadata and are thus suitable for provenance monitoring. Conceptual provenance provides high-level abstractions, such as the W3C RSPARQL protocol for streaming RDF, without requiring knowledge of the underlying storage mechanisms. Fit-for-purpose assessment is pragmatic and aims for the expected use cases rather than broad conformance to detailed standards.

2.6. Privacy, Security, and Risk Management

Privacy, confidentiality, and security-related issues are vital ingredients of public health information systems. Several standards exist for common risk management processes, including principles of de-identification, risk-benefit analysis, encryption, key management, and trust establishment in data sharing. De-identification techniques, adopting privacy models to protect patient confidentiality, providing mutual encrypted communication between stakeholders, and secure key management integration frameworks can ensure data sharing security and confidentiality.

De-identification procedures are techniques for modifying personal information about patients so that it cannot be easily associated with the patients. In this method, two types of distinct information are defined: direct identifiers must be removed from the data set and replaced with unique codes, while indirect identifiers must also be altered or suppressed. Confidentiality is the cornerstone of biomedical research; however, it often conflicts with openness and transparency principles, motivating proponents of a loss of confidentiality risk measure for evaluating the utility of sharing information openly without any de-identification of the data set. Moreover, additional attributes indicating ethnicity, religion, disability status, education level, and incomes are explicitly removed from the data set during the process.

A systematic analysis of de-identification discusses individual privacy risks for real data. A joint decomposition of the risk into membership probability, background knowledge

probability, and circumstances is proposed, covering several known de-identification privacy definitions. There is a need to characterize stronger privacy control mechanisms to mitigate de-identification and other privacy problems. Such techniques assist in keeping the data safe by ensuring the risk potential is acceptably low, thereby maintaining reasonable usability.

2.6.1. De-identification and Confidentiality Techniques

Every time when users, patients, or healthcare providers interact with a clinical data platform, they create sensitive information that can, in theory, be used against their interest. In many cases, the value of this information for social welfare (e.g., for business or research purposes) outweighs its potential damage to users' privacy. Therefore, privacy-preserving mechanisms allow for the exchange and exploitation of sensitive patient information while maintaining confidentiality and trust.

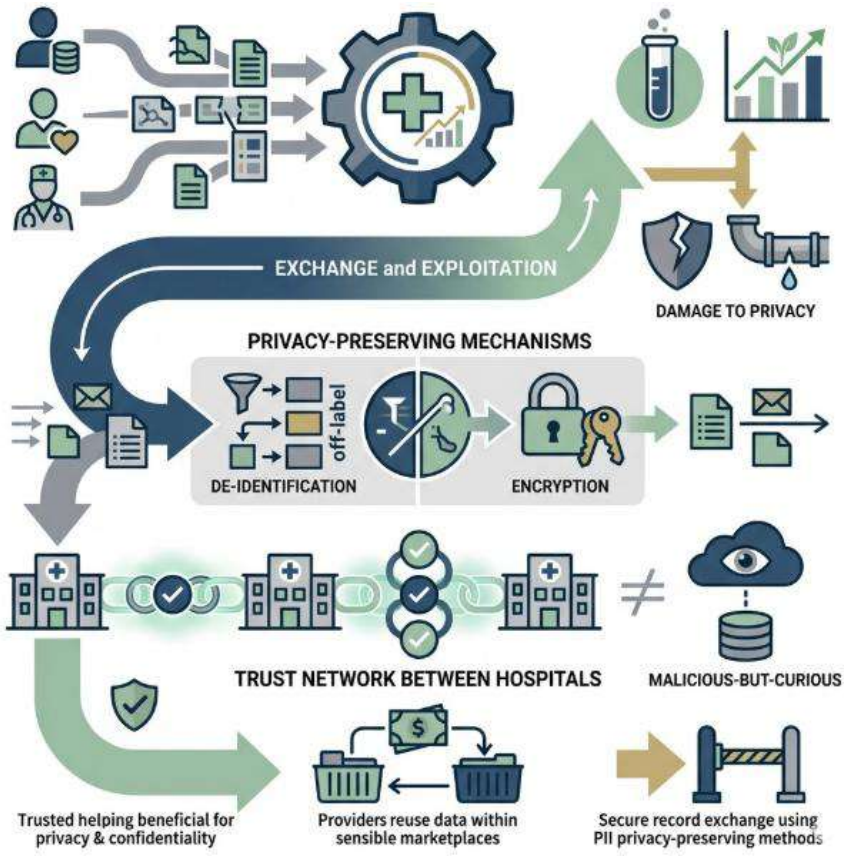

Fig 2.2: Secure Inter-Institutional Data Exchange: A Framework for Privacy-Preserving Mechanisms and Trust Networks in E-Health

These privacy-preserving mechanisms span from de-identification to encryption techniques. De-identification of Personally Identifiable Information (PII) is one of several main approaches to preserving privacy in data-sharing environments. PII may take several forms (e.g., patient names, addresses, Social Security Numbers). De-identification tools focus on removing such information. They modify a data object's identification fields (e.g., patient names) to prevent the association of the object with its original data subject. When leaks or data breaches occur, users' trust in platforms suffers. Past data contaminants (e.g., patient pathology images) can also lead to a loss of trust. Provider organizations work in sensible marketplaces, reusing PII and PHI data from others. Both, however, restrict public access between organizations. As a result, using mixed trust sources to exchange records between organizations while preserving full PII privacy becomes important with respect to trust networks and payment.

The use of privacy-preserving data-sharing and analysis approaches requires trust among all the data sources and sinks, involving both the patients and the hospitals concerned. Existing trust networks are not adapted for flexible and rapidly changing environments such as e-health data sharing and analysis. In specific cases, however, their messages or data may be more sensitive and should therefore only be made accessible to those who have a real need. Results indicate that a suitable trust network between hospitals is beneficial in terms of privacy and confidentiality during actual data sharing, compared to a malicious-but-curious model. In e-health, the use of second-party data sources is more privacy-aware than using a single public source.

2.6.2. Encryption, Key Management, and Trust Establishment

User's Privacy-Preserving Information Security Infrastructures (PIRPISI) concepts constitute valuable building blocks for the establishment of privacy-preserving cryptographic communication channels and databases. Infrastructure as a Service (IaaS) clouds, such as Amazon Web Services, typically employ Trusted Third Parties (TTPs) that store service providers' and clients' encrypted data and Kerberos servers for user identification, authentication, and de-identification. The IaaS cloud operator and storage service provider can have access to plaintexts because they store keys in plaintext. Confidentiality is provided to data stored in the cloud; however, it can be compromised by the TTP's leaks or by serving legally authorized government agencies.

The Pirpisi protocols allow the creation of a de-identification service and a cloud cloud-cryptography-as-a-service (CCAAS) for an IaaS cloud. The identification of users is done via a manager who associates a key in the private cloud IaaS with a user and the data is encrypted in empty boxes with a public key. Such data can only be decrypted with the right key in the key server. The approach is appropriate for egocentric networks. A trust manager is responsible for user identification, authentication, and key

assignment; an authentication server provides user authentication into a private cloud with a public key; and a TTP decrypts the encrypted content such-origin servers and authorized users can view the content in plaintext.

2.7. Integration Patterns and Data Pipelines

Three fundamental integration patterns have been established for interconnected data systems: batch-oriented extract, transform, load (ETL), real-time publish-subscribe messaging integrations employing intermediary brokers, and streaming micro-batch-based pipelines. Sequential ETL and its variants – extract-load-transform (ELT), extract-load-validate-transform (ELVT), and more – remain popular choices for producing high-quality data in a performant manner. In a prototypical ETL process, incremental ETL variants regularly extract data in bulk from source databases, apply transformations via scheduled jobs with data quality checks, and load the resultant datasets into the destination system for data sharing or analytic usage.

The function of receiving continuous data streams from sources and then filtering, enriching, and distributing the data to multiple sinks according to the publish-subscribe principle is implemented by messaging brokers. In contrast to sequential ETL pipelines, which operate in cycles of inactivity and activity, pipelines based on Apache Hadoop or similar technologies operate in micro-batches, gradually assimilating the incoming continuous data stream and executing discrete unit-of-work micro-batches at periodic intervals.

2.7.1. Extract, Transform, Load and Variants

Integrating clinical datasets can be based on the Extract–Transform–Load (ETL) approach or, more generally, the Extract–Load–Transform process. The ETL pattern involves the extraction of data from a variety of source systems, data preparation (cleaning, enrichment, validation, quality assessment), and loading into a single target repository (a data warehouse). A more generalized version of ETL is Extract–Load–Transform: the extracted data flows into a staging area or a data lake and the subsequent transformations are performed right before the data needs to be used.

In a Data Lake concept, storing all required data as close as possible to its source can be a viable alternative. End-user applications might apply whatever transformations necessary to the data at the moment of their actual usage. These transformations always entail a risk of introducing errors in the final result and often create a problem of persistent non-quality; therefore, transforming the data at the time of their usage is desirable, but should not be the only pattern applied. The Request/Response pattern,

where the Data Lake receives single queries, executes the required costly operations to create the needed result, and returns it to the requester when the same request is received at least twice in a reasonable amount of time, can also alleviate the non-quality problem.

2.7.2. Real-Time and Streaming Integrations

In addition to ETL and its variants, integrations that operate in real time or streaming mode represent a fundamental category within data interconnectivity architectures. In so-called Change Data Capture (CDC) integrations, data updates (inserts, updates, or deletes) are captured at source systems and forwarded to interested destinations as they occur—often with high granularity, i.e., per record. Equally critical are the orchestration and federation of data pipelines, involving the design of workflows that sequence, parallelize, or otherwise orchestrate the flow of data across diverse sources, intermediate transformations, and sinks.

CDC functionality is often provided out of the box by the underlying database management system (DBMS) and made accessible to external systems via a message bus or event streaming platform. More advanced scenarios can involve the deployment of dedicated CDC agents that capture commits in the DBMS's transaction log or pool the source systems' adaptation layer for changes. In contrast, support for orchestration, interaction, and data flow design is commonly available via external solutions. The predominant event-driven programming models supported by function-as-a-service platforms facilitate the execution of incremental, event-driven processing logic in close to real time.

2.8. Platform Infrastructure and Deployment Models

The infrastructure supporting medical data exchange across clinical platforms is an element of paramount importance. Decision-makers in the different organizations involved in the collaboration must decide on where the exchange and processing of the data will take place and deploy the necessary hardware or cloud services. The platforms may use cloud resources, on-premise servers, or a hybrid model combining both. By suitably configuring the data movement and transformation pipelines, the organization deciding on the infrastructure may choose between using a data lake or data warehouse as the target of the data transfers. Data mesh concepts enabling distribution of data across the data sources while ensuring the necessary service level objectives on the data transfers can also be considered.

Development, deployment, and configuration costs for the pipelines involved in the platform data-processing ecosystem need to be minimized to allow a bigger number of

clinical research projects to be executed with a data-economical approach. Scheduling and process execution complexity should also be kept in check. Thus, architectural paradigms emphasizing reusability and consolidation of continuous and frequent data transformations and movement should be prioritized.

2.8.1. Cloud, On-Premises, and Hybrid Deployments

Data platforms to support clinical data interconnectivity are implemented in a variety of infrastructures and deployments. Cloud-based infrastructures empower quick and agile development, scalability, and elasticity for platforms with unpredictable workloads. Additionally, these architectures increasingly assure cost control through pay-as-you-go pricing models. A downside is that some organisations prefer to keep their data under premises due to institutional policies or regulations or for strategic reasons, such as avoiding vendor lock-in. On-premises deployments are suitable for those customers with predictable workloads but limited and well-established usage patterns, thus reducing the costs associated with cloud infrastructure. Hybrid architectures exploit the benefits of cloud and on-premises infrastructures, thus allowing for burst capabilities on a need-basis without jeopardising data privacy.

In addition to the different cloud deployment options, interconnectivity architectures also exploit assured and well-tested approaches like the data lake and data warehouse paradigms, supported by a data mesh-like infrastructure concept. Data lakes accommodate data of different kinds, within and across clinical data linkages, without significant pre-processing. Data warehouses guarantee a proper design and integration of the data for a particular use case, assuring quality for a reduced subset of users. These models can also assist platforms potentially serving different user groups, with different levels of data quality and assurance policies, tackling also different yet overlapping use cases. A data mesh-like concept is proposed to accommodate these approaches with a federated infrastructure strategy. The deployment in clinical institutions of groups of logical data platforms, modelled through the described architectural patterns at the local level, and the federation of these platforms and groups of platforms constitutes a data mesh-like concept for the organisation and integration of clinical research data across the clinical institutions of a specific region.

2.8.2. Data Lakes, Data Warehouses, and Data Mesh Concepts

The data-holding architecture leverages the Cloud, its multiple-responsibility capabilities closely matching actual datastorage needs. Moving actual raw data-exchange activities to the cloud is a sound approach, particularly if testing Dynamic Product Data meshes. Datashare and other combinations show data-transmission costs

are bearable – unless transmission has to move large volumetric data at the origin of study (e.g., like medical images). Using the model for a Dynamic Product Data solution, engaged participants aim to provide answers to real operational and strategic business problems (e.g., acceleratory time-to-market, efficiency or quality improvement policy establishment, label volume cost reduction, and so on).

Where a Data Warehouse needs to fulfil a more traditional architectural setting, requirements could be focused on the ease of real-time connectivity with the Data Lake and on the latency of data availability (which should be lower than a Data Lake connection). Monitoring user requests can identify the real needs of the organization finally deciding the Data Warehouse architecture. It should be noted that, nowadays, documents can provide business operational knowledge, danger indicators, and product efficiency tracking; therefore, the effort can be directed towards making documents data consumable through appropriate pipelines.

2.9. Case Studies and Evaluation Metrics

The architectural elements and principles of interconnected clinical data platforms have been demonstrated and validated through several case studies that present different facets of clinical data interconnectivity, and their implications for clinical research.

The first set of case studies highlights the need for a practical interconnection architecture of clinical data platforms for data exchange in supporting clinical research, addressing the challenges posed by data-hungry machine learning models with limited datasets. It recounts the findings and post-launch of the FDA-CBER study on scientific and operational feasibility for connecting clinical data exchange systems across standards and governance. The study on truegenomic.com, the first genomic data exchange in conformity with the FDA's non-human primate COVID-19 studies, served to apply and validate the prevalent solutions mapped. The second set of case studies address the interconnection of clinical data platforms at a more mature level for a broader spectrum of data-hungry healthcare artificial intelligence. To this end, the pediatric and genomic clinical data collaboration between PHIS and BeID, including the underlying architecture, cloud premise and hybrid deployment alternatives, the clinical research data exchange facilitated by the associated Dapeliza-platform group, and the preclinical COVID-19 data exchange now also framing pediatric data in support of the curation of tgFAMsi.

Both sets echo a growing theoretical foundation that formally maps the architectural elements and principles of interconnection in clinical data platforms across an evolving landscape of cloud-based services, research interest, and governance mechanisms, all expressing interest for data-hungry healthcare AI. The formal concepts of

interconnection in clinical data platforms now enabled the identification of gaps and challenges at operational and research levels, and suggested directions to refine and advance experimental and pilot implementations.

2.9.1. Clinical Research Data Exchange

Clinical research organizations (CROs) frequently gather large amounts of clinical patient data to support the development of new medical products. Although these data sets are valuable to drug and device manufacturers, clinical data piracy and the cost of studies have led the CRO to seek ways to share these data sets more efficiently. Many FDA-regulated trials require the submission of separate clinical pharmacology studies, often at similar sites by different sponsors. CRO recognized this duplication of effort and conceived the Clinical Research Consortium to Share Data (CLINRADS) initiative. The goal was to build, expand, and operate a secure clinical data exchange infrastructure that enables the sharing of clinical data from completed FDA-regulated studies for non-commercial supporting research and advancement of science.

The initiative has developed virtualized clinical data exchange capability encompassing several therapies, and now provides an infrastructure to enable the more efficient exchange of completed data. The impetus was to demonstrate the feasibility of sharing these clinical data sets by providing a platform, procedure, and protocol to facilitate data requests and enable transparent process monitoring. The capability was built as a software-as-a-service offering and employs a service-oriented architecture. Currently, a petabyte of clinical data is held in a high-performance data warehouse and can be queried without data transfer, enabling rapid responses to data requests. The implementation is open-source, scalable, and potentially extensible to include data from non-CRO entities.

2.9.2. Pediatric and Genomic Data Integration

The integration of pediatric and genomic data from several clinical data research networks into a United States–based multi-network pediatric research effort is described. The multi-network effort leverages multiple centers to prospectively compile, in a heterogeneous-structured manner, population-based pediatric data for the discovery and validation of novel biomarkers and risk prediction models. The mother network employs a shared-individual-level-data approach via a central database, while the child networks focus on data harmonization for real-time data aggregation. Proposed are solution strategies for technical, operational, and governance aspects. A federated data-integration scheme that combines a central-individual-level-data repository with decentralized-meta-data repositories enables real-time discovery and validation of predictive models with stratified generalizability. Initial immature clinical application

continues to reveal biological principles for early prediction of skin outcomes, provides a platform for evaluating novel biomarkers, and creates a foundation for future data acquisition.

In the current scenario, with great advances in authentication and authorization technologies for clinical data sharing—especially for patient safety, liability, and ethical issues—real-time harmonization to support formal and informal multicenter studies is becoming more important than large-scale clinical trials needing reconciliation over a long timescale. Multi-network-based data integration affords advanced planning, such as harmonization methods with predictions of generalizability and clinical relevance of predictive outcomes, rather than just data aggregation. Population-based—book of clinical cases—using multi-network-based leverage is the future trend for clinical applications such as rare disease prediction, prediction of clinical severity and outcome, and disease prevention.

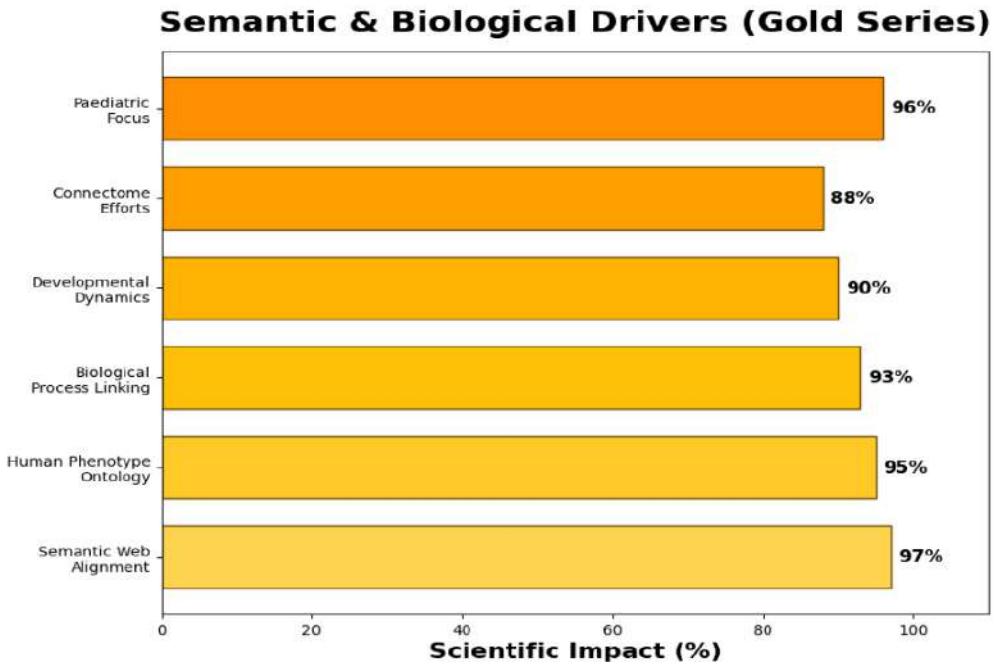

Fig 2.3: Semantic & Biological Drivers (Gold Series)

2.10. Challenges, Gaps, and Future Directions

The above considerations and efforts pave the way for clinically relevant applications through data interconnections, such as the Federated Research Network in The Cancer Data Alliance. However, many challenges remain, gaps still exist, and future directions can lead to more advances. Implementations have not yet reached all interconnection

types for all data types. Enabling higher-quality, real-time, on-demand, or value-driven answers still calls for scientific advances. The security, risk management, and privacy-potential approaches demand continual improvements. Some of the patterns have not yet been effectively implemented or evaluated.

An important point of focus involves enabling the clinical needs of biologically and developmentally driven research. For example, although the Semantic Web offers a sound basis for organizing connected data effectively, the results thus far have been better adapted to clinical rather than research environments. Efforts such as the connectome and the Human Phenotype Ontology have demonstrated alignment with the Semantic Web technology stack. A more systematic effort to identify, formally characterize, and comprehend the interconnections among children's data—especially among those linked to biological and developmental processes—could further drive advances in paediatric and developmental clinical research.

2.11. Conclusion

Interconnected clinical data platforms offer a compelling response to some of the major injustices and inefficiencies embedded within present health data practices. Whole genomes from thousands of children with cancer can now be sequenced, but the networks needed to facilitate full-scale genomic analysis or integrate these data with real-world clinical data remain nascent. Existing clinical research networks focus on data sharing of de-identified data rather than real-world data exchange. Data sharing is essential for advancing knowledge, but an architecture of interconnected clinical data platforms supporting data exchange can enable far more.

The pulling together of networks goes well beyond the exchange of accumulated research result data. In this mode of data integration, e.g. in a clinical data research share, these repositories have served as much as a depository, a long-term storage and archiving solution, as much as a place to enable the real-time use or integration of these research-centric clinical data in that it is designed to support only de-identified data, little of which is produced in actual real-world practice. Many of the currently operational data-sharing networks and networks in development are essentially clinical versions of the Internet Archive Repository, making key Buddhist sense through the connection of its propagating, first-in, first-out data-sharing philosophy with the data-sharing equivalent of an Internet-based proxy connecting to a search engine for quoranical access. Such functions are massively valuable yet insufficient and ignore nearly all dimensions of clinical research data beyond provenance.

References

Lehne, M., Sass, J., Essenwanger, A., Schepers, J., & Thun, S. (2019). Why digital medicine depends on interoperability. https://doi.org/10.1038/s41746-019-0158-1

Nagubandi, A. R. (2024). Breakthrough Real-Time AI-Driven Regulatory Intelligence for Multi-Counterparty Derivatives and Collateral Platforms: Autonomous Compliance for IFRS, EMIR, NAIC, SOX & Emerging Regulations. Journal of Information Systems Engineering and Management, 9.

Wesley, D. B., Blumenthal, J., Shah, S., et al. (2021). *A Novel Application of SMART on FHIR Architecture for Interoperable and Scalable Integration of Patient-Reported Outcome Data with Electronic Health Records*. Journal of the American Medical Informatics Association, 28(10), 2220–2225. https://doi.org/10.1093/jamia/ocab110

Gottimukkala, V. R. R. (2020). Energy-Efficient Design Patterns for Large-Scale Banking Applications Deployed on AWS Cloud. power, 9(12).

Kaushik, K. (2026). Scalable Cloud–AI Architecture for Synthetic Healthcare Data Generation and Anomaly Modeling: A Simulation-Based Study*. Discover Public Health, 23, 126. https://doi.org/10.1186/s12982-026-01464-6

Garapati, R. S., Adusupalli, B., Kaulwar, P. K., Gadi, A. L., Annapareddy, V. N., & Challa, K. (2025, December). The Evolution of Digital Payments: A Study on AI-Powered Transaction Monitoring Systems. In 2025 3rd International Conference on IoT, Communication and Automation Technology (ICICAT) (pp. 1-8). IEEE.

Tummers, J., Tobi, H., Catal, C., & Tekinerdogan, B. (2021). *Designing a Reference Architecture for Health Information Systems*. BMC Medical Informatics and Decision Making, 21, 210. https://doi.org/10.1186/s12911-021-01570-2

Kolla, S. H., Inala, R., & Kumar, M. V. K. (2026). Secure RAG Architectures with Small Language Models for Governance-Aligned LLM Deployment in Enterprise Service Management Platforms. International Journal of Economic Practices and Theories, 2026, 166-179.

Amershi, S., et al. (2023). Software engineering for machine learning. *ICSE Proceedings*.

Davuluri, P. N. (2020). Improving Data Quality and Lineage in Regulated Financial Data Platforms. Finance and Economics, 1(1), 1-14.

Doan, A., Halevy, A., & Ives, Z. (2012). Principles of data integration. Morgan Kaufmann.

Segireddy, A. R. (2021). Containerization and Microservices in Payment Systems: A Study of Kubernetes and Docker in Financial Applications. Universal Journal of Business and Management, 1(1), 1-17.

Dong, X. L., Rekatsinas, T., Gokhale, C., & Thirumuruganathan, S. (2020). Data integration and machine learning: A natural synergy. In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (pp. 2835–2838).

Bandi, V. D. V. K. Autonomous Data Platforms: Converging AI, MLOps, and Cloud Engineering for Digital.

Abedjan, Z., Golab, L., & Naumann, F. (2015). Profiling relational data: A survey. The VLDB Journal, 24(4), 557–581.

Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016). Data cleaning: Overview and emerging challenges. In Proceedings of the 2016 International Conference on Management of Data (pp. 2201–2206).

Amistapuram, K. (2021). Digital Transformation in Insurance: Migrating Enterprise Policy Systems to .NET Core. Universal Journal of Computer Sciences and Communications, 1(1), 1-17.

Rekatsinas, T., Chu, X., Ilyas, I. F., & Ré, C. (2017). HoloClean: Holistic data repairs with probabilistic inference. Proceedings of the VLDB Endowment, 10(11), 1190–1201.

Botlagunta Preethish Nandan. (2021). Enhancing Chip Performance Through Predictive Analytics and Automated Design Verification. Journal of International Crisis and Risk Communication Research , 265–285. https://doi.org/10.63278/jicrcr.vi.3040.

Chu, X., Morcos, J., Ilyas, I. F., Ouzzani, M., et al. (2015). KATARA: Reliable data cleaning with knowledge bases and crowdsourcing. Proceedings of the VLDB Endowment, 8(12), 1952–1955.

Kolla, S. H. (2026). Autonomous Enterprise Agents: Orchestrating Large and Small Language Models for Scalable Decision Automation in ITSM, HRSD, and CSM Platforms. INTERNATIONAL JOURNAL OF ADVANCES IN SIGNAL AND IMAGE SCIENCES,

Wang, J., Krishnan, S., Franklin, M. J., Goldberg, K., & Ré, C. (2021). Learning to clean data with data programming. Proceedings of the VLDB Endowment, 14(11), 2107–2120.

Nagabhyru, K. C. (2024). Data Engineering in the Age of Large Language Models: Transforming Data Access, Curation, and Enterprise Interpretation. Computer Fraud and Security.

Uday Surendra Yandamuri. (2023). An Intelligent Analytics Framework Combining Big Data and Machine Learning for Business Forecasting. International Journal Of Finance, 36(6), 682-706. https://doi.org/10.5281/zenodo.18095256

Mandel, J. C., Kreda, D. A., Mandl, K. D., Kohane, I. S., & Ramoni, R. B. (2016). SMART on FHIR: A standards-based, interoperable apps platform for electronic health records. [https://doi.org/10.1016/j.jbi.2016.03.004]