Enhancing Clinical Data Infrastructure for AI Research: A Comparative Evaluation of Data Management Architectures

Richard Gebler, Ines Reinecke, Martin Sedlmayr, Miriam Goldammer

Submitted to: Journal of Medical Internet Research
on: March 26, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Enhancing Clinical Data Infrastructure for AI Research: A Comparative Evaluation of Data Management Architectures

Richard Gebler[1]; Ines Reinecke[2]; Martin Sedlmayr[1]; Miriam Goldammer[1]

[1] Faculty of Medicine at University Hospital Carl Gustav Carus, TUD Dresden University of Technology Dresden DE
[2] Data Integration Center, Center for Medical Informatics, University Hospital Carl Gustav Carus at Dresden University of Technology Dresden DE

Corresponding Author:
Richard Gebler

Faculty of Medicine at University Hospital Carl Gustav Carus, TUD Dresden University of Technology
Fetscherstraße 74
Dresden
DE

Abstract

Background: The rapid growth of clinical data, driven by digital technologies and high-resolution sensors, presents significant challenges for healthcare organisations aiming to support advanced AI research and improve patient care. Traditional data management approaches may struggle to handle the large, diverse and rapidly updating datasets prevalent in modern clinical environments.

Objective: This study compares three clinical data management architectures - clinical data warehouses (cDWH), clinical data lakes (cDL) and clinical data lakehouses (cDLH) - by analysing their performance using the FAIR principles and the Big Data 5 Vs (Volume, Variety, Velocity, Veracity, Value). The aim is to provide guidance on selecting an architecture that balances robust data governance with the flexibility required for advanced analytics.

Methods: We developed a comprehensive analysis framework that integrates aspects of data governance with technical performance criteria. A rapid literature review was conducted to synthesise evidence from multiple studies, focusing on how each architecture manages large, heterogeneous and dynamically updating clinical data. The review assessed key dimensions such as scalability, real-time processing capabilities, metadata consistency, and the technical expertise required for implementation and maintenance.

Results: The results show that cDWHs offer strong data governance, stability and structured reporting, making them well suited for environments that require strict compliance and reliable analysis. However, they are limited in terms of real-time processing and scalability. In contrast, cDLs offer greater flexibility and cost-effective scalability for managing heterogeneous data types, although they may suffer from inconsistent metadata management and challenges in maintaining data quality. cDLHs combine the strengths of both approaches by supporting real-time data ingestion and structured querying; however, their hybrid nature requires high technical expertise and involves complex integration efforts.

Conclusions: The optimal data management architecture for clinical applications depends on an organisation's specific needs, available resources, and strategic goals. Healthcare institutions need to weigh the trade-offs between robust data governance, operational flexibility and scalability to build future-proof infrastructures that support both clinical operations and AI research. Further research should focus on simplifying the complexity of hybrid models and improving the integration of clinical standards to improve overall system reliability and ease of implementation.

(JMIR Preprints 26/03/2025:74976)
DOI: https://doi.org/10.2196/preprints.74976

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Enhancing Clinical Data Infrastructure for AI Research: A Comparative Evaluation of Data Management Architectures

Abstract

Background: The rapid growth of clinical data, driven by digital technologies and high-resolution sensors, presents significant challenges for healthcare organisations aiming to support advanced AI research and improve patient care. Traditional data management approaches may struggle to handle the large, diverse and rapidly updating datasets prevalent in modern clinical environments.

Objective: This study compares three clinical data management architectures - clinical data warehouses (cDWH), clinical data lakes (cDL) and clinical data lakehouses (cDLH) - by analysing their performance using the FAIR principles and the Big Data 5 Vs (Volume, Variety, Velocity, Veracity, Value). The aim is to provide guidance on selecting an architecture that balances robust data governance with the flexibility required for advanced analytics.

Methods: We developed a comprehensive analysis framework that integrates aspects of data governance with technical performance criteria. A rapid literature review was conducted to synthesise evidence from multiple studies, focusing on how each architecture manages large, heterogeneous and dynamically updating clinical data. The review assessed key dimensions such as scalability, real-time processing capabilities, metadata consistency, and the technical expertise required for implementation and maintenance.

Results: The results show that cDWHs offer strong data governance, stability and structured reporting, making them well suited for environments that require strict compliance and reliable analysis. However, they are limited in terms of real-time processing and scalability. In contrast, cDLs offer greater flexibility and cost-effective scalability for managing heterogeneous data types, although they may suffer from inconsistent metadata management and challenges in maintaining data quality. cDLHs combine the strengths of both approaches by supporting real-time data ingestion and structured querying; however, their hybrid nature requires high technical expertise and involves complex integration efforts.

Conclusions: The optimal data management architecture for clinical applications depends on an organisation's specific needs, available resources, and strategic goals. Healthcare institutions need to weigh the trade-offs between robust data governance, operational flexibility and scalability to build future-proof infrastructures that support both clinical operations and AI research. Further research should focus on simplifying the complexity of hybrid models and improving the integration of clinical standards to improve overall system reliability and ease of implementation.

Keywords: Clinical Data Management; Big Data in Healthcare; Data Architecture Evaluation; Data Warehouse; Data Lake; Data Lakehouse

Introduction

As digital technology advances, the healthcare sector is experiencing a considerable increase in data volume. Forecasts predict an average annual increase in the volume of data generated worldwide of approximately 24.37%, a trend that underscores the significance of data management of big data in the medical field [1]. The integration of modern sensors and wearables, the adoption of electronic health records, and advancements in medical imaging and genome sequencing have led to an increase in both semi-structured and unstructured data [2,3]. The adoption of HL7 Fast Healthcare Interoperability Resources (FHIR) as an international standard relying on semi-structured data formats, and its adoption by the European Union for cross-border healthcare, has changed the digital landscape by making it easier to integrate and exchange these complex datasets [4]. In order to achieve a sustainable and transparent secondary use of data, it is crucial that data governance and management practices address organisational requirements, ensuring data is securely available, traceable, and managed in compliance with established standards for data exchange and security.

The availability of large and well-prepared datasets is a key factor in the development of modern AI applications, such as predictive analytics and personalised medicine. These algorithms are able to learn patterns and insights more accurately when they have access to clean and consistent data [5]. In practice, this requires healthcare organisations to implement comprehensive data collection, integration, and quality control processes to ensure that the final datasets accurately reflect the truth from the complex clinical environment. Underscoring the importance of reliable, standardised data for training robust machine learning models and facilitating real-time clinical decision support, the selection of data architecture has a direct impact on the ability to apply AI systems and their subsequent performance.

Looking at the current state of data architectures in the clinical application highlights the legacy of clinical data warehouses (cDWHs), a data architecture that has been established for decades [6]. As central hubs, they have been used to bring together data from various sources, optimise data processing by ensuring interoperability standards and protect patient data. Historically, using cDWHs, data has been organized in table structures. This helps create one reliable source of truth for all data storage systems. To understand the resulting challenges from this state of the art, some key information about DWHs need to be considered. In DWHs, transactions follow ACID properties (atomicity, consistency, isolation, durability). Atomicity means transactions are all-or-nothing, consistency ensures data remains the same before and after transactions, isolation means transactions do not affect each other and durability ensures that once a transaction is completed, it stays saved. If there is a failure or a change in the data structure, the existing data is not affected. Also, changes are not made automatically. These properties make sure the data is valid and reflect the high priority of data validity in healthcare.

However, cDWHs now face growing challenges. The rapid increase in data volume from diverse sources, such as imaging, sensor data, and genomics, makes it difficult to maintain the traditional fixed schema approach [7]. In addition, AI-driven workflows require not only large and varied datasets but also near real-time data processing to handle streaming information and provide timely insights [8]. If data arrives in formats with inconsistent structures or with missing values, it can reduce the accuracy and reliability of AI outputs. This situation has led to a new set of technical demands on data management, including more flexible storage, advanced data governance strategies, and scalable infrastructure for large-scale analytics [9].

To meet these increasing requirements, more recent concepts of data lakes and data lakehouses have been adapted to the clinical context, resulting in clinical data lakes (cDLs) and clinical data lakehouses (cDLHs).

A cDL is an architectural approach that enables the storage of large amounts of raw data in its original format, including structured, semi-structured and unstructured data [10]. This is in contrast to traditional cDWHs, which store data in a highly structured, organized way. cDLs are designed to provide high scalability and flexibility in data storage, allowing organisations to store data without the need for a defined schema first.

The cDLH is a new, more complex hybrid approach that combines the scalable storage options of a cDL with the structured queries and performance optimisations of a cDWH [11]. This architecture aims to overcome the limitations of the two previous models by providing a platform on which raw data can be stored cost-effectively and processed in structured formats as required by applying open data formats and ACID-transactions as well as comprehensive data management techniques. However, this architecture is still relatively new and designed primarily for cloud-based infrastructures, which may not align with the strict on-premise requirements of many European healthcare institutions.

This paper aims to analyse and compare the data management architectures cDWH, cDL and cDLH.

We focus on aspects such as data variety, real-time analytics, governance, and regulatory requirements that are necessary for reliable and AI-enabling healthcare. By evaluating clinical data architectures against established technical and data governance requirements - guided by the FAIR Principles and the 5 V's of big data – we provide in-depth insights into their efficiency and usability for clinical research data management. Our goal is to develop practical guidance for healthcare institutions of different aims and sizes, offering support in choosing a future-oriented data architecture that improves both medical research and patient care.

In the following sections, we describe our methodology, present the comparative results, discuss the implications of our findings, and conclude with recommendations.

Methods

To perform a systematic analysis and comparison of data management architectures designed to manage clinical data, we applied a comprehensive methodological framework that integrates the 5 V's of Big Data with the FAIR Principles. The resulting requirements are specifically tailored to the needs of big data research in healthcare to enable the comparison of the data architectures.

Step 1: Defining Analysis Criteria

In order to enable a comprehensive comparison of the architectures for clinical data management, we decided to analyse two different types of requirements. On the one hand, we aim to analyse aspects that relate to the more fundamental handling and management of the data itself and can be summarised under the term data governance. The FAIR Principles are a suitable framework for this, which has been established in recent years [12,13]. On the other hand, we aim to cover technical requirements, e.g. the handling of large amounts of data or the speed of data updates. These are considered by deriving analysis criteria from the 5 Vs of Big Data. We then adapted both frameworks to the clinical context and specified their general definitions to highlight healthcare characteristics, e.g. interoperability standards in healthcare and typical complexity of clinical data. Moreover, we emphasise certain aspects that are directly relevant to AI applications, such as data quality and real-time processing. This underlines the importance of having a data architecture that can support advanced AI methods.

Step 2: Conducting a Rapid Literature Review

Based on the requirements defined in section 2.1, a rapid literature review was conducted in relevant databases on the PubMed, Scopus, ACM Digital Library and Web of Science databases.

We focused on literature that addressed the relationship of data architectures to the FAIR Principles or to at least three of the five Vs of Big Data (Volume, Velocity, Variety, Veracity, Value).

We applied inclusion and exclusion criteria as provided in Textbox 1.

Textbox 1. Inclusion and exclusion criteria for the rapid review.

Included if a paper:

- discussed at least one of the data architectures (cDWH, cDL, or cDLH).
- addressed the FAIR Principles (Findable, Accessible, Interoperable, Reusable) or covered at least three of the 5 Vs of Big Data in a meaningful way.
- connected these frameworks (FAIR and/or 5 Vs) with the chosen data architecture.
- focused on clinical, healthcare or biomedical contexts.

Excluded if a paper:

- was not available to us as a full text.
- was not written in English.

Following a PRISMA-like workflow, all records from the four databases were assembled using the combined search string, as provided in Textbox 2.

Textbox 2. Search string for the literature search.

("clinical" OR "medical" OR "healthcare" OR "biomedical") AND ("lakehouse" OR "data lake" OR "data warehouse") AND ("FAIR Principles" OR ("findable" AND "accessible" AND "interoperable" AND"reusable") OR "5V" OR (("volume" AND "velocity" AND "variety") OR ("volume" AND "velocity" AND "veracity") OR ("volume" AND "velocity" AND "value") OR ("volume" AND "variety" AND "veracity") OR ("volume" AND "variety" AND "value") OR ("volume" AND "veracity" AND "value") OR ("velocity" AND "variety" AND "veracity") OR ("velocity" AND "variety" AND "value") OR ("velocity" AND "veracity" AND "value") OR ("variety" AND "veracity" AND "value")))

Next, duplicate records were removed, followed by a title and abstract screening to quickly exclude papers that were clearly out of scope. Finally, a full-text review of the remaining papers was performed to determine whether they met all inclusion criteria. During this process, one person screened the titles, abstracts and full-paper.
Then, a detailed review of these studies and documentation of their key characteristics were carried out, resulting in a comparison table.

In the discussion, we append with step 3:

Step 3: Developing Recommendations for Healthcare Data Management

Based on the results of the literature review and supplemented with secondary literature, we developed a set of recommendations. Given that AI applications rely heavily on the underlying data infrastructure, it is essential that the recommendations also ensure that data is optimised for advanced AI analytics. These recommendations are aimed at healthcare executives and technical experts involved in the evaluation, selection and implementation of data management solutions. In addition, we considered the implications of each data architecture for AI applications. For example, the ability to support real-time data streaming, high-quality structured data and flexible schema management are key for AI-driven tasks such as clustering, classification, and predictive analytics. Detailed considerations on these recommendations will be presented in the Discussion section.

Therefore, in this paper, we present the complete process from requirements through comparison of variants to use-case specific recommendations for clinical data architectures.

Results

To arrive at our findings, we first present a set of comprehensive requirements derived from existing frameworks. Given the increasing role of AI in healthcare, these requirements are critical as they support the data quality and structure that AI systems demand. These requirements serve as the basis for our following analysis of clinical data architectures.

Defining Analysis Criteria

The FAIR Principles set out the most important governance requirements for data management. Specified for (clinical) data architectures, these aspects ensure that the data is [14]:

- **Findable**: Data must be identified by unique and persistent identifiers [14] (e.g., GUIDs). It should be indexed or catalogued so both humans and machines can locate relevant datasets easily. Descriptive metadata must be consistently provided [14]. While "Findable" might intersect with “Interoperable” in terms of using common standards, its main emphasis is on how easily data can be located rather than how data integrates with other systems. In a clinical context, patient data often resides in multiple departments (e.g. the main clinical information system or specialised systems for laboratory results, imaging data). Unique identifiers, such as the patient IDs, medical case IDs and lab result ids as well as well-structured metadata enable clinicians or researchers to quickly find and retrieve the needed information [13].

- **Accessible***:* Data should be retrievable via standard communication protocols and appropriate authentication or authorization methods [14]. Even if the data itself becomes unavailable over time, associated metadata should remain accessible [14]. “Accessible” emphasizes secure retrieval under clearly defined rules [14]. This differs from “Findable” (which focuses on locating data) and from “Interoperable” (which focuses on format and standardisation). In a clinical context, a clinician who e.g. needs a patient’s MRI scans from the radiology archive must be able to securely access the images through hospital-approved protocols (e.g., DICOM servers, controlled hospital networks [12]). This ensures that sensitive data is protected yet still available to authorized personnel.
- **Interoperable**: Data is stored in commonly applicable, machine-readable formats, using standardized vocabularies (e.g., SNOMED CT, LOINC) to maintain semantic consistency. It should link to related datasets and metadata, enabling systems to exchange and understand the information [14]. “Interoperable” focuses on seamless data exchange and shared definitions. It does not overlap with “Accessible,” which deals with how authorized users obtain data, or “Reusable,” which deals with how data can be used repeatedly. In a clinical context, when patient data is transferred between hospitals or systems, interoperability ensures that e.g. terms like “blood pressure” or “fever” are interpreted identically. Thus, a cardiology department in Hospital A can share vital signs data with Hospital B’s critical care unit without losing meaning, thanks to standardized formats like HL7 FHIR [12].
- **Reusable***:* Data must have detailed domain-specific metadata and clear usage licenses or conditions so that it can be effortlessly repurposed for new studies or applications [14]. “Reusable” highlights the potential for data to be employed in future projects. This does not repeat “Findable” or “Interoperable,” as it deals with the conditions enabling repeated use rather than how data is discovered or shared. In a clinical context, e.g. a research group conducting a clinical trial on diabetes may decide to use past laboratory measurements (e.g., glucose levels) if these values are documented in a standardised format. Clearly stated permissions and provenance records allow them to reuse the data responsibly, improving reproducibility and speeding up new investigations [13].

While the FAIR Principles address data management and governance aspects, ensuring that data can be effectively found, retrieved, combined and reused, the area of big data also presents significant technical challenges. These challenges are captured by the 5 V's of Big Data - Volume, Velocity, Variety, Veracity and Value - which serve as essential characteristics and requirements for modern data management systems:

- *Volume* addresses the challenge of storing, processing and managing large data sets from sources [15,16] like electronic health records, whole-genome sequencing, and high-resolution imaging. It includes handling structured clinical data and the costs of scaling storage and processing capabilities. This ensures data integrity and accessibility without sacrificing security. “Volume” purely addresses the size and storage challenges. It does not address different data formats (which are the focus of “Variety”), or data flow speed (which is the focus of “Velocity”).
- *Variety* refers to the management of different types of medical data [15]. This includes structured data from electronic medical records and laboratory results, semi-structured data

such as HL7 v2 messages or FHIR resources and unstructured data such as clinical notes, waveforms and radiology images. This ensures a comprehensive view of patient information. “Variety” is about handling different types of data. It does not address data quality (focus of “Veracity”), or the usefulness of data insights generated from this comprehensive view of the patient (evaluated by “Value”). While “Interoperable” focusses on exchanging the same data between systems, “Variety” addresses the combination of different kinds of data in one system.

- *Velocity* includes the ability to both efficiently batch process large volumes of mass data and handle speed-critical data updates in near real-time via streaming [15,16]. This duality supports time-critical medical decisions by providing up-to-date patient data and enables the rapid integration of new data sources to ensure data relevance and currency [16]. Velocity” centers on the timeliness of data processing, independent of the amount (“Volume”) or kind (“Variety”) of data.
- *Veracity* highlights the critical importance of data quality and integrity, including traceability and comprehensive data management to ensure data accuracy and reliability [15]. This includes the development of a semantic understanding that allows the meaning and context of data to be captured and interpreted, which is essential for making accurate clinical decisions and maintaining patient safety [16,17]. “Veracity” emphasizes trustworthiness and data quality. This differs from “Interoperable” in the FAIR Principles, even though interoperability supports better veracity of data by reducing sources of error through common standards.
- *Value* measures the ability to query and analyse data to generate insights for research and patient care [18]. This includes support for analytics tools such as business intelligence (BI), big data exploration, machine learning (ML) and AI, which are essential for extracting insights from complex data sets [18]. It also considers the complexity in maintenance, handling and management of data management tools and infrastructures to not only maximize the immediate value but also ensure the long-term sustainability and adaptability of data management solutions. “Value” focuses on the eventual benefits of data usage. However, medical research has high demands on “Veracity” of data as a prerequisite to create “Value” from it. Furthermore, bringing different data sources with high “Variety” together is often a key element of success in clinical research. Therefore, clinical “Value” will often benefit from high performance in these two Vs. Also, while the FAIR framework does not address the “Value” created by data, it aims to provide the data governance necessary to create “Value” from data in a sustainable way by “Reuse”.

Taken together, the FAIR Principles and the five V’s establish a robust set of both governance and technical requirements against which clinical data architectures can be systematically assessed.

Conducting a Rapid Literature Review

Searching for data architectures combined with the FAIR principles or the 5Vs, a total of 43 records were identified from the combined databases. After the removal of seven duplicates, the titles and abstracts of 36 papers were screened, leading to the exclusion of 22 that did not meet the inclusion criteria. Then, a full-text review was conducted on the remaining 14 papers, resulting in the exclusion of another five papers. These exclusions were primarily due to a lack of clear descriptions of data architecture or insufficient consideration of the FAIR Principles or the 5 Vs of Big Data. The final result set contained nine papers. This workflow is illustrated in figure 1.

As a result of the rapid literature review, the following characteristics of the data management architectures cDWHs, cDLs and cDLHs are shown in Table 1.

Table 1. Comparison of the data management architectures using the 5 V’s of Big Data [7,19–26].

Although certain criteria overlap, this evaluation differentiates between governance and metadata management (FAIR) and technical performance metrics (5V's). This means, a quality control system that includes automated validation and detailed provenance tracking can be listed under 'Veracity' to show trustworthiness and consistent schema enforcement. It can also be listed under 'Reusability' to show its capacity to support repeated use of the data.

Requirement	cDWH [7,19,21–23,25,26]	cDL [19,21,24–26]	cDLH [20,25]
Findable			
	Pros: Centralised metadata with unique persistent IDs, standardised schemas, integrated catalogues, detailed data lineage	Pros: Flexible schema-on-read enables dynamic cataloguing, supports diverse data formats	Pros: Hybrid metadata management combines centralised and federated approaches, global unique identifiers, standardised API interfaces enhance integration, supports structured and unstructured data
	Cons: Fixed structure limits adaptation to heterogeneous data, high maintenance effort, scalability challenges	Cons: Decentralised metadata can lead to inconsistencies, missing fixed data schema complicates cross-system retrieval, requires continuous harmonisation	Cons: Cons: High initial financial and resource investment (including hardware acquisition, infrastructure setup, and extensive personnel training), coupled with complex inter-organisational coordination
Accessible			
	Pros: Centralised repository, standardized SQL access, robust authorization controls, high data availability, user-friendly interfaces	Pros: Flexible REST API access, support for HL7 and FHIR data feeds, diverse access methods, versioned data access	Pros: Open and standardized APIs and interfaces (REST, SQL), scalability through distributed services, high availability, seamless integration of structured and unstructured data
	Cons: Limited flexibility for external access, expensive scaling for very large volumes	Cons: Variable performance under heavy loads, requires advanced technical skills, complex protocol management	Cons: Complex initial setup for multi-layered data access, requiring advanced technical expertise in configuring distributed systems, securing diverse interfaces (e.g. REST, SQL), and integrating heterogeneous APIs
Interoperable			
	Pros: Standardized SQL-based data models, centralized ETL ensuring traceable data harmonization, comprehensive metadata management	Pros: Supports semantic focus on data governance through metadata, flexible integration with external systems, flexible schema-on-read	Pros: Unified architecture leveraging shared services, controlled vocabularies with standardized API frameworks, hybrid real-time and batch processing, integrated metadata frameworks
	Cons: Rigid, batch-oriented schema limits rapid updates, high curation and maintenance effort	Cons: Decentralized structure complicates harmonization, continuous integration demands, potential lack of ACID consistency	Cons: Significant initial integration and governance effort, coordination challenges across federated systems
Reusable			

	Pros: Structured data model, detailed provenance, comprehensive metadata, robust quality controls, strong compliance	Pros: Flexible schema-on-read, accommodates diverse data types, supports curated data marts, enables versioning	Pros: Combines structured reliability with flexible data handling, semantic enrichment, integrated metadata and versioning
	Cons: Fixed schemas limit adaptability and exploration of new research questions	Cons: Inconsistent metadata, variable data quality	Cons: Integration and maintenance require significant resources
Volume			
	Pros: Optimised for large volumes of structured data with efficient indexing and query performance	Pros: Highly scalable distributed storage is cost effective for very large (petabyte-scale) raw data	Pros: distributed software architecture supports dynamic scaling, balanced support for structured and unstructured data with large volumes
	Cons: Requires expensive vertical scaling, relies on batch processing causing performance peaks		Cons: Complex resource management and integration required
Variety			
	Pros: Defined schemas maintain consistent quality for homogeneous data types from multiple data sources	Pros: Schema-on-read easily integrates structured, semi-structured and unstructured data, highly scalable for diverse formats	Pros: supports diverse data types and evolving standards, supports multiple data models
	Cons: Inflexible when facing rapidly changing or unstructured formats, integration of heterogeneous data types requires specialized solutions or blobs	Cons: Requires robust ontology mapping to maintain semantic consistency	Cons: Demands ongoing monitoring, requires strong governance to manage new formats effectively
Velocity			
	Pros: Established batch processing, consistent scheduled updates, ideal for retrospective analysis	Pros: Near real-time ingestion, rapid processing enabling timely analytics, support for continuous data streams	Pros: Integrates real-time and batch processing using advanced frameworks (e.g., Delta Lake), achieves low-latency pipelines
	Cons: Limited real-time capabilities, higher latency when current data is needed, scalability challenges	Cons: Complex resource optimisation, increased integration and harmonisation demands	Cons: Challenging configuration and scaling

Veracity			
	Pros: Strong data quality management, harmonised data and clear lineage build trust	Pros: Retains original integrity of data, versioning supports traceability	ACID transactions, enforced schemas and advanced version management ensure high data integrity and real-time validation
	Cons: Complex handling of unstructured data, basic versioning and error correction restrict comprehensive quality assurance	Cons: Complex harmonisation of heterogeneous data and limited lineage tracking can challenge consistency	Cons: Demands higher technical and governance efforts to maintain quality
Value			
	Pros: Centralised integration produces reliable research utilities, analytics and reporting	Pros: Enables flexible data exploration with real-time insights, supports diverse data types and exploratory research	Pros: Combines reliability of traditional warehousing with innovative data handling, high data quality, interoperability and scalability add overall value
	Cons: High implementation and maintenance costs, limited flexibility may restrict secondary use	Cons: Continuous management required to ensure long-term value	Cons: Specialist expertise and significant integration efforts needed for effective management

These results show that choosing the right data architecture for clinical research and healthcare depends on an institution’s needs, resources, and goals. Our literature review shows that each clinical data architecture has different requirements in terms of implementation effort, maintenance, technical expertise and coordination. The analysis shows that the complexity increases progressively from cDWH to clinical cDL and cDLH [25,27–29]. The following discussion examines the complexity associated with each data architecture and outlines the most appropriate use cases.

Discussion

Complexity

To verify and deepen these findings, we have supplemented our study with secondary literature. In the following sections, we explore four key aspects of complexity that are critical for healthcare organisations when selecting and implementing a data architecture: implementation effort, maintenance and scalability, required technical expertise, and coordination.

Implementation Effort

The implementation effort differs notably between the three architectures.

- **cDWHs** require detailed data models, extensive ETL processes and a stable database system[7,30]. This usually requires a fixed project plan with defined phases for schema design, ETL development, testing and go-live [22,30]. Although the initial effort is high, the resulting structure is clear and stable for consistent data management.
- **cDLs** benefit from a schema-on-read approach, which reduces the need for rigid upfront data modelling. However, continuous maintenance and data harmonisation demand a high level of expertise [31]. Additionally, an infrastructure for big data technologies-such as Hadoop and cloud object storage-must be established. This requires experts who can set up and manage distributed systems [27,29,31]. The inclusion of streaming frameworks like Kafka or Flume further increases the complexity [32]. In summary, the absence of a unified data model may be advantageous initially but leads to significantly more complex metadata management over time.
- **cDLH**’s hybrid nature combines elements from both cDWHs and cDLs. In addition to schema-on-read functionalities, cDLs require schema-on-write techniques and ACID transactions [27,32]. This dual requirement calls for the orchestration of complex ETL/ELT pipelines, robust security concepts and distributed compute environments [27,28]. Consequently, the initial integration and development effort is very high. Often, specialised applications and frameworks are needed-tools that are still in the early stages of evaluation within the medical domain.

Maintenance and Scalability

The maintenance and scalability of data management architectures are important factors for smooth operation.

- **cDWHs** require continuous maintenance. Changes in data sources, schema extensions, and performance tuning demand regular attention [7,30]. Vertical scaling to increase storage and compute becomes very expensive as data volumes grow [31]. In addition, the batch-oriented structure of cDWHs may become a bottleneck when there is an increasing need for real-time data processing [22]. cDWHs offer stability and clear data governance but may struggle with the costs and limitations of scaling for real-time applications.
- **cDLs** are maintained through active monitoring, strict data governance and continuous data quality management. These measures help to avoid the creation of data swamps. The schema-on-read method requires constant adaptation of ingestion pipelines and greater coordination during verification [33]. In terms of scalability, the use of distributed file systems allows for affordable horizontal scaling, making it easier to handle large and diverse datasets [33]. cDLs offer flexible maintenance and cost-effective scalability, but require careful coordination and robust data governance to maintain data quality.
- cDLHs combine components from cDWHs and cDLs. This integration leads to more complex maintenance tasks, as both cDL and cDWH elements must be managed by multiple granular applications for distributed systems [27]. Although cDLHs support dynamic scalability for both compute and storage capacity, this flexibility requires specific expertise in resource and cost management [32]. In summary, cDLHs offer dynamic and scalable solutions that bridge the gap between structured and unstructured data, but require a more complex maintenance management.

Required Technical Expertise

The technical expertise needed to implement and maintain clinical data management systems is a key consideration.

- **cDWHs** rely on relational database systems, SQL, ETL tools, and established data warehousing concepts including data modelling [7,21,30]. An additional technical stack and expertise is required for extended requirements such as near real-time processingd. cDWHs depend on well-established database technologies and practices, but advanced processing needs may require extra skills and tools.
- **cDLs** require expertise in big data frameworks such as Hadoop and Spark, cloud storage solutions, streaming tools like Kafka or Flume and technical expertise in distributed systems [19,33]. Users must also be skilled in analysing and modelling unstructured data [33]. Moreover, more extensive change management and monitoring are necessary because changes are only detected when data is read from the lake. cDLs demand specialised big data skills and a proactive approach to change management, making them suitable for organisations with robust technical capabilities.
- cDLHs combine traditional data warehousing with modern big data techniques. They require interdisciplinary expertise covering both classic DWH principles (e.g. schema-on-write, ACID transactions) and contemporary big data methods (e.g. schema-on-read, distributed systems, streaming). Often, these systems are cloud-native, which means additional skills in cloud technologies and DevOps

practices (such as Docker, Kubernetes, security, and identity access management) are needed even for on-premise solutions [27,32,34]. cDLHs require a broad range of technical expertise across both traditional and modern data management technologies, posing a challenge for organisations with limited specialised resources.

Complexity of Coordination

- **cDWHs** operate with a centralised governance model where responsibilities are clearly defined. Departments and IT teams must work closely to ensure continuous schema adjustments and maintain data quality within the data models [7,30,31]. cDWHs rely on largely internal coordination, as a single main instance, such as a clinic or data integration centre, typically manages the system.
- **cDLs** use a decentralised and flexible approach, which demands robust metadata strategies. Different domains, such as laboratory data, imaging, and genomics, must be coordinated effectively [31,33]. Coordination in cDLs is particularly challenging when integrating new data sources, establishing new governance standards or maintaining schema changes [33].
- **cDLHs** require synchronisation of governance and security updates across both the cDLs and cDWHs components [28,29,34]. They have to manage the complex requirements of big data systems alongside traditional data warehousing. This hybrid processing approach, which integrates both near real-time and batch processing workflows, requires careful coordination to ensure seamless operation and data integrity cDLHs result in increased coordination complexity due to their combined systems and the need to manage parallel processing methods.

Limitations and Additional Considerations

Our primary focus in this research was on comparing data management architectures regardless of local conditions. Still, our experience shows that factors such as cost implications and legacy system integration are critical. Furthermore, we excluded alternative data integration architectures and acknowledged the inherent limitations of frameworks such as the FAIR Principles and the Five Vs of Big Data. We briefly address these aspects here to provide a broader perspective.

General Limitations

Some essential requirements for clinical data management are not fully covered by the FAIR Principles or the five Vs of Big Data. These include privacy and confidentiality, audit and compliance mechanisms, management and documentation of ethical consent, long-term archiving, risk management, detailed user access management and cultural sensitivity. However, it should be noted that regulatory compliance and the protection of sensitive data must always be ensured, regardless of the chosen data architecture. Therefore, we did not include these aspects in the comparison of data architectures, as they should be addressed by separate legal and organisational measures. A comprehensive data management strategy must take into account specific legal and ethical requirements in addition to the FAIR Principles and the five Vs to ensure holistic and compliant data management.

It is important to note that this study does not consider data integration architectures such as Lambda (batch) and Kappa (stream). Similarly, no combination or generalisation of approaches such as Data Fabric or Data Mesh have been evaluated for several reasons. Most importantly, Data Fabric was not considered a viable option because in the clinical domain, production systems cannot be accessed for complex queries. Furthermore, Data Mesh represents an organisational approach that focuses on domain-specific data products rather than traditional architectural archetypes such as Data Lakes or Data Warehouses.

Cost Analysis and Resource Implications

In addition to technical and organisational considerations, cost analysis plays a critical role in the selection of a data management architecture. For example, while cDWHs may have high up-front costs due to extensive ETL development and schema design, they may offer lower long-term operational costs due to their stable nature. In contrast, cDLs offer cost-effective scalability for large datasets, but require ongoing investment in metadata management and system monitoring. cDLHs promise a balance between structure and flexibility, but require significant initial integration costs and ongoing maintenance. Future implementations should carefully weigh these cost implications against the performance and scalability benefits of each architecture.

Legacy System Integration

Integration with legacy clinical systems remains a critical factor. cDWHs generally offer smoother integration with existing relational databases and established clinical information systems. In contrast, cDLs may require additional data transformation layers and custom connectors to interface effectively with legacy systems. While cDLHs aim to combine the strengths of both, their hybrid nature can present challenges when integrating older data systems. Institutions should consider their current infrastructure and plan transitional strategies that bridge legacy systems with modern data architectures.

Application Scenarios and Recommendations

Depending on the type of data, analysis requirements and available resources, a different model is recommended in each case. In the following sections, we describe the scenarios in which each of the three architectures offers the greatest benefits.

cDWHs

cDWHs are ideal for environments that require consistent reporting, strong compliance and reliable analysis of structured data. They work best when data is mostly relational and fixed schemas ensure high data quality. Moreover, the structured and standardised character of cDWHs provides a stable foundation for AI applications, where reliable data is essential for training robust machine learning models. This is particularly useful for hospital governance and controlling, where clear and powerful analysis is required.

Scenarios, where cDWHs should be considered:

- **Structured Analysis and Compliance**: In scenarios where data is primarily in tabular form (for instance, as recorded in a Hospital Information System) and there is an urgent need for rigorous data protection and comprehensive traceability, the optimal solution is often a cDWH. Stable reporting, long-term trend analyses or controlling systems as well as business intelligence tools with regular audits and fixed key performance indicators benefit from a cDWH.

- **Resource-Limited IT Environments:** In settings where SQL expertise is available but there is limited capacity for complex big data or cloud solutions. Therefore, small or medium-sized clinics may prefer a cDWH. These institutions can rely on well-established ETL processes and structured reports.
- **Low to Moderate Data Diversity**: Departments that record mostly numerical and tabular data (as seen in standardised clinical management systems) benefit from the fixed schema of a cDWH. The simplicity of the data model helps maintain consistency and ease of reporting and research.
- **Legacy System Integration:** Institutions with established relational databases or enterprise resource planning systems may find it easier to integrate these with a cDWH. This reduces the need for extensive data transformation and minimises complexity.

Although cDWHs often involve high initial implementation effort, their structured nature ensures clear data governance and auditability. However, they may be less suitable for real-time processing and highly diverse data sources.

cDL

cDLs are designed for exploratory research and the management of heterogeneous data sets. They enable organisations to store large volumes of raw data in its original format, which is useful when data arrives in many different types and formats. In addition, this flexibility is a key enabler for AI research, as it allows researchers to experiment with diverse data types and develop innovative algorithms that can learn from unstructured and semi-structured information.

Scenarios, where cDLs should be considered:

- **Exploratory Research:** In projects where data requirements are not yet fully defined, a cDL offers the flexibility to store various data types. Researchers can work with unstructured data such as doctors' notes, images, and sensor outputs without a fixed schema.
- **Machine Learning Prototyping:** A cDL is ideal for experimental machine learning or data science projects. It allows users to quickly explore data, test different algorithms and develop without the need for complex ETL pipelines or strict data modelling. This supports innovative research.
- **Analysis of Log and Streaming Data**: When near-real-time data is needed-such as recordings from IoT devices or sensor networks-a cDL can efficiently handle fast data ingestion. For example, it can support automated anomaly detection in patient vital signs or system logs.
- **Cost-Effective Storage of High-Volume Data and Compute**: Due to the big data paradigm, cDLs offer scalable, distributed storage that can handle petabytes of data. This makes them well suited for institutions collecting large amounts of diverse data on a limited budget.
- **Future-Proofing Data:** By storing raw data, a cDL allows organisations to reprocess or model the data later, as new analytical methods emerge. This is

particularly valuable in fast-evolving research environments.

While cDLs offer cost-effective scalability and flexibility (via a schema-on-read approach), they demand disciplined metadata management and robust data governance. Without these measures, there is the risk of creating a data swamp. Consequently, managing the complexity of heterogeneous data may require specialised technical expertise.

cDLH

cDLHs combine the strengths of cDWHs and cDLs. They offer the robust governance and ACID compliance of a cDWH along with the flexibility and scalability of a cDL. Crucially, this hybrid approach supports advanced AI-driven analytics by ensuring that data remains both standardised and adaptable, which is vital for real-time processing and training of machine learning models. This hybrid approach is best suited to institutions that need both structured reporting and the capacity to manage unstructured data. However, this adds more complexity and requires high technical expertise.

Scenarios, where cDLs should be considered:

- **Comprehensive Data Integration**: In large, research-intensive hospitals, a cDLH can integrate classic tabular data with vast amounts of unstructured research data (such as images and omics data). This allows the same data set to be used for standard reporting and advanced AI applications.
- **Real-Time and Batch Processing**: Institutions that require both routine batch analyses and live data streaming for real-time decision support benefit from a cDLH. For example, care centres with real-time alarm systems in intensive care units can use a cDLH to support both immediate and historical analyses.
- **Multi-Stakeholder Research Platforms**: When several partners (such as clinics, research institutions, and industry) collaborate, a cDLH provides a unified platform. It supports strict access controls and data governance while offering the flexibility needed for varied analytical or machine learning methods.
- **Long-Term Investment in Flexibility**: Organisations planning for future growth may choose a cDLH to avoid redundant systems. By integrating reporting, data science, real-time analyses, and exploratory research in one platform, they can reduce long-term operational complexity and ensure users acceptability.

The main challenge with cDLHs is the high complexity of implementation and maintenance. They require a broad range of technical expertise and resources to manage both traditional data warehousing and modern big data processes. However, for large organisations with complex data landscapes, the benefits of a unified platform can outweigh these challenges.

Conclusions

In summary, each data management architecture offers distinct advantages and limitations, as illustrated in figure 2. cDWHs are best suited for stable, structured environments with strong compliance and legacy system integration. They are ideal when clear, predefined data models are needed and when audit trails are critical. cDLs provide a flexible, cost-effective solution for exploratory research and high-volume, heterogeneous data sets. They excel in environments where rapid prototyping and scalability are key. cDLHs combine the strengths of both approaches, delivering robust governance alongside flexibility. They are best suited for large, research-intensive institutions that require both real-time processing and traditional reporting.

The choice of architecture should consider immediate analytical needs, long-term scalability, technical expertise, and overall maintenance complexity. Healthcare organisations must balance these factors with their available resources and strategic goals.

This study emphasizes the importance of choosing the right data management architecture for medical research, patient care and advanced AI applications. The analysis shows that cDWHs, cDLs and cDLHs not only offer distinct strengths and weaknesses but also provide the essential foundation of reliable, standardised data that AI systems require for effective learning and decision support. The choice depends on the unique needs of each institution. By conducting this analysis, we've provided a clear framework to help healthcare institutions to identify the most suitable data architecture.

Using our comparison and recommendations, healthcare organisations can make focused decisions about which data architecture suits their needs to manage big data effectively. Focussing on the potential, cDLHs would be best suited for the diverse requirements of the healthcare sector, as they offer the most diverse functionalities and flexibility. However, cDLHs have the disadvantage of requiring high technical expertise and involving complex integration efforts. Future research should focus on fine-tuning cDLHs with the aim of reducing their complexity and integrating medical standards more effectively. This will expand their usability while meeting the high demands of the healthcare sector.

Acknowledgements

This publication was funded by the German Federal Ministry of Education and Research (BMBF) Network of University Medicine 2.0: “NUM 2.0", Grant No. 01KX2121, Project: Netzwerk der Datenintegrationszentren der Universitätsmedizin - NUM-DIZ.

This manuscript was initially drafted in German and translated into English using DeepL and DeepL Write. The authors then thoroughly reviewed and edited the translated text to ensure accuracy and clarity of expression.

Authors' Contributions

RG is the main author who conceived the study and defined the analysis criteria that linked the requirements to the medical context. Together with MG, RG developed the search strategy, which was subsequently approved by IR. RG performed the search screenings and conducted the rapid review-including the PRISMA workflow - while MG contributed to the validation of the literature selection and data extraction. The review results and the comparison table were verified and discussed with MG. The discussion section, particularly the additional considerations regarding complexity, was designed collaboratively by RG and MG, with the final discussion and recommendations on application scenarios proposed by RG and reviewed by both MG and IR. RG drafted the initial manuscript, and IR, MG, and MS contributed critical revisions. All authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Abbreviations

cDL: clinical Data Lake
cDLH: clinical Data Lakehouse
cDWH: clinical Data Warehouse

References

1. Daten - Volumen der weltweit generierten Daten bis 2028 [Internet]. Statista. [cited 2025 Mar 17]. Available from: https://de.statista.com/statistik/daten/studie/267974/umfrage/prognose-zum-weltweit-generierten-datenvolumen/
2. Shilo S, Rossman H, Segal E. Axes of a revolution: challenges and promises of big data in healthcare. Nat Med. 2020 Jan;26(1):29–38.
3. Baloch L, Bazai SU, Marjan S, Aftab F, Aslam S, Neo TK, et al. A Review of Big Data Trends and Challenges in Healthcare. Int J Technol. 2023 Oct;14(6):1320–33.
4. 23rd Meeting of the eHealth Network - European Commission [Internet]. [cited 2025 Mar 17]. Available from: https://health.ec.europa.eu/system/files/2023-05/ehealth_20230330_sr_en.pdf
5. Mendhe D, Dogra A, Nair PS, Punitha S, Preetha KS, Babu SBGT. AI-Enabled Data-Driven Approaches for Personalized Medicine and Healthcare Analytics. In: 2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM) [Internet]. 2024 [cited 2025 Mar 17]. p. 1–5. Available from: https://ieeexplore.ieee.org/document/10568722
6. Pavlenko E, Strech D, Langhof H. Implementation of data access and use

procedures in clinical data warehouses. A systematic review of literature and publicly available policies. BMC Med Inform Decis Mak. 2020 Jul 11;20:157.

7. Sebaa A, Chikh F, Nouicer A, Tari A. Medical Big Data Warehouse: Architecture and System Design, a Case Study: Improving Healthcare Resources Distribution. J Med Syst. 2018 Feb 19;42(4):59.

8. Babu H, Arivazhagn D, Kumar G, Patil S, Akhtar MN, Bakar EA. Examining the Potential Benefits and Challenges of Utilizing AI for Big Data Analysis. In: 2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON) [Internet]. 2023 [cited 2025 Mar 17]. p. 1–8. Available from: https://ieeexplore.ieee.org/document/10442332

9. Mungoli N. Scalable, Distributed AI Frameworks: Leveraging Cloud Computing for Enhanced Deep Learning Performance and Efficiency [Internet]. arXiv; 2023 [cited 2025 Feb 21]. Available from: http://arxiv.org/abs/2304.13738

10. El Aissi MEM, Benjelloun S, Loukili Y, Lakhrissi Y, Boushaki AE, Chougrad H, et al. Data Lake Versus Data Warehouse Architecture: A Comparative Study. In: Bennani S, Lakhrissi Y, Khaissidi G, Mansouri A, Khamlichi Y, editors. WITS 2020. Singapore: Springer; 2022. p. 201–10.

11. Zaharia M, Ghodsi A, Xin R, Armbrust M. Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. In: 11th Conference on Innovative Data Systems Research, CIDR 2021, Virtual Event, January 11-15, 2021, Online Proceedings [Internet]. www.cidrdb.org; 2021. Available from: http://cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf

12. Sinaci AA, Núñez-Benjumea FJ, Gencturk M, Jauer ML, Deserno T, Chronaki C, et al. From Raw Data to FAIR Data: The FAIRification Workflow for Health Research. Methods Inf Med. 2020 Jul 3;59:e21–32.

13. Holub P, Kohlmayer F, Prasser F, Mayrhofer MTh, Schlünder I, Martin GM, et al. Enhancing Reuse of Data and Biological Material in Medical Research: From FAIR to FAIR-Health. Biopreservation Biobanking. 2018 Apr 1;16(2):97–105.

14. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3(1):160018.

15. Bellazzi R. Big data and biomedical informatics: a challenging opportunity. Yearb Med Inform. 2014 May 22;9(1):8–13.

16. Mooney SJ, Westreich DJ, El-Sayed AM. Epidemiology in the Era of Big Data. Epidemiol Camb Mass. 2015 May;26(3):390–4.

17. Lee CH, Yoon HJ. Medical big data: promise and challenges. Kidney Res Clin Pract. 2017 Mar;36(1):3–11.

18. Ting DSJ, Deshmukh R, Ting DSW, Ang M. Big data in corneal diseases and cataract: Current applications and future directions. Front Big Data. 2023 Feb 1;6:1017420.

19. Natarajan K, Weng C, Sengupta S. A Model for Multi-Institutional Clinical Data Repository. Stud Health Technol Inform. 2023 May 18;302:312–6.

20. Bellandi V, https://orcid.org/0000-0003-4473-6258, View Profile, Ceravolo P, https://orcid.org/0000-0002-4519-0173, View Profile, et al. Data management for continuous learning in EHR systems. ACM Trans Internet Technol [Internet]. [cited 2025 Mar 20];0(ja). Available from: https://dl.acm.org/doi/10.1145/3660634

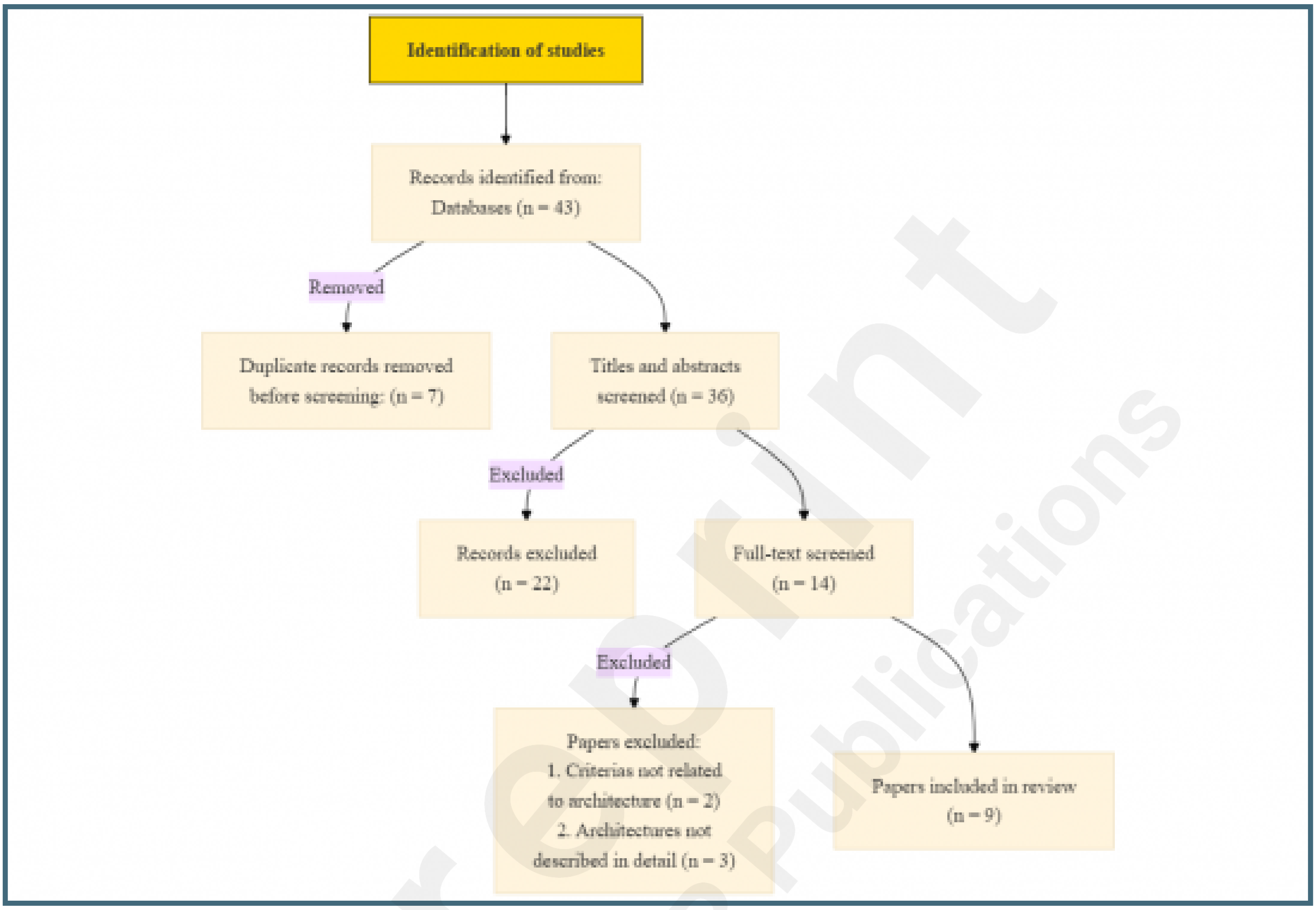
21. Parciak M, Suhr M, Schmidt C, Bönisch C, Löhnhardt B, Kesztyüs D, et al. FAIRness through automation: development of an automated medical data integration infrastructure for FAIR health data in a maximum care university hospital. BMC Med Inform Decis Mak. 2023 May 15;23(1):94.
22. Lacey JV Jr, Chung NT, Hughes P, Benbow JL, Duffy C, Savage KE, et al. Insights from Adopting a Data Commons Approach for Large-scale Observational Cohort Studies: The California Teachers Study. Cancer Epidemiol Biomarkers Prev. 2020 Apr 1;29(4):777–86.
23. Guilbert B, Rieu T, Delamarre D, Laurent F, Serre F, Leroy S, et al. ONCO-FAIR Project: Improving Data Interoperability in Oncology Chemotherapy Treatments for Data Reuse. In: Digital Health and Informatics Innovations for Sustainable Health Care Systems [Internet]. IOS Press; 2024. p. 1373–7. Available from: https://ebooks.iospress.nl/doi/10.3233/SHTI240667
24. Al-Hgaish A, Alzyadat W, Al-Fayoumi M, Alhroobis AM, Thunibat A. Preserve quality medical drug data toward meaningful data lake by cluster. Int J Recent Technol Eng. 2019;8(3):270–7.
25. Barnes C, Bajracharya B, Cannalte M, Gowani Z, Haley W, Kass-Hout T, et al. The Biomedical Research Hub: a federated platform for patient research data. J Am Med Inform Assoc. 2022 Apr 1;29(4):619–25.
26. Dhayne H, Haque R, Kilany R, Taher Y. In Search of Big Medical Data Integration Solutions - A Comprehensive Survey. IEEE Access. 2019;7:91265–90.
27. Harby AA, Zulkernine F. From Data Warehouse to Lakehouse: A Comparative Review. In: 2022 IEEE International Conference on Big Data (Big Data) [Internet]. 2022 [cited 2025 Feb 27]. p. 389–95. Available from: https://ieeexplore.ieee.org/document/10020719
28. Xiao Q, Zheng W, Mao C, Hou W, Lan H, Han D, et al. MHDML: Construction of a Medical Lakehouse for Multi-source Heterogeneous Data. In: Traina A, Wang H, Zhang Y, Siuly S, Zhou R, Chen L, editors. Health Information Science. Cham: Springer Nature Switzerland; 2022. p. 127–35. (Lecture Notes in Computer Science).
29. Čuš B, Golec D. Data Lakehouse: Benefits in small and medium enterprises. Mednar Inov Posl J Innov Bus Manag. 2022;14(2):1–10.
30. Gerbel S, Laser H, Schönfeld N, Rassmann T. The Hannover Medical School Enterprise Clinical Research Data Warehouse: 5 Years of Experience. In: Auer S, Vidal ME, editors. Data Integration in the Life Sciences. Cham: Springer International Publishing; 2019. p. 182–94. (Lecture Notes in Computer Science).
31. Nambiar A, Mundra D. An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management. Big Data Cogn Comput. 2022 Dec;6(4):132.
32. Oreščanin D, Hlupić T. Data Lakehouse - a Novel Step in Analytics Architecture. In: 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO) [Internet]. 2021 [cited 2025 Feb 27]. p. 1242–6. Available from: https://ieeexplore.ieee.org/document/9597091
33. Quix C, Hai R. Data Lake. In: Sakr S, Zomaya A, editors. Encyclopedia of Big Data Technologies [Internet]. Cham: Springer International Publishing; 2018 [cited 2025 Jan 13]. p. 1–8. Available from: https://doi.org/10.1007/978-3-319-63962-8_7-1
34. Begoli E, Goethert I, Knight K. A Lakehouse Architecture for the Management

and Analysis of Heterogeneous Data for Biomedical Research and Mega-biobanks. In: 2021 IEEE International Conference on Big Data (Big Data) [Internet]. 2021 [cited 2025 Mar 17]. p. 4643–51. Available from: https://ieeexplore.ieee.org/document/9671534

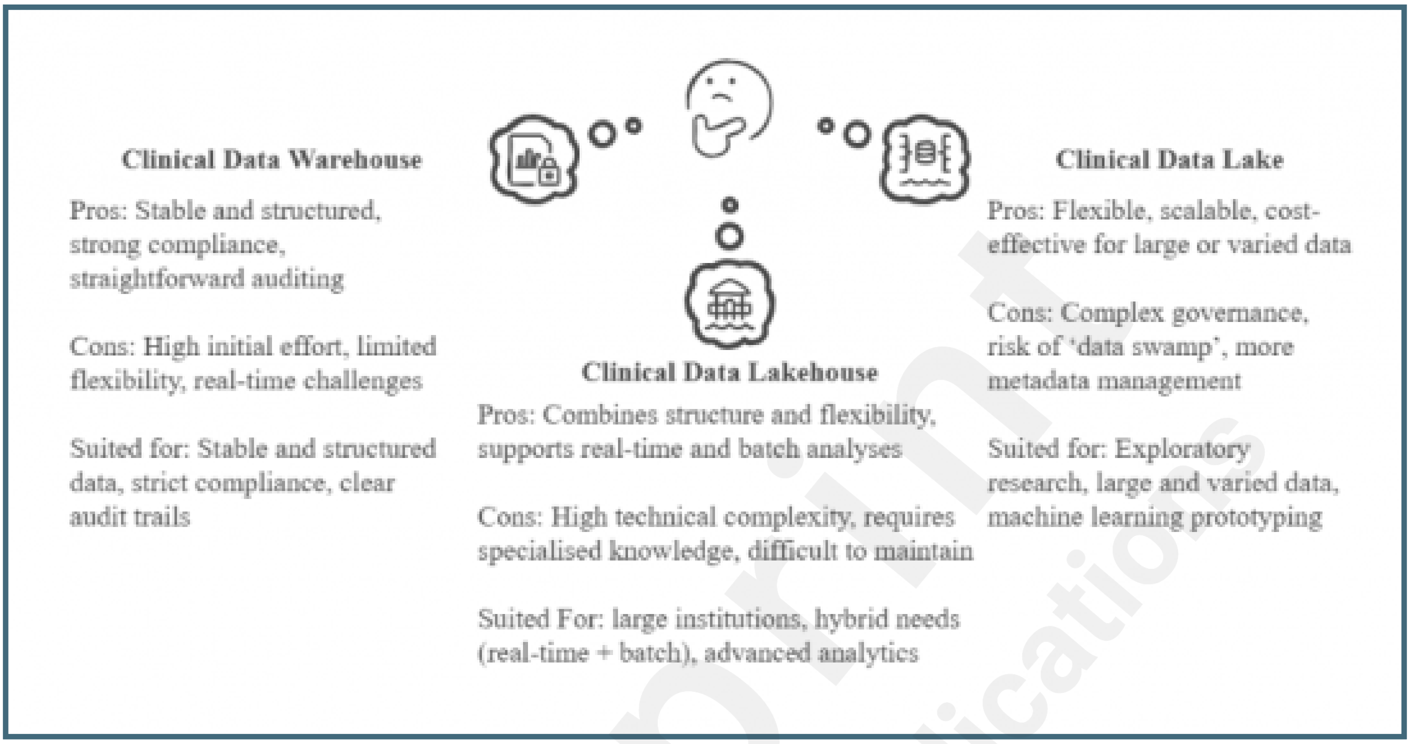
Supplementary Files

Figures

This figure illustrates the PRISMA-like workflow used in this study. It shows the identification of records from four databases, the removal of duplicates, the screening of titles and abstracts, and the full-text review leading to the final inclusion of nine articles.



A comparative summary of clinical data warehouses (cDWH), clinical data lakes (cDL), and clinical data lakehouses (cDLH). It highlights the strengths and weaknesses of each architecture based on the FAIR Principles and the Big Data 5 Vs, summarising key features, advantages, and limitations.



Clinical Data Warehouse

Pros: Stable and structured, strong compliance, straightforward auditing

Cons: High initial effort, limited flexibility, real-time challenges

Suited for: Stable and structured data, strict compliance, clear audit trails

Clinical Data Lakehouse

Pros: Combines structure and flexibility, supports real-time and batch analyses

Cons: High technical complexity, requires specialised knowledge, difficult to maintain

Suited For: large institutions, hybrid needs (real-time + batch), advanced analytics

Clinical Data Lake

Pros: Flexible, scalable, cost-effective for large or varied data

Cons: Complex governance, risk of 'data swamp', more metadata management

Suited for: Exploratory research, large and varied data, machine learning prototyping