

From Data Sources to User Screens: a Multimodal hEalth Data lakehouse for ITAlY (MEDITA)

Maria Paola Priola^{1,*†}

¹Department of Economics and Business, University of Cagliari, Italy

Abstract

Public health data are essential for understanding population health, monitoring trends, and informing evidence-based decision-making. Traditional Data Warehouses (DWs) have been central in standardizing and analyzing large-scale datasets, yet they struggle with fragmented sources, heterogeneous formats, and limited adaptability. **Multimodal hEalth Data lakehouse for ITAlY (MEDITA)** is a proof-of-concept platform built around the Data Lakehouse paradigm, combining the governance and reliability of DWs with the flexibility of Data Lakes. MEDITA ingests both tabular and textual data from Italian public health sources through an Extract-Transform-Load-Model-Deploy (ETLMD) pipeline, harmonizing schemas and enabling statistical and predictive analysis. The platform integrates interactive dashboards, time-series forecasting modules, and a Retrieval-Augmented Generation (RAG) chatbot that supports natural-language queries in Italian. By offering a unified and user-friendly analytical environment, MEDITA improves accessibility for researchers and citizens, reducing barriers to evidence-based insight and public engagement. Future extensions include support for additional modalities such as audio, video, and electronic health records, further advancing multimodal analytics in public health.

Keywords

Public Health Data, Data Lakehouse, Data Lakes, Retrieval-Augmented Generation, Web Interface, Natural Language Processing, Large Language Models

1. Introduction

Public health data provide a critical foundation for population surveillance and policy evaluation. However, their integration remains challenging due to fragmentation, structural heterogeneity, and limited interactivity across existing infrastructures.

Data Warehouses (DWs) have historically addressed part of this challenge by providing “subject-oriented, integrated, time-dependent, non-volatile, and non-normalized” repositories [1]. They consolidate information across organizations [2, 3, 4] and enable reporting, interoperability, and historical analysis. However, DWs rely on rigid schema-on-write architectures, require costly development, and involve complex Extract-Transform-Load (ETL) processes [5, 6]. These limitations are particularly acute in healthcare, where multimodal inputs and rapidly evolving schemas are the norm [7].

Data Lakes (DLs) were later introduced to improve scalability and flexibility through schema-on-read ingestion [8]. Despite their adaptability, DLs frequently degrade into “data swamps” due to insufficient governance, lack of harmonization, and weak support for advanced analytics [9, 10]. More recently, the *Data Lakehouse* paradigm has emerged, combining the reliability and governance of DWs with the scalability of DLs [11, 12, 13]. Data Lakehouse maintain open formats, enforce consistency, and support machine learning workloads without sacrificing flexibility.

Alongside integration and governance, democratizing access to health data requires new interaction paradigms. Large Language Model (LLM)-based conversational agents are one such avenue, translating natural language queries into actionable insights. While LLMs are increasingly used in personal health, their application to public health remains rare. A key challenge across all domains is the risk of hallucinations, essentially defined as factually incorrect or unverifiable outputs [14, 15, 16]. Retrieval-Augmented Generation (RAG) has shown promise in mitigating this risk by grounding answers in external knowledge [17].

ITADATA2025: The 4th Italian Conference on Big Data and Data Science, September 9–11, 2025, Turin, Italy

✉ mariap.priola@unica.it (M. P. Priola)

🌐 <https://priolap.github.io/portfolio/> (M. P. Priola)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The relevance of an infrastructure that combines interoperability with accessibility is underscored by international initiatives such as the World Health Organization (WHO) Health for All database [18] and the European Health Data Space (EHDS)¹. In Italy, valuable datasets are published by the Ministry of Health² and the Italian National Institute of Statistics (ISTAT)³, yet they often lack harmonization and integrated analytical support [19]. Existing aggregators like DatiOpen⁴ and the national open data portal⁵ mainly offer download-based exploration and provide no predictive analytics or natural language querying.

This paper introduces the **Multimodal hEalth Data lakehouse for ITALy (MEDITA)**, a prototype platform that operationalizes the Data Lakehouse design for public health. MEDITA ingests both structured and unstructured sources through an end-to-end Extract–Transform–Load–Model–Deploy (ETLMD) pipeline. It applies schema-on-read classification, attribute harmonization, and record standardization to ensure interoperability across heterogeneous datasets. Beyond integration, it embeds a modeling stage that supports statistical analysis, time-series forecasting, and conversational querying through the Health Advisor for Needs and Knowledge (HANK), a domain-specific RAG chatbot in Italian.

By combining structured governance, multimodal ingestion, and AI-powered interactivity, MEDITA addresses persistent gaps in the Italian context, where open health data are increasingly available but remain fragmented and difficult to exploit. This architecture aligns with the FAIR (Findable, Accessible, Interoperable, and Reusable) principles [20] and offers a unified interface for researchers and the general public. To the best of my knowledge, no current platform combines multimodal integration, Artificial Intelligence (AI)-powered analytics, and natural language interaction in this way.

The contributions of this paper are threefold. First, it advances the Data Lakehouse paradigm by introducing adaptive transformation routines, including: (i) a typology for mapping heterogeneous inputs, (ii) fuzzy matching to resolve attribute inconsistencies, and (iii) record harmonization through standardized geographic codes and deduplication. Second, it extends classical analytics by embedding statistical modeling and predictive capabilities within the same environment, accessible through interactive dashboards. Third, it integrates HANK, a domain-specific RAG chatbot for Italian public health data, enabling natural language queries for both researchers and citizens. Together, these features demonstrate MEDITA as a proof-of-concept platform that unifies heterogeneous data integration, governance, and AI-powered interactivity into a web-based infrastructure.

The remainder of the paper is structured as follows: Section 2 reviews related work; Section 3 outlines data sources; Section 4 details the methodology; Section 5 presents experimental results; and Section 6 concludes with implications and future directions.

2. Related Works

Existing research in health data management, ETL processes, web-based health interfaces, and conversational agents provides the foundation for MEDITA’s design. In this section, I highlight prior contributions and outline the specific gaps this work aims to address.

2.1. Health Data Management

Health data infrastructures have traditionally relied on DWs, whose implementation varies widely between countries, shaped by governance models, technological infrastructure, and regulatory environments. Core challenges, such as data fragmentation, system complexity, privacy, and scalability, influence three main dimensions: strategic governance, architecture, and integrative capacity.

¹For details, https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en.

²For details, <https://www.dati.salute.gov.it/dati/homeDataset.jsp>.

³For details, <https://www.istat.it/it/archivio/14562>.

⁴For details, <http://www.datiopen.it/>.

⁵For details, <https://www.dati.gov.it/>.

Several studies evaluate national and regional health data systems for research and clinical use. Burgun et al. [21] highlight how decentralization in Italy and France inhibits data reuse, citing bureaucratic obstacles in France and fragmented infrastructure in Italy. By contrast, centralized registries in Nordic countries facilitate cross-institutional access and data sharing. In the United States, Campion Jr et al. [22] point to coordination and staffing challenges in Enterprise DWs for Research, underscoring the resource demands of scalable infrastructure.

Architecture strongly influences adaptability. Examples include Unified Modeling Language (UML)-based schema standardization [23], health data cubes supporting Online Analytical Processing (OLAP)-style queries [24], and iterative DWs for communicable disease monitoring at the BC Centre for Disease Control [25]. In resource-limited or data-rich contexts, integration and scalability are essential. Bangladesh's national DW facilitates remote healthcare monitoring [26], while the Dutch system integrates millions of records across providers [27]. HRADIS [28] improves accessibility using Microsoft SQL Server's analytics layer.

Despite these advances, most infrastructures are tied to classical DW paradigms, optimized for internal or institutional research rather than multimodal public health analytics. They often lack mechanisms for unstructured data ingestion or AI-based interactivity, motivating hybrid approaches such as Data Lakehouses.

MEDITA addresses this gap by contributing a modular, scalable Data Lakehouse proof-of-concept tailored for Italy, harmonizing multimodal public health data and boosting both researchers and citizen knowledge.

2.2. Data Integration in Public Health

Data integration traditionally relies on the ETL paradigm, originally conceived for relational databases and later extended to health infrastructures. Early developments were driven by enterprise database systems [29, 30, 31, 32], logical rule languages [33, 34], and schema modeling standards [35, 36].

Over time, ETL practices were adapted to meet the specific requirements of healthcare, where interoperability standards such as HL7 FHIR⁶ have become key drivers of innovation. For example, Gruendner et al. [19] store FHIR resources in JavaScript Object Notation (JSON), while Hong et al. [37] propose NLP2FHIR for extracting structured content from clinical notes. At the national level, India implements cloud-based ETL for real-time analytics [38], while Germany translates HL7 into Observational Medical Outcomes Partnership (OMOP) Common Data Model to support interoperability [39].

Yet, healthcare ETL frameworks continue to face persistent challenges of fragmentation, heterogeneity, and privacy [40]. These limitations have accelerated the shift toward architectures that unify ingestion, harmonization, and modeling. Recent proposals for Data Lakehouse architectures show how these stages can be integrated within a single pipeline [11, 41, 10].

Within this domain, the relevance of Data Lakes and Data Lakehouses is increasingly evident. Aziz [42] propose an Open Data Lake Framework that combines flexible ingestion with strong governance to support advanced analytics and AI workloads. Similarly, Begoli et al. [12] describe a Data Lakehouse for biomedical research and mega-biobanks, emphasizing privacy, FAIR compliance, and multimodal integration of genomic and clinical data. These studies demonstrate how adaptability and governance can be effectively combined to manage heterogeneous health data, offering concrete evidence of their suitability for health ecosystems.

Building on these principles, MEDITA contributes a proof-of-concept Data Lakehouse through an ETLMD pipeline that embeds schema harmonization, modeling, and deployment as native stages, enabling integration of heterogeneous Italian public health data.

⁶FHIR is a standard developed by HL7 for exchanging healthcare data: <https://www.hl7.org/index.cfm>.

2.3. Web-Based Interfaces in Healthcare

While data ingestion frameworks shape internal data flow, front-end interfaces define how users engage with health data. Web-based platforms have emerged as essential tools for clinical decision support, research, education, and public health planning.

At the patient level, systems like MyHealthPortal [43] offer real-time self-care tools by integrating wearables with symptom tracking. Interoperability with electronic health records improves longitudinal care but requires robust privacy safeguards.

For researchers, platforms such as BMI Investigator support multimodal queries across genomic, imaging, and clinical data [44], while Hadoop-based systems demonstrate scalability for hospital disease detection [45]. Visualization dashboards are also prominent: Chen et al. [46] use Uniform Manifold Approximation and Projection (UMAP) to explore rare diseases, and Geoportal integrates geospatial and health indicators via OLAP [47].

Public health platforms often operate at macro level. GIS-based dashboards track accessibility [48], while Hawaii's IBIS-PH centralizes surveillance indicators [49]. The COVID-19 pandemic accelerated demand for timely data: Covid-Warehouse aggregates Italian regional metrics for crisis response [50], while Turcan and Peker [51] propose a multidimensional DW for pandemic monitoring.

Nevertheless, most interfaces are siloed by domain and rarely integrate multimodal data or advanced analytics. MEDITA addresses this gap by coupling interactive dashboards with predictive modeling and conversational AI, making public health data usable across heterogeneous users.

2.4. User-Agent Chatbots

Conversational agents in healthcare have evolved from rule-based systems to machine learning and, more recently, LLM-powered chatbots. Early systems supported reminders or triage but offered limited contextual reasoning. Recent work demonstrates broader potential, showing their effectiveness for patient engagement and diagnostics [52], multimodal interaction combining text, voice, and images [53], and supervised assistants such as MEDIC [54] and Medibot [55].

Generative models are now reshaping the field. Med-PaLM achieves expert-level performance on USMLE-style questions [56], while MultiMedQA benchmarks reasoning and factuality across multiple medical tasks [57].

Nonetheless, many chatbots remain domain-specific or ungrounded, raising risks of misinformation. LLMs often display overconfidence [58], failing to recognize when they lack sufficient knowledge and generating hallucinations [14, 16], which pose significant risks in clinical use.

These limitations underscore the importance of grounding strategies. This work contributes HANK, a retrieval-augmented medical chatbot designed for Italian public health, which anchors responses in curated evidence and abstains when context is insufficient. This approach aims to minimize hallucinations and enhance transparency, supporting safe, data-grounded interaction at national scale.

3. Data Sources

MEDITA integrates structured and unstructured open-access health data to support multimodal analysis of Italy's healthcare landscape. These datasets enable longitudinal, socio-economic, and thematic analyses across various domains, providing a representative foundation for public health insight spanning policy, service delivery, and population behavior.

Structured data is sourced from three national institutions and covers 12 public health domains, including hospital care, pharmaceuticals, nutrition, mental health, demographics, and socio-economic factors. Indicators were selected for their relevance to public health monitoring and for their consistency across institutional datasets. Specifically, the Ministry of Health provides national and regional healthcare records, the Italian National Institute of Statistics (ISTAT) supplies health, demographic, and socio-economic data, and the Italian Medicines Agency (AIFA) contributes pharmaceutical datasets on drug consumption and pricing. All structured data is released under national open data licenses.

Unstructured textual data consists of Italian-language health news articles covering policy, research, service delivery, and epidemics. This corpus enriches the semantic context and powers the natural language querying features of the platform within the HANK chatbot, enhancing both the quality of interaction and the relevance of responses.

4. Proposed Architecture

MEDITA adopts a five-stage ETLMD (Extract, Transform, Load, Model, Deploy) pipeline to manage the full data lifecycle with modularity, automation, and interoperability. As shown in Figure 1, the workflow integrates data ingestion, harmonization, AI-based modeling, and deployment. Key methodological innovations occur in the Transform and Model stages.

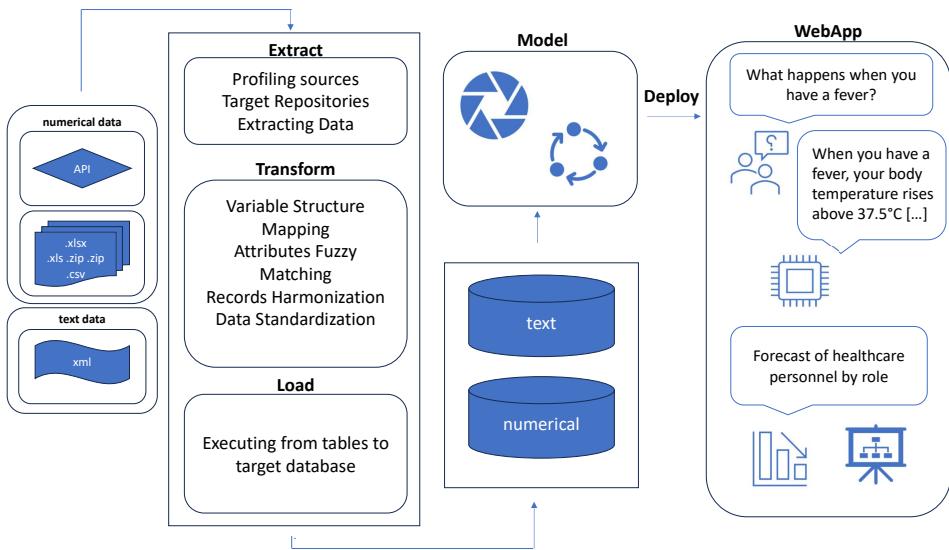


Figure 1: MEDITA Pipeline

The **Extract** stage acquires raw data from heterogeneous sources while preserving structural integrity. Metadata mapping is guided by operational flags; when enabled, the system parses source feeds, extracts schema information, and generates structured JSON metadata to inform downstream transformation routines. A system logger monitors all pipeline activity and records events at multiple severity levels.

In the **Transform** phase, data are cleaned, normalized, and harmonized using format-aware routines. Transformation logic includes schema detection, fuzzy attribute matching, record harmonization, and data standardization, as described in detail in Section 4.1.

The **Load** stage persists processed data into a structured relational database. A custom wrapper built on SQLAlchemy connects to a local SQLite instance through Object-Relational Mapping (ORM), enabling efficient Create, Read, Update, and Delete (CRUD) operations and ensuring alignment between the database schema and the application codebase [59, 60, 61].

The **Model** stage integrates both statistical modeling and conversational AI. Time-series forecasting and regression analysis are supported natively, while a zero-shot prompt engineering template anchors the RAG framework for natural language interaction with the data. The design and evaluation of the user-agent chatbot are detailed in Sections 4.2.1–4.2.2.

The **Deploy** phase manages reproducibility and accessibility. Git-based automation [62] supports continuous integration and version control, while the Streamlit interface is deployed on Streamlit Cloud,

making the prototype globally accessible without dedicated infrastructure.

4.1. Transform Phase

The Transform phase prepares extracted data for integration through format-aware routines applied dynamically according to input type. Its purpose is to resolve inconsistencies and align variable semantics across datasets using a reproducible, rule-based approach.

Textual sources are processed using lightweight cleaning modules tailored for downstream transformer models. Preprocessing includes removal of HTML tags and non-semantic tokens while preserving linguistic context to retain meaning.

Tabular data requires more extensive normalization due to inconsistencies in delimiters, quoting conventions, nested structures, and encodings. Inputs may be delivered in plaintext, CSV, or XML-based objects, with numeric formats varying by locale. All outputs are standardized as pipe-delimited files with UK-style formatting, consistent headers, and sanitized characters to ensure compatibility with modeling tools and databases.

The Transform phase comprises four core steps: (i) *Indicator Structure Mapping* identifies, for each indicator within a domain, the full topology of source objects and embedded tables that must be ingested together, and detects hierarchical subtypes within the same indicator; (ii) *Attributes Fuzzy Matching* resolves inconsistencies by aligning attribute names across instances of the same indicator using approximate string-matching techniques; (iii) *Records Harmonization* standardizes geographic and categorical variables in accordance with European NUTS standards, and derives a superkey to uniquely identify records while preserving data granularity; and (iv) *Data Standardization* enforces schema consistency and formatting rules to ensure cross-dataset uniformity. Together, these procedures support the robust integration of heterogeneous health data into a unified analytical framework.

4.1.1. Indicator Structure Mapping

A key challenge in standardizing input data is determining the number and layout of tables within each object associated with target indicators, especially in the absence of prior schema knowledge. While plain-text formats typically represent a single table structure, binary and Open XML formats often contain multiple tables embedded within a single file. In XML-based formats, each table is encapsulated as a separate object within a compressed archive and may exhibit hierarchical relationships.

Such hierarchies are typically structured around a general indicator with associated subtypes. For instance, a *Personnel* indicator may include subtypes such as *Doctors* and *Nurses*, linked via shared attributes. To address this, MEDITA introduces a classification framework that uses a flag-based system to detect and map object-table structures per indicator. This ensures reliable schema identification and supports harmonization across heterogeneous formats.

Formally, let $\mathcal{F} = \{f_1, f_2, \dots, f_m\}$ denote the set of data objects, where each object $f \in \mathcal{F}$ contains a set of tables $\mathcal{S}_f = \{s_1, s_2, \dots, s_n\}$, and $|\mathcal{S}_f|$ is its cardinality. Let $\mathcal{I} = \{i_1, i_2, \dots, i_p\}$ be the set of indicators, where each $i \in \mathcal{I}$ may include a main type and one or more subtypes. Each indicator i is associated with a subset $\mathcal{F}_i \subseteq \mathcal{F}$, where relevant data are distributed across objects and tables. Subtypes represent more granular components of an indicator and may appear within a single object or be distributed across instances.

To support schema harmonization, I define a classification function δ over object-table pair (s, f) as follows:

$$\delta(s, f) = \begin{cases} \delta_{SS} & \text{if } f = 1 \wedge s = 1, \\ \delta_{SM} & \text{if } f = 1 \wedge s > 1, \\ \delta_{MS} & \text{if } f > 1 \wedge s = 1 \quad \forall f \in \mathcal{F}_i, \\ \delta_{MM} & \text{if } f > 1 \wedge \exists f \in \mathcal{F}_i : s > 1. \end{cases} \quad (1)$$

This yields four structural categories: Single–Single (SS), Single–Multi (SM), Multi–Single (MS), and Multi–Multi (MM). The classification supports adaptive transformation and automatic subtype detection as new table configurations appear, enhancing the flexibility and robustness of the integration pipeline.

Table 1
Examples of Indicator Structure Mapping

Source / Indicator	Input Format	Analysis	Outcome	
Hospital Admissions	Single CSV file with one table	Ingested as a single standardized table	SS	
Semi-residential Health Services	Mental	Multiple CSV files	Merged longitudinally into one harmonized table	MS
Healthcare Workforce		XML-based workbook with multiple tables	Multiple tables with hierarchical subtypes (e.g., <i>Doctors</i> , <i>Nurses</i>), resulting in two distinct tables harmonized under common keys	MM

To illustrate, Table 1 presents real-world cases of indicator structure mapping applied to data. Cases of SM were not present in the current release of MEDITA and are therefore not illustrated.

4.1.2. Attributes Fuzzy Matching

When structural configurations such as MS or MM are encountered, inconsistencies often arise across tables belonging to the same indicator. These include slight changes, additions, or removals of attributes, which hinder interoperability. To resolve this, MEDITA applies fuzzy string matching to align attributes across all instances of an indicator.

Formally, let $\mathcal{T}_i = \{t_1, t_2, \dots, t_m\}$ denote the set of tables associated with a given indicator i . Each table t_j has a set of attributes $\mathcal{C}_j = \{c_1^j, c_2^j, \dots, c_n^j\}$. The pairwise similarity between any two attribute names $c_a^j \in \mathcal{C}_a$ and $c_b^k \in \mathcal{C}_b$ is computed using the normalized Levenshtein distance:

$$\text{Similarity}(c_a^j, c_b^k) = \left(1 - \frac{d(c_a^j, c_b^k)}{\max(|c_a^j|, |c_b^k|)} \right) \times 100 \quad (2)$$

Attribute alignment between t_j and t_k is accepted when $\text{Similarity}(c_a^j, c_b^k) \geq 90$. The threshold has been empirically tuned through testing to optimize matching accuracy across heterogeneous tables. The optimization objective minimizes cumulative dissimilarity across all attribute pairs of the indicator:

$$\min \sum_{j=1}^{m-1} \sum_{k=j+1}^m \sum_{c_a^j \in \mathcal{C}_j} \sum_{c_b^k \in \mathcal{C}_k} (100 - \text{Similarity}(c_a^j, c_b^k)) \quad (3)$$

The Levenshtein distance is zero for identical strings and bounded by the maximum string length. For equal-length strings, it is upper bounded by the Hamming distance [63]. It also satisfies the triangle inequality, ensuring consistent approximate matching [64]. This enables robust attribute alignment across heterogeneous schemas representing the same indicator.

In practice, fuzzy matching is particularly valuable for handling temporal or typographic variations that would otherwise fragment schemas. For example, yearly workforce tables labeled *Doctors_2021*, *Doctors_2022*, etc., are automatically recognized as belonging to the same attribute family. Likewise, inconsistencies such as pluralization or typographical errors are resolved through approximate string similarity.

4.1.3. Records Harmonization

The harmonization step ensures consistency in geographic and categorical attributes across indicators. A key distinction is made between *statistical units* (e.g., regions) and *administrative units* (e.g., *Provincia Autonoma di Bolzano*, *Trento*), which is critical when analyses require both regional aggregation and

administrative specificity. For example, hospital admissions may be aggregated at the regional level for comparability, while workforce allocation is often managed at the provincial level.

Region names are standardized, and codes are aligned with the NUTS level 2 classification⁷. Indicators are labeled as *geographic* if they include regional identifiers, or as *registry-based* otherwise, guiding the transformation routines accordingly.

To prevent information loss during aggregation—such as collapsing distinct categories when grouping by region—each harmonized indicator table is assigned a *superkey*, the minimal set of attributes uniquely identifying each row. Formally, given a dataset $D \in \mathbb{R}^{m \times n}$ with attributes C , the superkey is defined as the smallest subset $C_{\text{sub}} \subseteq C$ such that:

$$\min_{C_{\text{sub}} \subseteq C} |C_{\text{sub}}| \quad \text{subject to} \quad r_i \neq r_j \quad \forall i \neq j, \text{ where } r_i, r_j \in D[C_{\text{sub}}]. \quad (4)$$

If no such subset exists, the procedure returns \emptyset . This routine is applied automatically, enabling fine-grained analysis while avoiding aggregation bias.

4.1.4. Data Standardization

The final step addresses technical inconsistencies across heterogeneous datasets, such as delimiters, encodings, numerical formats, and XML structures. MEDITA implements adaptive parsing utilities that dynamically adjust to the format of each input source, ensuring that all indicators are transformed under a common framework.

Schema consistency is enforced through operations such as removing empty rows, correcting misaligned indices, trimming whitespace, and normalizing attribute names. Data type integrity is also standardized: locale-dependent number formats (e.g., commas as decimal separators) are converted into a consistent style, leading-zero identifiers are stored as strings, and integer-like values (e.g., 2020) are explicitly cast as integers.

This automated normalization framework produces schema-compliant, analysis-ready datasets aligned with relational database best practices. By minimizing manual intervention while preserving data fidelity, it provides the foundation for reproducible, multidimensional analysis in later stages of the pipeline.

4.2. Model Phase

Generative models alone often struggle with factuality and verifiability, issues that are particularly critical in healthcare. To address this, the Model phase of MEDITA integrates HANK, a domain-specific chatbot based on RAG, into the ETLMD pipeline. HANK delivers context-aware, Italian-language responses grounded in a curated healthcare corpus. A key element of its reliability is the use of structured prompt templates that constrain answers to retrieved evidence, thereby minimizing hallucinations and reducing reliance on model priors.

The underlying framework is developed in a companion study [65], which focuses on benchmarking generative models for public health question answering. Here we summarize the components of this setup and refer the reader to prior work for further details.

The modeling workflow comprises three core components. First, a dense vector store is constructed from Italian health-related news articles. Documents are segmented into chunks of 512 tokens with a 64-token overlap and embedded into high-dimensional space using the *Paraphrase Multilingual MiniLM L12 Sentence Transformer* [66, 67], which balances retrieval accuracy and efficiency. Second, user queries are parsed and transformed into a zero-shot prompt engineering template that restricts responses to retrieved evidence. Third, the retriever identifies the most relevant documents, which are then passed to a generator to synthesize answers.

For the generation stage, HANK relies on Gemma-2 [68], integrated within MEDITA as the default language model⁸. Performance is evaluated using standard metrics such as Exact Match (EM), ROUGE

⁷See: <https://ec.europa.eu/eurostat/web/nuts> for details.

⁸See Priola [65] for a broader evaluation of alternative models.

[69], BLEU [70], METEOR [71], and BERTScore [72], alongside the Negative Missing Information Scoring System (NMISS) [65], which focuses on contextual completeness and spurious hallucination detection.

To reduce memory footprint and latency during inference, HANK employs post-training quantization, converting both model weights and activations to lower-precision representations. Quantization significantly accelerates inference and reduces resource demands while maintaining accuracy [73].

By combining quantization with retrieval and controlled prompting, this efficiency-oriented design ensures that HANK produces responses that are both computationally efficient and grounded in Italian public health data, extending beyond the limitations of conventional generative models.

4.2.1. Retrieval-Augmented Generation

The RAG framework consists of two components: a retriever that identifies relevant documents c for a given query q , and a generator that produces the response r token by token, conditioned on q and c . Italian-language news articles are embedded using Sentence Transformers [74], enabling high-quality semantic retrieval.

This work adopts the **RAG-Sequence** variant, which conditions the entire response on a fixed retrieved document set. Let $\mathcal{P}_\eta(c | q)$ denote the retriever and $\mathcal{P}_\theta(r_i | q, c, r_{1:i-1})$ the generator. The full generation objective is:

$$\mathcal{P}_{\text{RAG}}(r | q) = \sum_{c \in \text{top-}k(\mathcal{P}_\eta(c | q))} \mathcal{P}_\eta(c | q) \prod_{i=1}^N \mathcal{P}_\theta(r_i | q, c, r_{1:i-1}) \quad (5)$$

where $r = (r_1, r_2, \dots, r_N)$ is the output sequence.

To ensure factual accuracy, MEDITA uses a prompting scheme in which user queries are embedded within structured instructions. The structure and rationale for these prompts are described next.

4.2.2. Prompt Engineering Template

Prompt engineering refers to the process of crafting input instructions that guide a language model to perform a specific task. As noted by Liu et al. [75], the structure and wording of prompts significantly influence both model behavior and performance. In healthcare contexts, carefully designed prompts help reduce hallucinations and improve factual accuracy [76, 77]. Manual prompts have shown strong empirical results across a range of tasks, including question answering and reasoning [78].

The structured prompt adopted by HANK follows a simple question–answer format anchored to the retrieved context illustrated in Table 2. This zero-shot design relies on pretrained knowledge while enforcing contextual grounding [79, 80].

Table 2

Prompt engineering template

Component	Example
Instruction	You are a chatbot that loves helping people! Given the following context section, answer the question using only the provided context. If you are unsure and the answer is not explicitly written in the documentation, respond by saying, 'I'm sorry, I can't help you based on the information I have.'
User	[User Question Here]
Context	[Context Retrieved Here]
System	[System Answer here]

4.3. Implementation Stack

MEDITA follows a layered design that balances reproducibility and ease of deployment. At the storage layer, the platform uses SQLite as a lightweight relational backend, enabling rapid ingestion and query without server overhead, appropriate for a proof-of-concept Data Lakehouse.

The data ingestion and preparation process is handled by an ETLMD pipeline written entirely in Python. This combines widely used libraries such as *pandas* and *SQLAlchemy* for tabular manipulation, *Dask*, *Multiprocessing* for parallelization, and *fuzzywuzzy* for fuzzy logic, together with utilities like *BeautifulSoup* for parsing semi-structured formats and JSON metadata for schema-on-read structural classification. This design enables heterogeneous file formats to be harmonized under a consistent framework.

Analytical functionality combines classical statistics and machine learning models. Econometric models are provided via *statsmodels*, while *scikit-learn* supports supervised learning and diagnostics. For embedding and neural components, the stack uses *HuggingFace*, *Transformers* and *Keras*.

The conversational interaction uses Sentence-BERT embeddings via *Langchain* and a vector store built with *Chroma* backed by FAISS indexing. Generation uses HuggingFace *Transformers*, with 4-bit quantization through *BitsAndBytesConfig* to reduce latency and memory usage during inference. This RAG stack is integrated end-to-end within the pipeline.

The user interface is implemented as a *Streamlit* web application, combining lightweight deployment with interactivity. Visualizations are produced using *Plotly*, while authentication and access control are managed through YAML-based configuration files.

Finally, deployment is realized on Streamlit Cloud, which provides global accessibility in the absence of a dedicated infrastructure. More advanced scenarios, such as Docker-based containerization or cloud-native deployments, are left for future extensions.

4.4. Target Users and Use Cases

MEDITA is designed to serve two main communities of users: researchers, who require reproducible access to harmonized indicators for analysis, and citizens, who benefit from simplified interaction with public health information.

For researchers, the platform provides tools to consolidate fragmented datasets and conduct comparative analyses across time and geography. A researcher can search for indicators by keyword, inspect metadata, and filter attributes much like in a search engine. Once selected, the system automatically checks for missing values, data integrity, and historical consistency. Indicators of interest can then be subjected to diagnostic statistical tests and used in forecasting models. For example, an epidemiologist studying cardiovascular disease may merge hospitalization data with demographic indicators, validate the combined dataset, and forecast future admissions. The FAIR principle of *Findability* is ensured because every indicator remains traceable back to its official source, with standardized metadata reporting the structure of the dataset.

Beyond expert users, MEDITA also democratizes access for citizens, who often find open datasets technically available but practically inaccessible. Through HANK, non-specialists can query the system in natural language—for instance, asking “What is Asperger syndrome?” and receive grounded responses. Citizens may also explore health indicators across regions without requiring statistical expertise, making public health evidence both transparent and actionable.

5. Experimental Results

5.1. Data Collection

The structured corpus comprises 1,403 features distributed across 12 public health domains, as shown in Table 3. After harmonization, these were consolidated into 110 integrated indicators, totaling nearly half a billion observations.

Table 3
Topic Distribution Statistics

Topic	No. Features	No. Obs	No. Records	NAs (%)
Education and Disability	36	9,216	528	5.99
Food	106	116,674,075	8,869,732	4.00
Health and Mortality	185	178,960,360	8,070,354	2.36
Hospital Data	111	4,450,239	199,814	2.56
Medical Devices	141	61,473,812	6,484,224	27.07
Mental Health	109	472,707	41,450	0.46
NSS Personnel	153	1,135,756	58,855	0.23
National Health System	136	305,919	23,277	7.48
Pharmaceuticals	265	42,725,248	3,083,274	17.82
Social Security	54	165,528	9,196	-
Sociodemographic Indicators	69	11,371,407	520,260	7.93
Violence	38	43,472	1,144	-
Total	1,403	417,787,739	27,362,108	15.27

Note: This table summarizes data showing the total number of attributes, records, observations, and ratio of missing data percentages.

Overall missingness averages about 8% across variables, although the total reported in Table 3 is higher (15.3%) due to the concentration of gaps in domains such as *Medical Devices* and *Pharmaceuticals*. In contrast, domains such as *Mental Health*, *Social Security*, and *Violence* are nearly complete. Data complexity is uneven: some domains exceed 150 features, while others remain under 40.

The dataset includes a diverse range of attribute types, such as boolean, categorical, date, numerical, and free-text. This heterogeneity reflects the richness of Italian public health data but also highlights the need for rigorous harmonization procedures. Prioritizing the reduction of missing values in critical domains will further enhance overall dataset quality.

The temporal distribution shown in Figure 2 illustrates observation counts across domains for the indicators with temporal granularity. The top contributors by volume include *Health and Mortality*, *Medical Devices* and *Pharmaceuticals*. Several indicators exhibit fluctuating reporting intensity over time, reflecting shifts in institutional priorities, policy interventions, or changes in reporting standards.

In parallel, 126,470 Italian-language health news articles were collected from 2010 to 2024. A preprocessing step removed HTML tags and artifacts. Articles average 348 tokens (SD = 264), with lengths ranging from 6 to 5,594 tokens. The median length is 284 tokens, with an interquartile range of 195–421, indicating considerable variation in reporting styles—a factor that complicates downstream semantic retrieval.

The news corpus spans a broad set of themes, as shown in Figure 3. The most represented is *regional affairs and governance* (25%), followed by *biomedicine* (20%), *media commentary* (13%), *national health governance* (13%), and *health professions* (12%). Furthermore, 11% of the articles are *uncategorized* due to missing metadata rather than the lack of thematic classification. The least represented categories include topics such as *COVID-19* and *Local health governance*. This thematic diversity supports future classification, topic modeling, and retrieval tasks.

5.2. ETL Pipeline Performance

To evaluate efficiency, I benchmark the ETL pipeline during ingestion and harmonization. All experiments were conducted on a 12th Gen Intel(R) Core(TM) i7-12700H CPU 2.70 GHz, with 14 cores, 20 logical processors, and 16 GB of DDR4 RAM.

Overall, the system processed the 12 thematic domains described in Section 5.1, resolving more than 600 attribute mismatches and 1,100 inconsistencies across a total of 268 input tables. Examples include the harmonization of misaligned headers and the automatic detection of malformed rows.

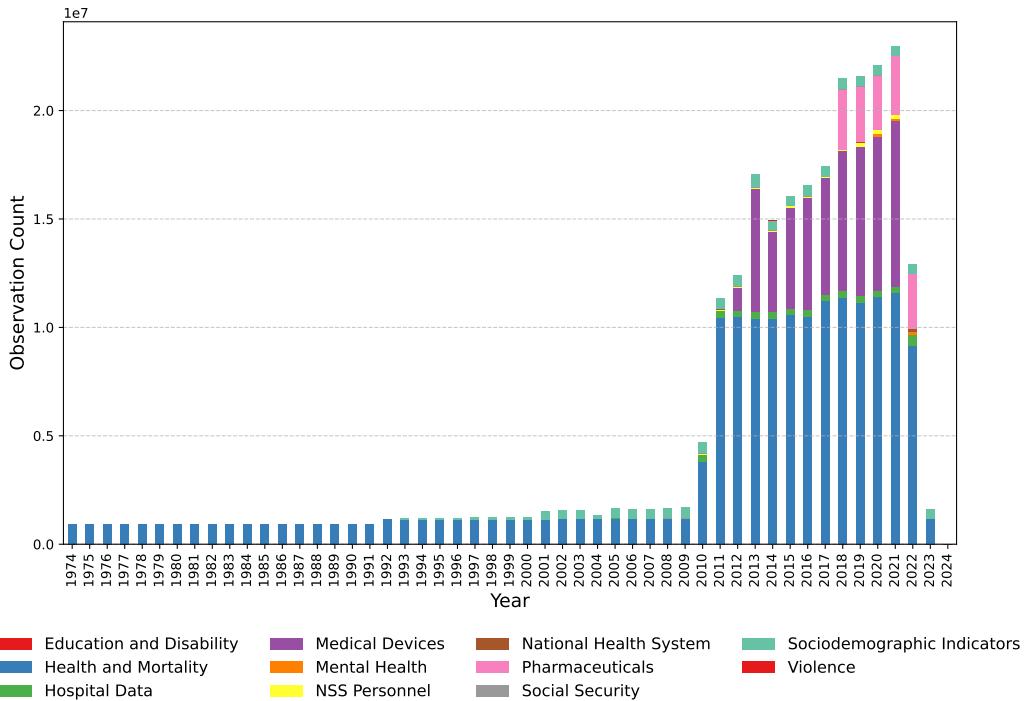


Figure 2: Data Topic Distribution Over Time (Observation Counts)

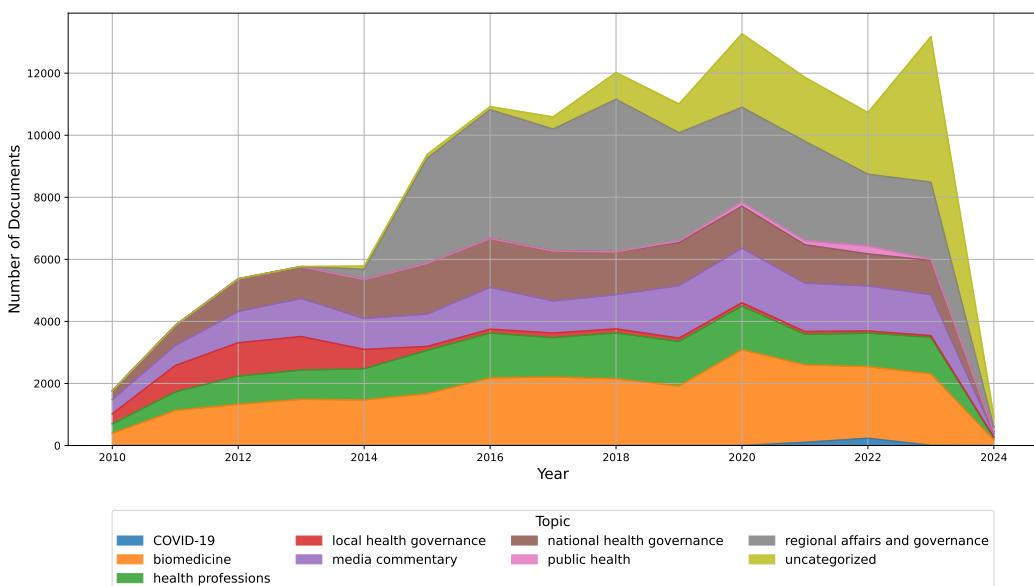


Figure 3: News Topic Distribution Over Time

For structured data, the extraction phase required a total of 1 hour 29 minutes, with durations ranging from about 15 seconds for AIFA, to roughly 35 minutes for MINIS, and nearly 1 hour for ISTAT. Transformation consumed in total 1 hour and 37 minutes, varying from about 4 minutes (AIFA), to nearly 24 minutes (MINIS), and more than 1 hour for ISTAT. For unstructured sources, each news article requires 0.09 seconds to be ingested and transformed. The loading phase required about 9 minutes overall.

5.3. User-Agent Chatbot

HANK combines semantic search with transformer-based response generation, ensuring both factual accuracy and contextual relevance. It supports user-centered exploration across a broad spectrum of general health topics, offering concise answers grounded in publicly available sources. Table 4 presents illustrative examples of HANK’s behavior across different query types.

Table 4

Examples of Context-Grounded Responses Generated by HANK

Question	Retrieved Context (excerpt)	Chatbot Response
What are the new findings about the gut microbiome?	“Gut microbiota varies by diet, environment, and health. Some microorganisms, like <i>Bacteroidetes</i> , are linked to obesity. Metagenomics targets gut-related diseases.”	Metagenomics enables sequencing of gut microorganisms to study diseases like liver cirrhosis and IBD, offering insights into microbiome-related conditions.
How can epigenetics help us understand diseases?	“Results can support protocols for precision medicine, adapting to patient characteristics.”	I am sorry, I cannot help you based on the information I have.

Note: Retriever extracts content chunks; Generator uses these to compose answers using the RAG framework.

The examples show how HANK synthesizes retrieved content into informative responses when evidence is sufficient, and abstains when it is not. In the second case, for instance, the model refrains from answering because the retrieved documents lack content relevant to the user’s query.

To evaluate HANK within the MEDITA platform, I rely on a benchmark of 100 general health questions stratified by complexity (low, mid and high), as well as multiple evaluation metrics. Reference answers are first manually drafted, then validated by a domain expert. Model responses are annotated with a binary label to indicate the presence or absence of hallucinations. NMISS is applied specifically to non-hallucinated responses and further restricted to borderline cases, thereby filtering out trivial matches or complete failures. This ensures NMISS targets semantically context-sensitive answers where classical metrics often fall short.

Outperformance under NMISS indicates cases where classical metrics assign lower scores to non-hallucinated answers that remain semantically faithful but diverge lexically from the reference. In such instances, NMISS successfully captures contextual adequacy that classical token-based metrics underestimate due to their surface-level overlap criteria. The full breakdown of results across levels and annotation protocol is detailed in Priola [65].

Across classical metrics, **Gemma-2** demonstrates strong and stable performance. Under NMISS, Gemma-2 achieves major gains in more than 80% of valid BLEU and ROUGE-1 cases, with additional improvements under ROUGE-L, confirming its tendency to generate semantically faithful but lexically divergent responses.

Table 5

Performance of Gemma-2

	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	METEOR	EM	BERTScore
Gemma-2	0.722 (82.51)	0.663 (5.59)	0.698 (45.45)	0.607 (82.43)	0.702 (12.14)	0.370	0.880

Note. Values outside parentheses are averages across all question levels (low, medium, high) under classical metrics. Numbers in parentheses indicate the percentage of valid cases where NMISS outperformed the corresponding metric. NMISS is not applicable to EM and BERTScore.

Given its strong performance in both standard and hallucination-sensitive evaluation, Gemma-2 is adopted as the default generation model in MEDITA. It offers the best balance between factual accuracy, contextual relevance, and efficiency.

5.4. Statistical and Multidimensional Analysis

MEDITA enables interactive statistical analysis, supporting tasks such as exploratory data analysis, predictive modeling, and diagnostic validation. Users can select numerical variables and apply standard checks, including stationarity [81], multicollinearity, and heteroscedasticity [82]. Supported regression models include OLS, GLS, ARIMAX, and SARIMAX [83]; Logistic Regression is available for classification. The system automatically handles missing data and flags issues such as high null rates or low sample-to-feature ratios. Results are returned with diagnostic summaries and interactive visualizations, using standard metrics such as MAE, MSE, and R^2 for regression, or accuracy, precision, recall, F1, and ROC-AUC for classification.

Beyond statistical models, MEDITA structures health data according to OLAP-inspired principles, though without implementing a full cube engine. Temporal and geographical filters enable *slice* operations, while combining multiple conditions supports *dice*-like subsetting. Forecasting modules perform a system-driven *roll-up* of time series to the yearly level. This is a design choice, since annual reporting represents the most consistent level of granularity across Italian public health datasets. While some indicators exist at monthly resolution, they are not consistent across sources. Conversely, *drill-down* below the regional (NUTS-2) level is intentionally disabled, as subregional data are not systematically available. These restrictions preserve interoperability and comparability across indicators in this proof-of-concept release.

In this setting, *facts* correspond to quantitative indicators, while *measures* are basic computations such as totals, averages, or medians. *Dimensions* define the axes of analysis, currently restricted to temporal and geographical attributes. Predictions generated by forecasting modules can also be treated as new facts, linking descriptive exploration with predictive analytics and moving the platform beyond static reporting toward proactive decision support. Table 6 summarizes how these OLAP-inspired operations are implemented in the platform.

Table 6
Mapping of OLAP-inspired operations

Operation	Platform Layer	Implementation in MEDITA
<i>Slice</i>	Data Explorer	Temporal and geographical filters restricting the scope of analysis.
<i>Dice</i>	Data Explorer Forecaster	Multiple combined filters (dataset, time period, attributes) to extract analytical subcubes.
<i>Roll-up</i>	Forecaster	Automatic aggregation of temporal series to the yearly level, ensuring interoperability.
<i>Drill-down</i>	—	Not supported: regional (NUTS-2) level is the lowest consistent granularity; finer detail is left for future work.

5.5. Web Application Interface

The MEDITA platform features an intuitive, Streamlit-based dashboard⁹, designed to guide users through the analytical workflow via dedicated pages for data exploration, modeling, and assistance. The interface prioritizes usability and accessibility, with Figure 4a illustrating the homepage.

Users begin with the **Infographics** section (Figure 4b), which provides aggregated summaries by year, topic, or region to support initial exploration. The **Data Explorer** (Figure 4c) enables deeper investigation through interactive visualizations, variable-level details, and dynamic querying.

⁹<https://docs.streamlit.io/>

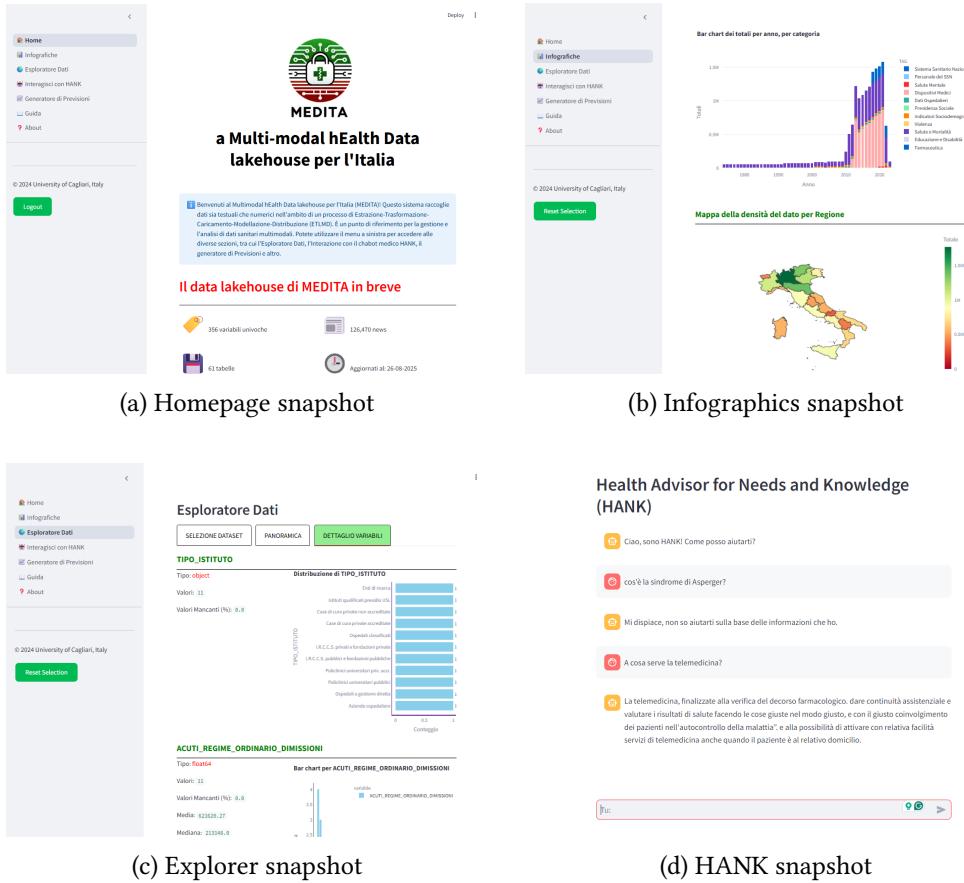


Figure 4: User interface of the MEDITA platform

Next, the **HANK Chatbot** (Figure 4d) offers natural language access to health-related insights, assisting non-technical users with contextualized responses grounded in the platform’s knowledge base. The **Forecast Generator** enables predictive modeling of selected trends, supporting scenario analysis and decision-making. Additional sections include **Help**, offering practical guidance and FAQs, and **About**, outlining platform objectives and references.

6. Conclusions

The growing complexity of healthcare data demands infrastructures that integrate, harmonize, and analyze heterogeneous sources at scale. While Data Warehouses offer governance and Data Lakes flexibility, only the emerging Data Lakehouse paradigm reconciles these trade-offs. Existing institutional platforms for public health, however, remain mostly limited to query-based dashboards or filter-based access to static structured data, offering little in terms of harmonization, interactivity, or predictive modeling.

This paper introduces the **Multimodal Health Data lakehouse for ITALy (MEDITA)**, the first Italian proof-of-concept of a Data Lakehouse for public health. By combining a five-stage ETLMD pipeline with schema harmonization, advanced analytics, and a Retrieval-Augmented Generation chatbot, MEDITA enables both researchers and citizens to interact with multimodal health data through reproducible, FAIR-aligned workflows.

Looking forward, this work aims to extend the prototype to additional modalities such as electronic health records, audio, and video, and to explore cloud-native deployment for scalability. Interactivity will also be enhanced by adding notebook-style environments, enabling advanced users to directly experiment with data and models within the same ecosystem.

In this way, the platform is positioned to evolve into a next-generation framework for public health

intelligence, supporting scientific inquiry and evidence-based policy.

Acknowledgments

This publication was produced while attending the PhD program in Economics and Business at the University of Cagliari, Cycle 38, with the support of a scholarship financed by the Ministerial Decree no. 351 of 9th April 2022, based on the NRRP - funded by the European Union - NextGenerationEU - Mission 4 “Education and Research”, Component 1 “Enhancement of the offer of educational services: from nurseries to universities” - Investment 3.4 “Advanced teaching and university skills”.

Declaration on Generative AI

During the preparation of this work, the author used generative AI to check grammar and spelling, and to improve readability. The content and conclusions of the paper are the sole responsibility of the author.

References

- [1] S. H. A. El-Sappagh, A. M. A. Hendawi, A. H. El Bastawissy, A proposed model for data warehouse etl processes, *Journal of King Saud University-Computer and Information Sciences* 23 (2011) 91–104.
- [2] W. H. Inmon, *Building the data warehouse*, John wiley & sons, 2005.
- [3] H. Jiawei, K. Micheline, P. Jian, *Data preprocessing, Data Mining: Concepts and Techniques*. 3rd edition. Waltham, MA: Morgan Kaufmann-Elsevier (2012).
- [4] R. Kimball, M. Ross, *The data warehouse toolkit: The definitive guide to dimensional modeling*, ed. wiley (2019).
- [5] N. Stolba, A. M. Tjoa, The relevance of data warehousing and data mining in the field of evidence-based medicine to support healthcare decision making, *International Journal of Computer Systems Science and Engineering* 3 (2006) 143–148.
- [6] J. A. Lyman, K. Scully, J. H. Harrison Jr, The development of health care data warehouses to support data mining, *Clinics in laboratory medicine* 28 (2008) 55–71.
- [7] S. Nugawela, Data warehousing model for integrating fragmented electronic health records from disparate and heterogeneous clinical data stores, Ph.D. thesis, Queensland University of Technology, 2013.
- [8] J. Dixon, Pentaho, hadoop, and data lakes, <https://jamesdixon.wordpress.com>, 2010. Blog post.
- [9] A. A. Harby, F. Zulkernine, From data warehouse to lakehouse: A comparative review, in: 2022 IEEE international conference on big data (big data), IEEE, 2022, pp. 389–395.
- [10] A. A. Harby, F. Zulkernine, Data lakehouse: a survey and experimental study, *Information Systems* 127 (2025) 102460.
- [11] M. Armbrust, A. Ghodsi, R. Xin, M. Zaharia, et al., Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics, in: *Proceedings of CIDR*, volume 8, 2021, p. 28.
- [12] E. Begoli, I. Goethert, K. Knight, 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), IEEE, 2021, pp. 4643–4651.
- [13] R. Gebler, I. Reinecke, M. Sedlmayr, M. Goldammer, Enhancing clinical data infrastructure for ai research: Comparative evaluation of data management architectures, *Journal of Medical Internet Research* 27 (2025) e74976.

- [14] N. M. Guerreiro, D. M. Alves, J. Waldendorf, B. Haddow, A. Birch, P. Colombo, A. F. Martins, Hallucinations in large multilingual translation models, *Transactions of the Association for Computational Linguistics* 11 (2023) 1500–1517.
- [15] N. Varshney, W. Yao, H. Zhang, J. Chen, D. Yu, A stitch in time saves nine: Detecting and mitigating hallucinations of llms by actively validating low-confidence generation, arXiv preprint arXiv:2307.03987 (2023).
- [16] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, *ACM Computing Surveys* 55 (2023) 1–38.
- [17] P. Lewis, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, arXiv (2020). URL: <https://arxiv.org/abs/2005.11401>.
- [18] H. Mahler, The meaning of " health for all by the year 2000", in: *World Health Forum*, volume 2, 1981, pp. 5–22.
- [19] J. Gruendner, C. Gulden, M. Kampf, S. Mate, H.-U. Prokosch, J. Zierk, et al., A framework for criteria-based selection and processing of fast healthcare interoperability resources (fhir) data for statistical analysis: design and implementation study, *JMIR medical informatics* 9 (2021) e25645.
- [20] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, *Scientific data* 3 (2016) 1–9.
- [21] A. Burgun, E. Bernal-Delgado, W. Kuchinke, T. van Staa, J. Cunningham, E. Lettieri, C. Mazzali, D. Oksen, F. Estupiñan, A. Barone, et al., Health data for public health: towards new ways of combining data sources to support research efforts in europe, *Yearbook of medical informatics* 26 (2017) 235–240.
- [22] T. R. Campion Jr, C. K. Craven, D. A. Dorr, B. M. Knosp, Understanding enterprise data warehouses to support clinical and translational research, *Journal of the American Medical Informatics Association* 27 (2020) 1352–1358.
- [23] M. G. Ayadi, R. Bouslimi, J. Akaichi, A framework for medical and health care databases and data warehouses conceptual modeling support, *Network Modeling Analysis in Health Informatics and Bioinformatics* 5 (2016) 1–21.
- [24] S. I. Khan, A. S. M. Latiful Hoque, Development of national health data warehouse for data mining, *Database Systems Journal* 6 (2015).
- [25] S. A. M. Rizi, A. Roudsari, Development of a public health reporting data warehouse: lessons learned, in: *MEDINFO 2013*, IOS Press, 2013, pp. 861–865.
- [26] S. I. Khan, A. Hoque, M. Ullah, National health data warehouse bangladesh for remote health monitoring: Features, problems and privacy issues, in: *Remote health monitoring workshop*, volume 6, 2016.
- [27] L. M. Fleuren, T. A. Dam, M. Tonutti, D. P. de Bruin, R. C. Lalisan, D. Gommers, O. L. Cremer, R. J. Bosman, S. Rigter, E.-J. Wils, et al., The dutch data warehouse, a multicenter and full-admission electronic health records database for critically ill covid-19 patients, *Critical Care* 25 (2021) 1–12.
- [28] B. Ozaydin, F. Zengul, N. Oner, S. S. Feldman, Healthcare research and analytics data infrastructure solution: a data warehouse for health services research, *Journal of medical Internet research* 22 (2020) e18579.
- [29] E. F. Codd, A relational model of data for large shared data banks, *Communications of the ACM* 13 (1970) 377–387.
- [30] IBM, Ibm data warehouse manager, 2009. Available at: <http://www-306.ibm.com/software/data/integration/datastage/>.
- [31] Microsoft, Sql server 2005 integration services (ssis), 2009. Available at: <http://technet.microsoft.com/enus/sqlserver/bb331782.aspx>.
- [32] Oracle, Oracle warehouse builder 10, 2009. Available at: <http://www.oracle.com/technology/products/warehouse/>.
- [33] S. Tsur, C. Zaniolo, Ldl: a logic-based data-language, in: *VLDB*, volume 86, 1986, pp. 33–41.
- [34] S. Naqvi, S. Tsur, A logical language for data and knowledge bases, *Computer Science Press, Inc.*, 1989.

- [35] O. Omg, Q. F. A. Specification, Object management group, Download of the UML Specification (2005).
- [36] S. Luján-Mora, P. Vassiliadis, J. Trujillo, Data mapping diagrams for data warehouse design with uml, in: Conceptual Modeling—ER 2004: 23rd International Conference on Conceptual Modeling, Shanghai, China, November 8–12, 2004. Proceedings 23, Springer, 2004, pp. 191–204.
- [37] N. Hong, A. Wen, F. Shen, S. Sohn, C. Wang, H. Liu, G. Jiang, Developing a scalable fhir-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data, *JAMIA open* 2 (2019) 570–579.
- [38] R. Kavitha, E. Kannan, S. Kotteswaran, Implementation of cloud based electronic health record (ehr) for indian healthcare needs, *Indian Journal of Science and Technology* (2016).
- [39] Y. Peng, E. Henke, I. Reinecke, M. Zoch, M. Sedlmayr, F. Bathelt, An etl-process design for data harmonization to participate in international research with german real-world data based on fhir and omop cdm, *International Journal of Medical Informatics* 169 (2023) 104925.
- [40] W. H. Organization, The new digital destination for open health data, 2023. Available at: <https://www.who.int/news-room/feature-stories/detail/who-releases-data.who.int/>.
- [41] D. Mazumdar, J. Hughes, J. Onofre, The data lakehouse: Data warehousing and more, *arXiv preprint arXiv:2310.08697* (2023).
- [42] F. Aziz, Next-generation healthcare analytics: The open lakehouse framework, *International Journal of Research and Analytical Reviews (IJRAR)* 11 (2024) 94–105.
- [43] S. K. Tanbeer, E. R. Sykes, Myhealthportal—a web-based e-healthcare web portal for out-of-hospital patient care, *Digital Health* 7 (2021) 2055207621989194.
- [44] S. Mullin, J. Zhao, S. Sinha, R. Lee, B. Song, P. L. Elkin, Clinical data warehouse query and learning tool using a human-centered participatory design process, in: *Data, Informatics and Technology: An Inspiration for Improved Healthcare*, IOS Press, 2018, pp. 59–62.
- [45] G. R. Banu, P. Kuppuswamy, N. Sasikala, Implementation of big data in health information systems: sample approaches in saudi hospital, *International Journal of Computer Applications* 160 (2017).
- [46] X. Chen, J. Wang, C. Faviez, X. Wang, M. Vincent, R. Tsopra, A. Burgun, N. Garcelon, An integrated pipeline for phenotypic characterization, clustering and visualization of patient cohorts in a rare disease-oriented clinical data warehouse, in: *Digital Health and Informatics Innovations for Sustainable Health Care Systems*, IOS Press, 2024, pp. 1785–1789.
- [47] A. Karabegovic, M. Ponjovic, Geoportal as interface for data warehouse and business intelligence information system, in: *Advances in Business ICT*, Springer, 2014, pp. 27–40.
- [48] N. Schuurman, M. Leight, M. Berube, A web-based graphical user interface for evidence-based decision making for health care allocations in rural areas, *International Journal of Health Geographics* 7 (2008) 1–12.
- [49] J. Chosy, K. Benson, D. Belen, R. Starr, T. L. St John, R. R. Starr, L. K. Ching, Insights in public health: For the love of data! the hawaii ‘i health data warehouse, *Hawai’i Journal of Medicine & Public Health* 74 (2015) 382.
- [50] G. Agapito, C. Zucco, M. Cannataro, Covid-warehouse: A data warehouse of italian covid-19, pollution, and climate data, *International Journal of Environmental Research and Public Health* 17 (2020) 5596.
- [51] G. Turcan, S. Peker, A multidimensional data warehouse design to combat the health pandemics, *Journal of Data, Information and Management* 4 (2022) 371–386.
- [52] A. Prasad, B. S. S. S. Jennifer, D. Ghsoh, H. Busschetty, D. T. J.T, Chatbot in healthcare, *International Journal of Engineering Research in Computer Science and Engineering* (2022). URL: <https://api.semanticscholar.org/CorpusID:253281361>.
- [53] B. S. Garimella, H. S. Garlapati, S. Choul, R. Cherukuri, P. Lanke, Advancing healthcare accessibility: Development of an ai-driven multimodal chatbot, in: 2024 4th International Conference on Intelligent Technologies (CONIT), 2024, pp. 1–10. doi:10.1109/CONIT61985.2024.10626795.
- [54] S. Patil, J. Darji, S. Hingu, A. Thakkar, Medic: Smart healthcare ai assistant, in: *Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021)*, 2021.
- [55] M. Nahala, K. Roshan, F. Cruz, T. Vivek, Medibot: A medical assistant chatbot (????).

- [56] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al., Palm: Scaling language modeling with pathways, *Journal of Machine Learning Research* 24 (2023) 1–113.
- [57] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al., Large language models encode clinical knowledge, *Nature* 620 (2023) 172–180.
- [58] M. Griot, C. Hemptonne, J. Vanderdonckt, D. Yuksel, Large language models lack essential metacognition for reliable medical reasoning, *Nature communications* 16 (2025) 642.
- [59] R. Copeland, *Essential sqlalchemy*, ” O'Reilly Media, Inc.”, 2008.
- [60] J. Martin, *Managing the data base environment*, Prentice Hall PTR, 1983.
- [61] T. Halpin, Object-role modeling (orm/niam), in: *Handbook on architectures of information systems*, Springer, 2006, pp. 81–103.
- [62] S. Chacon, B. Straub, S. Chacon, B. Straub, Git in other environments, *Pro Git* (2014) 389–399.
- [63] R. W. Hamming, Error detecting and error correcting codes, *The Bell system technical journal* 29 (1950) 147–160.
- [64] D. Fisman, J. Grogin, O. Margalit, G. Weiss, The normalized edit distance with uniform operation costs is a metric, *arXiv preprint arXiv:2201.06115* (2022).
- [65] M. P. Priola, Addressing hallucinations with rag and nmiss in italian healthcare llm chatbots, *arXiv preprint arXiv:2412.04235* (2024).
- [66] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2019. URL: <http://arxiv.org/abs/1908.10084>.
- [67] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2020. URL: <https://arxiv.org/abs/2004.09813>.
- [68] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, et al., Gemma 2: Improving open language models at a practical size, *arXiv preprint arXiv:2408.00118* (2024).
- [69] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: *Text summarization branches out*, 2004, pp. 74–81.
- [70] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
- [71] S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.
- [72] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, *arXiv preprint arXiv:1904.09675* (2019).
- [73] G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, S. Han, Smoothquant: Accurate and efficient post-training quantization for large language models, in: *International conference on machine learning*, PMLR, 2023, pp. 38087–38099.
- [74] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* 30 (2017).
- [75] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, *ACM computing surveys* 55 (2023) 1–35.
- [76] S. Tonmoy, S. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, A. Das, A comprehensive survey of hallucination mitigation techniques in large language models, *arXiv preprint arXiv:2401.01313* (2024).
- [77] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, D. C. Schmidt, A prompt pattern catalog to enhance prompt engineering with chatgpt, *arXiv preprint arXiv:2302.11382* (2023).
- [78] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam,

G. Sastry, A. Askell, et al., Language models are few-shot learners, *Advances in neural information processing systems* 33 (2020) 1877–1901.

[79] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, *Advances in neural information processing systems* 35 (2022) 22199–22213.

[80] S. Qiao, Y. Ou, N. Zhang, X. Chen, Y. Yao, S. Deng, C. Tan, F. Huang, H. Chen, Reasoning with language model prompting: A survey, *arXiv preprint arXiv:2212.09597* (2022).

[81] G. Elliott, T. J. Rothenberg, J. H. Stock, Efficient tests for an autoregressive unit root, 1992.

[82] G. James, *An introduction to statistical learning*, 2013.

[83] C. Chatfield, H. Xing, *The analysis of time series: an introduction with R*, Chapman and hall/CRC, 2019.