iSys - Journal of Information Systems, 2026, 19:1
ISSN: 1984-2902 • doi: 10.5753/isys.2026.8404
This work is licensed under a Creative Commons Attribution 4.0 International License.

Research Paper

HERO: An Architecture for Integrating Health Data from Multiple Data Sources

Izabel Yale Neves Nascimento [Universidade Federal do Agreste de Pernambuco / Universidade de Pernambuco | *izabel.yalenn@upe.br*]
Tiago Brasileiro Araújo [Instituto Federal de Educação, Ciência e Tecnologia da Paraíba |*tiago.brasileiro@ifpb.edu.br*]
Luis Filipe Alves Pereira [Universidade Federal do Agreste de Pernambuco |*luis-filipe.pereira@ufape.edu.br*]
Veruska Borges Santos [Universidade Federal de Campina Grande |*veruska.santos@ccc.ufcg.edu.br*]
Dimas Cassimiro do Nascimento Filho [Universidade Federal do Agreste de Pernambuco |*dimas.cassimiro@ufape.edu.br*]

Universidade Federal do Agreste de Pernambuco, Avenida Bom Pastor, Boa Vista, Garanhuns, PE, CEP 55292-278, Brazil.

Abstract. Public health management in Brazil benefits from the availability of large volumes of public data on financing, service delivery, infrastructure, and population coverage. Existing approaches support interoperability and multi-source integration through ETL/ELT pipelines, semantic standards, and scalable architectures such as Data Lakes and Data Lakehouses. However, in public health management contexts, these approaches often remain focused on clinical ecosystems, domain-specific workflows, or isolated interoperability requirements, leaving open challenges related to governed integration, reproducibility, and continuous analytical consumption of heterogeneous public data. To address this gap, this work proposes the HERO architecture, a modular and scalable framework that extracts, transforms, integrates, and makes available public health data from multiple sources through a structured ETL pipeline and a Data Lakehouse. HERO builds upon principles of the HEALER architecture, introduced in 2023, and extends them with a focus on systems integration, maintainability, extensibility, and reproducibility in public health management scenarios. As a proof of concept, this study applies Data Envelopment Analysis (DEA) to data consolidated by HERO in order to evaluate the technical efficiency of municipalities in the state of Pernambuco, Brazil. The results reveal temporal and spatial efficiency patterns, persistent inequalities, and relevant correlations across different data sources, demonstrating HERO's usefulness for reproducible analytical workflows and decision-support tools. Overall, the proposed architecture not only automates and reliably consolidates heterogeneous public health data, but also enhances their analytical value while remaining compatible with broader analytical uses in evidence-based public health management.

Keywords: Health Management; Data Integration; ETL; Data Lake; Efficiency Analysis; DEA

Received: 06 June 2026 • **Accepted:** 20 October 2026 • **Published:** 06 November 2026

1 Introduction

Brazilian public health management is supported by a comprehensive national health system and a wide availability of public data covering financing, service delivery, infrastructure, and population coverage. In particular, Primary Health Care (PHC), the main gateway to Brazil's Unified Health System (SUS), plays a central role in health promotion, disease prevention, continuity of care, and territorial equity, concentrating a large share of healthcare services and responsibilities [Silva *et al*., 2023]. This context creates significant opportunities for transparency, performance monitoring, and evidence-based decision-making.

However, the effective use of this potential is constrained by persistent challenges related to resource management and information integration. Scarcity of resources, fragmentation of care systems, frequent staff replacement, and pronounced regional inequalities continue to affect public health management [Cancian *et al*., 2023]. Despite the growing availability of public health data and national initiatives aimed at strengthening digital health infrastructure [SAÚDE, 2020], information systems remain fragmented across multiple national databases that operate in isolation, with heterogeneous data models, distinct update frequencies, and limited interoperability [Machado and Tavares, 2023]. As a result, public managers face significant difficulties in producing consolidated indicators, performing longitudinal analyses, and systematically monitoring the efficiency of resource allocation across municipalities and regions.

In recent years, governmental initiatives such as the InovaSUS Digital Laboratory, a Brazilian program aimed at fostering digital innovation, interoperability, and transparency in the public health sector, have reinforced institutional interest in digital transformation [Sousa, 2022]. These initiatives reflect a broader movement toward data-driven public management and the adoption of digital health strategies. Nevertheless, despite these efforts, the absence of integrated and governed data architectures remains a major barrier to transforming raw, heterogeneous data into actionable knowledge. Existing approaches often focus on isolated systems, manual data extraction, or purely analytical methods, without addressing the architectural challenges required to ensure automation, scalability, reproducibility, and long-term maintainability.

From an architectural perspective, effective public health data integration requires more than data availability. It demands solutions capable of supporting automated data ingestion, standardized transformations, traceability across processing stages, and reliable storage structures that enable consistent analytical use. Without such architectures, efficiency assessment methods and decision-support tools remain limited in scope, difficult to reproduce, and costly to maintain, particularly in federative contexts characterized by institu-

tional and territorial heterogeneity.

Existing work addresses interoperability and multi-source data integration through different solution paradigms, ranging from ETL/ELT pipelines [Ferré *et al*., 2020; Inácio *et al*., 2016] to semantic interoperability standards such as openEHR [Mello *et al*., 2021] and scalable data management paradigms such as Data Lakes and Data Lakehouses [Guntupalli, 2023]. Within this broader space, specific architectural approaches have also been proposed for healthcare data ecosystems, most notably HEALER [Manco *et al*., 2023]. While these solution paradigms and architectures have supported the integration of heterogeneous datasets in clinical and domain-specific applications [Machado and Tavares, 2023; Araújo *et al*., 2025], important challenges remain regarding governed integration, periodic data updates, reproducible analytical consumption, and extensibility across different analytical methods. These challenges are particularly relevant to public health management settings characterized by fragmented governmental information systems and limited interoperability across public databases [SAÚDE, 2020; Sousa, 2022]. Section 3 discusses these solution paradigms and architectural approaches in detail, including a direct comparison between HERO and HEALER.

In this context, this study investigates the following research question: **how can a governed and extensible data architecture support the integration of heterogeneous public health data sources and enable reproducible analytical workflows for performance evaluation in Primary Health Care?**

To answer this question, this study pursues the following objectives: **(O1)** to design a governed, ETL-based architecture that redefines the zone structure and processing flow proposed by HEALER to fit the constraints and periodicity of public health data publication; **(O2)** to implement a reference instantiation of this architecture capable of integrating heterogeneous Brazilian governmental health data sources into a unified and traceable Data Lakehouse; and **(O3)** to evaluate the applicability of the proposed architecture, both in terms of its architectural qualities and its usefulness as a foundation for analytical workflows, through a real-world case study using Data Envelopment Analysis (DEA) as a representative analytical instance. Section 4 details the research design adopted to pursue these objectives.

Motivated by this gap, this work proposes the Health Efficiency Resource Oriented (HERO) architecture, a modular and scalable framework that integrates and governs public health data from multiple governmental sources. The SmartAPSUS[1] project, an initiative focused on developing software solutions to support data monitoring and decision-making in public health management, originated the conception of the architecture [Nascimento *et al*., 2026]. Within this project, the team identified and investigated a set of public health indicators related to PHC performance using publicly available data sources. Initially, the team manually extracted these indicators from different governmental systems, integrated them, and explored them to enable preliminary analytical studies [Pereira *et al*., 2025a]. This early stage enabled the application of DEA to evaluate the operational efficiency of Brazilian municipalities in allocating resources for Primary Health Care, serving as a proof of concept for efficiency analysis based on integrated public health data [Pereira *et al*., 2025b].

This initial study highlighted the challenges associated with manual data integration, limited scalability, and the lack of automated data governance. These limitations motivated the design of the HERO architecture as a structured, automated, and reproducible solution for integrating heterogeneous public health data sources. Discussions with professionals from the public health and information technology domains involved in ongoing data initiatives within the Brazilian health system further informed the conception of the architecture and helped align the proposed pipeline with practical requirements related to data governance, interoperability, and performance monitoring. Although this study applies the architecture to Brazilian public health information systems, the proposed solution is not restricted to the Brazilian context. Instead, it provides a generic and reusable framework for public health data integration in environments characterized by decentralized information systems, heterogeneous data models, and periodic reporting structures.

This work contributes (i) the HERO architecture for integrating periodically published public health data through a governed ETL pipeline; (ii) a reference implementation that consolidates heterogeneous governmental data and exposes the resulting datasets through REST APIs; and (iii) an evaluation combining evidence on integration, maintainability, extensibility, and reproducibility with a DEA-based case study in PHC. The architectural contribution concerns data preparation and access, while the DEA results demonstrate the use of the integrated dataset in a specific analytical workflow.

HERO adopts modern interoperability standards, such as Fast Healthcare Interoperability Resources (FHIR), promoting integration with the National Health Data Network (Rede Nacional de Dados em Saúde - RNDS), Brazil's national health data exchange infrastructure, and compliance with the National Health Information and Informatics Policy (Política Nacional de Informação e Informática em Saúde - PNIIS). Beyond its technical contribution, the architecture demonstrates social relevance by supporting transparency, equity-oriented analysis, and more informed public health management.

The remainder of this paper is organized as follows. Section 2 presents the theoretical background related to public health data management, data architectures, and efficiency evaluation. Section 3 discusses related work on data integration architectures and DEA-based performance assessment, including a direct comparison between HERO and HEALER. Section 4 presents the research design and methods adopted in this study. Section 5 details the HERO architecture and its components. Section 6 describes the case study and experimental setup. Section 7 presents and discusses the results. Finally, Section 8 concludes the paper and outlines directions for future work.

2 Theoretical Basis

This section presents the theoretical foundations that support the proposed architecture and the analytical approach adopted

[1] https://smartapsus.ufape.edu.br/home.html

in this study. It discusses key concepts related to Primary Health Care, public health data integration, data engineering architectures, and efficiency evaluation, establishing the basis for the HERO architecture and its application.

2.1 Primary Health Care and Public Health Management

PHC services are responsible for a large share of healthcare delivery and are directly associated with population health outcomes, equity, and efficiency in resource use [Macinko *et al.*, 2003]. Due to its territorial organization and capillary reach, PHC is also a strategic level for monitoring public health performance and evaluating the effectiveness of health policies.

However, PHC management faces structural challenges, including limited resources, frequent staff replacement, and significant regional disparities. These challenges increase the need for reliable indicators capable of supporting planning, monitoring, and evaluation at the municipal and regional levels [Cancian *et al.*, 2023; Costa *et al.*, 2025]. In addition, limitations related to data integration and interoperability across health information systems further hinder effective management and decision-making processes [Silva, 2022]. In this context, data-driven approaches become essential for improving governance, identifying inefficiencies, and supporting evidence-based decision-making in public health systems.

2.2 Public Health Data Integration and Interoperability

Public health data management in the Brazilian SUS is historically characterized by fragmentation and information silos, which hinder interoperability and integrated analysis [Machado and Tavares, 2023]. Multiple national information systems collect complementary data on financing, service delivery, infrastructure, and coverage. However, these systems were developed independently, resulting in heterogeneous data models, distinct update frequencies, and limited semantic alignment.

The implementation of the RNDS represents an important institutional effort toward standardization and secure data exchange [Falcão *et al.*, 2024]. Nevertheless, technical and organizational challenges persist, including uneven adoption of interoperability standards and limited synchronization across systems. As a result, local managers and researchers still face difficulties in integrating data from multiple sources to generate consolidated indicators and perform longitudinal analyses.

In this scenario, architectural solutions focused on data integration and governance play a complementary role. Such architectures enable the ingestion, transformation, and consolidation of heterogeneous data, even in the absence of full interoperability at the national infrastructure level. By providing standardized processing pipelines and unified data representations, these solutions support local decision-making, performance monitoring, and analytical applications in public health.

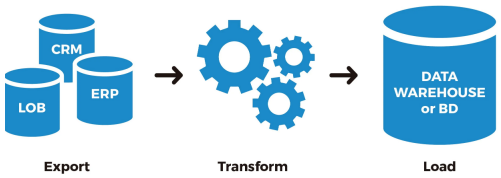


Figure 1. ETL Process; **Source:** `https://blog.bismart.com/en/what-do-we-do-etl`



Figure 2. Conceptual illustration of the Data Lakehouse model as a combination of Data Warehouse and Data Lake characteristics. **Source:** Authorship

2.3 Data Architectures for Health Systems: ETL, ELT, and the Data Lakehouse

Public health data integration involves both a processing strategy and a storage model for heterogeneous governmental datasets [SAÚDE, 2020].

Extract, Transform, and Load (ETL) validates and transforms data before loading them into the target repository, whereas Extract, Load, and Transform (ELT) loads data before transformation. Figure 1 illustrates ETL. Both strategies require validation and traceability to address differences in coding schemes, reporting periods, and data completeness [Phanikanth and Sudarsan, 2016].

Regarding storage, Data Warehouses organize structured analytical data under a *schema-on-write* approach [Nambiar and Mundra, 2022], while Data Lakes accommodate heterogeneous data under *schema-on-read*. The Data Lakehouse combines flexible storage with structured querying, governance, and transactional capabilities [Guntupalli, 2023], as illustrated in Figure 2.

HERO combines governed ETL processing with a Lakehouse-based storage layer. Section 3.1 explains the rationale for these choices, and Section 5 describes their implementation in the proposed architecture.

2.4 Efficiency Evaluation Using Data Envelopment Analysis

Operational efficiency in this study is assessed through Data Envelopment Analysis (DEA), a widely used non-parametric technique for evaluating relative efficiency among comparable units; its formal treatment, including the specific model adopted, is presented in Section 6.5 as part of the case study [Charnes *et al.*, 1978].

One of the main advantages of DEA is its ability to handle multiple inputs and outputs without requiring predefined functional forms. This characteristic makes it particularly suitable for evaluating complex systems such as PHC, where resource allocation and service delivery involve multiple dimensions. DEA models may be input-oriented or output-oriented, depending on the managerial control context, allowing different perspectives on efficiency assessment. When combined with integrated and governed data infrastructures, DEA becomes a powerful tool for supporting evidence-based resource allocation, performance benchmarking, and longitudinal monitoring in public health systems.

3 Related Work

The literature on health data management has evolved through at least three complementary lines of work: (i) scalable data processing architectures based on ETL/ELT, Data Lakes, and Data Lakehouses; (ii) semantic interoperability approaches based on standardized data models and exchange mechanisms; and (iii) analytical studies that consume integrated or manually prepared datasets for performance evaluation. Positioning HERO requires considering these three perspectives together, since the proposed contribution lies not only in data integration itself, but in combining governed integration, analytical readiness, and extensibility within a reusable framework.

Recent advances in data engineering have enabled the development of large-scale architectures for integrating heterogeneous healthcare datasets. Data Lake and Data Lakehouse approaches have been increasingly adopted to support the ingestion, storage, and processing of structured and unstructured health information originating from multiple systems. These architectures typically organize the data lifecycle into layered environments responsible for ingestion, processing, refinement, and analytical consumption, enabling scalable and flexible data management pipelines.

In the Brazilian health informatics context, other studies have also addressed data processing and integration workflows, even when not explicitly framed as reusable architectural proposals. For example, prior work proposed an open "situation room" platform for SUS administrative data, implementing an automated ETL process with extraction from public repositories, transformation into structured repositories, and support for analysis and visualization [Ferré *et al*., 2020]. Similarly, integrated epidemiological analysis platforms have combined ETL, preprocessing, indexing, and visualization components to enable the retrieval and georeferenced exploration of structured and semi-structured health data [Inácio *et al*., 2016]. In addition, studies based on openEHR, an open standard for the semantic modeling and exchange of electronic health records, have explored interoperability through standardized templates and archetypes, the main role of formal models in structuring and sharing healthcare information [Mello *et al*., 2021]. Together, these studies show that the literature already contains relevant efforts on health data workflows, interoperability, and analytical processing. However, they are often domain-specific, visualization-oriented, or centered on semantic modeling, without providing a governed and reusable architecture that integrates heterogeneous public health data sources, automates the full data lifecycle, and directly supports continuous analytical consumption across different evaluation methods.

3.1 Positioning HERO with Respect to Other Architectures

Among the architectural proposals in the healthcare domain, the HEALER architecture [Manco *et al*., 2023] represents an important conceptual reference for health data ecosystems and the direct architectural predecessor of HERO. HEALER adopts a zone-based data lake structure composed of ingestion, raw storage, processing, refinement, and consumption layers and follows the ELT paradigm, in which data are first preserved in their original format and later transformed according to analytical needs. This approach prioritizes flexibility and scalability and is particularly suitable for clinical environments characterized by large volumes of heterogeneous data and evolving analytical requirements.

However, the characteristics of public health information systems differ from those of clinical data environments. Public health datasets are typically structured, periodically published, and primarily used to generate standardized indicators for monitoring and policy evaluation. In such contexts, maintaining large raw data repositories and postponing transformations may increase operational complexity and storage costs without providing proportional analytical benefits. HERO builds upon the conceptual principles of HEALER while extending them to address the specific requirements of public health data integration: it adopts an ETL-based approach in which data are validated, normalized, and transformed before persistence, storing only semantically consistent datasets, and it introduces design elements suited to public-sector environments, including microservice-based ingestion components, batch-oriented orchestration aligned with periodic data publication, and extensible connectors for integrating new data sources.

Table 1 synthesizes the main architectural differences among HERO, HEALER, and a third related proposal, across the dimensions that motivate the proposed redesign.

Rather than merely transferring HEALER to a new application domain, HERO reconfigures these architectural decisions to better support periodic data publication, cross-source indicator derivation, and reproducible analytical consumption in public health management. This comparison, together with the related work discussed below, positions HERO as an architectural redesign rather than an incremental parameterization of HEALER.

A third related proposal further illustrates the diversity of architectural approaches to health data integration. Lima *et al*. [2021] introduce the concept of a *data pool*, a lightweight intermediate model that allows multiple institutions to organize heterogeneous clinical records, electronic health records, and Picture Archiving and Communication System (PACS) imaging data into interoperable pools, which can later be combined into a research-oriented Data Lake. Unlike HEALER and HERO, which define a single governed pipeline operated within one institution or governmental scope, the data pool model distributes interoperability across institutional boundaries, prioritizing multicentric research reproducibility over continuous, periodic indicator generation. This comparison reinforces that HERO's contribution lies specifically in providing centralized governance and standardized, periodic ETL processing tailored to public health management, a requirement that neither the raw-data-centered HEALER nor the multicentric data pool model was designed to address.

3.2 DEA-Based Efficiency Evaluation in Health Systems

Parallel to architectural research, a large body of literature applies DEA to evaluate efficiency in healthcare systems. DEA is used to measure the relative efficiency of comparable decision-making units by analyzing the relationship between resource inputs and service outputs. In the context of PHC, several studies have applied DEA to evaluate the

Table 1. Synthesis of architectural approaches for health data integration: HERO, HEALER, and the data pool model

Characteristic	HEALER	HERO	Data Pool
Processing strategy			
Governed ETL with pre-persistence validation	❍	✓	❍
Raw-data preservation via ELT (schema-on-read)	✓	❍	❍
No fixed ETL/ELT pipeline; lightweight pooling of pre-structured records	❍	❍	✓
Target domain			
Oriented to clinical/exploratory research use cases	✓	❍	✓
Designed for periodic, structured governmental publication	❍	✓	❍
Targets multicentric data sharing across independent institutions	❍	❍	✓
Storage emphasis			
Centralized, single-repository storage model	✓	✓	❍
Refined, analytically-ready storage prioritizing reuse	❍	✓	❍
Distributed, interoperable intermediate pools (no single repository)	❍	❍	✓
Orchestration			
Continuous, large-scale, evolving ingestion	✓	❍	❍
Batch-oriented, aligned with periodic publication cycles	❍	✓	❍
Ad hoc, agreement-based construction across institutions	❍	❍	✓
Consumption model			
Generic consumption layer for diverse exploratory use	✓	❍	❍
Explicit separation enabling reuse by multiple analytical methods	❍	✓	❍
Oriented to multicentric statistical/scientific research reuse	❍	❍	✓

performance of municipalities or health units using financial, structural, and service indicators [Varela *et al*., 2012; Cabral *et al*., 2019; Capeletti *et al*., 2024; Ferrera *et al*., 2014]. These studies demonstrate the methodological suitability of DEA for evaluating public health performance but often rely on static datasets and manual data collection processes.

Other works have explored DEA applications in different sectors of the public and private economy, including public safety [Castro and outros, 2024; Silva *et al*., 2024], digital banking [Loureiro *et al*., 2024], and energy systems [Marques and outros, 2024]. In these contexts, methodological extensions such as Dynamic DEA and Network DEA have been proposed to capture temporal dynamics and multi-stage production processes. Although these approaches enhance analytical modeling, they are often developed independently of scalable data infrastructures capable of supporting automated and continuous data integration.

A preliminary study reported in [Pereira *et al*., 2025b] demonstrated the feasibility of applying DEA to evaluate the efficiency of Brazilian municipalities in Primary Health Care and introduced a probabilistic perspective through Markov Chains. However, the study relied on manual data extraction and limited sampling, which constrained reproducibility and the ability to update efficiency estimates over time. The present work advances this line of research by implementing an automated, batch-oriented pipeline within a governed lakehouse architecture, enabling systematic data integration and periodic estimation of efficiency indicators.

Table 2 summarizes representative studies identified in the literature, highlighting their domains, analytical methods, and limitations with respect to data integration and automation.

4 Research Design and Methods

This study follows an applied research approach, grounded in a quantitative perspective [Wazlawick, 2014], whose outcome is a reusable software architecture rather than a purely theoretical model. More specifically, the study is framed within the Design Science Research (DSR) paradigm [Hevner *et al*., 2004], which centers the production of scientific knowledge on the construction and rigorous evaluation of innovative artifacts intended to solve identified problems, as opposed to the behavioral-science paradigm, which seeks primarily to explain or predict organizational phenomena. This framing is consistent with broader reflections on the epistemological and methodological diversity of Information Systems research published in this journal [Araujo *et al*., 2017]. Consistent with this classification, the research process combines the construction of an architectural artifact (HERO) with its evaluation through a real-world instantiation, following the objectives O1–O3 stated in Section 1.

4.1 Research Stages

The research process comprises four main stages, executed sequentially but with iterative refinement between the first two:

1. **Architectural design.** Building on the HEALER architecture [Manco *et al*., 2023] and on the limitations identified in a prior, manually integrated study [Pereira *et al*., 2025b], the HERO architecture was designed to redefine the processing strategy (ETL instead of ELT) and the zone structure required to support periodic public health data publication, as detailed in Section 5.
2. **Reference implementation.** Two complementary applications, *health-data-etl* and *health-data-client*, were developed to instantiate the architecture, respectively

Table 2. Summary of related works on efficiency evaluation and health data architectures

Reference	Domain	Method	Main Limitation
[Varela *et al*., 2012]	PHC municipalities	DEA	Static dataset, manual collection
[Cabral *et al*., 2019]	PHC services	DEA	Limited sample and temporal scope
[Capeletti *et al*., 2024]	Municipal health systems	DEA	Lack of automated data integration
[Ferrera *et al*., 2014]	PHC efficiency	DEA	Study-specific data preparation
[Castro and outros, 2024]	Public safety	DEA	Sector-specific dataset
[Silva *et al*., 2024]	Public safety	Dynamic DEA	Focus on modeling, not infrastructure
[Loureiro *et al*., 2024]	Digital banking	Network DEA	Domain-specific analysis
[Marques and outros, 2024]	Energy sector	Network DEA	Limited integration architecture
[Pereira *et al*., 2025b]	Brazilian PHC	DEA + Markov Chains	Manual data extraction
[Manco *et al*., 2023]	Health data ecosystems	Data lake architecture	Designed for raw-data preservation and exploratory clinical workflows

covering data ingestion/storage/exposure and analytical consumption, as detailed in Section 6.

3. **Data integration.** The reference implementation was executed against three heterogeneous Brazilian governmental data sources, SIOPS, SISAB, and e-Gestor, covering municipalities in the state of Pernambuco between 2021 and 2025.
4. **Analytical evaluation.** The integrated dataset was used to compute public health indicators and to apply DEA as a representative analytical instance, allowing the joint evaluation of (i) the architecture's ability to support a real analytical workflow and (ii) the substantive efficiency results obtained through that workflow.

4.2 Materials

This study relies on two types of materials, the public data sources integrated by the architecture and the software artifacts developed and used to operationalize it:

- **Data sources**: three Brazilian national public health information systems, SIOPS (financial data, via a documented REST API), SISAB (service-delivery data, via web scraping and CSV downloads, since no public API is available), and e-Gestor (structural/coverage data, via reverse-engineered API endpoints). These systems were selected because, together, they cover the financial, structural, and service-delivery dimensions required for the efficiency evaluation described in Section 6.
- **Software artifacts**: the two reference applications (*health-data-etl* and *health-data-client*), a PostgreSQL relational database used as the Refined Zone of the Data Lakehouse, and the R `Benchmarking` package used to compute DEA scores.

4.3 Architectural Evaluation Dimensions

To evaluate HERO as an architecture, and not only through the analytical results it enables, this study adopts four complementary evaluation dimensions, used later in Section 7 to structure the architectural evaluation:

- **Integration**: the ability to ingest, standardize, and consolidate heterogeneous public health data from multiple governmental systems, with distinct access mechanisms, into a unified and governed pipeline.
- **Maintainability**: the modular organization of the architecture into well-defined components and layers, facilitating isolated changes, evolution of connectors, and adaptation of processing rules with limited impact on the remaining pipeline.
- **Extensibility**: the capacity to incorporate new data sources, derived indicators, and analytical consumers without redesigning the architectural core.
- **Reproducibility**: the support for repeatable analytical workflows through governed ETL execution, traceability mechanisms, persistent refined storage, and standardized access to processed data.

Together, these dimensions operationalize how HERO extends HEALER beyond a change of application domain, and they provide the criteria against which the architectural evaluation in Section 7 is organized.

5 HERO Architecture

HERO organizes public health data integration into components covering acquisition, processing, storage, access, and analytical consumption. The architectural choices underlying this organization and their differences from HEALER are discussed in Section 3.1 and summarized in Table 1. This section details how these choices are operationalized.

HERO is structured into six main components: Data Sources, Ingestion Component, Processing Component, Storage Component, Interaction Component, and Data Consumers. These components organize the entire data lifecycle, from acquisition to analytical consumption, following a zone-based architectural model inspired by HEALER [Manco *et al*., 2023], as illustrated in Figure 3.

The **Data Sources Component** represents the heterogeneous external environments from which data originate, encompassing structured and unstructured repositories such as APIs, public databases, web portals, and systems that require web scraping or direct SQL queries. This layer is intentionally decoupled from the internal architecture, enabling flexibility and extensibility, while serving as the initial entry point for raw data into the pipeline.

The **Ingestion Component** is responsible for acquiring data and forwarding it to the internal pipeline through the Transient Landing Zone, where no prior transformations are

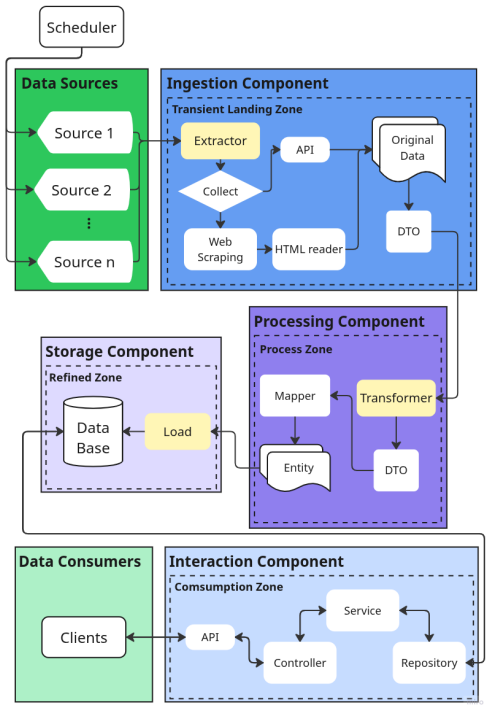


Figure 3. HERO (Health Efficiency Resource Oriented): Proposed Architecture; **Source:** Authorship

applied. Data extraction is performed via APIs or web scraping mechanisms and encapsulated in intermediate Data Transfer Objects (DTOs), which are lightweight data structures used to transfer data between system components without embedding business logic. This approach ensures decoupling between source structures and the internal architectural model, thereby promoting modularity and facilitating maintenance.

The **Processing Component** executes the transformation, normalization, and enrichment of data through the Process Zone. Domain-specific business rules are applied to ensure semantic consistency and analytical reliability, including validation of mandatory fields, standardization of formats, cross-source verification, and enrichment through derived attributes. The processed data are then mapped into persistent Entity objects via a Mapper, preparing them for structured storage and analytical use.

The **Storage Component** consolidates the core of the Data Lakehouse, represented by the Refined Zone, where data are stored in a relational database with well-defined primary and foreign keys, enabling cross-source integration, traceability, version control, and complex analytical queries. This structure supports data integrity, auditability, and reliability for decision-making and efficiency assessment processes.

The **Interaction Component** exposes refined data through a microservices-based API layer, structured into Repository, Service, and Controller modules, ensuring scalable, modular, and efficient access to the processed information. This layer operationalizes the Consumption Zone, acting as the formal interface between the architecture and its users.

Finally, the **Data Consumers** component comprises analytical scripts, visualization platforms, and decision-support systems that access refined data through the Interaction Component. This separation allows analytical consumers to evolve independently of data acquisition and transformation. In the

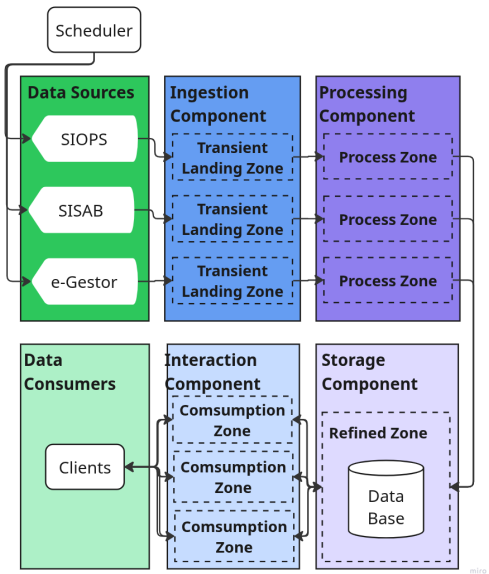


Figure 4. Running Case; **Source:** Authorship

evaluated implementation, the *health-data-client* application uses DEA to assess municipal PHC efficiency, providing a concrete example of analytical consumption.

5.1 Running Example of Data Processing

To illustrate how the HERO architecture operates in practice, Figure 4 presents a simplified running example of the data processing workflow. The example considers three public health datasets collected for a given municipality from the Brazilian governmental systems SIOPS, SISAB, and e-Gestor.

Suppose that the SIOPS system reports that a municipality invested BRL 12 million in Primary Health Care during a given period. In parallel, SISAB records that the municipality performed approximately 95,000 health visits, while e-Gestor indicates the presence of 48 PHC teams responsible for providing services to the local population.

Within the Data Sources component, these datasets exist independently in their respective governmental systems. Data collection is triggered by a scheduler that periodically retrieves updated information from each source. In the Ingestion Component, the datasets are acquired through different mechanisms depending on the characteristics of each system: SIOPS data are extracted through the documented REST API, SISAB data are collected from CSV files generated by the platform, and e-Gestor information is retrieved through API requests. The retrieved datasets are temporarily stored in the Transient Landing Zone, which acts as a staging area where raw data are buffered before processing.

The data then moves to the Processing Component, where a series of transformations are applied within the Process Zone. This stage includes validation procedures to verify data integrity and completeness, normalization steps to standardize formats across heterogeneous sources, and the derivation of analytical indicators required for performance evaluation. Examples of derived indicators include APS budget per capita, productivity measures based on health visits, and team density indicators based on the number of primary care teams.

After processing, the structured datasets are stored in the

Refined Zone within the Storage Component. In the case study implementation, this zone is materialized as a PostgreSQL relational database where entities such as municipalities, health visits, coverage indicators, and financial variables are organized using well-defined relationships. This refined storage layer ensures data consistency and enables cross-source analytical queries.

The processed data are then exposed through the Interaction Component, which provides multiple consumption services. In the running example, one of these services performs efficiency analysis based on Data Envelopment Analysis, generating efficiency indicators for municipalities. Additional services include dashboards for visualization and an audit API for structured access to the processed datasets.

Finally, these services are consumed by external applications represented in the Data Consumers component. Examples include Python analytical scripts used for research experiments and external client applications that access the data through the exposed APIs. Although DEA is used in this study as a representative analytical application, the architecture itself remains flexible and can support other analytical methods or decision-support tools.

6 Case Study

This section describes the experimental environment used to operationalize and validate the HERO architecture. The experiment used real-world data from Brazilian public health information systems and executed through an end-to-end pipeline including automated data collection, integration, storage, and analytical processing.

The case study focuses on municipalities in the state of Pernambuco, located in northeastern Brazil. Pernambuco comprises 185 municipalities and a population of approximately 9.6 million inhabitants, making it an appropriate context for evaluating large-scale public health data integration and efficiency analysis.

To operationalize the architecture, this work implements two complementary applications: *health-data-etl* and *health-data-client*. Together, these applications compose the analytical pipeline used in the case study, instantiating the research stages described in Section 4.

6.1 Applications and Analytical Pipeline

To execute the experiment, this work developed two applications for the proposed architecture: one responsible for integrating, storing, and exposing public data, and another dedicated to the analytical processing and interpretation of these data.

The *health-data-etl* application[2] is responsible for the ingestion, transformation, storage, and structured exposure of public health data through REST APIs. This component implements an ETL-based pipeline designed to integrate heterogeneous data from three Brazilian governmental systems: SIOPS, SISAB, and e-Gestor. The application follows a modular architecture that separates responsibilities across different layers. A common layer centralizes shared components such as configuration files, data models, DTOs, and utility classes used throughout the application. An API layer exposes the processed data through REST endpoints, allowing external services and analytical applications to access the integrated datasets in a structured manner. The ETL layer implements the extraction, transformation, and loading processes, adapting the data acquisition strategy according to the characteristics of each source system.

The transformation rules applied within the ETL layer follow a consistent validation and standardization protocol across the three source systems, implemented as reusable validation classes within the common layer. Mandatory fields, such as municipality identifier, reference year, and bimester, are validated for non-null values and type consistency before being mapped to persistent entities. Municipality identifiers are standardized against the official IBGE municipality code table, ensuring consistent joins across SIOPS, SISAB, and e-Gestor records despite their distinct native identifier formats. Missing values in numeric fields are flagged and excluded from indicator computation rather than imputed, avoiding the introduction of bias into the efficiency analysis. Duplicate records, arising primarily from repeated API requests or overlapping CSV exports, are identified and removed based on a composite key combining municipality, year, and bimester. The pipeline currently targets a fixed multi-year window (2021–2025) and a fixed set of reporting periods, defined as constants at the application level rather than through external configuration; extending this window to additional years or periods requires updating these definitions in code, a maintainability consideration discussed further in Section 7.

To support monitoring and reproducibility of the pipeline execution, the implementation also records structured log events for the main ETL and analytical stages, including extraction start and end, processed temporal scope (e.g., year and bimester), transformation steps, file persistence, and DEA computation milestones. These logs provide process-level traceability and support the verification of the execution flow across the different stages of the architecture.

All processed data are stored in a PostgreSQL relational database designed to support cross-referenced analysis across financial, structural, and service delivery indicators. In addition to providing access to raw and processed data, the API also exposes endpoints responsible for computing derived indicators required for the DEA efficiency analysis.

The *health-data-client* application[3] corresponds to the analytical layer of the HERO architecture and is responsible for performing efficiency analysis using the integrated datasets. Implemented in Python, this application consumes the REST endpoints exposed by the *health-data-etl* service and executes automated analytical workflows. The analytical process begins with the automated collection of indicators through API requests, retrieving financial, structural, and service delivery data for each municipality. These data are then preprocessed and consolidated by municipality and year in order to construct the analytical datasets used in the efficiency evaluation. Based on these datasets, the application generates several analytical outputs, including efficiency rankings, correlation analyses, spatial visualizations, and comparative

[2] Source code available at: `https://github.com/izabelnascimento/health-data-etl`

[3] Source code available at: `https://github.com/izabelnascimento/health-data-client`

graphs designed to support the interpretation of the results. The modular structure of the analytical scripts allows the experiments to be easily replicated across multiple years and enables adjustments in analytical parameters or indicator configurations.

6.2 Data Sources and Integration

The case study integrates data from three Brazilian national public health information systems that provide complementary information on financial resources, service delivery, and structural capacity in Primary Health Care. These systems play a key role in monitoring and managing health policies at the municipal level, and this study selected them because together they capture different dimensions of the public health system.

- **SIOPS (Public Health Budget Information System)** provides financial information on municipal health expenditures. This system records detailed budgetary data related to public health investments reported by municipalities, including operational and capital expenses. SIOPS offers a documented REST API, which enabled the direct extraction of financial indicators used in the analysis. The availability of this API facilitated the automation of data collection and ensured consistent retrieval of financial information across multiple municipalities and years.
- **SISAB (Primary Health Care Information System)** provides records of health services delivered within the Primary Health Care network. The system consolidates data on healthcare visits and other service activities performed by primary health teams across the country. Since SISAB does not offer a public API, data extraction required the implementation of web scraping techniques combined with automated downloads of CSV files generated by the platform. These files were subsequently processed and integrated into the ETL pipeline to derive indicators related to healthcare productivity.
- **e-Gestor** provides information on PHC coverage and the distribution of health teams across municipalities. This platform supports the monitoring of the Family Health Strategy and other primary care programs implemented by the Brazilian SUS. Although the system exposes REST API endpoints, they are not formally documented. Therefore, the data extraction process required reverse engineering techniques through network traffic inspection in order to identify the request patterns used by the platform.

Despite the heterogeneity of access mechanisms across these systems, the HERO architecture integrates these datasets through a unified ETL pipeline. The architecture then structures the integrated dataset in a relational database, enabling cross-referenced analysis between financial inputs, structural capacity indicators, and service delivery outputs used in the efficiency evaluation.

6.3 Data Storage and Data Model

The integrated dataset produced by the HERO architecture is stored in a PostgreSQL. The database schema consolidates information collected from the SIOPS, SISAB, and e-Gestor systems and organizes it in a structure that enables efficient analytical queries and traceability of the data processing pipeline.

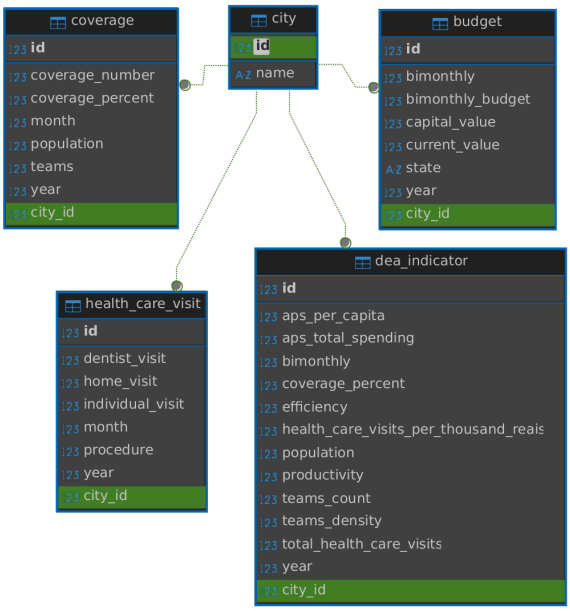


Figure 5. Conceptual Entity–Relationship model of the integrated dataset; **Source:** Generated by Dbeaver

Figure 5 presents the conceptual Entity–Relationship (ER) model used in the storage layer. The schema is centered on the `city` entity, which represents municipalities and serves as the main reference for linking financial, operational, and coverage indicators. Additional entities store specific information extracted from the different public systems, including budgetary data, healthcare service records, and Primary Health Care coverage indicators.

This relational model enables the integration of heterogeneous datasets and supports the derivation of analytical indicators required for efficiency evaluation. By structuring the data in this way, the architecture facilitates the computation of DEA inputs and outputs while preserving data provenance and enabling reproducibility of the analytical process.

6.4 Indicator Derivation

The indicators used in the efficiency analysis were derived from the integrated dataset produced by the HERO architecture. Raw data collected from SIOPS, SISAB, and e-Gestor were transformed into standardized indicators representing financial inputs, structural capacity, and service outputs.

The study computed the following indicators:

$$\text{APS Budget per Capita} = \frac{\text{Total APS Budget}}{\text{Municipal Population}} \quad (1)$$

$$\text{Team Density} = \frac{\text{Number of PHC Teams}}{\text{Population}} \times 1000 \quad (2)$$

$$\text{Productivity} = \frac{\text{Total Health Visits}}{\text{Number of PHC Teams}} \quad (3)$$

$$\text{Health Visits per Thousand BRL} = \frac{\text{Total Health Visits}}{\text{APS Budget}} \times 1000 \quad (4)$$

These indicators capture complementary dimensions of PHC performance. Financial indicators represent the level of public investment in health services, structural indicators represent the availability of operational resources, and service indicators represent the volume of care delivered to the population. The processing pipeline automatically computed and stored all indicators in the relational database, ensuring traceability and reproducibility of the analytical process.

6.5 DEA Model and Efficiency Computation

Municipal efficiency is evaluated using Data Envelopment Analysis, a non-parametric technique widely used to measure the relative efficiency of comparable entities known as DMUs. In this study, each municipality was considered a DMU representing a local Primary Health Care system responsible for transforming financial and structural resources into health service outputs [Charnes *et al.*, 1978].

DEA evaluates efficiency by comparing each DMU with a frontier constructed from the best observed performance among all units in the dataset. Unlike econometric approaches, DEA does not assume a predefined functional relationship between inputs and outputs. Instead, it uses linear programming to construct an efficiency frontier and measure the distance of each unit from this frontier. As a result, the efficiency score obtained for each municipality reflects its performance relative to the best practices observed among the municipalities included in the analysis [Charnes *et al.*, 1978].

In formal terms, the efficiency of a municipality j can be expressed as the ratio between a weighted combination of outputs and a weighted combination of inputs:

$$\theta_j = \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \tag{5}$$

where θ_j represents the efficiency score, y_{rj} represents the quantity of output r produced by municipality j, and x_{ij} represents the quantity of input i consumed by municipality j. The weights u_r and v_i are determined through linear programming in order to maximize the efficiency score for each unit while respecting feasibility constraints for all municipalities in the dataset.

DEA models can be formulated with different orientations depending on the managerial context. Input-oriented models seek to minimize resource consumption while maintaining output levels, whereas output-oriented models aim to maximize outputs given fixed inputs. In the context of public health management, municipal administrators typically have greater control over the allocation of financial and structural resources than over the demand for health services. For this reason, an input-oriented model was adopted in this study.

The efficiency computation follows the classical Charnes, Cooper, and Rhodes (CCR) model with constant returns to scale proposed by [Charnes *et al.*, 1978]. Under this formulation, efficiency scores range between 0 and 1. A value equal to 1 indicates that the municipality operates on the efficiency frontier, while values below 1 indicate potential inefficiencies relative to the best observed practices.

The implemented DEA computation used the `Benchmarking` package in the R statistical environment, and it was integrated into the analytical pipeline of the HERO architecture. The resulting efficiency scores were subsequently stored in the analytical dataset and used to generate the visualizations and comparative analyses presented in the following section.

6.6 Threats to Validity

Some limitations should be considered when interpreting the results of this study. First, the efficiency analysis depends on the selection of input and output indicators used in the DEA model. Although the selected variables capture relevant dimensions of PHC performance, alternative combinations of indicators could lead to different efficiency estimates.

Second, the reliability of the results depends on the quality of the underlying public datasets. Potential inconsistencies, publication delays, missing values, or errors in the original systems may affect the accuracy of the integrated dataset. In addition, the analysis focuses on the first semester of each year, which ensures consistency across data sources but may not fully capture seasonal variations in health service demand and delivery.

Third, the study relies on aggregated municipal indicators. While this level of aggregation is appropriate for comparative analysis across municipalities, it may hide important intra-municipal differences in service performance. Moreover, contextual factors such as socioeconomic conditions, demographic characteristics, geographic accessibility, and epidemiological profiles were not explicitly modeled, although they may influence efficiency outcomes.

Another limitation concerns the high correlation observed between some structural indicators, particularly team density and coverage. This relationship reflects the design of the Brazilian PHC model, but it may also introduce redundancy into the analytical model. Finally, the indicators used in this study primarily capture quantitative aspects of service provision and do not include qualitative dimensions such as service quality, patient satisfaction, continuity of care, or clinical outcomes.

Despite these threats to validity, the case study demonstrates the feasibility of integrating heterogeneous governmental data sources and applying analytical methods to support large-scale monitoring of Primary Health Care performance. In addition, the HERO architecture shows strong potential to support reproducible, scalable, and data-driven evaluations in public health management.

7 Results and Discussion

This section evaluates HERO through two complementary types of evidence. The architectural evaluation in Section 7.6 examines integration, maintainability, extensibility, and reproducibility using the reference implementation and its execution artifacts. The DEA case study examines the analytical use of the integrated dataset for PHC efficiency assessment in Pernambuco between 2021 and 2025.

After describing the dataset, the section presents temporal, spatial, input–output, and correlation analyses, followed by the architectural evaluation. The closing synthesis distinguishes what these results demonstrate about HERO from the municipal efficiency patterns identified through DEA.

7.1 Dataset Description

The proposed architecture and the efficiency analysis use an integrated dataset derived from Brazilian public health information systems, as described in Section 6.3. The automated data ingestion pipeline implemented in the HERO architecture builds this dataset by integrating heterogeneous governmental data sources and consolidating them in a relational analytical database.

The dataset covers all municipalities in the state of Pernambuco, totaling 185 DMUs included in the efficiency analysis. This study collected data for the period from 2021 to 2025, focusing on the first semester of each year to ensure consistent availability across all data sources.

Table 3 summarizes the main characteristics of the datasets used in the validation process, including their source systems, the type of information they provide, their temporal coverage, and their analytical role in the efficiency model.

The integration of these datasets allowed the construction of a consistent analytical dataset combining financial resources, structural capacity, and healthcare service delivery indicators. This integrated dataset constitutes the empirical basis for the DEA efficiency evaluation and the analyses presented in the following sections.

7.2 Temporal Evolution of Efficiency

Figure 6 illustrates the trajectories of the ten most efficient (*Top 10*) and ten least efficient (*Bottom 10*) municipalities during the first semester of selected years. The visualization highlights patterns of stability, improvement, and persistence in technical efficiency scores over time.

At the upper end of the distribution, several municipalities consistently maintain efficiency scores above 0.8, indicating stable managerial capacity and effective conversion of resources into service delivery. Municipalities such as *Buíque*, *Flores*, and *Serra Talhada* frequently appear among the most efficient across bimesters, suggesting relatively balanced relationships between financial resources, team structure, and service outputs. These municipalities appear to operate closer to the efficiency frontier, meaning they are able to transform available inputs into health services more effectively than others.

By contrast, municipalities in the *Bottom 10* group tend to remain clustered below efficiency scores of 0.2. Municipalities such as *Santa Filomena*, *Abreu e Lima*, and *Cortês* repeatedly appear among the lowest performers, indicating persistent difficulties in converting available inputs into effective health service outputs. These patterns may reflect operational inefficiencies, structural limitations, or managerial challenges that limit the ability of local systems to fully utilize available resources.

Although some municipalities experience trajectory changes over time, the general structure of the distribution remains relatively stable: a small group of municipalities consistently operates near the efficiency frontier, while a larger group remains substantially below optimal performance levels. This persistence suggests that efficiency differences are not merely random fluctuations, but reflect organizational capacity and contextual conditions that tend to remain stable over time.

From a practical perspective, these results provide valuable insights for public health management. Municipalities that consistently demonstrate high efficiency can serve as important benchmarks for the health system. By examining how these municipalities organize their services, allocate resources, and manage primary health care teams, policymakers may identify strategies that could potentially be replicated or adapted in other regions. In parallel, municipalities that consistently appear among the lowest performers deserve closer analytical attention, as persistent low efficiency may indicate structural problems, underutilization of resources, or operational challenges within local health systems.

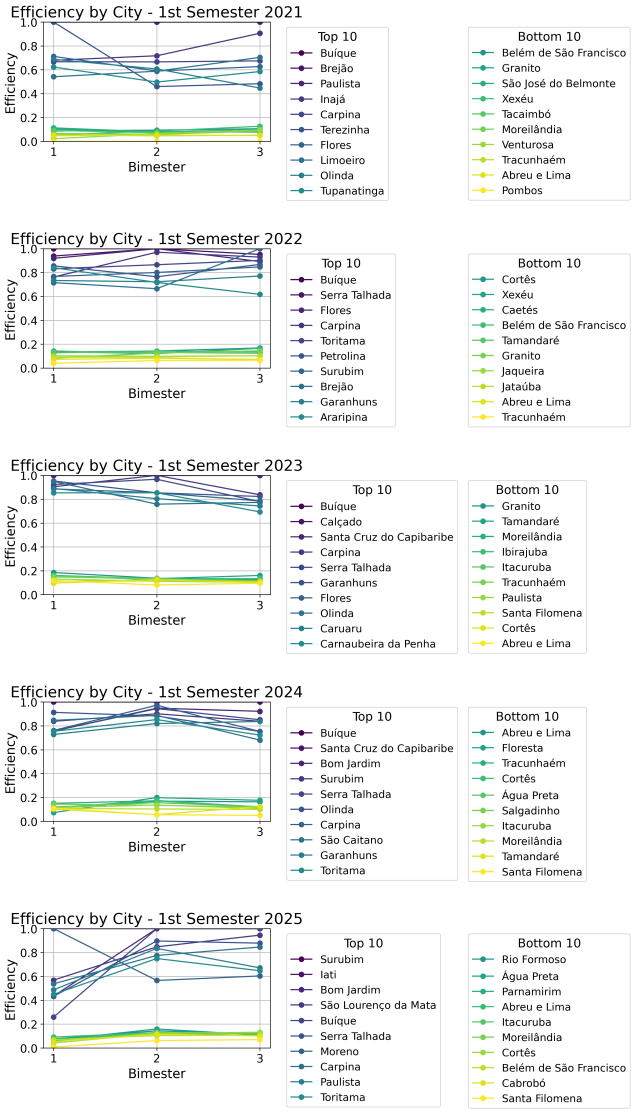


Figure 6. Efficiency trajectories of Top 10 and Bottom 10 municipalities across bimesters; **Source:** Authorship

7.3 Spatial Distribution of Efficiency in Pernambuco

Spatial visualization provides an important perspective for understanding how efficiency patterns are distributed across

Table 3. Summary of datasets used in the efficiency analysis

System	Type of Data	Indicator Derived	Period	Role in DEA Model
SIOPS	City health budget	APS budget per capita	2021–2025	Financial input
SISAB	Primary health care visits	Productivity; health visits per thousand BRL	2021–2025	Service outputs
e-Gestor	PHC coverage and team distribution	Team density	2021–2025	Structural input

the territory. Figure 7 presents the spatial distribution of municipal efficiency scores in Pernambuco from 2021 to 2025.

The maps reveal a persistent pattern of territorial heterogeneity. In 2021, municipalities across the state mostly show low efficiency scores, with a mean value of approximately 0.26 and only a limited number of municipalities approaching the efficiency frontier. Several clusters of low efficiency emerge, particularly in the interior of the state, where municipalities show lower productivity levels relative to their available structure.

From 2022 to 2023, the spatial distribution shows a gradual improvement, with mean efficiency increasing to approximately 0.38 and 0.41 respectively. During this period, more municipalities move into intermediate efficiency categories, suggesting improvements in resource utilization or service delivery capacity. In 2024, the spatial pattern remains relatively similar, with efficiency levels stabilizing around a mean of approximately 0.40. High-efficiency municipalities remain spatially scattered rather than concentrated in a single region, suggesting that efficiency is more closely related to local management practices than to geographic location alone.

However, in 2025, the mean efficiency declines again to approximately 0.29, indicating a possible deterioration in performance or changes in the underlying indicators during that period. Despite these fluctuations, the maps consistently reveal significant spatial inequalities, with municipalities exhibiting markedly different levels of performance even within the same macro-regions.

Beyond their analytical value, these spatial visualizations are particularly useful for public health managers and policymakers. By presenting efficiency patterns in a geographic format, the maps allow decision-makers to quickly identify regions where efficiency levels are consistently lower and where targeted interventions may be required. At the same time, municipalities that consistently appear in higher efficiency categories can be more easily recognized, enabling managers to investigate successful local practices that may serve as benchmarks for other regions.

In this context, spatial efficiency maps function not only as analytical tools but also as practical instruments for planning and monitoring public policies. They help managers prioritize actions, allocate technical support, and design territorially oriented strategies aimed at improving the overall performance of Primary Health Care services across the state.

7.4 Relationship Between Inputs, Outputs and Efficiency

Figures 8 and 9 present scatter plots relating per capita APS expenditure (input) and productivity measured through health service outputs for the first semesters of 2021 and 2025. These two years represent the temporal boundaries of the dataset and allow a clearer observation of how the relationship between

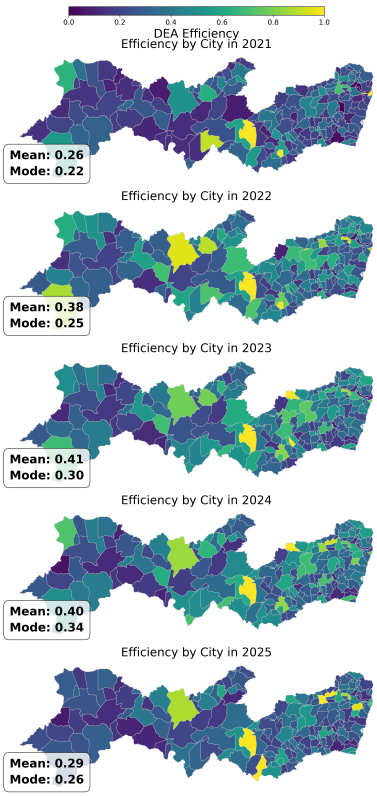


Figure 7. Spatial distribution of technical efficiency — Pernambuco, 2021 to 2025; **Source:** Authorship

financial resources and service productivity evolved over the analyzed period. Each point represents a municipality and is colored according to its DEA efficiency score.

When comparing the two years, the visualizations reveal a considerable expansion in the range of per capita expenditures. In 2021, most municipalities are concentrated within a relatively narrow spending interval, whereas in 2025, several municipalities appear with substantially higher per capita investments. Despite this increase in spending variability, the relationship between inputs and outputs remains heterogeneous, indicating that higher financial investment does not necessarily translate into higher productivity or efficiency.

For example, municipalities such as *Brejão* appear in both years, maintaining high productivity levels with moderate financial inputs, positioning themselves close to the DEA efficiency frontier. These municipalities represent examples of efficient resource utilization, where the available budget is effectively converted into service delivery.

In contrast, some municipalities with substantially higher per capita expenditures in 2025, such as *Santa Filomena* and *Orobó*, remain located in regions of the plot associated with lower productivity levels. This pattern suggests that increases in financial investment alone do not guarantee improvements in service output, highlighting potential inefficiencies in the

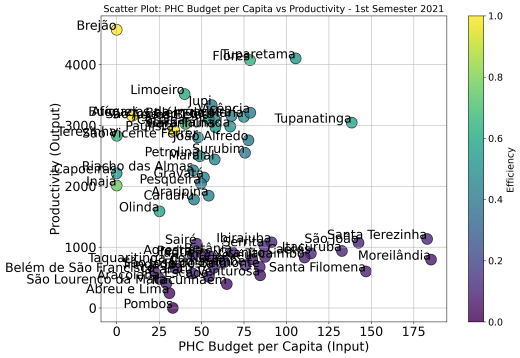


Figure 8. Dispersion: per capita APS budget vs. productivity - 1st semester of 2021; **Source:** Authorship

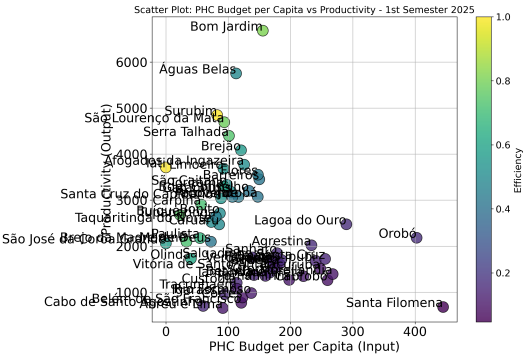


Figure 9. Dispersion: per capita APS budget vs. productivity - 1st semester of 2025; **Source:** Authorship

allocation or management of resources.

Another important case is *Bom Jardim*, which in 2025 appears with very high productivity levels relative to its budget, positioning itself among the municipalities closest to the efficiency frontier. Such cases illustrate how improvements in service organization or operational practices may significantly impact performance even without proportionally higher investments.

From a manager perspective, scatter plots such as these provide a powerful analytical tool for understanding different performance profiles among municipalities. Municipalities located in the upper-left region of the plot represent cases where moderate spending is associated with high productivity, indicating efficient resource utilization and potential benchmark cases. Conversely, municipalities located in the lower-right region combine high expenditure with lower productivity, suggesting the presence of operational bottlenecks or inefficiencies that may require further investigation.

When interpreted together with the spatial maps and temporal efficiency trajectories presented in previous sections, this type of visualization offers a comprehensive diagnostic perspective for policymakers. It allows managers not only to identify where efficiency differences occur geographically, but also to understand how financial inputs are being transformed into service outputs over time. Such insights can support more informed decision-making, helping policymakers identify successful management practices, investigate persistent performance gaps, and design targeted interventions aimed at improving the overall efficiency of PHC services across municipalities.

7.5 Correlation Analysis Between Variables

Figure 10 presents the correlation matrix between the indicators used in the DEA model for the first semester of 2025. The analysis includes financial inputs (per capita APS budget), structural inputs (team density and coverage), service outputs (health visits per thousand reais and productivity), and the resulting efficiency scores.

One of the most notable patterns in the matrix is the strong positive correlation between productivity and efficiency ($r = 0.72$). This relationship indicates that municipalities with higher service productivity tend to achieve higher efficiency levels within the DEA model. In practical terms, this suggests that the ability to convert available resources into health service outputs plays a central role in determining municipal performance.

In contrast, structural indicators such as team density and coverage show moderate negative correlations with efficiency (both $r = -0.51$). This result suggests that simply expanding the structural capacity of the system, for example, by increasing the number of primary health care teams or coverage levels, does not necessarily translate into more efficient service delivery. In some cases, higher structural capacity may even indicate underutilization of resources if the available teams are not generating proportional service outputs.

Another important pattern is the near-perfect correlation between team density and coverage ($r = 0.99$). This strong relationship reflects the structural characteristics of the Brazilian PHC model, where the expansion of Family Health Strategy teams directly drives increases in population coverage. While this relationship is expected from a policy perspective, it also indicates that these two indicators capture closely related dimensions of the system.

The financial input variable, represented by per capita APS budget, shows a moderate negative correlation with efficiency ($r = -0.37$). This finding reinforces an important insight also observed in the dispersion analysis: higher financial investment alone does not guarantee better performance. Instead, efficiency appears to depend more strongly on how effectively municipalities organize their services and transform available resources into health outputs.

From a managerial perspective, these correlations provide useful signals for public health decision-makers. The results suggest that improving productivity and service organization may have a greater impact on efficiency than simply expanding financial or structural resources. Therefore, policies aimed at strengthening management practices, optimizing team workflows, and improving the utilization of existing resources may play a key role in enhancing the overall efficiency of Primary Health Care services.

7.6 Architectural Evaluation

This subsection evaluates HERO against the four dimensions defined in Section 4.3. Integration is examined through source access mechanisms, schema integration, and record completeness; maintainability and extensibility through code organization and documented changes; and reproducibility through execution artifacts and setup documentation. These observations, obtained from the *health-data-etl* repository and its

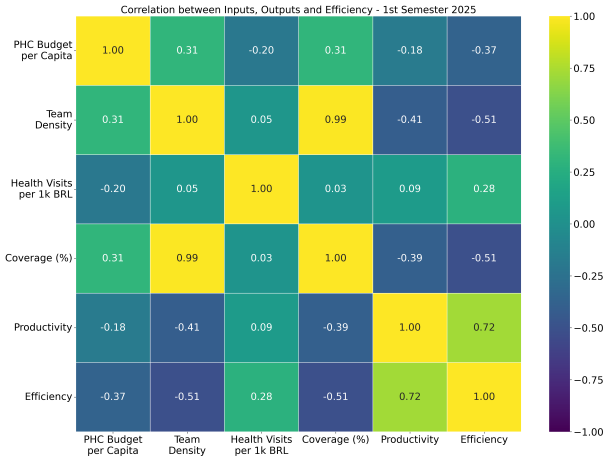


Figure 10. Correlation between inputs, outputs and efficiency indicators; **Source:** Authorship

execution and commit history, assess the reference implementation separately from the municipal efficiency scores.

Regarding **Integration**, the three source systems are accessed through distinct, confirmed mechanisms, documented REST for SIOPS, web scraping and CSV parsing for SISAB, and undocumented REST discovered through reverse engineering for e-Gestor, and are unified in a five-entity relational schema through a shared `City` foreign key. A full SISAB collection cycle produced 11,100 measured records in the `health_care_visit` table (185 municipalities $\times$ 12 months $\times$ 5 years), with no missing values across the 44,400 underlying data points.

Regarding **Maintainability**, the codebase comprises 63 classes distributed across common (32), API (11), and ETL (19) layers. The commit that added the third data source touched only five new, source-specific files, with no changes to shared model, mapper, or loader components, indicating effective isolation between sources. Two limitations are reported transparently: extraction and transformation classes do not share a common interface, and reporting periods are hardcoded rather than externally configurable, a design choice confirmed by a commit that required editing and recompiling a constant list to extend data collection to a new year.

Regarding **Extensibility**, two extensions occurred after the core pipeline was already operational: a DEA-based analytical layer, added approximately one month after the three data sources were functional, and an additional collection year, added six months later through a single isolated commit. A representative extractor/transformation pair (SISAB) comprises 114/127 lines of code, and adding the third source required only one new data-transfer object, with no changes to shared model, repository, or configuration components.

Regarding **Reproducibility**, ease of replication in a new environment or by a different team was assessed alongside execution traceability. The *health-data-etl* and *health-data-client* repositories include a README documenting environment setup (description, dependencies, and database configuration) and step-by-step execution instructions, so that reproducing the pipeline does not require access to the original authors beyond the publicly available source code. Reproducing a full SISAB collection cycle end-to-end, from a clean environment to persisted records, was evaluated by providing a concrete estimate of the effort required to replicate this stage. However, reproducibility is not equally guaranteed across all three sources: while SIOPS is accessed through a documented, stable public API, SISAB and e-Gestor depend, respectively, on scraping a government web form and on an undocumented API accessed through reverse engineering, both of which may change without notice and are outside the architecture's control, representing an external dependency risk for future re-execution by third parties.

Table 4 summarizes these findings. Taken together, this evidence rebalances the evaluation of HERO toward its architectural qualities, complementing the DEA-based analytical results presented earlier in this section, while transparently reporting current limitations as opportunities for future refinement rather than claims of full architectural maturity.

Table 4. Summary of architectural evaluation evidence

Dimension	Highlight
Integration	3 confirmed access mechanisms; 5-entity schema; 11,100 records measured
Maintainability	63 classes; new-source commit isolated to 5 files; hardcoded periods
Extensibility	2 post-launch extensions (DEA layer, +1 year); new source = +1 DTO
Reproducibility	Full cycle: 11m29s; cross-validated via SISAB's own timestamp

7.7 Synthesis of Results and Implications

The architectural findings show that the reference implementation integrates three governmental sources with distinct access mechanisms into a shared relational schema. The full SISAB collection cycle yielded 11,100 records, while source-specific changes and subsequent extensions provide evidence of modular evolution in the evaluated implementation. Together with execution logs and setup documentation, these observations support the feasibility of the integration and analytical access workflow described in Section 7.6.

The DEA case study demonstrates how this integrated dataset can support temporal comparisons, spatial visualizations, and the identification of potential municipal benchmarks. The analyses reveal persistent performance differences, territorial heterogeneity, and a positive association between productivity and efficiency scores. These findings concern relative efficiency under the selected DEA model and indicators; they do not establish the causes of the observed differences.

For public health management, HERO provides a common data preparation and access workflow from which analytical applications can be executed. The DEA results illustrate one such application, while the architectural evidence concerns the integration and delivery of the data on which it depends.

8 Conclusion and Future Work

Brazilian public health management still faces significant limitations resulting from the fragmentation of information systems, the heterogeneity of data formats, and the lack of interoperability among different databases, such as SIOPS,

SISAB, and e-Gestor. This scenario compromises the production of consistent indicators and restricts the ability of managers to make evidence-based decisions [Costa *et al*., 2025; Silva, 2022]. Recent studies reinforce that the absence of standardization and integration hinders both access to and the efficient sharing of health information, directly affecting the quality of management.

In response to this problem, this work proposed, implemented, and evaluated the HERO architecture, designed as a scalable, secure, and analytically useful framework for consolidating public health data. The architecture extends HEALER by incorporating advances focused on modularity, maintainability, and the integration of an ETL pipeline that enables ingestion, transformation, and storage within a structured *Data Lakehouse*. This design also supports the creation of modern interfaces, such as APIs, which promote transparency and structured analytical consumption of integrated data, aligned with the guidelines of the PNIIS and the national Digital Health Strategy [Jesus, 2021; Sousa, 2022].

The validation of HERO is grounded not only in its technical implementation, but also in the practical context in which it was conceived. The architecture emerged within the SmartAPSUS project, an applied initiative aligned with the Brazilian SUS that brought together information technology specialists and public health professionals around real problems of public health data monitoring and decision support. This context helped define the practical requirements of the framework and ensured that its design responded to concrete demands related to data governance, interoperability, traceability, and analytical use in public health management.

From an architectural and operational perspective, the case study demonstrated that HERO successfully supports an end-to-end pipeline, from data ingestion to analytical consumption. The framework integrated heterogeneous data from three Brazilian governmental systems, SIOPS, SISAB, and e-Gestor, despite their distinct access mechanisms, update frequencies, and data structures. By executing the ETL pipeline over multiple years, HERO automated data extraction, transformation, integration, and storage within a structured Data Lakehouse, enabling consistent cross-source analysis and longitudinal monitoring. In addition, the implementation records structured execution logs and stores processed indicators in a relational database, supporting process-level traceability and verification of pipeline execution.

From an analytical perspective, this study adopted DEA as a representative use case to evaluate the technical efficiency of the municipalities of Pernambuco in allocating health resources. The generated indicators, originally explored in a previous study on municipal PHC efficiency [Pereira *et al*., 2025b], supported rankings, spatial maps, dispersion plots, and correlation analyses, revealing coherent temporal and territorial patterns across municipalities. In the present work, however, HERO enabled these indicators to be produced through a governed, automated, and reproducible pipeline. Therefore, the results validate not only the feasibility of the proposed framework, but also its ability to provide standardized, governed, and analytically ready datasets for downstream evaluation methods.

Taken together, these findings indicate that HERO works as intended: it integrates heterogeneous public health data, preserves traceability throughout the pipeline, and supports reproducible analytical consumption over the refined dataset. More broadly, the proposed solution provides a technological foundation for monitoring, evaluation, and decision-support workflows in public health management, while remaining compatible with analytical approaches beyond the specific DEA use case adopted in this study.

Based on the results obtained and the limitations encountered throughout the research, several directions emerge for deepening and expanding the contributions of this work. On the analytical side, future work may include expanding the temporal scope of the analysis, incorporating additional health outcome indicators, applying second-stage statistical models such as cluster analysis, performing qualitative case studies of municipalities with extreme performance, extending the application of HERO to other Brazilian states, improving noise detection and data quality assessment, exploring opportunities for resource reallocation, and investigating the temporal behavior of health indicators using predictive techniques based on Markov Chains or *Machine Learning*. On the architectural evaluation side, future work should include an empirical, benchmarked comparison between HERO and related architectures such as HEALER and the data pool model, under comparable datasets and conditions; a formal specification of the requirements and expectations of intended end users, such as public health managers; an experimental protocol involving domain experts to evaluate the quality of the generated indicators and of the integrated data; and a dedicated scalability evaluation, stress-testing the pipeline under substantially larger data volumes or a higher number of concurrent data sources.

Authors' Contributions

The authors contributed to this work as follows, according to the CRediT (Contributor Roles Taxonomy) framework. **Izabel Yale Neves Nascimento**: Conceptualization; Methodology; Software; Formal Analysis; Investigation; Data Curation; Writing – Original Draft; Visualization. **Tiago Brasileiro Araújo**: Supervision; Validation; Project Administration; Writing – Review & Editing; Resources. **Luis Filipe Alves Pereira**: Supervision; Project Administration; Writing – Review & Editing. **Veruska Borges Santos**: Writing – Review & Editing; Validation. **Dimas Cassimiro do Nascimento Filho**: Writing – Review & Editing; Validation.

Competing Interests

The authors declare no competing interests.

Acknowledgments

We acknowledge the support of the National Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq) and the funding provided by the Department of Science and Technology of the Secretariat of Science, Technology, Innovation and Health Complex of the Ministry of Health (Decit/SECTICS/MS) under research grant No. 445259/2023-0.

Availability of Data and Materials

The source code of both applications implementing the HERO architecture is publicly available on GitHub: *health-data-etl* at `https://github.com/izabelnascimento/health-data-etl` and *health-data-client* at `https://github.com/izabelnascimento/health-data-client`.

Both repositories include README files describing environment setup, database configuration, and execution of the ETL pipeline and analytical scripts. Reproducing the reported case study requires the 2021–2025 collection window, the Pernambuco municipality selection, and the first-semester analytical scope. As described in Section 6, reporting years and periods are currently defined as application constants and must be checked before execution. The *health-data-client* scripts consume the REST APIs and compute efficiency using the input-oriented CCR model described in Section 6.5. These instructions and the structured execution logs support re-execution of the reported workflow.

The datasets used in the study are derived from publicly available Brazilian governmental systems (SIOPS, SISAB, and e-Gestor) and can be reproduced by executing the ETL pipeline provided in the repository.

Ethical Considerations

This study uses exclusively publicly available, aggregated, and anonymized data from Brazilian governmental health information systems (SIOPS, SISAB, and e-Gestor). No individual-level data, human subjects, or sensitive personal information were collected or processed. Therefore, no ethics committee approval was required.

AI Disclosure

The authors used ChatGPT (OpenAI) to assist with language editing, concision, and the organization of selected passages of the manuscript. All suggested edits were critically reviewed and approved by the authors. AI tools were not used to produce the study data, perform the analyses, or generate the reported results or references.

References

Araujo, R., Fornazin, M., and Pimentel, M. (2017). Uma análise sobre a produção de conhecimento científico nas pesquisas publicadas nos primeiros 10 anos da isys (2008-2017). *iSys - Brazilian Journal of Information Systems*, 10(4):45–65. DOI: 10.5753/isys.2017.351.

Araújo, T. B., Efthymiou, V., and Stefanidis, K. (2025). Fairness and explanations in entity resolution: an overview. *IEEE Access*.

Cabral, K. F. D., Gama, C. F., and da Silva, F. T. (2019). Atenção primária à saúde: uma análise à luz da eficiência técnica dos recursos no estado de minas gerais. *Revista de Gestão em Sistemas de Saúde*, 8(2). DOI: https://doi.org/10.5585/rgss.v8i2.14923.

Cancian, M., Cavalcante, W. T., and Pinho, S. T. d. (2023). Desafios na gestão pública no processo de gestão em saúde: Uma revisão de literatura. *Brazilian Journal of Implantology and Health Sciences*, 5(4):2697–2715. DOI: 10.36557/2674-8169.2023v5n4p2697-2715.

Capeletti, N. M., Amado, C. A. F., and Santos, S. P. (2024). Performance assessment of primary health care services using data envelopment analysis and the quality-adjusted malmquist index. *Journal of the Operational Research Society*, 75(2):361–377. DOI: https://doi.org/10.1080/01605682.2023.2197928.

Castro, G. and outros (2024). Avaliação da eficiência dos centros integrados de segurança pública de alagoas: Uma aplicação da análise envoltória de dados. *Anais do LVI Simpósio Brasileiro de Pesquisa Operacional*.

Charnes, A., Cooper, W. W., and Rhodes, E. (1978). Measuring the efficiency of decision making units. *European journal of operational research*, 2(6):429–444. DOI: https://doi.org/10.1016/0377-2217(78)90138-8.

Costa, M. V. d. S., Camargos, M. C. S., Viana, S. M. N., and Mendes, U. V. d. S. (2025). Avanços e desafios da interoperabilidade no sistema Único de saúde. *Journal of Health Informatics*, 17(1):1112. DOI: https://doi.org/10.59681/2175-4411.v17.2025.1112.

Falcão, M. Z., Rachid, R., and Fornazin, M. (2024). Ai innovation in healthcare and state platforms under a rights-based perspective: the case of brazillian rnds. *Data & Policy*, 6:e70. DOI: https://doi.org/10.1017/dap.2024.70.

Ferrera, J. M. C., Cebada, E. C., and Zamorano, L. R. M. (2014). The effect of quality and socio-demographic variables on efficiency measures in primary health care. *The European Journal of Health Economics*, 15:289–302. DOI: https://doi.org/10.1007/s10198-013-0476-1.

Ferré, F., de Oliveira, G., de Queiroz, M., and Gonçalves, F. (2020). Sala de situação aberta com dados administrativos para gestão de protocolos clínicos e diretrizes terapêuticas de tecnologias providas pelo sus. In *Anais do XX Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 392–403, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbcas.2020.11530.

Guntupalli, B. (2023). Data lake vs. data warehouse: Choosing the right architecture. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 4(4):54–64.

Hevner, A. R., March, S. T., Park, J., and Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1):75–105. DOI: 10.2307/25148625.

Inácio, A., Savaris, A., Alves, J., and von Wangenheim, A. (2016). Sistema integrado de análise e visualização epidemiológica georreferenciada online para telemedicina. In *Anais do XVI Workshop de Informática Médica*, pages 2547–2556, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbcas.2016.9901.

Jesus, T. (2021). Ministério da saúde atualiza a política nacional de informação e informática em saúde.

Lima, D. M., Moreno, R. A., Pires, F. A., and Gutierrez, M. A. (2021). Uma proposta de data lake para pesquisa em saúde a partir de data pools multicêntricos interoperáveis. In *Anais do XXXVI Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 367–372, Rio de Janeiro, RJ, Brasil. SBC. DOI: 10.5753/sbbd.2021.17900.

Loureiro, V. A., da Silva, P. U., and Almeida, M. (2024).

Análise da eficiência dos bancos digitais no brasil utilizando a metodologia network data envelopment analysis (ndea). *Anais do LVI Simpósio Brasileiro de Pesquisa Operacional*.

Machado, M. L. and Tavares, S. (2023). Governança e interoperabilidade de dados para a saúde.

Macinko, J., Starfield, B., and Shi, L. (2003). The contribution of primary care systems to health outcomes within oecd countries, 1970–1998. *Health Services Research*, 38(3):831–865. DOI: 10.1111/1475-6773.00149.

Manco, C., Dolci, T., Azzalini, F., Barbierato, E., Gribaudo, M., Tanca, L., *et al*. (2023). Healer: A data lake architecture for healthcare. *Workshops of the EDBT/ICDT Joint Conference, EDBT/ICDT-WS 2023*, 3379:N–A.

Marques, A. C. and outros (2024). Análises de eficiência em rede para o setor elétrico brasileiro. *Anais do LVI Simpósio Brasileiro de Pesquisa Operacional*.

Mello, B., Donida, B., Costa, C. A., Rigo, S., Alvez, F., Petry, L., Rodrigues, H., Foletto, L., Antunes, R., Ramos, G., and Righi, R. (2021). Sars event notification: template openehr brasileiro para interoperabilidade semântica de dados com foco na covid-19. In *Anais do XXI Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 434–439, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbcas.2021.16088.

Nambiar, A. and Mundra, D. (2022). An overview of data warehouse and data lake in modern enterprise data management. *Big data and cognitive computing*, 6(4):132. DOI: https://doi.org/10.3390/bdcc6040132.

Nascimento, I., Santos, V., Pereira, L. F., Araújo, T., Rocha, C. G., Filho, F. M., Andrade, R., Vanderlei, I., and Filho, D. N. (2026). Primary healthcare efficiency indicators for analysis and decision-making support for managers. In *Anais do XXII Simpósio Brasileiro de Sistemas de Informação*, pages 556–575, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbsi.2026.248568.

Pereira, L., Ferreira, L., Nascimento, D., Vanderlei, I., Silva, D., Santos, V., Gomes, M., and Melo, I. (2025a). Design, integration, and evaluation of a demand forecasting service in the context of primary healthcare. In *Anais do XXI Simpósio Brasileiro de Sistemas de Informação*, pages 506–514, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbsi.2025.246552.

Pereira, L., Nascimento, I., Santos, V., Pinheiro, R., and Filho, D. N. (2025b). Avaliação da evolução da eficiência operacional de municípios brasileiros na gestão de recursos para atenção primária à saúde. *Anais do XXV Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 737–747. DOI: 10.5753/sbcas.2025.7743.

Phanikanth, K. V. and Sudarsan, S. D. (2016). A big data perspective of current etl techniques. *2016 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, pages 330–334. DOI: 10.1109/ICACCE.2016.8073770.

SAÚDE, M. D. (2020). Ministério da saúde lança três publicações sobre a estratégia de saúde digital para o brasil 2020-2028.

Silva, A. P. S. e., Silva, d., Júnior, O. V. P., Ramos, V. d. B. F., Santos, C. G. d., Bezerra, L. d. A. F., Fachini, M., Gurgel, H. F., Carvalho, L. S. G. d., Oliveira, G. P. d., Souza, I. R. d. S. M. d., and Barbosa, M. J. L. (2023). Desafios na atenção básica à saúde no brasil: Enfoque na saúde pública. *Brazilian Journal of Implantology and Health Sciences*, 5(5):5065–5073. DOI: 10.36557/2674-8169.2023v5n5p5065-5073.

Silva, J. (2022). Sistemas de informação em saúde: um estudo dos processos de registro, integração e recuperação da informação das bases sisab e e-sus ab. Master's thesis, Universidade Federal de Alagoas.

Silva, M., Rabello, H., and Nepomuceno, T. (2024). Análise de eficiência do policiamento nas Áreas integradas de segurança em pernambuco: Uma abordagem dea com dados em painel. *Anais do LVI Simpósio Brasileiro de Pesquisa Operacional*.

Sousa, A. (2022). Governança da estratégia de saúde digital.

Varela, P. S., de Andrade Martins, G., and Fávero, L. P. L. (2012). Desempenho dos municípios paulistas: uma avaliação de eficiência da atenção básica à saúde. *Revista de Administração*, 47(4):624–637. DOI: https://doi.org/10.5700/rausp1063.

Wazlawick, R. S. (2014). *Metodologia de Pesquisa para Ciência da Computação*. Elsevier, Rio de Janeiro, 2 edition.