The 12th International Conference on E-Health and Bioengineering - EHB 2024
Grigore T. Popa University of Medicine and Pharmacy Iasi, November 14-15, 2024, Iasi, ROMANIA

Addressing Complex Data Integration and Harmonization Scenarios

Marina Loulaki, Athanasios Kiourtis, Argyro Mavrogiorgou, Dimosthenis Kyriazis
Department of Digital Systems, University of Piraeus,
Piraeus, Greece, {mloulaki, kiourtis, margy, dimos}@unipi.gr

***Abstract*—In recent years, the increasing volume and complexity of data, especially in domains like healthcare, has emphasized the need for improved use of data to extract valuable insights, that can positively impact society in multiple perspectives (e.g., economic, societal, life quality). The importance of data integration has grown, since the combination of multiple datasets can drastically enhance the value of analysis, compared to individual dataset analysis. However, significant challenges remain, including compatibility issues among data sources and lack of standardization, restricting the full potential of such analyses. This manuscript explores the purpose and importance of data integration, focusing on the healthcare domain. A methodology is proposed, outlining the steps for harmonizing and integrating diverse datasets to ensure consistency and compatibility. This aims to enable advanced analysis, resulting in more accurate outcomes than those derived from using a single dataset alone.**

Keywords—data harmonization; data cleaning; data merging; data integration; healthcare

I. INTRODUCTION

The amount of data that has been generated over the last decade has drastically increased. By the end of 2024 it is expected to grow by 1076%, compared to 2014. In the following years, it is expected to continue to grow, reaching even higher levels [1]. Especially in the healthcare industry, this seems to be relevant, since data production is rapidly evolving. The form of the data itself often comes in various forms and formats [2]. It is worth noting that different healthcare facilities tend to collect different types of information, which may be stored under different variable names, even when the underlying data is quite similar. Understanding the role of data merging and integration is essential, particularly when managing divergent data sources that represent similar content. Data merging involves combining data from multiple sources into a single dataset, often eliminating duplicate records or redundant information. The main goal is to create a unified, consolidated view of the data by removing redundancies and standardizing records [3]. Data integration on the other hand, focuses on bringing together autonomous and heterogeneous data sources in a way that enables uniform access, despite the differences in how these sources are structured, managed, or controlled [4]. Both processes are equally important in the healthcare domain, because they work together to create a cohesive data ecosystem. By ensuring that patient data is comprehensive, accurate, and accessible, data merging and integration improve healthcare delivery and enable better clinical decision-making, research, and patient outcomes.

Despite the progress that has been accomplished in both data merging and data integration, challenges still need to be addressed. Taking into consideration the ever-growing volume and complexity of healthcare data, the processes should be scaled accordingly to handle them. Existing solutions often are unable to handle issues, such as data interoperability or data quality, across diverse systems, especially when there is a high number of datasets involved. This creates deficiencies in managing and analyzing data, as incomplete or inconsistent information reduces the reliability of insights and limits the potential for advancing data-driven healthcare solutions. This paper aims to address these gaps by introducing a comprehensive approach that integrates data harmonization techniques with current data merging and integration methods. Our approach, compared to existing ones, provides a more scalable and accurate solution for the diverse healthcare datasets challenges, enhancing interoperability and quality.

The rest of the manuscript is structured as follows. Section II reviews Related Work, discussing existing data integration and merging approaches in healthcare. Section III presents the Proposed Approach, outlining the steps of problem identification, data profiling, and data harmonization. Section IV provides the Preliminary Evaluation, where the proposed methodology is applied to real-world healthcare datasets. Finally, Section V concludes with a discussion of the results and suggestions for future work.

II. RELATED WORK

A. Data Actions

Data Actions refer to processes that involve handling and manipulation of data, such as integration, merging, and transformation. These actions are crucial for combining datasets from multiple sources into a unified form. As described in [5], data integration involves the combination of data from multiple sources, resulting in the creation of a unified dataset, typically through a global schema that abstracts the complexities which originate from the individual data sources. A key part to this process is the mapping of the relationships between the global and the source schemas. Those two schemas facilitate smooth interactions between the datasets.

979-8-3315-3214-7/24/$31.00 ©2024 IEEE

Unlike simple data pooling, model-based data integration, as highlighted in [6], it explicitly manages to confront the differences between the data sources. With this, the unique strengths are preserved, and potential biases are addressed. For example, data integration might include patient records that have been integrated from various healthcare providers, merged into a single Electronic Health Record (EHR) system. Data coming from different providers will probably use different schemas, focusing on diagnosis codes or on patient treatments. With model-based integration, these differences are mapped into a global schema, providing healthcare professionals with a harmonized view of patient histories. The definition of data merging is challenging, as there are few relevant works than other topics, and even those that do exist often do not provide a clear or consistent definition [7]. Data merging refers to combining two or more datasets or objects into a single, unified output. For example, in health data records, merging could involve combining patient records from multiple hospitals, where some records may include full medical histories (complete) while others contain only recent visits (incomplete). The result could involve adding new patients or updating existing records with missing data. The merging process unifies data from different sources, producing a result that can shrink, stay the same, or grow, based on the data structure. The order of merging typically does not affect the outcome, ensuring efficient integration of partial updates (deltas) and complete datasets, maintaining data integrity.

B. Importance of Data Integration & Related Challenges

Unified data is critical across industries, especially in today's data-driven world where organizations depend on diverse data sources [11]. Integrating data from multiple platforms allows for more accurate analysis, helping overcome the barriers posed by heterogeneous data structures. Unified data addresses interoperability issues by enabling smooth data exchange and analysis across different systems. Isolated data can hinder progress and efficiency, while unified information significantly improves decision-making and outcomes [8]. In the corporate world, the potential of big data is fully realized when advanced tools and skilled personnel are in place to process it effectively, driving better performance [10]. In healthcare, unified data offers substantial benefits. Integrating data from sources like EHRs, clinical trials, and public health databases enhances research and patient outcomes. Unified data helps overcome inconsistencies or quality issues, providing standardized views that lead to more effective clinical trials and accurate diagnoses. Moreover, it supports large-scale public health initiatives, such as disease outbreak tracking or monitoring treatment efficacy, by making data more accessible and actionable [9]. Unified data systems ensure that healthcare providers have access to complete and up-to-date information, reducing errors or treatment delays. By resolving interoperability challenges, these systems enable a more intelligent, data-driven approach to patient care, improving outcomes and enhancing public health responses.

C. Data Integration Solutions

As discussed earlier, the growing complexity of data sources has led to an increased need for efficient integration solutions. Organizations now rely on data integration to uncover insights and drive decision-making. Without proper integration, challenges can hinder both data analysis and decision-making processes. This is particularly critical in healthcare, where effective data integration is essential. Several tools address this issue, such as data standardization, data warehouses, AI/ML, and data migration tools. Standardizing data formats ensures consistency across sources, facilitating sharing and interoperability. Data warehouses centralize data into a single repository, ensuring easier access, analysis, and support for historical data and large-scale analytics, which are crucial for healthcare studies. Data warehouses also improve organizational agility by enabling efficient querying and reporting. AI and ML solutions automate data cleaning, transforming, and unifying processes across different systems. These technologies help detect patterns, fill missing data, and standardize inconsistencies, enhancing scalability and efficiency while reducing manual data harmonization efforts. Data migration tools further facilitate the secure transfer of data, ensuring integrity and minimizing risks of data loss or corruption. AI-based methods, such as ML and NLP, help standardize unstructured clinical data and improve integration, as discussed in [12]. Protocols like HL7 and FHIR also ensure semantic interoperability. Despite these advancements, challenges like data heterogeneity and system interoperability persist. Real-time data integration, as noted in [13], emphasizes synchronization, which is crucial in healthcare. The solution builds on current techniques, addressing healthcare data challenges like heterogeneity and interoperability through AI/ML methods to improve data accuracy and accessibility.

III. PROPOSED APPROACH

The proposed approach's mechanism focuses on integration, merging, and ultimately data harmonization. The integration and merging of healthcare datasets ensures consistency, interoperability and completeness among the data. These processes lead the way towards data harmonization, where the unified datasets are not only combined, but also aligned through a common structure. Our approach consists of two main steps: Data Profiling, and Data Harmonization. By completing these steps, the unification of two or more healthcare-related datasets is accomplished. The details of each step in the approach are outlined in Figure 1.

Data Profiling: The data profiling step focuses on identifying and addressing inconsistencies and data quality issues across multiple healthcare datasets. Key problems include inconsistent measurement units, such as when one dataset records glucose levels in mg/dL while another uses mmol/L, varying schema structures where column names or data types differ for similar variables, and data quality concems like missing values and duplicates. To resolve these issues, schema-matching algorithms, such as fuzzy matching, are employed to align equal fields between datasets, ensuring consistent variable naming. Unit standardization is achieved by applying predefined conversion factors, for example, converting glucose levels from mg/dL to mmol/L using the Pint package. The Pandas library is used for data profiling and to identify errors like missing values or duplicates. Additionally, Scikit-learn's imputation techniques, such as KNN, are applied to handle missing data effectively.

Data Harmonization: The data harmonization process aims to address several key issues, including semantic errors, conflicting data formats, inconsistent measurement units, and incomplete records. The first step involves a thorough analysis of the datasets to identify these problems. Tools such as Pandas or Dask, which are optimal for handling large datasets and distributed computing, are used to automatically detect inconsistencies such as missing values, duplicated entries, and incorrect data types. Once the issues are identified, various standardization techniques are applied to harmonize the data. Pandas' built-in functions manage other standardization tasks, such as aligning date formats or categorical variables across datasets. For missing data, Scikit-learn offers a range of imputation techniques, such as K-nearest neighbors (KNN), mean or median imputation, depending on the type and pattern of missing data. Continuous variables are normalized using Z-score normalization to ensure uniformity across datasets. After standardization, Pandas is employed to merge the datasets, combining them based on unique keys while maintaining data integrity. Every step, including data transformation and modifications, is logged using Python's logging module to ensure full traceability. The resulting harmonized dataset, along with a detailed log of changes, is prepared for further analysis.

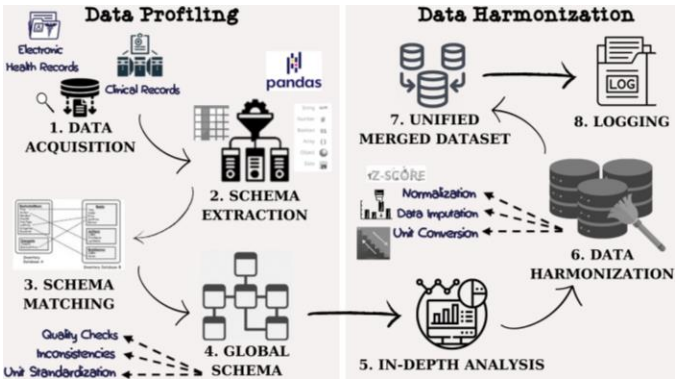

Figure 1. Steps of Data Harmonization

IV. PRELIMINARY EVALUATION

A. Working Environment & Evaluation Results

The data harmonization process was implemented in a Python 3.10.3 environment, utilizing key libraries such as Pandas 2.0.2 for data manipulation and Scikit-learn 1.3.0 for machine learning tasks. The development was conducted using the Visual Studio Code IDE, configured for Python development. A logging mechanism was set up from Python's logging module (version 0.5.1.2), ensuring tracking of the harmonization steps, error handling, and imputation actions. The environment was configured to handle datasets in formats like CSV and Excel, with logging outputs stored in separate log files for detailed and imputation-specific operations. To showcase the analyzed methodology, the proposed harmonization approach was applied to two datasets derived from Kaggle [14][15], with the identified problem being the Diabetes disease. The goal was to harmonize and unify these datasets for a more comprehensive analysis of diabetes risk.

The **(i) Diabetes Prediction Dataset (Dataset 1) [14]** contains 100,000 instances and nine key attributes commonly used to predict diabetes risk. It is primarily sourced from EHRs and provides a longitudinal view of patient data from multiple healthcare providers. The key features in this dataset include: **(1) Gender**, **(2) Age**, **(3) Hypertension:** A binary variable indicating whether the patient has hypertension, **(4) Heart disease:** A binary variable indicating the presence of heart disease, **(5) Smoking history**: A categorical variable with the following categories: not current, former, No Info, current, never, and ever, **(6) BMI**, **(7) HbA1c level**: Hemoglobin A1c, measuring average blood sugar levels over the past 2-3 months, **(8) Blood glucose level**, **(9) Diabetes**: The target variable, indicating whether the patient has diabetes.

The **(ii) Healthcare Diabetes Dataset (Dataset 2) [15]**: It is sourced from the National Institute of Diabetes and Digestive and Kidney Diseases, contains 2,769 instances and 10 diagnostic attributes designed to predict diabetes. The key features in this dataset include: **(1) Id**: Unique identifier, **(2) Pregnancies**: Number of pregnancies the patient has had **(3) Glucose:** Plasma glucose concentration from a 2-hour oral glucose tolerance test **(4) BloodPressure:** Diastolic blood pressure in mm Hg **(5) SkinThickness**: Triceps skinfold thickness in mm **(6) Insulin**: 2-Hour serum insulin (mu U/ml) **(7) BMI**, **(8) Diabetes Pedigree Function**: The likelihood of diabetes based on family history **(9) Age**, **(10) Outcome**: The target variable, indicating whether the patient has diabetes. Figure 3 depicts the harmonization process results.

Before Harmonization		After Harmonization	
Dataset 1	**Dataset 2**	**Unified Dataset**	**Source**
Gender	-	Gender	Dataset 1
Age	Age	Age	Both datasets
Hypertension	-	Hypertension	Dataset 1
Heart disease	-	Heart disease	Dataset 1
Smoking history	-	Smoking history	Dataset 1
BMI	BMI	BMI	Both datasets
HbA1c level	-	HbA1c level	Dataset 1
Blood glucose level	Glucose	Blood glucose level	Both datasets
Diabetes	Outcome	Diabetes	Both datasets
-	Pregnancies	Pregnancies	Dataset 2
-	Blood Pressure	Blood Pressure	Dataset 2
-	Skin Thickness	Skin Thickness	Dataset 2
-	Insulin	Insulin	Dataset 2
-	Diabetes Pedigree Function	Diabetes Pedigree Function	Dataset 2

Figure 3. Datasets Before and After Harmonization

The outcomes of the proposed harmonization approach are presented in two stages: Data Profiling and Data Harmonization. In the profiling step, the datasets were acquired and examined, identifying key differences, such as inconsistent column names, missing values, and dataset-specific variables. The data schemas for both datasets were created, and a global schema was generated based on identified differences. For instance, while Age and BMI were present in both datasets different data types were standardized, variables like Hypertension and Pregnancies were dataset specific. Additionally, an assumption was made for the "blood glucose level" variable, treating it as equivalent to the "Glucose" variable from the second dataset, despite not having

detailed information on the specific glucose test procedures used in both datasets. The variable Id from the second dataset was dropped due to its irrelevance. In the profiling process equivalent fields were aligned and schemas were standardized to ensure consistency. In the harmonization step, additional transformations were applied. The datasets had limited overlapping attributes, resulting in a significant number of missing values. To address this, Scikit-learn's K-nearest neighbors (KNN) and median imputation methods were used to fill in the missing values based on the nature of the data. Continuous variables were standardized to ensure consistency across the datasets. To evaluate the effectiveness of these preprocessing steps, a Logistic Regression model from Scikit-learn, was applied to different versions of the dataset, both before and after harmonization, allowing a straightforward comparison. Results indicated that models trained on the harmonized and imputed dataset showed improved accuracy compared to those trained on the incomplete dataset. While further experiments could explore additional imputation techniques, these initial findings suggest that harmonization and imputation enhance the dataset's reliability. Following these adjustments, the final harmonized dataset, consisting of 102,768 records, was created, with all transformations logged and prepared for further analysis. This ensured that the dataset was complete and ready for robust analysis of diabetes risk.

V. Discussion & Concluding Remarks

The healthcare domain benefits from diverse data, but simply having data is not enough. Merging and harmonizing them leads to improved diagnostic accuracy and personalized treatment plans. Large-scale data integration also minimizes errors and supports public health policies. However, challenges like system interoperability and data quality persist. In this paper, we addressed these challenges by implementing a harmonization mechanism using tools for schema matching, unit standardization, and data imputation. The selected preprocessing steps allowed us to overcome many obstacles, ensuring that the final harmonized dataset was ready for analysis. The methods analyzed in this paper can also be adapted to other fields, such as finance, where improvements can be made towards risk management, by integrating datasets from various financial institutions. Similarly, in education, harmonizing student data from different sources can support the development of personalized learning plans by providing a unified view of academic performance and learning needs. The paper's main goal is to enhance healthcare data integration, offering short-term gains in patient care and long-term benefits like interoperable systems. Looking forward, we plan to explore more advanced AI/ML techniques, such as deep learning-based imputation to improve the handling of missing data and reduce potential biases. Addressing interoperability remains a key focus, with plans to integrate standards like HL7 FHIR for better system compatibility across different healthcare platforms [16][17][18]. This study focused on healthcare, the methodology can be applied to other sectors, including finance [19] and policy making [20], for further harmonization.

Acknowledgment

The SmartCHANGE project has received funding from the Horizon Europe R&I programme under the GA No. 101080965.

References

[1] "Data growth worldwide 2010-2025|Statista," *Statista*, https://www.statista.com/statistics/871513/worldwide-data-created/

[2] A. H. Ramirez, *et al.*, "The All of Us Research Program: Data quality, utility, and diversity", Patterns, **8**, 2022, p. 100570.

[3] Data Merging Essentials, https://www.astera.com/.

[4] A. Doan, A. Halevy, Z. Ives, "Principles of Data Integration", 2012.

[5] G. Cima, et al., "Abstraction in Data Integration", ACM/IEEE Symposium on Logic in Computer Science, 2021, pp. 1-11.

[6] N. J. Isaac, et al., "Data Integration for Large-Scale Models of Species Distributions", *Trends in Ecology & Evolution*, ***35*(1)**, 2021, pp. 56–67.

[7] P. Konopka, B. Von Haller, "Exploring data merging methods for a distributed processing system", Journal of Physics Conference Series, **2438(1)**, 2023, p. 012038.

[8] W. Xiaojin, et al., "Research on data standardization and unified data interface based on digital station system", Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), **5**, 2022, pp. 1372-1376.

[9] Z. Kaoudi, J. Quiané-Ruiz, "Unified data analytics", VLDB Endowment, **15(12)**, 2022, pp. 3778–3781.

[10] M. Ghasemaghaei, "Improving Organizational Performance Through the Use of Big Data", Journal of Computer Information Systems, **60(5)**, 2012, pp. 395–408.

[11] M. Janssen, et al., "Data governance: Organizing data for trustworthy Artificial Intelligence", Government Information Quarterly, **37(3)**, 2020, p. 101493.

[12] N. Pushadapu, "Artificial Intelligence for Standardized Data Flow in Healthcare: Techniques, Protocols, and Real-World Case Studies", Journal of AI-Assisted Scientific Discovery**, 3(1)**, 2023, pp. 435-474.

[13] C. R. Sahara, A. M. Aamer, "Real-time data integration of an internet-of-things-based smart warehouse: a case study", International Journal of Pervasive Computing and Communications, **18(5)**, 2021, pp. 622–644.

[14] Diabetes prediction dataset, https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data

[15] Healthcare Diabetes Dataset, https://www.kaggle.com/datasets/nanditapore/healthcare-diabetes

[16] A. Mavrogiorgou, et al., "beHEALTHIER: A microservices platform for analyzing and exploiting healthcare data", 34th International Symposium on Computer-Based Medical Systems, 2021, pp. 283-288.

[17] D. Kyriazis, et al., "The CrowdHEALTH project and the hollistic health records: Collective wisdom driving public health policies", Acta Informatica Medica, **27(5)**, 2019, p.369.

[18] N. Reščič, et al., "SmartCHANGE: AI-based long-term health risk evaluation for driving behaviour change strategies in children and youth", International Conference on Applied Mathematics & Computer Science, 2023, pp. 81-89.

[19] A. Mavrogiorgou, et al., "FAME: federated decentralized trusted data marketplace for embedded finance", International Conference on Smart Applications, Communications and Networking, 2023, pp. 1-6.

[20] O. Biran, et al., "PolicyCLOUD: A prototype of a cloud serverless ecosystem for policy analytics", Data & Policy, **4**, 2022, p. e44.