

Towards a ubiquitous health system through integration of heterogeneous data sources using relevant attributes without identification keys

Khaled Kechida^a , Hamoudi Kalla^a , Mourad Brik^b 

^a Computer Science Department, University of Batna 2, Batna, Algeria

^b University of M'Sila, University Pole, Bordj Bou Arreridj Road, M'Sila 28000 Algeria

ABSTRACT

Considerable advances in software and hardware technologies now enable information systems to integrate heterogeneous data from multiple sources. Such integration helps address data scarcity by combining related datasets. However, the absence of standard formats and the unstructured nature of many datasets limit automated processing, interoperability, and practical analysis. Heterogeneity also arises because each facility manages its own data using different information systems, creating a major challenge for integration. The objective of this study is to integrate heterogeneous data sources using relevant attributes (e.g., name, date of birth, e-mail), which distinguishes our approach, rather than relying on identification keys that make integration manual and hard to generalise. The purpose of our method is to integrate all heterogeneous data for each individual into a unified database. For patients, it centralises medical data, supports care continuity, and cuts redundant examination costs. This approach can also be applied to education (academic records), commerce (customer data), and epidemiological studies. A case study illustrates the merging of data into a standardised XML (eXtensible Markup Language) format via a data processing pipeline, demonstrating the applicability of the proposed approach, and discusses the main challenges encountered and possible solutions. This methodology enhances efficiency, optimises resources, and informs decision-making.

Keywords XML schema, mapping relation, interoperability, standardisation, data personalisation, Data sharing

Categories

• Information systems → Data management systems → Information integration

Email

Khaled Kechida – khaled.kechida@univ-msila.dz (CORRESPONDING)

Hamoudi Kalla – hamoudi.kalla@univ-batna2.dz

Mourad Brik – mourad.brik@univ-msila.dz

Article history

Received: 4 September 2025

Accepted: 24 January 2026

Online: 31 July 2026

1 INTRODUCTION

Over the past two decades, the integration of heterogeneous data sources has become a critical challenge across many sectors that rely on heterogeneous and distributed sources, in domains such as digital health, finance, e-commerce, and administrative services. Organisations must consolidate information originating from diverse systems, formats, and structures to enable

Kechida, K., Kalla, H., Brik, M. (2026). Towards a ubiquitous health system through integration of heterogeneous data sources using relevant attributes without identification keys. *South African Computer Journal* 38(1), 43–78. <https://doi.org/10.18489/sacjv38i1/23806>

Copyright © the author(s); published under a [Creative Commons NonCommercial 4.0 License](https://creativecommons.org/licenses/by-nc/4.0/) 
SACJ is a publication of *SAICSIT*. ISSN 1015-7999 (print) ISSN 2313-7835 (online)

consistent access and efficient decision-making. This challenge is particularly critical in digital health, where medical institutions handle large volumes of information originating from disparate systems such as medical records, electronic health records (EHR), or imaging platforms. Users often require knowledge of the underlying structure of each data source, whether structured, semi-structured, or unstructured, and the appropriate querying languages to access relevant information (Baum et al., 2023; Liu et al., 2023). Consequently, accessing consistent and reliable information across distributed healthcare systems remains difficult. In practice, healthcare professionals collaborate, share data, and make decisions across geographically dispersed environments, emphasising the need for shared decision-making and improved communication regarding patient preferences to enhance the overall efficiency of healthcare delivery and avoid costly, repetitive examinations (Rognan et al., 2023).

Nonetheless, integrating such multi-origin clinical data is challenging due to the significant heterogeneity of formats, structures, naming conventions, and identification keys. Healthcare data typically exhibits substantial volume and diversity, and its integration requires sophisticated methods capable of extracting, normalising, and consolidating information to support knowledge sharing, informed decision-making, and patient-centred care (d’Ussel et al., 2024; Rambau et al., 2025; Welbotha & van den Berg, 2024). XML remains widely used as a medium for document exchange and storage, making it a suitable candidate for addressing multi-source integration issues (Dhayne et al., 2019; Li et al., 2019). Yet conventional join operations, typically based on strictly identical keys, prove insufficient in real-world medical contexts, where patient records may vary in structure, attribute naming, or even value formatting (Phan & Phan, 2022). Record linkage in the absence of deterministic identifiers remains a significant challenge in many information systems, particularly when patient, client, or citizen records must be consolidated across heterogeneous sources. In real-world settings, especially in emergency, multilingual, or resource-constrained environments, only partial information, such as a name, date of birth, or approximate location, is often available. Integrating heterogeneous data with different identification keys into a unified database requires a framework capable of performing string-similarity-based joins, enabling linkage with a minimum amount of identifying information and without relying on each source’s non-uniform identification key. Studies have explored similarity-based joins, schema alignment, or hybrid approaches that combine different similarity measures (Xu & Lu, 2019). Still, existing research often falls short of simultaneously addressing structural heterogeneity, attribute variability, and duplicate record detection, which are essential to ensure data quality and consistency (Albayrak et al., 2022; Z. Chen et al., 2019). Our approach directly addresses this limitation by combining attribute normalisation, dynamic synonym expansion, relevance-based attribute selection, and similarity-driven clustering to infer potential matches even when identifiers are incomplete or inconsistent. This allows fragmented digital ecosystems to recover coherent identities without depending on strict, unique identifiers.

To address these challenges, this research proposes a two-phase approach to integrate heterogeneous data more accurately and robustly. The first phase ensures harmonisation of sources through normalisation and attribute standardisation, producing structurally aligned

XML documents. The second phase applies several similarity measures to non-key attributes to compute a similarity matrix. The primary key is not used, as it varies across sources, shows no beneficial correlation, and lacks any discriminative power: it does not genuinely reflect the identity or characteristics of the entity (e.g., patient), and therefore cannot serve as a reliable attribute for grouping. A clustering algorithm will then use this matrix to group together similar entity records. This approach allows for better detection of similar cases, reduces duplication, and strengthens the overall reliability of the integration. In this context, implementing an information system capable of integrating health data from heterogeneous sources and enriching the new database with medical data that may be missing in one facility but available in another, and sharing it effectively between practitioners and patients, can significantly accelerate decision-making and reduce redundant examinations.

The objectives of this research are threefold:

1. to design a robust process for normalising and standardising XML documents from diverse sources into a unified target schema;
2. to develop a similarity-computation model that integrates multiple measures for resilient record matching; and
3. to implement clustering techniques to aggregate similar records, clean duplicates, and maintain a single representative entity per group.

The approach incorporates mathematical models for attribute selection, similarity computation, and clustering (Z. Chen et al., 2019), as well as filtering and cleaning procedures that merge similar records and remove duplicates (Albayrak et al., 2022; Alluhaidan et al., 2023; Ormesher, 2024). After clustering and refinement, the processed instances are consolidated into a single, coherent XML output. The remainder of this paper is organised as follows. **Section 2** reviews related work. **Section 3** describes the proposed methodology, including the automated normalisation pipeline, similarity modelling, clustering algorithm, and XML integration workflow. **Section 4** describes the application of the proposed mathematical approach for integrating heterogeneous data into shared EHRs. **Section 5** presents a case study, while **Sections 4** and **6** provide a discussion, conclusion and summary of the contributions

2 RELATED WORK

This research aims to propose a complete pipeline for the integration of heterogeneous data. This review is structured around the main objectives of the study: the integration of heterogeneous sources, the normalisation of entities, multi-attribute similarity computation, duplicate detection, clustering, and the fusion of records to construct unified profiles that can be exploited for decision support and epidemiological analysis, as well as for the merging of user profiles (patients, clients, students, companies, etc.). Although the existing literature addresses some of these aspects, no prior work combines all of these steps within a unified

framework, a gap that the present contribution seeks to address. Across sectors such as finance, e-commerce, e-health, particularly in resource-limited settings, the absence of a universal unique identifier forces organisations to rely on quasi-identifiers (name, date of birth, sex, address) for record linkage. Financial institutions must reconcile inconsistent account identifiers (Steinkamp et al., 2022), government agencies face similar difficulties when merging citizen records (Mutemaringa et al., 2023), and e-commerce platforms must link transactions despite varying personal (Adapa, 2025). Genealogical studies likewise match historical records with incomplete or inconsistent information (Zhang et al., 2018). These cases illustrate that heterogeneous formats, missing identifiers, and inconsistent data are pervasive challenges, underscoring the broad relevance of our approach. Healthcare is similarly affected: patient data remain fragmented across hospitals, laboratories, and clinics, with limited data exchange, hindering longitudinal reconstruction (Chuma, 2025). This fragmentation leads to duplicate or mis-linked records, producing false negatives, false positives, and incomplete histories, which in turn cause repeated tests, diagnostic delays, and treatment errors, ultimately compromising patient safety, care quality, and the reliability of public-health data (Grannis et al., 2022).

Recent work has explored various strategies for standardising heterogeneous data sources. Westermann et al. (2026) shows that domain-specific XML can be transformed automatically into semantic knowledge graphs, improving interoperability and queryability, although its approach remains limited to AutomationML and does not generalise to diverse XML structures. Iglesias et al. (2024) provides an efficient engine for converting heterogeneous inputs into RDF through declarative mappings, yet several advanced cases remain only partially supported. Ling et al. (2025) demonstrates the value of automated generation of structured XML from relational databases, but full compliance with established markup standards is still incomplete, and the output quality depends heavily on the accuracy of the underlying metadata. The integration of heterogeneous data has been the focus of extensive research, particularly in the medical field, where information system interoperability is a critical challenge. Some of these approaches aim to unify data using medical ontologies represented by reference models such as OpenEHR (open Electronic Health Record), HL7 FHIR (Health Level Seven), and MIMIC-III (Medical Information Mart for Intensive Care-III (Roehrs et al., 2019) to ensure seamless interoperability across systems. While these standards define structural interoperability, they do not provide mechanisms for detecting duplicated patients or merging fragmented clinical records. Other studies leverage mediator/wrapper-based middleware techniques to integrate semantically heterogeneous databases, applying paradigms such as Global-as-View (GaV) and Local-as-View (LaV) (Adeyelu & Anyebe, 2019; Buron et al., 2020). Converting data structures into XML relies on methodologies and algorithms that enable seamless mapping between heterogeneous formats and XML, ensuring interoperability, compatibility, and efficient exchange across diverse systems. For example, the XML-REG approach uses an efficient mapping schema to convert XML data into relational database (RDB) storage. Yet, DTD files are often not provided alongside XML documents (Song & Haw, 2020). Although DTDs and XML Schemas (XSD) define valid structures, they remain rigid when data models evolve or when sources use differing syntactic conventions. XSLT transformations provide greater flexibility, but they gen-

erally cannot resolve lexical inconsistencies such as accents, extra spaces, or case variations. Putrama and Martinek (2024) highlight the challenges of integrating heterogeneous data and proposes semantic integration, machine learning, and distributed approaches as potential solutions. However, these methods remain complex to deploy and have limitations in terms of standardisation and interoperability. The JSON-to-XML paradigm enables bidirectional conversion while preserving the original JSON structure, helping to address challenges posed by heterogeneous data formats. However, no standard method exists to generate an XML schema directly from JSON data, and existing conversion libraries offer varying advantages and limitations (Mukhitova et al., 2024). Methodological studies on XML-based database modelling have sought to enhance interoperability and data exchange across platforms. Tools such as ETL (Extraction, Transformation, and Loading) facilitate data extraction and transformation, while HL7 standards and DICOM (Digital Imaging and Communications in Medicine) frameworks support information exchange within healthcare systems (Oemig & Blobel, 2022). WordNet assists in text normalisation, but does not resolve XML structural mapping (Kafe, 2023). The Unified Medical Language System (UMLS) harmonises diverse biomedical terminologies into unified concepts, enabling semantic interoperability in clinical knowledge representation (L. Chen et al., 2023). Nevertheless, UMLS does not address structural transformations or cross-tag mappings, and its coverage of non-English terminologies remains limited. This highlights the need for a dynamic and adaptable synonym table to support XML integration, which constitutes a distinctive feature of our approach.

Attribute selection aims to enhance approach performance while reducing the dimensionality for computing overall entity similarity. Saelee et al. (2024) apply formal concept analysis and decision trees to identify a subset of significant attributes within a concept hierarchy, maintaining comparable performance. Jeong et al. (2022) present a supervised method for mixed-attribute data, selecting the most discriminative variables to optimise classification and handle data heterogeneity. A key limitation of both approaches is their focus on specific experimental contexts, without addressing the integration of heterogeneous sources. Culbertson et al. (2017) analysed a convenience sample of nine healthcare systems across different geographic sites. Results showed high completeness of patient demographic data for surname (99.96 %), given name (99.95 %), date of birth (98.82 %), sex (99.73 %), postal code (94.71 %), and complete address (94.65 %), with these key demographics consistently collected across all sites. This essential demographic data strengthens our research to select relevant attributes for patient classification. Several studies address duplicate detection in multi-source data; recent research has increasingly focused on probabilistic and de-identified matching techniques. Cardinal et al. (2023) examine de-identified data linkage in the absence of a common unique identifier, focusing on methods capable of tolerating recording errors. The system incorporates date of birth, given name, surname, and sex, and shows that anonymous Bayesian matching can reach high accuracy. However, its performance remains sensitive to data quality and may decline with very large or highly heterogeneous datasets. Forbes et al. (2024) employ Jaro-Winkler similarity to match records, though performance drops with poor or heterogeneous metadata. Wibowo et al. (2025) and Inan (2020) employ Levenshtein, Jaro-Winkler, and cosine similar-

ity measures to detect duplicate academic titles; however, these methods often fail to capture deeper semantic meanings.

Clustering is an unsupervised learning technique used to group patients with similar attributes (e.g., name, date of birth, email) based on pairwise similarity scores. Several methods can be applied to a patient similarity matrix, including hierarchical clustering (iteratively merging clusters until a predefined cluster count or similarity threshold is reached) (Lakshmi, 2025). DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based algorithm that identifies clusters based on data point density and effectively handling noise (Ramapo et al., 2024). In contrast, K-means is computationally efficient but less effective for non-linear clusters due to its assumption of spherical clusters (Bellavita et al., 2025).

3 PROPOSED METHOD FOR STANDARDISING AND INTEGRATING HETEROGENEOUS DATA

In this study, a systematic method for integrating heterogeneous data is employed, which is essential to ensure that analyses are conducted on a unified, coherent, and interpretable dataset. Data are collected from diverse sources and processed through multiple stages: initially, the data undergo cleaning and structural harmonisation (schemas, attribute names, table organisation), followed by value normalisation (formats, units, codings, handling of missing values) and attribute standardisation using a dynamic synonym table, enabling unification of names and values even when sources employ different terminologies or languages. To identify the most relevant variables for entity recognition, statistical analyses of attributes (variance and correlations) are performed, and similarity between entities is assessed using textual measures (Jaro-Winkler and Levenshtein distances), yielding global matching scores. Following these transformations, quality checks are conducted, including consistency, uniqueness, completeness, and adherence to the standard schema, before loading the data into a unified repository ready for analysis. This methodological approach forms the foundation of the study, ensuring comparability, reproducibility, and reliability of the results. Nevertheless, certain limitations are acknowledged, notably the risk of semantic information loss during transformations, dependency on the quality of the synonym table, and challenges posed by highly heterogeneous sources, which may influence the final interpretation of results.

3.1 Structure Normalisation Process

The process ensures XML data standardisation by correcting inconsistencies, removing redundancies, and verifying structural integrity through syntax parsing. **Algorithm 1** illustrates the transformation of heterogeneous XML files into normalised tree structures, with tag correction and encoding conversion, before storage via the XmlDocument DOM (Document Object Model) interface and formatting with the minidom library (**Figure 1**).

Algorithm 1 Normalise_structure_XML**Require:** XML files (input directory)**Ensure:** Normalised XML files (output directory)

```

1: for each file in input directory do
2:   if file extension is .xml then
3:     Parse(file)
4:     Format(file)
5:     Write(normalised file, output directory)
6:   end if
7: end for

```

▷ using standard encoding
 ▷ using *XmlDocument*, *minidom*

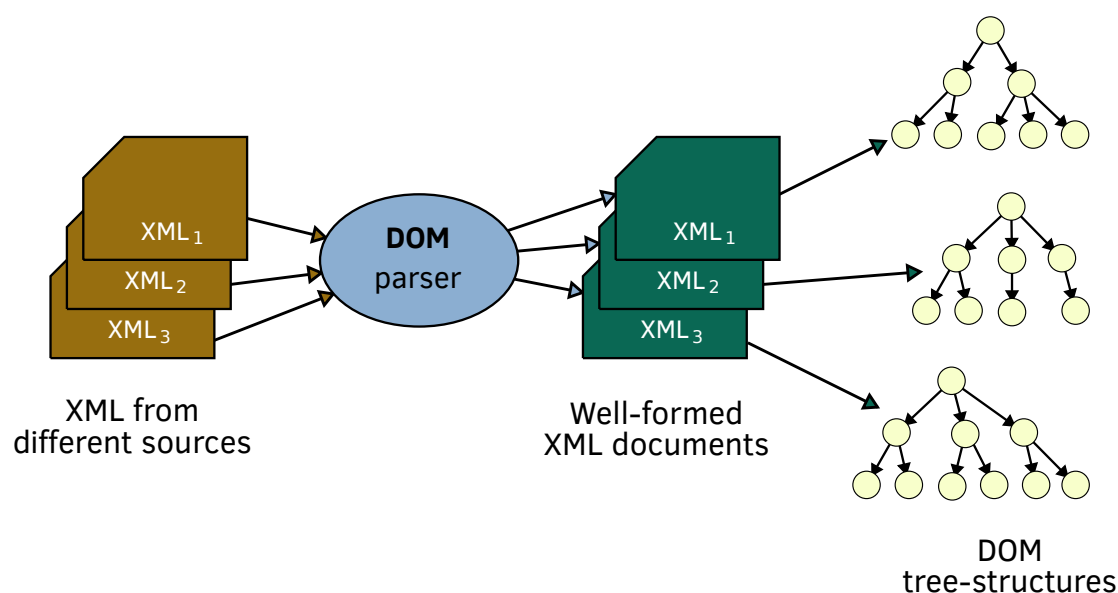


Figure 1: DOM tree structure

3.2 Value Normalisation Process

This process standardises XML data values as outlined in [Section 3.1](#). Text is converted to lowercase, extraneous whitespace and special characters are removed, and accented characters are replaced by unaccented forms. These steps ensure consistent formatting suitable for automated processing. [Figure 2](#) shows an example of the normalisation procedure.

3.3 Attributes Standardisation Process via a Dynamic Alternative Approach

A key element of our approach is the dynamic synonym table, which facilitates the standardisation of attribute names and their values across multilingual sources. Unlike fixed synonym lists,

```

<malades>
  <malade>
    <nin>?123@45</nin>
    <nom>DupôN?t_Je an</nom>
    <mail>dupont.jean@gmail.com</mail>
    <sexe>#Malé</sexe>
    <date_nais>19/11/1993</date_nais>
    <adresse>
      <rue>9 Anys W#eg</rue>
      <ville>Virginia</ville>
      <code_post>9430_</code_post>
    </adresse>
  </malade>
</malades>

```

Input: non-normalised XML

Output: normalised XML

```

<?xml version='1.0' encoding='utf-8'?>
<malades>
  <malade>
    <nin>12345</nin>
    <nom>dupont jean</nom>
    <mail>dupont.jean@gmail.com</mail>
    <sexe>male</sexe>
    <date_nais>1993-11-19</date_nais>
    <adresse>
      <rue>9 anys weg</rue>
      <ville>virginia</ville>
      <code_post>9430</code_post>
    </adresse>
  </malade>
</malades>

```

Figure 2: XML normalisation

this table is progressively enriched during the integration process. For example, the name attribute may appear as `<name>` (English), `<naam>` (Afrikaans), `<nom>` (French), or `<nombre>` (Spanish), and its values may differ across languages or conventions. By associating all these variants with a single canonical form, the table ensures consistent interpretation and improves the accuracy of similarity-based record linkage. To address the lack of a universal schema and synonym table for XML in healthcare, attributes normalised in Sections 3.1 and 3.2 are further standardised into a uniform XML format. A custom, flexible synonym table was developed

to map heterogeneous tags. Externally maintained (CSV, JSON, XML), it supports evolving structures and ensures adaptability. Unlike static approaches like XSLT, this method allows continuous updates, as illustrated in Table 1 and in Figure 3.

Table 1: Synonym table

Synonyms	Standard	Behaviour
malade, person, customer, ...	patient	standard
num, nin, mat, nss, ...	id	standard
naam, nom, name, ...	name	standard
naiss, dob, né, ...	birthDate	:%Y-%m-%d
mail, e-mail, ...	email	standard
seks, sexe, sex, sexo, ...	gender	standard
...	...	...

3.4 Customisation of heterogenous XML data

The previous sections addressed normalisation of XML structures and values, as well as standardisation of element and attribute names using a synonym table (e.g., patient, client, malade, person). This pre-processing step is essential before computing the correspondences between elements. In this research, the data sources to be integrated are heterogeneous: deterministic identifiers (e.g., national identification numbers or unique patient identifiers) are not consistent across the different XML sources originating from several institutions. Some files do not include them, others label them differently, and others still populate them incompletely or inaccurately. Consequently, a method relying solely on a unique identifier for record linkage would lead to false negatives (the same patients not identified) or false positives (identifiers shared by multiple patients). Employing a well-defined set of relevant attributes offers a credible, rigorous, and appropriate alternative to the absence of homogeneous deterministic identifiers, while preserving guarantees of quality and reproducibility in record linkage.

3.4.1 Selected attributes

The choice of attributes is therefore essential. It may be performed manually, according to the user's needs, or dynamically, through functions that automatically identify the most relevant attributes for entity recognition. Based on their availability in the heterogeneous XML files and after an initial assessment of data quality, the relevant attributes (e.g., name, date of birth, email, etc.) were selected for similarity computation.

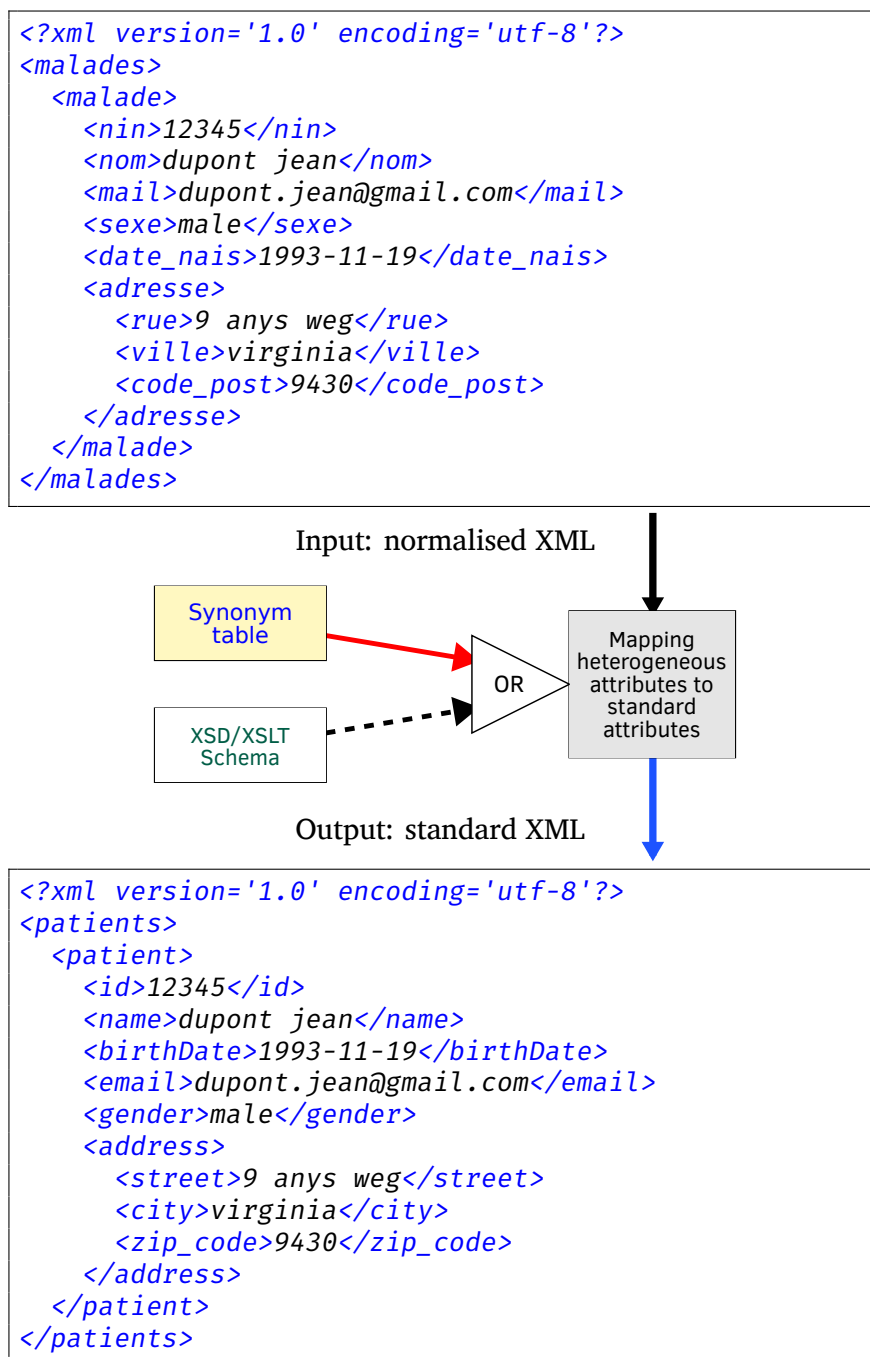


Figure 3: Attribute standardisation process

3.4.2 Relevant Attributes for Identification

Relevant attributes such as name, date of birth, and email address act as primary identifiers for a person due to their strong discriminatory power. Combined, they improve the accuracy of the match and make it more robust. Even if these fields contain spelling variations or typing errors, these discrepancies can be corrected using approximate matching techniques (fuzzy matching, Jaro distance, etc.). Based on this, each pair of records from different sources is compared on these attributes, a similarity score is calculated for each. Then these partial scores are weighted and combined into an overall score. If this overall score exceeds a set threshold, the records are considered identical. Finally, a clustering algorithm can group several equivalent records, allowing the merging of records for the same patient from separate sources.

3.4.3 Limitations and Partial Re-identification

The approach based on relevant attributes has inherent limitations, particularly when data are incomplete, erroneous, or ambiguous, which may lead to false negatives or false positives in record linkage. In situations where automatic matching remains uncertain, partial re-identification is employed: records can be temporarily flagged for manual verification or cross-checked against complementary sources before final merging. This strategy ensures the integrity of patient records while maintaining the efficiency of the automated linkage process.

3.4.4 Mathematical approach for dynamic attribute selection

After validating the normalisation, structuring, and standardisation processes, we identify the most relevant attributes for similarity calculation by applying appropriate similarity measures. Based on statistical and machine learning principles, this process selects the discriminative attributes by maximising variance and minimising correlation. This optimisation process is illustrated in Figure 4, which ensures that only the most informative attributes are retained.

Variance Calculation for Attributes

Variance is used to select the attributes that are relevant for computing similarity between patients. A high variance indicates that an attribute is discriminative and helpful in distinguishing patients from heterogeneous sources. In contrast, a low variance shows that the attribute provides little information and can be excluded. Combined with the other formulas, it contributes to calculating the distance between entities (patients). Mathematically, the variance of an attribute measures the spread of its values around their mean.

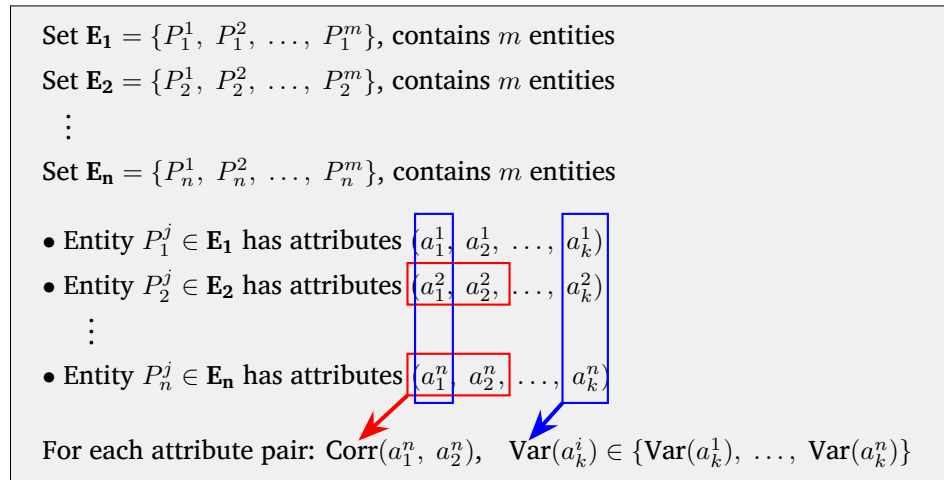


Figure 4: The relevance criteria of an attribute

Let a_i^j denote the value of the i -th attribute for entity j , over a set of n entities. The variance of attribute a_i is given by:

$$\text{Var}(a_i) = \frac{1}{n} \sum_{j=1}^n (a_i^j - \bar{a}_i)^2 \quad (1)$$

where:

- n : total number of entities
- a_i^j : value of attribute i for entity j
- $\bar{a}_i$: mean of attribute a_i across all entities, calculated as:

$$\bar{a}_i = \frac{1}{n} \sum_{j=1}^n a_i^j \quad (2)$$

Attribute Correlation

Attribute correlation measures the covariation between two attributes across entities, for example, age and weight of patients. In our heterogeneous data integration approach, this measure helps identify highly correlated attributes ($|\rho| \approx 1$), indicating redundant information. Such redundant attributes can be removed to reduce dimensionality and improve the efficiency of similarity computation between patients. Combined with variance and other indicators, this analysis ensures that only the most relevant attributes are used to group similar patients from different sources correctly.

The linear correlation between attributes a_i and a_k is:

$$\text{Corr}(a_i, a_k) = \frac{\sum_{j=1}^n (a_i^j - \bar{a}_i)(a_k^j - \bar{a}_k)}{\sqrt{\sum_{j=1}^n (a_i^j - \bar{a}_i)^2} \cdot \sqrt{\sum_{j=1}^n (a_k^j - \bar{a}_k)^2}} \quad (3)$$

where:

- n : total number of entities
- a_i^j, a_k^j : values of attributes a_i and a_k for entity j
- $\bar{a}_i, \bar{a}_k$: means of attributes a_i and a_k across all entities

$\text{Corr} = 1$ (perfect positive), $\text{Corr} = -1$ (perfect negative), $\text{Corr} = 0$ (no linear relation).

Example

For three patients with ages 25, 30, and 45, and corresponding weights 70, 75, and 90 kg, the mean age is:

$$\bar{a}_{\text{age}} = \frac{25 + 30 + 45}{3} \approx 33.33$$

and the variance is:

$$\text{Var}(a_{\text{age}}) = \frac{(25 - 33.33)^2 + (30 - 33.33)^2 + (45 - 33.33)^2}{3} \approx 72.2$$

This high variance indicates that the “age” attribute is informative for distinguishing patients. Computing the correlation between age and weight yields $|\rho| \approx 0.99$, showing near-perfect redundancy. In this case, one of the two attributes could be excluded to reduce redundancy while preserving discriminative power for similarity computation. Attributes with low variance or high correlation are disregarded, ensuring only the most informative attributes are used to compare patients from heterogeneous sources.

The proposed model selects relevant attributes from sets P_1 and P_2 by combining two criteria: maximising variance and minimising correlation with other attributes. For each attribute a_i , a relevance score $S(a_i)$ is computed as a weighted combination of these criteria:

$$S(a_i) = \alpha \cdot \text{Var}(a_i) - \beta \sum_{k=1, k \neq i}^n \text{Corr}(a_i, a_k) \quad (4)$$

where α and β are weights adjusted based on the relative importance of each criterion. This score quantifies the attribute’s usefulness for distinguishing patients: a high $S(a_i)$ indicates a

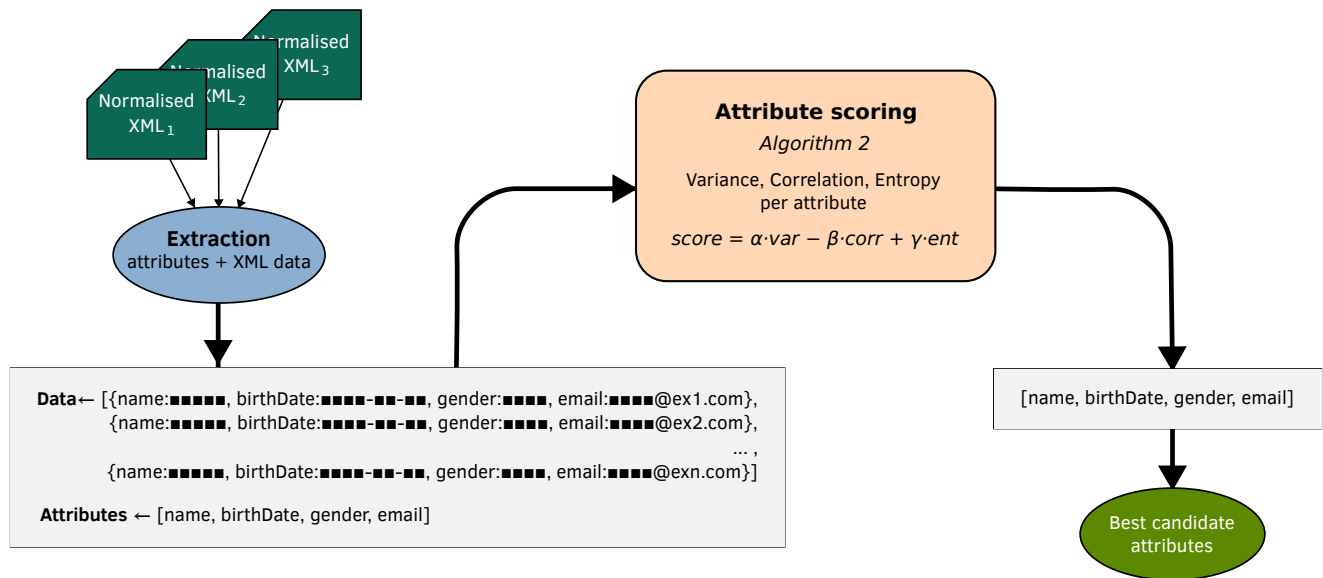


Figure 5: Select relevant attribute (details in Algorithm 2)

discriminative attribute that is retained for similarity computation, while attributes with low scores are excluded. This model can be implemented following the steps illustrated in Figure 5 and detailed in Algorithm 2.

Algorithm 2 Attribute Selection

Require: data, attributes

Ensure: Best candidate attributes

```

1: preprocess_data ← preprocess_data(data, attributes)           ▷ convert to usable format
2: for each attr ∈ attributes do                                  ▷ select candidate attributes
3:   values ← processed_data[attr]
4:   variance ← calculate_variance(values)
5:   entropy ← calculate_entropy(values)
6:   correlation_sum ← 0
7:   for each other_attr ∈ attributes do
8:     if attr ≠ other_attr then
9:       correlation_sum + ← calculate_correlation(values, processed_data[other_attr])
10:    end if
11:  end for
12:  scores[attr] ← α · variance − β · correlation_sum + γ · entropy  ▷ attribute score
13: end for
14: return arg maxattr scores[attr]
  
```

3.5 Global Similarity Model

The proposed approach computes the global similarity between entities by applying attribute-specific similarity functions to the relevant attributes identified in Section 3.4.1. The weighted measures are aggregated into an overall similarity score, which is then compared against a predefined threshold T to determine whether two entities are considered similar.

3.5.1 Calculate the similarity between relevant attributes

To compute the similarity $S_k(a_k^1, a_k^2)$ between each pair of candidate attributes, specific similarity functions are applied to each relevant attribute pair. Among these functions are:

Exact Distance

The Exact Distance function strictly compares two values of the same attribute (e.g., date of birth), returning 1 if they are identical and 0 otherwise. It is applied to stable attributes to ensure reliable matches between patients from different sources, thereby improving the accuracy of the overall similarity computation. The Exact Distance function is formally defined as:

$$S_{\text{dob}}(a_k^1, a_k^2) = \begin{cases} 1 & \text{if } a_k^1 = a_k^2 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

Levenshtein Distance Similarity

Levenshtein similarity is a normalised string metric ranging from 0 (dissimilar) to 1 (identical), defined as:

$$S_{\text{lev}}(a_k^1, a_k^2) = \left(1 - \frac{D_{\text{Lev}}(a_k^1, a_k^2)}{\max(\text{len}(a_k^1), \text{len}(a_k^2))} \right) \times 100 \quad (6)$$

where $D_{\text{Lev}}(a_k^1, a_k^2)$ is the minimum number of insertions, deletions, or substitutions required to transform one string into another. It compares textual attributes (e.g., name, address, email) across sources, tolerating minor differences and typos, enabling reliable matching of similar records.

Jaro-Winkler Distance

The Jaro-Winkler distance measures string similarity, emphasising shared prefixes (up to 4 characters). It is defined as:

$$S_{\text{JW}}(a_k^1, a_k^2) = \text{Jaro}(a_k^1, a_k^2) + \ell \cdot p \cdot (1 - \text{Jaro}(a_k^1, a_k^2)) \quad (7)$$

where ℓ is the common prefix length ($\ell \leq 4$) and p is a scaling factor ($0 \leq p \leq 1$). Values close to 1 indicate strong similarity. In heterogeneous data integration, Jaro-Winkler is

particularly effective for matching textual attributes such as patient names, capturing minor variations while giving additional weight to common prefixes, thereby enhancing the accuracy of cross-source record linkage.

3.5.2 Automatic Weighting of Relevant Attributes

An adaptive weighting function enhances the similarity model by reflecting each attribute's discriminative power based on its variance and correlation between entities $P_1^i \in E_1$ and $P_2^j \in E_2$:

$$w_k = \frac{\alpha}{1 + \text{Var}(a_k^1, a_k^2)} - \beta \cdot \text{Corr}(a_k^1, a_k^2) \quad (8)$$

where α and β are scaling parameters. This approach prioritises attributes that better distinguish records, improving overall matching accuracy.

3.5.3 Deriving Global Similarity Scores for Entity Matching

The mathematical model for calculating similarity between attributes proposed in Section 3.5.1 can be adapted to various datasets and attribute types. Let us consider two sets, E_1 and E_2 , corresponding to XML₁ and XML₂ in our research. These sets contain n and m entities (e.g., patients) respectively, such that:

$$\begin{aligned} E_1 &= \{P_1^1, P_1^2, \dots, P_1^n\} \\ E_2 &= \{P_2^1, P_2^2, \dots, P_2^m\} \end{aligned}$$

Each entity P_1^i or P_2^j is defined by a set of attributes such as name, date of birth, email, or other medical identifiers, denoted a_k^1 and a_k^2 . Each entity in E_1 or E_2 is characterised by a vector of attributes:

$$\begin{aligned} P_1^i &= (a_1^1, a_2^1, \dots, a_k^1) \\ P_2^j &= (a_1^2, a_2^2, \dots, a_k^2) \end{aligned}$$

where k denotes the total number of attributes, and each a_k^1 or a_k^2 corresponds to an attribute value within its entity (e.g., name = 'xxx', email = 'xxx@ex.com').

The overall similarity score between two entities $P_1^i \in E_1$ and $P_2^j \in E_2$ is computed as a weighted combination of the individual attribute similarities:

$$S_{\text{global}}(P_1^i, P_2^j) = \sum_{k=1}^N w_k S_k(a_k^1, a_k^2) \quad (9)$$

where N is the total number of relevant attributes, w_k is the weight assigned to attribute a_k according to its importance in identifying an entity (computed via Equation (8)), and $S_k(a_k^1, a_k^2)$ is the similarity score between the values of each relevant attribute pair (computed

via Equations (5) to (7)). The weights w_k reflect the discriminative power and reliability of each attribute, and are normalised such that:

$$\sum_{k=1}^N w_k = 1$$

Once the global similarity score has been calculated, a decision threshold T is applied to determine whether two entities P_1^i and P_2^j should be considered a match. The threshold T is defined empirically, typically based on validation experiments or domain-specific constraints. For example, if $T = 0.85$, only entity pairs satisfying $S_{\text{global}}(P_1^i, P_2^j) \geq 0.85$ are regarded as matches.

4 APPLICATION OF THE PROPOSED MATHEMATICAL APPROACH

The mathematical similarity model proposed in Section 3.4.1 can be adapted to diverse datasets and attribute types, including multi-patient XML files. Figure 6 illustrates how relevance criteria are adapted for each attribute, which then informs the computation of attribute-specific similarity metrics. These metrics are combined, using adjustable weights, into a global similarity score, enabling clustering of patients across multiple XML files. Both weights and thresholds can be tuned to meet specific application requirements.

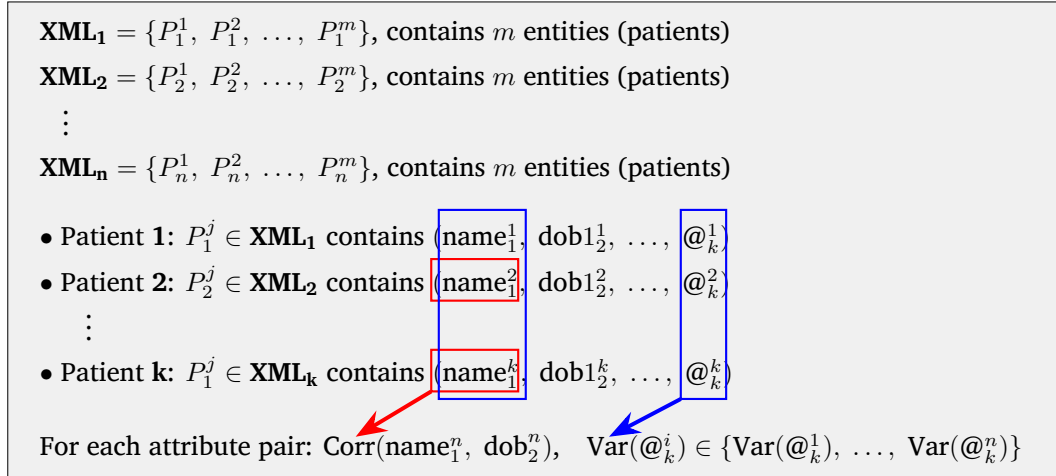


Figure 6: Relevance criteria of attributes on a set of patients

4.1 Implementation of a Mathematical Method for Discriminating Patient Attributes

The objective is to select a subset $A' \subseteq A$ of candidate attributes that maximises the scoring function $S(a_i)$ and enables the grouping of patients into “homogeneous subgroups”. Attribute

relevance depends on three criteria: discriminative power, redundancy, and information content, measured respectively by variance, correlation, and entropy. The steps of this selection are detailed in [Algorithm 2](#).

4.2 Implement the Mathematical Similarity Model

After selecting the relevant attributes, the similarity model described in [Section 3.5](#) can be applied to multiple XML files to compute the similarity between key attributes (such name, date of birth, gender, and email) of each patient in one XML file and all patients across the other XML files.

4.2.1 Similarity Between Two Attributes

For each pair of attributes a_k from two patients P_1^1 and P_2^1 , we define a similarity function $S_k(a_k^1, a_k^2)$ tailored to the specific attribute type. In our study, the selected attributes include name, date of birth, and email. Several similarity measures, as described in [Section 3.5.2](#), are applied:

- **Jaro-Winkler** for names: $S_{\text{name}}(a_1^1, a_1^2)$
- **Exact or tolerant matching** for dates of birth: $S_{\text{dob}}(a_2^1, a_2^2)$
- **Levenshtein distance** for emails: $S_{\text{email}}(a_3^1, a_3^2)$

Each function returns a similarity score in the range $[0, 1]$, where:

$$S_k(a_k^1, a_k^2) = \begin{cases} 1 & \text{identical values} \\ 0 & \text{entirely dissimilar values} \end{cases}$$

4.2.2 Global Similarity Between Two Patients

Global patient similarity is computed as a weighted sum of attribute similarities, for example:

$$S_{\text{global}}(P_1^1, P_2^1) = w_1 S_{\text{name}}(a_1^1, a_1^2) + w_2 S_{\text{dob}}(a_2^1, a_2^2) + w_3 S_{\text{email}}(a_3^1, a_3^2)$$

where w_1 , w_2 , and w_3 denote the weights for name, date of birth, and email respectively, and each w_k adjusts the corresponding attribute's influence (e.g., $w_3 > w_1$ if email is more discriminative than name). [Algorithm 3](#) details the computation of overall patient similarity.

Algorithm 3 Similarity Between Two Patients**Require:** XML documents, selected attributes (from Algorithm 2)**Ensure:** $S_{\text{global}}(P_1^i, P_2^j)$ ▷ Global similarity

```

1: Extract patients  $P_1^i \in E_1$  and  $P_2^j \in E_2$ 
2: for each attribute pair  $(a_k^1, a_k^2) \in (P_1^i, P_2^j)$  do
3:    $S_1(a_1^1, a_1^2) \leftarrow S_{\text{JW}}(\text{name}_1^1, \text{name}_1^2)$  ▷ Jaro-Winkler
4:    $S_2(a_2^1, a_2^2) \leftarrow S_{\text{exact}}(\text{dob}_1^1, \text{dob}_1^2)$  ▷ exact match
5:    $S_3(a_3^1, a_3^2) \leftarrow S_{\text{lev}}(\text{email}_1^1, \text{email}_1^2)$  ▷ Levenshtein
6: end for
7:  $S_{\text{global}}(P_1^i, P_2^j) \leftarrow 0$ 
8: for  $k \leftarrow 1$  to  $K$  do ▷  $K$ : number of selected attributes
9:    $S_{\text{global}}(P_1^i, P_2^j) \leftarrow S_{\text{global}}(P_1^i, P_2^j) + w_k \cdot S_k(a_k^1, a_k^2)$ 
10: end for
11: return  $S_{\text{global}}(P_1^i, P_2^j)$ 

```

Given a threshold $T = 0.8$, if $S_{\text{global}}(P_1^i, P_2^j) > T$, then $P_1^i \sim P_2^j$, and the two entities are considered a match.

4.3 Construction of extended similarity relationships

The proposed mathematical model can be extended to multiple datasets $\{E_1, E_2, E_3, \dots, E_n\}$, each containing multiple entities, using a similarity matrix. Suppose we have several data sources $\{\text{XML}_1, \text{XML}_2, \dots, \text{XML}_n\}$, each containing multiple patient records to be compared. Each source XML_i contains a set of patient records:

$$\text{XML}_i = \{P_i^1, P_i^2, \dots, P_i^m\}$$

Each patient P_i^j from XML_i is described by a set of attributes $\{a_k^i\}$, such as name, date of birth, email, and other medical identifiers. The problem is formalised in terms of distance and similarity computation across attributes. The proposed approach employs weighted similarity functions adapted to each relevant attribute to identify identical or similar patients and construct consistent subsets.

4.3.1 Construction of an Extended Multidimensional Matrix

Let us consider a set of patients originating from multiple XML sources XML_i . Specifically:

$$\begin{aligned}
 \text{XML}_1 &= \{P_1^1, P_1^2, \dots, P_1^m\} \\
 \text{XML}_2 &= \{P_2^1, P_2^2, \dots, P_2^m\} \\
 &\vdots \\
 \text{XML}_n &= \{P_n^1, P_n^2, \dots, P_n^m\}
 \end{aligned}$$

Each pair of patients (P_i, P_j) drawn from two sources XML_i and XML_j is described by two attribute vectors:

$$P_i = (a_1^i, a_2^i, \dots, a_n^i)$$

$$P_j = (a_1^j, a_2^j, \dots, a_m^j)$$

To construct subsets of similar patients, a global similarity score is computed for every possible pair across all XML sources, resulting in a total of $|XML_i| \times |XML_j|$ similarity scores to be calculated.

4.3.2 Similarity Matrix

The similarity matrix is built by computing a global score for each patient pair (P_i, P_j) using relevant attributes and similarity functions. For n patients, the number of comparisons is $\frac{n(n-1)}{2}$, ensuring that each patient is matched against all others across XML files. A weighted score is computed for each attribute $S_k(a_k^i, a_k^j)$, where $a_k^i \in P_i$ and $a_k^j \in P_j$. Figure 7 details the process for generating the similarity matrix.

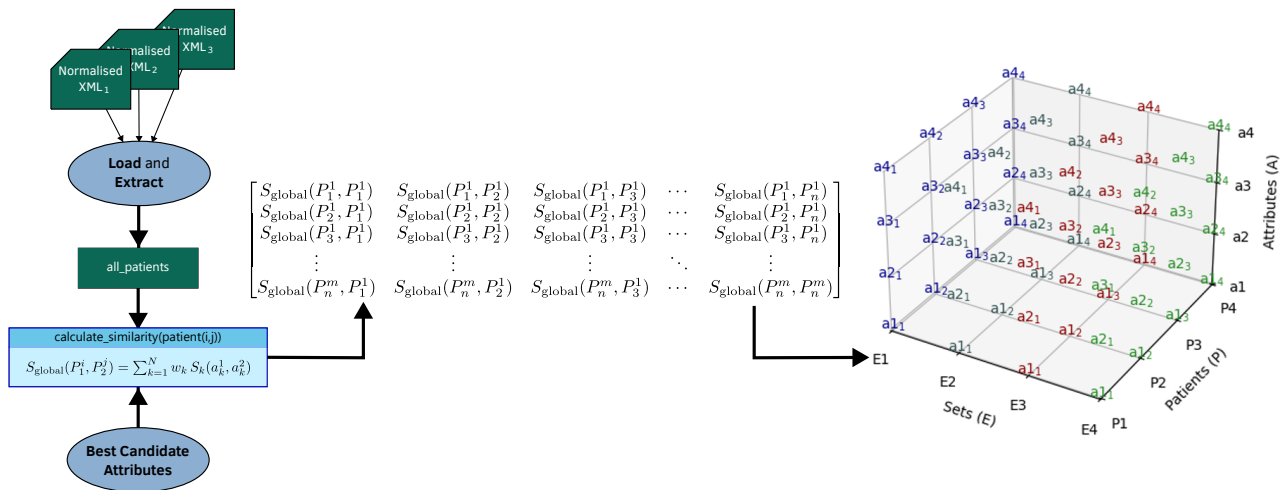


Figure 7: Process of Creating a Similarity Matrix using the S_{global} scores

The global similarity score between patients P_i and P_j is defined as:

$$S_{\text{global}}(P_i, P_j) = \sum_{k=1}^n w_k S_k(a_k^i, a_k^j)$$

where $S_{\text{global}} \in [0, 100]$, w_k is the weight of attribute a_k , and $S_k(a_k^i, a_k^j)$ is the similarity function between the values of attribute a_k for patients P_i and P_j .

4.4 Similar Patients Grouping Process

Patient data from multiple XML files can be integrated by clustering similar records using unsupervised learning. A global similarity score is computed for each patient pair. Density-based methods such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise) are applied to group patients with shared attributes without requiring a predefined number of clusters.

4.4.1 Cluster the similar data

DBSCAN is a density-based algorithm for identifying non-spherical clusters in similarity data. Applied to a similarity matrix of patient records, it groups instances according to their degree of similarity using two parameters: ϵ (neighbourhood distance) and $MinPts$ (the minimum number of points required to form a cluster). Similarity scores are converted into distances via:

$$d = 100 - S_{\text{global}}$$

before clustering. As shown in Figure 8, this process enables the automatic merging of records referring to the same patient despite minor attribute variations.

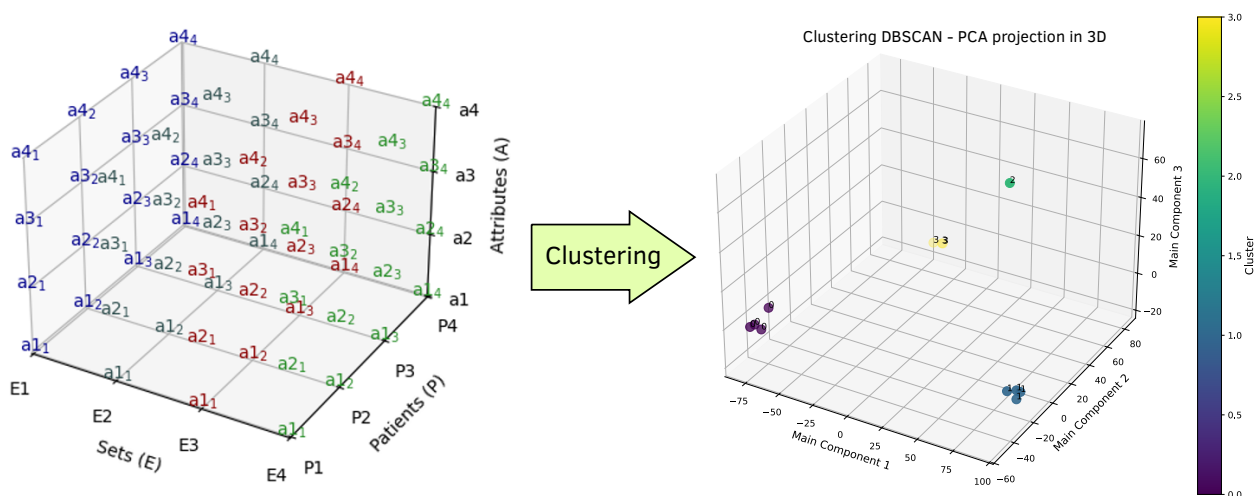


Figure 8: Groups of similar patients

4.4.2 Merge Uniform Clusters into A Single File

This step consolidates cluster-level XML files into a single unified file by assigning unique IDs to each cluster, merging similar patient records, standardising the XML structure, resolving conflicts by retaining the most complete data, filling in missing values from other records, and removing duplicates based on key attributes, ensuring that each patient appears only once in the final output (see Figure 9).

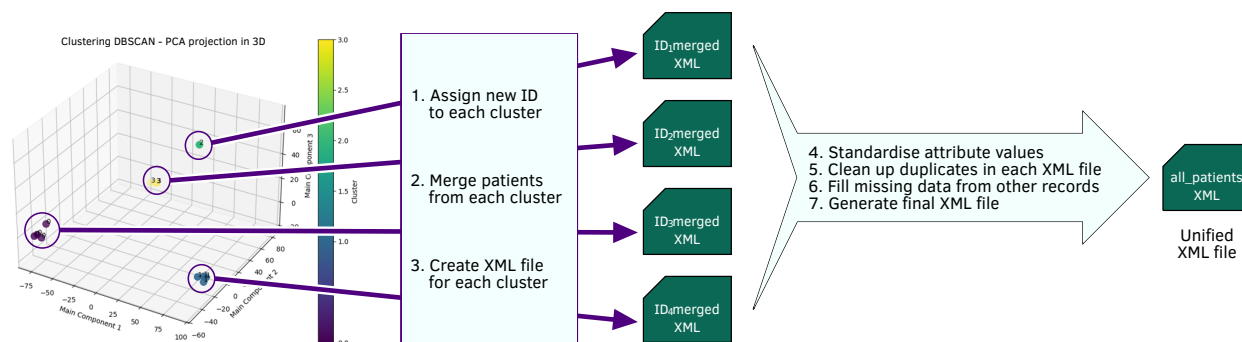


Figure 9: All clusters into a single final XML file

5 CASE STUDY

The study was conducted over a period of approximately nine months, during which the authors carried out all stages of data extraction, integration, and analysis independently. To address the objectives of this study, we adopted a quantitative paradigm focused on the systematic integration of patient data from multiple heterogeneous healthcare sources. A descriptive/experimental design was employed to evaluate the effectiveness of our integration approach and the similarity-based grouping of patient records. The study population consisted of electronic patient records from four healthcare facilities. One hundred records per source were used, ensuring sufficient diversity to assess the integration performance. Convenience sampling was applied to select patient records from each source's database. Data were collected by extracting the XML files provided. No direct interaction with patients occurred. The analysis performed during the pre-processing phase involved normalising both the structure and the values of the source files, followed by standardising the attributes through synonym matching based on a predefined synonym table. In the subsequent processing and integration phase, candidate attributes were selected to enable record linkage across heterogeneous sources, after which an algorithmic grouping procedure was applied. Based on calculated similarity scores applied to 10,000 comparisons between records, a similarity matrix was generated, and the DBSCAN clustering algorithm was employed to identify duplicate or similar patient records, eliminate redundancies, and complete missing data. All cleaned and updated record groups were then consolidated into a single standardised XML file. Ethical considerations were rigorously taken into account: all identifiers that could determine patients were anonymised. The overall system architecture is illustrated in Figure 10.

5.1 Phase 1: Pre-treatment

Phase 1 comprises Processes 0–4, including XML structuring, normalisation, semantic mapping using a synonym table, and validation, resulting in unified and harmonised XML files (Figure 11) for use in Phase 2.

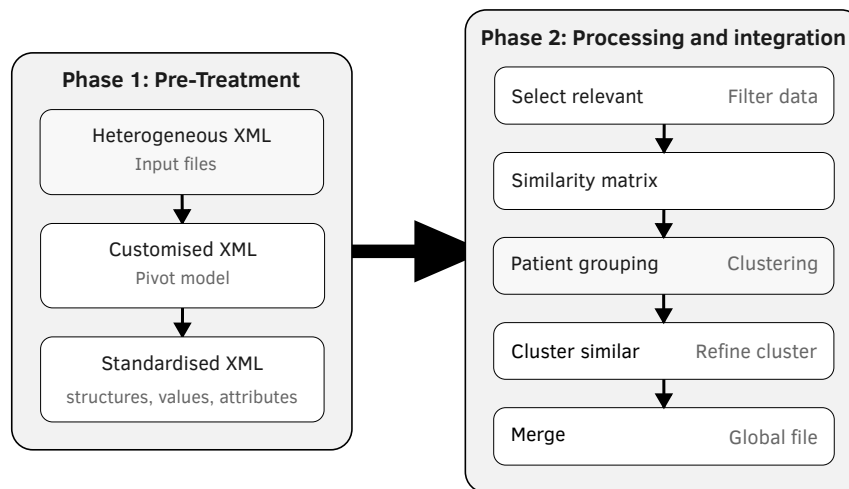


Figure 10: Global System Architecture Diagram

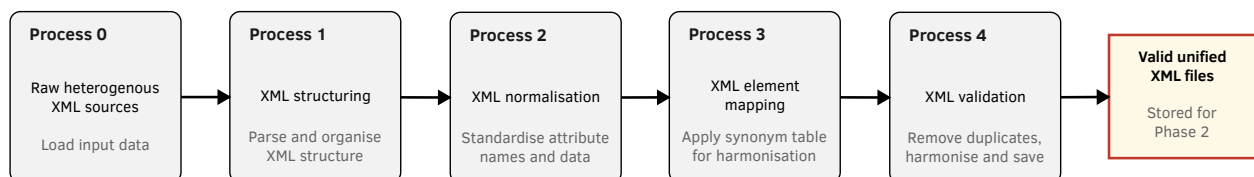


Figure 11: Phase 1: Pre-processing process for heterogeneous XML sources

5.2 Phase 2: Data processing and integration

Following Phase 1, in which the XML files were structured and the attributes and their values were standardised and normalised, Phase 2 comprises Processes 5–10 and is illustrated in [Figure 12](#) complemented by selected representative examples. In particular, Processes 6–8 are

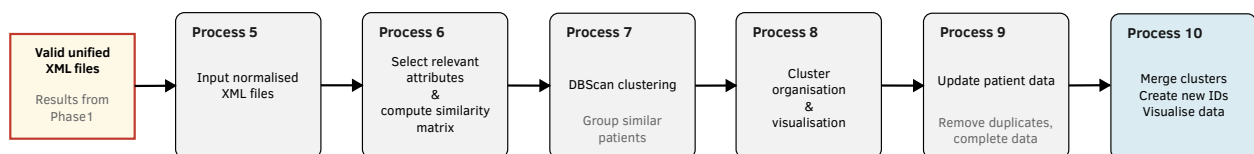


Figure 12: Phase 2: Data Processing and integration workflow

dedicated to the global similarity score computation (100×100), which uses the normalized XML files to construct the similarity matrix based on relevant attributes ([Figure 13\(a\)](#)). Subsequently, DBSCAN clustering groups similar patients from different XML sources based on a predefined similarity threshold ([Figure 13\(b\)](#)).

Processes 9–10: Update and standardise patient data by merging clustered nodes into a single XML file, retaining one occurrence per patient, completing missing information, and removing duplicates ([Figure 14\(a\)](#)). Subsequently, merge all standardised clusters into a

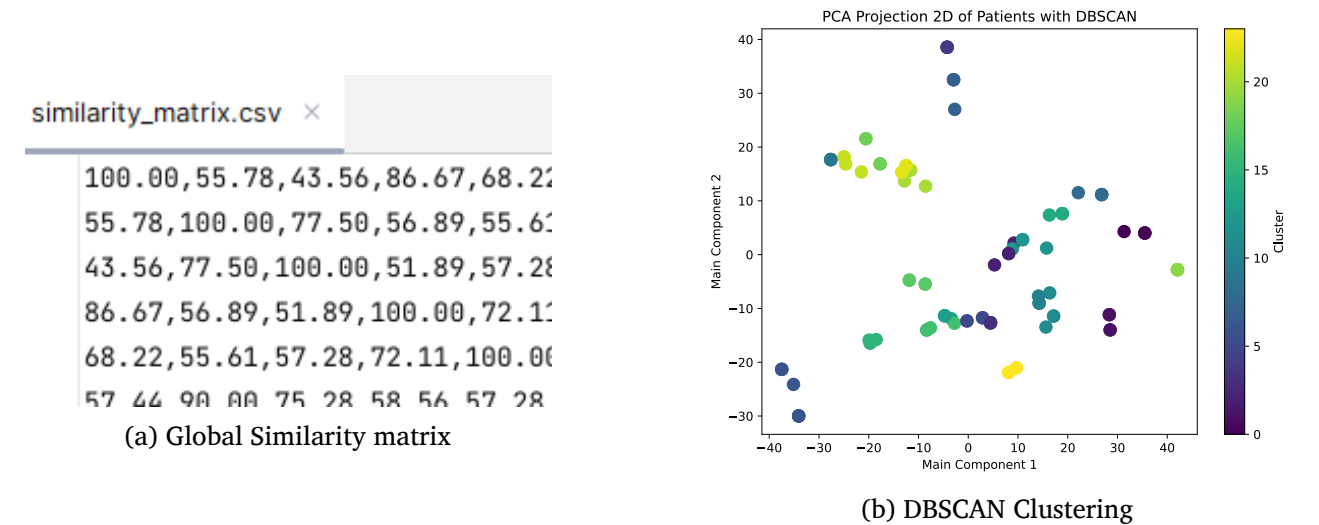


Figure 13: Processes 6–8

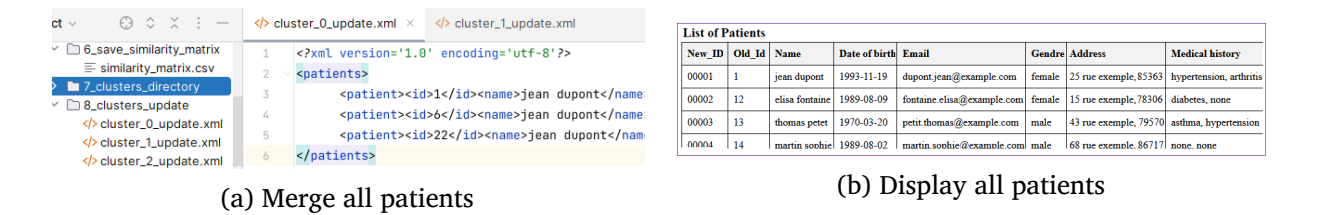


Figure 14: Processes 9–10

comprehensive XML file containing all identified patients, assign unique identifiers, and apply the display stylesheet (Figure 14(b)).

5.3 Analysis Method

Our method evaluates clustering and similarity through false positive/negative analysis. The similarity study was based on:

- 4 XML files (10,000 pairwise comparisons)
- Ground Truth matrix (Figure 15(a)). 9600 pairs are truly different, and 400 pairs truly similar.

The application of the analysis method is shown in Algorithm 4.

Algorithm 4 Evaluation Metrics Computation**Require:**

S_{truth} (ground-truth matrix),
 S_{global} (global similarity matrix),
 T (similarity threshold)

Ensure:

Precision,
Recall,
F1-score,
Accuracy,
Confusion matrix

```

1: for each  $(i, j) \in S_{\text{global}}$  do                                ▷ Binarise similarity matrix using threshold  $T$ 
2:   if  $S_{\text{global}}(i, j) \geq T$  then
3:      $\hat{S}(i, j) \leftarrow 1$                                        ▷ similar
4:   else
5:      $\hat{S}(i, j) \leftarrow 0$                                        ▷ different
6:   end if
7: end for
8: Flatten  $S_{\text{truth}}$  and  $\hat{S}$  into vectors
9: Compute confusion matrix entries:  $TP, FP, FN, TN$ 
10: Precision  $\leftarrow TP / (TP + FP)$                                 ▷ Calculate evaluation metrics
11: Recall  $\leftarrow TP / (TP + FN)$ 
12: F1-score  $\leftarrow (2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$ 
13: Accuracy  $\leftarrow (TP + TN) / (TP + FP + FN + TN)$ 
14: Construct the confusion matrix
15: Analyse errors ( $FP, FN$ )

```

5.3.1 Test Report 1

The first test report is based on the following parameters:

- Clustering (DBSCAN):
 - $\varepsilon = 20$ (i.e., $T \geq 80\%$)
 - $MinPts = 1$ (each patient can form its own group)
- Attribute weighting:
 - name: 0.50
 - date of birth: 0.25
 - email: 0.15
 - address: 0.10
- Ground truth matrix (Figure 15(a))

- Binary prediction matrix (Figure 15(b))

The class-wise performance of the model is summarised in Table 2.

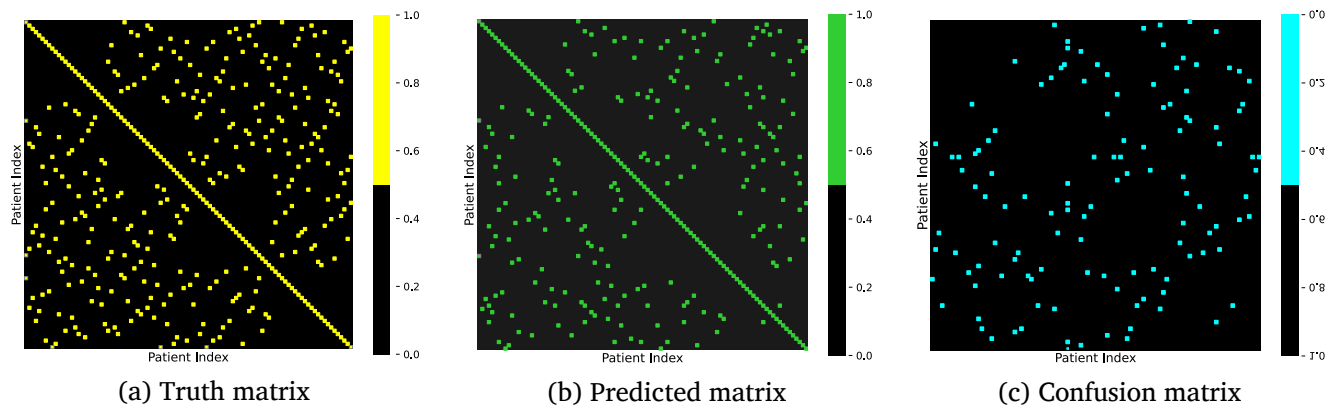


Figure 15: Report 1 (colours indicate similarity)

Table 2: Class-wise Performance (Report 1)

Class	Precision	Recall	F1-Score	Support
0 (different patients)	0.99	1.00	0.99	9600
1 (similar patients)	0.95	0.73	0.83	400

The confusion matrix is presented in Table 3 and illustrated in Figure 15(c). A recall of 73% for Class 1 means 106 truly similar patient pairs were not detected.

Table 3: Confusion Matrix (Report 1)

	Predicted Different	Predicted Similar
Actually Different	9584	16
Actually Similar	106	294

Evaluation of Test Report 1

The model achieves high overall accuracy (99%); however, its performance in detecting similar patients (Class 1) is limited, with a recall of 0.73 and 26.5% of true matches missed. Although precision is strong (95%), this comes at the expense of recall, resulting in missed duplicate records. The 80% similarity threshold ensures reliable avoidance of false matches but may cause true duplicates to be missed, possibly due to the high weight assigned to names

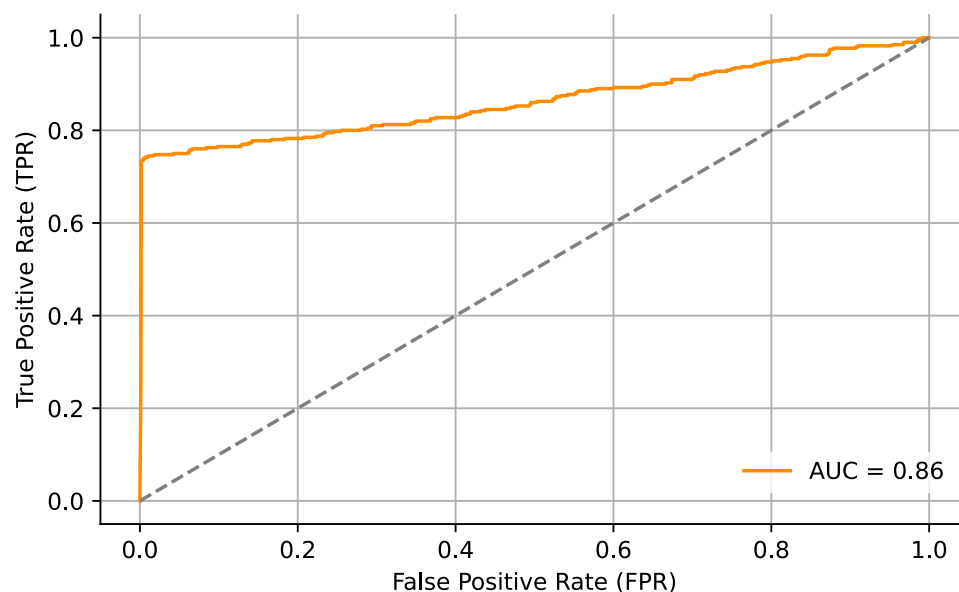


Figure 16: ROC curve, Report 1

(50%). **Figure 16** confirms an overall good performance (87%). Precision can be enhanced by cleaning and normalising data, applying higher similarity thresholds, and using adaptive attribute weighting.

5.3.2 Test Report 2

Clustering (DBSCAN): $\varepsilon = 10$ (i.e., $T \geq 90\%$), $MinPts = 1$ (each patient can form its own group). The parameters are as follows:

- Attribute weighting: name: 0.10, date of birth: 0.40, email: 0.40, address: 0.10
- Ground truth matrix (**Figure 15(a)**)
- Binary prediction matrix (**Figure 17(a)**)

The class-wise performance of the model is summarised in **Table 4**. The confusion matrix is

Table 4: Class-wise Performance (Report 2)

Class	Precision	Recall	F1-Score	Support
0 (different patients)	1.00	1.00	1.00	9600
1 (similar patients)	0.98	1.00	0.99	400

presented in **Table 5** and illustrated in **Figure 17(b)**.

Table 5: Confusion Matrix (Report 2)

	Predicted Different	Predicted Similar
Actually Different	9592	8
Actually Similar	0	400

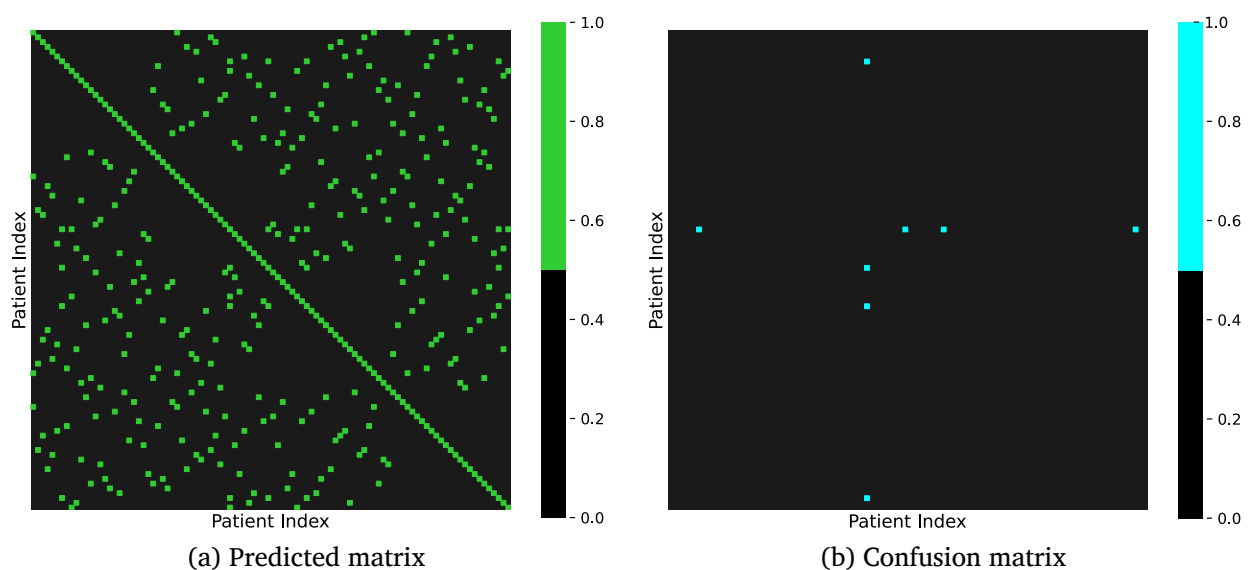


Figure 17: Report 2 (colours indicates similarity)

Evaluation of Test Report 2

In Report 2, satisfactory results were obtained by increasing the DBSCAN threshold to $\varepsilon = 10$ ($T \geq 90\%$) and adjusting the attribute weighting to name: 0.10, date of birth: 0.40, email: 0.40, address: 0.10. This yielded perfect recall (100%) for Class 1 with only eight false positives, resulting in Class 1 precision of 98%. Class 0 was perfectly predicted (100% precision and recall), offering greater reliability in detecting duplicates, as illustrated in Figure 18.

5.3.3 Comparative Summary of Report 1 and Report 2

Overall, the DBSCAN-based clustering system, with a strong emphasis on email and date of birth, demonstrated high reliability in grouping similar patients from heterogeneous sources while minimising false positives. These findings confirm its suitability for medical applications, particularly for patient deduplication and the integration of multi-source hospital databases, as further supported by the perfect performance curve shown in Figure 19 and the comparative summary presented in Table 6.

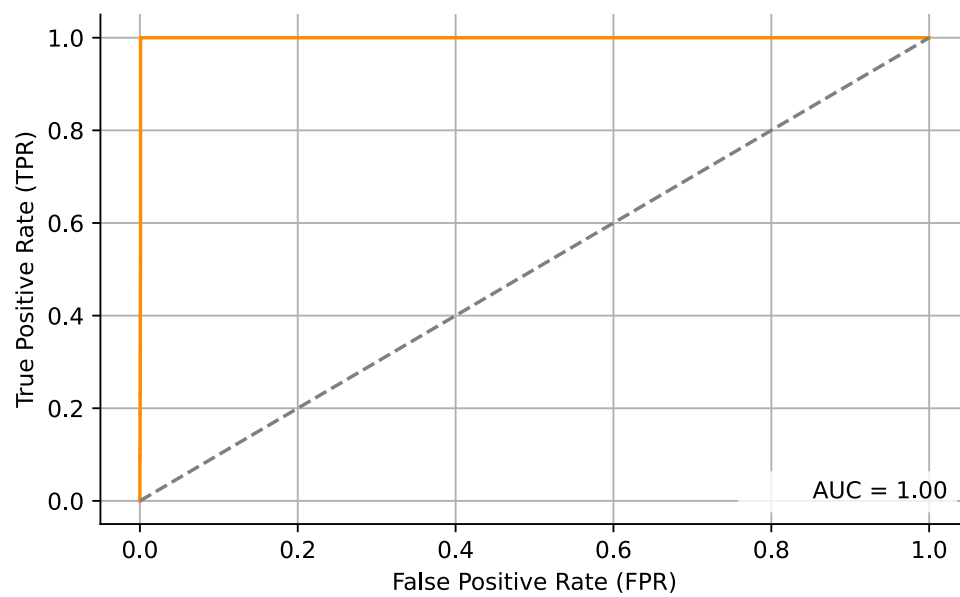


Figure 18: ROC curve, Report 2

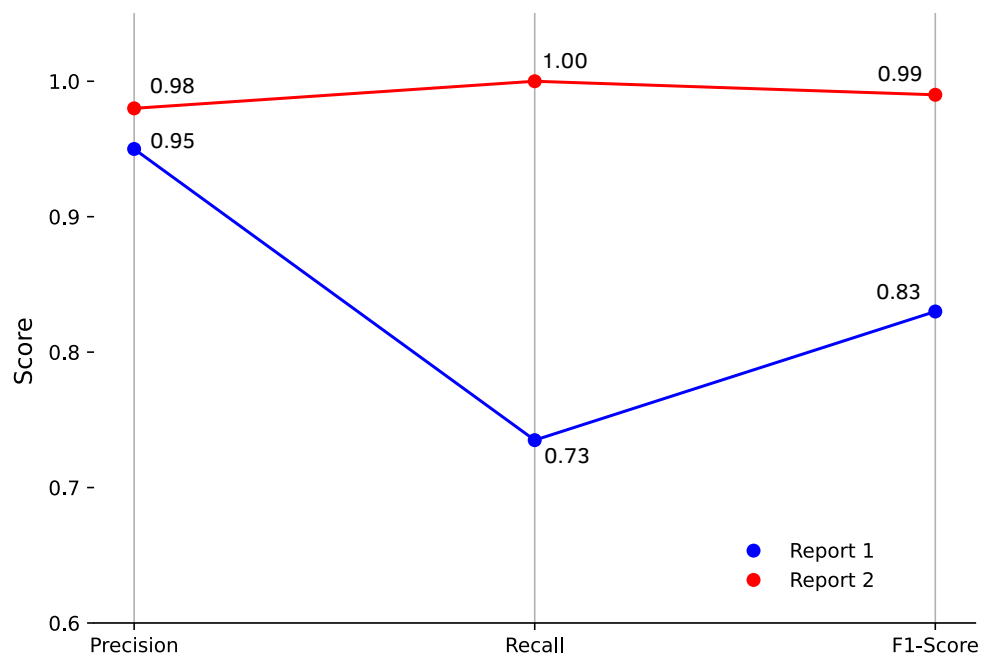


Figure 19: Comparative curve. Class 1 – Precision, Recall, FI-score

Table 6: Comparative Summary of Report 1 and Report 2

Criterion	Report 1 ($T \geq 80\%$, $\varepsilon = 20$)	Report 2 ($T \geq 90\%$, $\varepsilon = 10$)
Attribute weights	Name dominant (50%)	DoB & email dominant (40% each)
Precision (class 1)	0.95	0.98
Recall (class 1)	0.73	1.00
F1-score (class 1)	0.83	0.99
Overall accuracy	99%	100%
False negatives	106 false negatives	0 false negatives

6 DISCUSSION AND CONCLUSION

Analysis of false positives and negatives identified the primary sources of clustering errors and guided model adjustments. Minimising false positives requires increasing the similarity threshold, while reducing false negatives depends on adjusting attribute weights. Alternative clustering methods (e.g., K-means, agglomerative) were also considered to assess consistency. Overall, the model performs best when false positives are minimised, although recall can be improved by prioritising duplicate detection.

Jaro-Winkler is preferred for attributes such as name and address due to its emphasis on prefix similarity. In contrast, Levenshtein treats all edits equally and is less efficient for long strings, despite its effectiveness with short ones. To enhance match detection, we propose combining both metrics for the same attribute (e.g., name or address), as these fields often exhibit syntactic variations. A weighted combination of the two scores is computed as:

$$S_{\text{combined}}(a_k^1, a_k^2) = \alpha \cdot S_{\text{JW}}(a_k^1, a_k^2) + \beta \cdot S_{\text{Lev}}(a_k^1, a_k^2)$$

where $\alpha + \beta = 1$, and S_{JW} and S_{Lev} denote the Jaro-Winkler and Levenshtein scores respectively. An attribute pair is considered similar when:

$$S_{\text{combined}}(a_k^1, a_k^2) \geq T$$

The proposed approach has proven effective in grouping heterogeneous patient records; however, several challenges remain. To overcome these challenges, future work will focus on:

1. Integrating a machine learning-based normaliser or semi-supervised techniques to suggest corrections or standardised values based on validated examples.
2. Enriching synonym resources using domain-specific ontologies (e.g., SNOMED CT: Systematised Nomenclature of Medicine Clinical Terms (Stearns et al., 2001); LOINC: Logical Observation Identifiers Names and Codes (McDonald et al., 2003), introduce a dynamic synonym-expansion (Word2Vec (Mikolov et al., 2013), FastText (Bojanowski et al., 2017)) module that automatically learns new lexical and regional variants from data or domain ontologies, reducing dependence on static lists and limiting false negatives.

3. Exploring embedding-based models to capture semantic similarity and address multilingual challenges. Unsupervised clustering algorithms (K-means and DBSCAN) are used to automatically group similar terms. Missing data could be addressed through advanced imputation techniques, while neural network-based methods may optimise similarity computation. Alternative clustering algorithms, such as HDBSCAN or hybrid models, should also be explored to improve grouping performance.

To mitigate confidentiality and regulatory constraints while enabling, at least partially, an effective multi-source data fusion, several directions may inform future research or methodological enhancements to our approach. The combined use of federated learning and encryption or secure computation techniques would allow data from multiple institutions to be exploited without sharing raw information, an approach already shown to be viable in certain medical contexts, particularly in imaging. Adopting a privacy-by-design strategy would further strengthen protection by enforcing restricted access policies, clearly defined roles, and informed consent. Finally, the generation of synthetic data would provide a complementary alternative by reproducing the statistical structure of real datasets without containing identifiers, thereby reducing privacy risks. Together, these orientations would support partial or secure multi-source integration while limiting regulatory constraints and the risk of re-identification. By incorporating these enhancements, the methodology could be extended beyond healthcare to other domains requiring efficient integration of heterogeneous data sources.

This study demonstrates that, through mathematical modelling, it is possible to overcome the traditional challenges posed by heterogeneity in health data. Within this framework, we show that an electronic medical record no longer needs to be restricted to a single data type or a specific format; instead, it can evolve into a comprehensive, flexible, and scalable platform that integrates clinical information, biological data, sensor outputs, medical histories, and more, without relying on strict identification keys. This development paves the way for a truly ubiquitous health system, more integrated, more intelligent, more secure, and more respectful of data diversity, while preserving confidentiality. What distinguishes our approach is its generalisability: it is not limited to a particular use case. Still, it can be adapted to numerous healthcare settings, diverse data sources, and, potentially, extended to national or cross-institutional scales.

7 CONTRIBUTION

This paper proposes a structured methodology for integrating and deduplicating heterogeneous XML-based healthcare records lacking unique identifiers. The study introduces an XML normalisation framework to standardise files by unifying attribute structures and values, supported by a synonym table to handle lexical variations (e.g., “customer,” “person,” “patient”), ensuring semantic consistency across sources. A similarity computation strategy is developed to identify relevant attributes and construct a similarity matrix, which serves as the foundation for grouping similar patient records originating from different sources. The DBSCAN

algorithm is employed to cluster patient records without requiring a predefined number of clusters, enabling the detection of groups of varying sizes, with parameter tuning to control clustering quality and reduce over-segmentation or dispersion. A post-clustering harmonisation step refines cluster consistency by resolving redundancies and correcting minor attribute inconsistencies, thereby improving the overall quality of grouped records. The resulting clusters are then consolidated into a single unified XML file, reducing duplication and enhancing data structuring while maintaining compatibility with downstream applications. An end-to-end data processing pipeline orchestrates all stages of the methodology, from input data selection and pre-processing to similarity computation, clustering, harmonisation, and final data representation, ensuring reproducibility and facilitating practical application. In summary, the proposed methodology improves interoperability and data quality by consolidating heterogeneous XML sources into a unified and consistent database, reducing redundancy, enhancing structural coherence, and providing a reliable foundation for information exchange and advanced analysis.

References

- Adapa, C. S. R. (2025). Eliminating duplicate medical records: How modern solutions are revolutionizing healthcare data management. *Journal of Computer Science and Technology Studies*, 7(2), 496–505. <https://doi.org/10.32996/jcsts.2025.7.2.52>
- Adeyelu, A. A., & Anyebe, E. O. (2019). Implementing an improved mediator wrapper paradigm for heterogeneous database integration. *Nigerian Annals of Pure and Applied Sciences*, 1, 224–235. Retrieved January 25, 2026, from <https://napas.org.ng/index.php/napas/article/view/29>
- Albayrak, O. S., Aytekin, T., & Kalaycı, T. A. (2022). Duplicate product record detection engine for e-commerce platforms. *Expert Systems with Applications*, 193, 116420. <https://doi.org/10.1016/j.eswa.2021.116420>
- Alluhaidan, S. A., Pushparaj, Subbappa, A., Mishra, P. V., Chandrika, V., Vaish, A., & Sengupta, S. (2023). An automatic threshold selection using ALO for healthcare duplicate record detection with reciprocal neuro-fuzzy inference system. *Computers, Materials & Continua*, 74(3), 5821–5836. <https://doi.org/10.32604/cmc.2023.033995>
- Baum, L., Johns, M., Poikela, M., Möller, R., Ananthasubramaniam, B., & Prasser, F. (2023). Data integration and analysis for circadian medicine. *Acta Physiologica*, 237(4), e13951. <https://doi.org/10.1111/apha.13951>
- Bellavita, J., Pasquali, T., Del Rio Martin, L., Vella, F., & Guidi, G. (2025). Popcorn: Accelerating kernel k-means on GPUs through sparse linear algebra. *Proceedings of the 30th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming*, 426–440. <https://doi.org/10.1145/3710848.3710887>
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with sub-word information. *Transactions of the Association for Computational Linguistics*, 5, 135–146. https://doi.org/10.1162/TACL_A_00051

- Buron, M., Goasdoué, F., Manolescu, I., & Mugnier, M.-L. (2020). Obi-wan: Ontology-based RDF integration of heterogeneous data. *Proceedings of the VLDB Endowment*, 13(12), 2933–2936. <https://doi.org/10.14778/3415478.3415512>
- Cardinal, R. N., Moore, A., Burchell, M., & Lewis, J. R. (2023). De-identified Bayesian personal identity matching for privacy-preserving record linkage despite errors: Development and validation. *BMC Medical Informatics and Decision Making*, 23(1), 85. <https://doi.org/10.1186/s12911-023-02176-6>
- Chen, L., Qi, Y., Wu, A., Deng, L., & Jiang, T. (2023). Mapping chinese medical entities to the unified medical language system. *Health Data Science*, 3, 0011. <https://doi.org/10.34133/hds.0011>
- Chen, Z., Wang, Y., Narasayya, V., & Chaudhuri, S. (2019). Customizable and scalable fuzzy join for big data. *Proceedings of the VLDB Endowment*, 12(12), 2106–2117. <https://doi.org/10.14778/3352063.3352128>
- Chuma, K. G. (2025). ‘A tough nut to crack’: Inconsistent standards as roadblocks to data interoperability of health information systems in public hospitals in the Gauteng province of South Africa. *Oxford Open Digital Health*, 3, oqaf013. <https://doi.org/10.1093/oodh/oqaf013>
- Culbertson, A., Goel, S., Madden, M., Safaeinili, N., Jackson, K., Carton, T., Waitman, R., Liu, M., Krishnamurthy, A., Hall, L., Cappella, N., Visweswaran, S., Becich, M., Applegate, R., Bernstam, E., Rothman, R., Matheny, M., Lipori, G., Bian, J., ... Kho, A. (2017). The building blocks of inter-operability: A multisite analysis of patient demographic attributes available for matching. *Applied Clinical Informatics*, 08(2), 322–336. <https://doi.org/10.4338/ACI-2016-11-RA-0196>
- Dhayne, H., Haque, R., Kilany, R., & Taher, Y. (2019). In search of big medical data integration solutions – A comprehensive survey. *IEEE Access*, 7, 91265–91290. <https://doi.org/10.1109/ACCESS.2019.2927491>
- d’Ussel, M., Sacco, E., Moreau, N., Nizard, J., & Durand, G. (2024). Assessment of decision-making autonomy in chronic pain patients: A pilot study. *BMC Medical Ethics*, 25(1), 97. <https://doi.org/10.1186/s12910-024-01096-y>
- Forbes, C., Greenwood, H., Carter, M., & Clark, J. (2024). Automation of duplicate record detection for systematic reviews: Deduplicator. *Systematic Reviews*, 13(1), 206. <https://doi.org/10.1186/s13643-024-02619-9>
- Grannis, S. J., Williams, J. L., Kasthuri, S., Murray, M., & Xu, H. (2022). Evaluation of real-world referential and probabilistic patient matching to advance patient identification strategy. *Journal of the American Medical Informatics Association*, 29(8), 1409–1415. <https://doi.org/10.1093/jamia/ocac068>
- Iglesias, E., Vidal, M.-E., Collarana, D., & Chaves-Fraga, D. (2024). Empowering the SDM-RDFizer tool for scaling up to complex knowledge graph creation pipelines¹ (M. Alam & R. Verborgh, Eds.). *Semantic Web*, 1–28. <https://doi.org/10.3233/SW-243580>

- Inan, E. (2020). SimiT: A text similarity method using lexicon and dependency representations. *New Generation Computing*, 38(3), 509–530. <https://doi.org/10.1007/s00354-020-00099-8>
- Jeong, D. H., Jeong, B. K., Leslie, N., Kamhoua, C., & Ji, S. (2022). Designing a supervised feature selection technique for mixed attribute data analysis. *Machine Learning with Applications*, 10, 100431. <https://doi.org/10.1016/j.mlwa.2022.100431>
- Kafe, E. (2023). Mapping Wordnets on the fly with permanent sense keys. *Proceedings of the 12th Global Wordnet Conference*, 67–76. <https://doi.org/10.18653/V1/2023.GWC-1.8>
- Lakshmi, G. V. (2025). Implementation of agglomerative hierarchical clustering using single linkage. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 11(1), 724–727. <https://doi.org/10.32628/CSEIT251115>
- Li, C., Zeng, Q., Duan, H., & Xie, N. (2019). A mapping-based sharing approach for heterogeneous XML data supporting user personalization. *IEEE Access*, 7, 76789–76805. <https://doi.org/10.1109/ACCESS.2019.2922310>
- Ling, C., Ng, K., & Santoso, H. A. (2025). Mapping relational database to full-text XML for open journal system cross-platform article distribution. *Journal of Informatics and Web Engineering*, 4(3), 259–277. <https://doi.org/10.33093/jiwe.2025.4.3.16>
- Liu, S., Wang, X., Hou, Y., Li, G., Wang, H., Xu, H., Xiang, Y., & Tang, B. (2023). Multimodal data matters: Language model pre-training over structured and unstructured electronic health records. *IEEE Journal of Biomedical and Health Informatics*, 27(1), 504–514. <https://doi.org/10.1109/JBHI.2022.3217810>
- McDonald, C. J., Huff, S. M., Suico, J. G., Hill, G., Leavelle, D., Aller, R., Forrey, A., Mercer, K., DeMoor, G., Hook, J., Williams, W., Case, J., & Maloney, P. (2003). LOINC, a universal standard for identifying laboratory observations: A 5-year update. *Clinical Chemistry*, 49(4), 624–633. <https://doi.org/10.1373/49.4.624>
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 3111–3119.
- Mukhitova, A., Yerimbetova, A., & Barakhnin, V. (2024). Model of JSON to XML format conversion based on XSD data structure extraction. *TRUDY UNIVERSITETA*, 4(97). https://doi.org/10.52209/1609-1825_2024_4_394
- Mutemaringa, T., Heekes, A., Smith, M., Boulle, A., & Tiffin, N. (2023). Record linkage for routinely collected health data in an African health information exchange. *International Journal of Population Data Science*, 8(1). <https://doi.org/10.23889/ijpds.v8i1.1771>
- Oemig, F., & Blobel, B. (2022). Modeling digital health systems to foster interoperability. *Frontiers in Medicine*, 9, 896670. <https://doi.org/10.3389/FMED.2022.896670>
- Ormesher, I. (2024). *Duplicate detection with GenAI* (tech. rep.). ZEISS Digital Partners. arXiv. <https://doi.org/10.48550/ARXIV.2406.15483>
- Phan, A. C., & Phan, T. C. (2022). Similarity algorithms for fuzzy join computation in big data processing environment. *Journal of Computer Science and Cybernetics*, 39(2), 101–124. <https://doi.org/10.15625/1813-9663/17589>

- Putrama, I. M., & Martinek, P. (2024). Heterogeneous data integration: Challenges and opportunities. *Data in Brief*, 56, 110853. <https://doi.org/10.1016/j.dib.2024.110853>
- Ramapo, Kim, Y., Vasarhelyi, M., & Rutgers University. (2024). Anomaly detection with the density based spatial clustering of applications with noise (DBSCAN) to detect potentially fraudulent wire transfers. *The International Journal of Digital Accounting Research*, 55–91. https://doi.org/10.4192/1577-8517-v24_3
- Rambau, T. M., Munyoka, W., & Mashau, N. L. (2025). Identifying factors affecting the adoption of big data analytics in South African rural-based hospitals to improve service delivery. *South African Computer Journal*, 37(1). <https://doi.org/10.18489/sacjv37i1/17830>
- Roehrs, A., Da Costa, C. A., Da Rosa Righi, R., Rigo, S. J., & Wichman, M. H. (2019). Toward a model for personal health record interoperability. *IEEE Journal of Biomedical and Health Informatics*, 23(2), 867–873. <https://doi.org/10.1109/JBHI.2018.2836138>
- Rognan, S. E., Jørgensen, M. J., Mathiesen, L., Druedahl, L. C., Lie, H. B., Bengtsson, K., Andersson, Y., & Sporrang, S. K. (2023). ‘The way you talk, do I have a choice?’ Patient narratives of medication decision-making during hospitalization. *International Journal of Qualitative Studies on Health and Well-being*, 18(1), 2250084. <https://doi.org/10.1080/17482631.2023.2250084>
- Saelee, J., Wetchapram, P., Wanichsombat, A., Intarasit, A., Muangprathub, J., Boongasame, L., & Choopradit, B. (2024). Enhanced feature selection via hierarchical concept modeling. *Applied Sciences*, 14(23), 10965. <https://doi.org/10.3390/app142310965>
- Song, E., & Haw, S. (2020). XML-REG: Transforming XML into relational using hybrid-based mapping approach. *IEEE Access*, 8, 177623–177639. <https://doi.org/10.1109/ACCESS.2020.3026006>
- Stearns, M. Q., Price, C., Spackman, K. A., & Wang, A. Y. (2001). SNOMED clinical terms: Overview of the development process and project status. *Proceedings of the AMIA Symposium*, 662. <https://pmc.ncbi.nlm.nih.gov/articles/PMC2243297/>
- Steinkamp, J., Kantrowitz, J. J., & Airan-Javia, S. (2022). Prevalence and sources of duplicate information in the electronic medical record. *JAMA Network Open*, 5(9), e2233348. <https://doi.org/10.1001/jamanetworkopen.2022.33348>
- Welbotha, D., & van den Berg, C. (2024). Big data analytics capabilities and the organisational performance of South African retailers. *South African Computer Journal*, 36(2). <https://doi.org/10.18489/sacj.v36i2/17937>
- Westermann, T., Ramonat, M., Hujer, J., Gehlhoff, F., & Fay, A. (2026). Automatic mapping of automationml files to ontologies for graph queries and validation. *IEEE Open Journal of the Industrial Electronics Society*, 7, 589–607. <https://doi.org/10.1109/OJIES.2026.3672995>
- Wibowo, F. M., Nafan, M. Z., Gustalika, M. A., Fernando, H., Hussain, M., & Sahadun, N. A. (2025). Lightweight string similarity approaches for duplicate detection in academic titles. *Journal of Informatics and Web Engineering*, 4(3), 416–426. <https://doi.org/10.33093/jiwe.2025.4.3.25>

- Xu, P., & Lu, J. (2019). Towards a unified framework for string similarity joins. *Proceedings of the VLDB Endowment*, 12(11), 1289–1302. <https://doi.org/10.14778/3342263.3342268>
- Zhang, J., Kowsari, K., Harrison, J. H., Lobo, J. M., & Barnes, L. E. (2018). Patient2Vec: A personalized interpretable deep representation of the longitudinal electronic health record. *IEEE Access*, 6, 65333–65346. <https://doi.org/10.1109/ACCESS.2018.2875677>