SACJ 38(1) July 2026
Research Article

A system for linking and integrating heterogeneous datasets in health and demographic surveillance systems

Joseph Tlouyamma[a, e], Sello Mokwena[a], Cairo Bruce Ntimana[e], Masenyani Oupa Mbombi[c], Kagiso Peace Seakamela[e], Peter Mphekgwana[d], Katlego Mothapo[e], Reneilwe Given Mashaba[e], Eric Maimela[b]

[a] Department of Computer Science, University of Limpopo, Polokwane, South Africa
[b] Department of Public Health, University of Limpopo, Polokwane, South Africa
[c] Department of Nursing Science, University of Limpopo, Polokwane, South Africa
[d] Research Administration and Development, University of Limpopo, Polokwane, South Africa
[e] DIMAMO Population Health Research Centre, University of Limpopo, Polokwane, South Africa

ABSTRACT
To overcome the constraints of data silos within data-driven organizations, the linking and integration of data from several sources becomes even more crucial. Data linking and integration not only improve the precision of the analysis, but also discover hidden trends, patterns, and relationships that can be overlooked when working with separate datasets. Taking this into account, this study is aimed at linking and integrating data across disparate data sources in the Health and Demographic Surveillance System (HDSS). Data linkage was achieved by implementing a fuzzy pattern-matching algorithm (Levenshtein distance) written in structured query language (SQL). A user-friendly graphical user interface (GUI) was developed using the C# programming language to facilitate the process of data matching and linking across various datasets. The Pentaho data integration (PDI) tool facilitated the integration of data from various data sources into a common database. These open-source technologies facilitated efficient linking and integration of data, enabling researchers to holistically discover patterns and trends in the data. The proposed algorithm achieved 98.1% precision and recall of 88.6% while maintaining 93.0% of good balance between avoiding false matches and missed matches.

Keywords data linking, data integration, fuzzy pattern matching, data heterogeneity, structured query language, similarity index

Categories
- **Information systems** → **Data management systems** → **Information integration**
- *Applied Computing* → *Life and Medical Sciences* → *Healthcare information systems*

Email
Cairo Bruce Ntimana – cairo.ntimane@ul.ac.za (CORRESPONDING)

Article history
Received: 22 October 2024
Accepted: 30 November 2025
Online: 31 July 2026

Tlouyamma, J., Mokwena, S., Ntimana, C.B., Mbombi, M.O., Seakamela, K.P., Mphekgwana, P., Mothapo, K., Mashaba, R.G., Maimela, E. (2026). A system for linking and integrating heterogeneous datasets in health and demographic surveillance systems. *South African Computer Journal* 38(1), 22–42. https://doi.org/10.18489/sacjv38i1/20408

Copyright © the author(s); published under a Creative Commons NonCommercial 4.0 License
SACJ is a publication of *SAICSIT*. ISSN 1015-7999 (print) ISSN 2313-7835 (online)

1 INTRODUCTION

The proliferation of most recent data collection methodologies in Health and Demographic Surveillance Systems (HDSS) has led to the generation of a substantial number of fragmented and heterogeneous datasets. The semantic and syntactic heterogeneity of datasets and data fragmentation across multiple data sources present a significant challenge for analysis, especially with each dataset having distinct identifiers that refer to the same participant (Che Fauzi et al., 2021; Taki & Ardakani, 2021; Ubdaidillah & Ahmad, 2020; H. Wang & Wang, 2022). Leveraging data to conduct a comprehensive analysis of disease progression or multimorbidity is significantly impeded by the presence of uncommon identifiers for the same participant. For this reason, it is essential to establish a link between heterogeneous data sources (Salvadori et al., 2020) to maintain a global health perspective on a particular participant. It involves distinguishing and linking data across multiple databases, facilitating comprehensive data consolidation and integration for holistic analysis, thus simplifying the investigation and modelling of phenomena (Canalle et al., 2020; Thomas et al., 2020). Data linkage involves the acquisition, manipulation, and fusion of data from distinct structures safely and efficiently, and supplies data in a standardised format for the development of identical identifiers (Schildhauer, 2017). Linking data sources with disparate identifiers into a unified data file facilitates analysis and enhances overall service performance (Canalle et al., 2020).

Unlinked data pose a multitude of issues of bias, erroneous results, and a decline in the study's credibility (Presutti et al., 2011). To minimise such issues, the application of common identifiers across various data sources becomes significant. The lack of a common identifier that connects many pieces of data on a certain subject makes linking that data impractical and imprecise (Han & Lahiri, 2018). Connected data applications may encounter quality problems with unlinked data, including inconsistency, improperness, obsolescence, and incompleteness (Rula et al., 2016). The quality of linkage is critical in record linkage studies as it determines the extent to which linked records represent the same individual across different databases. The use of data linkage methods provides sensitivity analyses to evaluate the robustness of linked results, which can further help assess the impact of potential errors. There are several data linking methods, such as key matching, record linkage, entity resolutions, geospatial matching, machine learning-based linkage, probabilistic and deterministic matching (Comber & Arribas-Bel, 2019; Shlomo, 2018; Zhang et al., 2018). The predominant methodologies employed in health surveillance research encompass both probabilistic and deterministic matching linkage. Harmony of application of these methods enhances the efficiency and precision of record linking, thus contributing to a more effective and accurate integration of health-related data (Harron, 2022). It is necessary to evaluate the efficiency of record linking methodologies to effectively deal with challenges associated with the analysis of syntactic and semantic heterogeneous data.

Although prior research has demonstrated the efficacy of data linkage in improving the quality of health research (Harron et al., 2016), there is still a gap in the literature regarding the use of the Levenshtein distance algorithm to link multiple population-level health

https://doi.org/10.18489/sacjv38i1/20408

datasets. Most studies concentrated on the results of data linking, such as the precise identification of health trends or the estimation of disease patterns, rather than the performance of the data linkage process itself (Dusetzina et al., 2014). Consequently, there is a dearth of literature that investigates the effectiveness of fuzzy matching algorithms in linking and integrating data into a unified source of truth. In addition, the lack of consistent identification complicates cross-source participant identification, reducing the effectiveness of research efforts aimed at assessing health trends from heterogeneous data sources. This study implements and evaluates the Levenshtein distance algorithm in linking disparate health datasets. The approach seeks to provide methodological insights to influence future health research and practice. We hypothesize that the Levenshtein distance algorithm can effectively and precisely link records from semantic and syntactic heterogeneous datasets into a single source of truth from which a comprehensive data analysis on multimorbidity can be performed. This study aims to contribute innovative solutions to the challenges associated with linking heterogeneous HDSS datasets, which could ultimately advance the capabilities of health research and decision-making processes.

The remainder of the paper is structured as follows: The subsequent section delves into the literature regarding data integration and linking methodologies. Following this, the methodology is presented, which delineates the methodological procedures that were implemented during the research. Subsequently, the results were analysed to determine the effectiveness of the Levenshtein distance algorithm in designating scores and facilitating data linking. Afterward, the discussion, implications of the study, future research, and conclusion are presented.

2 LITERATURE REVIEW

Health data sometimes exhibits anomalies, including typographical mistakes, spelling variances, missing fields, or disparate data formats. Fuzzy matching algorithms tackle these issues by permitting approximate matches instead of perfect ones, therefore enhancing the probability of accurately connecting data that pertains to the same people or thing despite slight inconsistencies (Lauzanne et al., 2022). Fuzzy matching may be calibrated for precision, balancing the likelihood of false positives and negatives based on the application's requirements (Lauzanne et al., 2022). String matching for data linkage often requires extensive preparation (e.g., normalising names, addressing non-unique records) to minimise false positives and enhance match rates (McCormack & Smyth, 2017).

Several fuzzy matching algorithms exist, including Damerau distance, Dice coefficient, cosine similarity, longest common sequence (LCS), and Levenshtein distance (Pikies & Ali, 2021), and many others. Table 1 compares various fuzzy matching algorithms based on their key features and strengths in matching records. The implementation of these algorithms has the potential to enhance the keyword categorization ratio by 69% to 78% (Pikies & Ali, 2021).

Table 1 indicates that fuzzy logic-based algorithms, weighted fuzzy pattern matching, fuzzy preference relations, fuzzy graph matching, and hybrid fuzzy and machine learning are inadequate for structured text matching with typological errors, whereas the string similarity

Table 1: Comparison of main fuzzy matching algorithm types, features, and applications

Algorithm Type	Approach or Key Feature	Application or strength
Fuzzy logic-based algorithms (Lin et al., 2024; X. Wang et al., 2023)		
	Employ fuzzy logic principles and membership functions to address uncertainty and imprecision.	The process of making decisions based on ambiguous data, including map matching
Weighted fuzzy pattern matching (Dubois et al., 1993)		
	Allocates weights to pattern components and employs fuzzy set operations for matching	Expert systems, retrieval with incomplete or vague data
String similarity algorithm (Dubois et al., 1993)		
	Algorithms like Levenshtein, LCS, Dice, Cosine, and Damerau for string comparison	Record linkage, text matching, error-tolerant search
Fuzzy preference relations (Lin et al., 2024; Pikies & Ali, 2021)		
	Incorporates self-confidence and preference degree in relations	Decision-making, two-sided matching
Fuzzy graph matching (Khalili et al., 2021; Li et al., 2019)		
	Achieve optimisation by utilising fuzzy graphs and matching numbers	Graph-based matching problems, network optimization
Hybrid fuzzy and machine learning (Rudwan & Fonou-Dombeu, 2023)		
	Integrates fuzzy string matching with machine learning models	Semantic similarity detection, ontology matching

algorithm, such as Levenshtein distance, excels in this area. The Levenshtein distance is often favoured for its simplicity, clarity, and efficacy in quantifying string similarity. The Levenshtein distance quantifies the least number of single-character modifications (insertions, deletions, replacements) required to convert one string into another, making its outcomes straightforward and comprehensible (Alsmadi et al., 2024). According to these authors, the Levenshtein distance algorithm is extensively used in text clustering, translation memory, and record linkage, often surpassing other string-matching methodologies in efficacy and precision. Numerous studies use Levenshtein distance as a benchmark to assess the efficacy of more intricate or domain-specific algorithms (Kaur et al., 2022). To evaluate the quality of the linking algorithm, many metrics are taken into account, including specificity, sensitivity, false positive rate, false negative rate, positive predictive value, and negative predictive value (Doğan, 2019).

Diverse approaches and technologies have been developed to allow data linking and integration (Mountantonakis, 2021). These technological tools streamline the processes of extracting, transforming, and loading data, particularly when the data originates from several dissimilar sources. However, it is critical to acknowledge, meanwhile, that most of the current methodologies and strategies have been refined only for a restricted set of data sources, which limits their relevance (Kruse et al., 2020). In addition, strategies exist that encompass the creation of linked data records (Murray & Rivas, 2020) through the application of matching criteria, the generation of pairs of records, and the execution of partitioning procedures. Moreover, graphical user interfaces (GUIs) have been created to facilitate and simplify the linking and combination of data sets (Doukari et al., 2017). This enables users to merge data sets

using various types of connections, including their relationships. A graphical user interface (GUI) may contribute to a more precise and effective merging procedure (Alsubhi & Munson, 2016). The integration of a GUI helps optimise the merging procedure through the provision of user-friendly interfaces and tools that enable efficient manipulation and merging of data from various data sources.

For data integration in HDSS, the utilisation of business intelligence tools, such as Statistical Analysis Software (SAS), has been emphasised (Amek et al., 2015). SAS facilitates HDSS data integration by performing data cleansing, transformation, and integration operations (Dasari & Suresh, 2013). Another approach to data integration involves the use of the Pentaho data integration (PDI) tool. PDI facilitates the coordination and execution of jobs, as well as the development of transformations. PDI streamlines the procedures involved in integrating and transforming data obtained from diverse data sources (Barahama & Wardani, 2021). The tool offers a graphical user interface with a drag-and-drop environment that facilitates data capture, manipulation, cleaning, transfer, integration, and loading (Roldán, 2013). By loading data into a data warehouse using PDI, it is possible to facilitate planning, decision-making, and assessment (Salaki et al., 2016).

In preclinical research, where correlations must be established between multimodal data from many sources, relational databases facilitate the correlation of results across distinct data sets (Dasararaju & Taori, 2019). This capability is very critical. The integration of data sets generated from many tables enables analysts to obtain a holistic perspective of the data under examination. Integration efforts require novel and more intricate methodologies and strategies (Karabekiroglu et al., 2006). The linked and integrated data stored in a relational database can facilitate efficient data management and analysis. Implementing a relational database as a pre-analysis storage solution presents several benefits that improve the effectiveness, consistency, and adaptability of data management and analysis procedures (Wesley, 2000). Additionally, data can be easily reconstructed and analyzed to uncover hidden trends or patterns.

3 METHODOLOGICAL PROCEDURES

3.1 Study Setting

This research utilized data sourced from the Dikgale, Mamabolo, and Mothiba tribes, collectively constituting the DIMAMO HDSS. Established in 1995, the HDSS is located in the Capricorn district of Limpopo Province, South Africa. The HDSS collects data on socio-economic, demographic, verbal autopsies, vital events, and health concerns from rural communities. Other studies were also conducted within the HDSS surveillance area to collect additional data on mental health, cardiovascular risk factors, genomics, patient clinic data, sexual reproductive health, and dementia. The surveillance area encompasses 59 villages, classified as rural and semi-rural, and is inhabited by individuals with low socioeconomic status, high unemployment rates, limited educational background, and a notable prevalence of hypertension

and other chronic diseases.

3.2 Data Sources

The HDSS retains a collection of data sources consisting of numerous datasets that were previously acquired through various methods. These include, among others, data from projects that investigated mental health, clinic patient health, sexual reproductive health, dementia, environmental and genetic risk factors, and health and demographics (refer to Table 2). These datasets offer a diverse array of health-related information that could yield significant benefits if linked. The datasets encompass extensive health-related information that, when linked and integrated, could provide significant insights into multimorbidity and other health-related issues.

Table 2: Description of data sources

Projects	**Duration**	**Size of data**
DIMAMO database – Health and Socio-demographic surveys	Twice a year, telephonic data collection. Once a year, physical data collection (1995 – ongoing)	∼115,000 individuals
Sexual Reproductive Health and Rights (SRHR)	Six months (Jul – Dec 2022)	900 individuals (550 females, 350 males)
Africa Wits in-depth study of genetic and environmental factors for cardio-metabolic risk (AWI-Gen)	2015 and 2022	∼1400 individuals $\geq$ 40 years. Cardio-metabolic risk, genomics.
Strengthening Responses to dementia in developing countries (STRIDE)	Oct – Dec 2021	∼400 elderly individuals $\geq$ 60 years
Mental health study (Investigating the factors associated with mental health needs for pregnant and parenting adolescents)	Jan 2022 -– Dec 2023	500 participants
Clinic patient health (collection of data from the clinic patient files)	Mar 2023 to date (ongoing)	8000 patients

https://doi.org/10.18489/sacjv38i1/20408

3.3 How data linkage was achieved

3.3.1 System workflow

This section explains the data linking methodologies employed in this study. Figure 1 illustrates the processes involved in linking heterogeneous data sources. The C#-based Windows application was created to enhance data linkage functionalities through an intuitive user interface. The application facilitated the simultaneous comparison and assessment of records in two datasets. Embedded within the application was a fuzzy matching algorithm (Levenshtein distance) that computed and determined the degree of similarity between records from different datasets. The records from the external data source were compared with the records in the DIMAMO operational database to ascertain the necessity of a linkage. Linkage was conducted for entries in both datasets that were verified to belong to the same participant. The DIMAMO database served as a benchmark for linking, containing record identifiers utilised to connect records from external data sources. The fuzzy matching algorithm compared records from an external data source with those in the DIMAMO operational database to assess their similarity. This algorithm generated a similarity score to determine the degree of similarity.

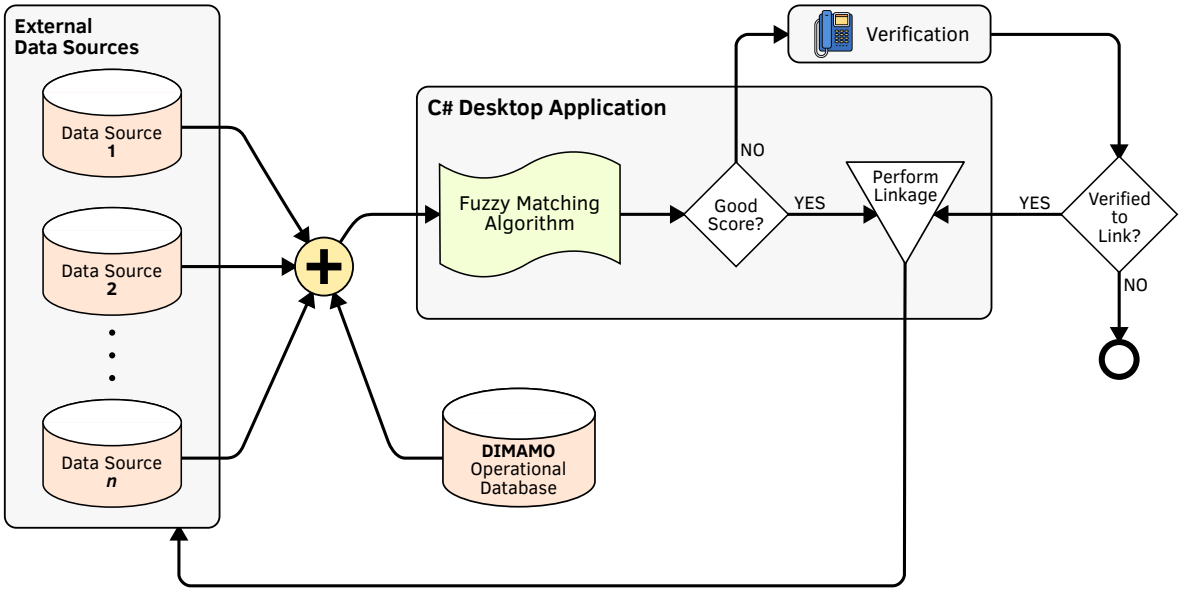

Figure 1: Schematic representation of the data link processes

Record linkage was conducted exclusively for records yielding scores between 80% and 100%. Records at the borderline (70% to 79%) underwent telephonic verification. The verification process involved directly contacting the participants whose records showed some degree of similarity, necessitating confirmation from the participants. Upon proper verification, the records were linked; otherwise, they remained unlinked. The linking and verification process was made possible through a Windows desktop application, which assigned a unique common identifier to the records that were successfully merged or linked. As determined by a higher

similarity index or verified via a telephonic call, a unique identifier lookup was conducted using the DIMAMO database to update external data sources, for records to maintain a common identifier designated as DSID.

3.3.2 A Fuzzy Logic Algorithm (Levenshtein distance)

Health and demographic data are distinguished by typographical errors and formatting variations, which result in unmatched data when exact string matching is performed. The traditional methods that match or link data based on equality operations are unable to accurately detect variations in the data resulting from manual data entry. This study employed a fuzzy matching algorithm known as Levenshtein distance to link heterogeneous data sources, thereby accounting for variations or errors in the data. The algorithm computes the minimum number of substitutions, deletions, and insertions necessary to convert one string into another by quantifying the difference between the two strings (s_1 and s_2). Fuzzy matching algorithms are specifically engineered to compare strings that possess some degree of similarity but do not match precisely.

Consider two strings denoted by s_1 and s_2 with lengths m and n, respectively. The edit distance between these two strings can be expressed as:

$$\text{Dist}(l,k) = \begin{cases} \max(l,k), & \text{if } \min(l,k) = 0 \\ \min \begin{cases} \text{Dist}(l-1,k)+1 \\ \text{Dist}(l,k-1)+1 \\ \text{Dist}(l-1,k-1)+1_{s_1[l] \neq s_2[k]} \end{cases}, & \text{otherwise} \end{cases}$$

where $\text{Dist}(l,k)$ is the distance between the first l characters of a string s_1 and the first k characters of a string s_2. The normalized similarity index or score can be written as:

$$\text{Sim}(s_1,s_2) = 1 - \frac{\text{Dist}(|s_1|,|s_2|)}{\max(|s_1|,|s_2|)}$$

The similarity score between s_1 and s_2 has a range of 0 to 1, with 1 indicating an exact match and 0 indicating dissimilarity between the two strings. To ascertain the likelihood of a match between s_1 and s_2, a threshold of 0.80 was selected. The decision was made to maintain a balance between true and deceptive matches and to enable the identification of ambiguous cases for manual review.

3.3.3 Matching on personal details

Personal details (names, date of birth, and gender) were compared across heterogeneous datasets for record linkage. The names variable included *surname*, *name1*, and *name2*. Date of birth was standardised by formatting it to *YYYY-MM-DD*. The gender values in the datasets were standardised as *Female* and *Male*. The relative descriptive power of each variable and

exploratory testing were used to assign a weight to each variable. For example, $w_{\text{name}} = 0.6$, $w_{\text{dob}} = 0.3$ and $w_{\text{gender}} = 0.1$. The composite similarity score was computed as follows:

$$\text{ComSim} = w_{\text{name}} \cdot \text{Sim}_{\text{name}}(s_1, s_2) + w_{\text{dob}} \cdot \text{Sim}_{\text{dob}}(s_1, s_2) + w_{\text{gender}} \cdot \text{Sim}_{\text{gender}}(s_1, s_2)$$

3.3.4 Matching on phone number and national identification number

The data was standardized by removing white spaces, country codes, and non-numeric characters between the numbers before comparison. This process was necessary to avoid a mismatch. Since these numbers serve as unique identifiers, the equality match was performed for s_1 and s_2 containing numbers (phone or national ID numbers) with a threshold of 90%. This threshold was selected to account for phone and national ID numbers with minor variations. The perfect match between two records would provide a definitive detail of the same individual. A definitive determination of the same individual would be provided by a precise match between two records. $\text{Sim}_{\text{phonenumber}}(s_1, s_2)$ and $\text{Sim}_{\text{IDnumber}}(s_1, s_2)$ were also taken into account, but separately, due to the possibility of errors in the capture of these numbers.

3.3.5 Data Extraction, Transformation, and Loading (ETL)

The datasets shown in Table 2 originate from several sources, complicating the execution of substantive data analysis. Therefore, it was essential to consolidate and combine the data into a unified source of truth. The Pentaho data integration (PDI) tool was used to extract, transform, and load the data from a variety of sources into an SQL server-based database that provided an enterprise-wide perspective of the data. The database consisted of tables used to store the data from various sources and can be mathematically defined as:

$$\text{DB} = \{\text{R}_1, \text{R}_2, \text{R}_3, \ldots, \text{R}_n\}$$

For each R_i representing the relation or a table with a set of columns C_i.

For every relation R_i in DB, there exists a primary key ($\text{PK}_i \subseteq \text{C}_i$) such that

$$\forall t_1, t_2 \in R_i : \left(\bigwedge_{c \in PK_i} t_1 \cdot c = t_2 \cdot c \right) \implies t_1 = t_2$$

This means that, in a particular table, if two rows have the same primary key values, then those two records should be merged into one row, to represent an entry for one individual. The tables are related to each other using a pair of keys. These are primary (PK_i) and foreign keys (FK_i). So, if $\text{FK}_{i \to j} \subseteq \text{C}_i$ in R_i, which references PK_j in a different table (R_j), then $\forall t \in \text{R}_i,\ \exists k \in \text{R}_j$ so that

$$\bigwedge_{c \in \text{FK}_{i \to j}, d \in \text{PK}_j} (t \cdot c = k \cdot d)$$

https://doi.org/10.18489/sacjv38i1/20408

$FK_{i\to j}$ reads as the foreign key from the table i to table j. This ensures that, for every foreign key ($FK_{i\to j}$) in the child table, there should exist a corresponding primary key (PK_j), maintaining referential integrity. Linking two tables (R_i and R_j) can be represented as

$$R_i \bowtie_{FK_{i\to j}=PK_j} R_j$$

This ensures that, all tables holding data from difference data sources can reconstructed to extract meaningful data for analysis. The overall dataset schema for holding integrated data from variety of sources is defined as:

$$DB_{multimorbidity} = \{(R_i, PK_i, \{FK_{i\to j}\}_j) | i = 1, 2, \ldots, n\}$$

Next, we seek define the process of extracting, transforming and loading the data into $DB_{multimorbidity}$. The data sources included CVS, Excel files and relational databases, which can be defined as a set: $S = \{S_1, S_2, S_3, \ldots, S_n\}$. Let $m \in \{1, 2, 3, \ldots, M\}$ denote the domain of heterogeneous data sources. The set of data sources belonging to domain m is defined as $S_m = \{S_{m,1}, S_{m,2}, S_{m,3}, \ldots, S_{m,k_m}\}$ where k_m represents the number of heterogeneous data sources (CVS, Excel files and relational databases) in the domain m. It should be noted that the data merging or linking using fuzzing matching algorithm was performed prior to integration and load into the $DB_{multimorbidity}$. The integration function that maps different data source (S_m) and loads data into relations (R_i) in the $DB_{multimorbidity}$ is defined as:

$$\mu_i : \bigcup_{m=1}^{n} S_m \to R_i$$

This allows for the selection of data from data sources, transforming and loading into a relation S, while ensuring that the integrity of PK_i and $FK_{i\to j}$ is not compromised. The entire ETL process using Pentaho data integration (PDI) into $DB_{multimorbidity}$ can be represented as:

$$DB_{multimorbidity} = \{(R_i, PK_i, \{FK_{i\to j}\}_j) | R_i = \mu_i(S), i = 1, 2, \ldots, n\}$$

where $\mu_i$is the integration function.

3.4 Validation

The system was subjected to rigorous testing prior to its implementation, thereby ensuring the precision of data linkage. The algorithm was subjected to a comprehensive evaluation on the test datasets that were categorized into three groups: perfect match, near-perfect match, and non-matching. The condition in which both input sequences are identical is referred to as a flawless or perfect match. Scores of 100% were obtained on this assessment. There were slightly smaller variations between the input sequences when the match was near-perfect (nearly flawless). These minor variations represented inconsistencies in data due to typographical errors and formatting variations. The algorithm considered these subtle variations and

assigned scores from 80% to below 100%. It was necessary to verify the veracity of each record; however, it was determined that the records belonged to the same individuals. The non-matching category comprises string variables that were either entirely or partially dissimilar. The algorithm generated scores between 0 and less than 80% when the input strings were fed through it. In this instance, the scores are deemed inadequate; consequently, the linking process is not feasible unless it is confirmed via a telephone call. The verification process was applied for records with a borderline score of 70% to 79%.

3.4.1 Performance Evaluation

The fuzzy matching algorithm's capacity to accurately identify the similarities between records from a diverse array of data sources was assessed. The algorithm was evaluated using performance metrics, including precision, recall, and F1 score. Precision quantifies the algorithm's effectiveness in avoiding false matches, whereas recall assesses its capacity to minimise missed matches. Maintaining an optimal equilibrium between precision and recall is essential, which is accomplished through the F1 score.

4 INTERPRETATION OF RESULTS

The results of the data linking from a diverse array of sources are summarized in Table 3. The application-level deployment of the fuzzy matching algorithm facilitated the efficient linkage or merging of heterogeneous datasets. When linking records, three methodologies were considered and correspondingly assigned weights. Most datasets produced 100% true matches when the combination of *Names*, *DoB*, and *Gender* was weighted at 0.6, 0.3, and 0.1. For this combination, a threshold of at least 80% acceptability was used in the decision to merge the datasets. This is equivalent to asserting that there is at least 80% probability of similarity between the merging fields from different datasets. The algorithm consistently performed well across most datasets, achieving 100% true matches. However, the true matches for the combination (*Names* + *DoB* + *Gender*) in the Awi-Gen datasets were below 87%, which indicates that approximately 13% of the records were not merged. This is because a portion of the individuals in this dataset came from outside the surveillance area, resulting in some records being discontinuous between the Awi-Gen and baseline datasets. A similar rationale applies to clinic patient-level data, which resulted in a 15% incidence of non-matching records.

In contrast to the previous instance, the threshold was established at a minimum of 90% to only consider trivial errors. From Table 3, we observe that ID and phone numbers, both weighted at 1.0, yielded 100% accurate matches for some datasets. It is important to note that, while ID and phone numbers were present in the datasets, some of the entries for these variables contained null values. Therefore, this 100% match represents the rows that contained values for these variables. The algorithm produced 22% false matches of phone numbers between the Awi-Gen and DIMAMO datasets, while 7% false matches were attained when the Clinic patient was linked to the DIMAMO datasets. The primary reason for this is the

Table 3: Linking diverse data sources to the DIMAMO baseline database

Data Source	Field Combination	Weighting scheme	Match Acceptance Threshold	True Matches	False Matches
DIMAMO	–	–	–	–	–
SRHR	*Names* + *DOB* + *Gender*	0.6:0.3:0.1	≥80%	100%	0%
	ID Number	–	–	–	–
	Phone Number	1.0: - : -	≥90%	100%	0%
Awi-Gen	*Names* + *DOB* + *Gender*	0.6:0.3:0.1	≥80%	87%	13%
	ID Number	–	–	–	–
	Phone Number	1.0: - : -	≥90%	78%	22%
STRIDE	*Names* + *DOB* + *Gender*	0.6:0.3:0.1	≥80%	100%	0%
	ID Number	–	–	–	–
	Phone Number	1.0: - : -	≥90%	100%	0%
Mental Health	*Names* + *DOB* + *Gender*	0.6:0.3:0.1	≥80%	100%	0%
	–	–	–	–	–
	Phone Number	1.0: - : -	≥90%	100%	0%
Clinic Linkage	*Names* + *DOB* + *Gender*	0.6:0.3:0.1	≥80%	85%	15%
	ID Number	1.0: - : -	≥90%	100%	0%
	Phone Number	1.0: - : -	≥90%	93%	7%

little discrepancies in phone numbers, which enable the algorithm to award a higher score; however, following verification, it was shown that the two phone numbers truly belonged to distinct people. However, there were instances of small discrepancies due to typographical errors; following investigation, it was determined that while two numbers differed, they were to the same individual. The algorithm effectively matched data based on elevated similarity but failed to comprehend the semantic distinction between the two records. This highlights the significance of including personal information (*Names*, *DOB*, *Gender*) in the linking process, in addition to the use of ID and phone numbers to enhance the algorithm's accuracy.

Figure 2 shows the distribution of scores generated by the fuzzy matching algorithm. The scores in the range 75–79 contributed 5% of the matches. This is the overall percentage score of the records that matched but required verification due to huge variations between the input strings. These variations were mainly because of typing errors across various datasets. Manual verification confirmed that the records with the similarity scores below 70% were indeed disjoint. The score ranges of 80–85, 86–90, and 91–95 resulted in fuzzy matching of 17%, 23% and 30% respectively. The verification was performed for records in this range since minor typos in capturing the dates and names were discovered.

https://doi.org/10.18489/sacjv38i1/20408

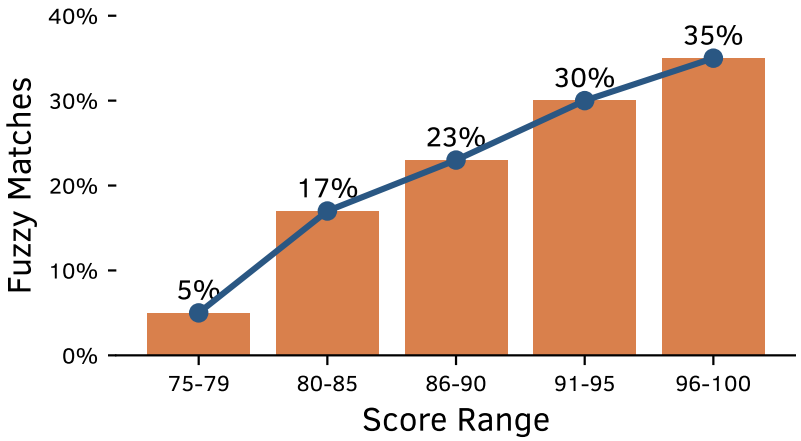

Figure 2: Distribution of similar scores generated by the algorithm

The score range (96–100) had a higher percentage (35%) of fuzzy matching compared to other ranges. This indicates better data quality, as this suggests minor or no typos for the records in this range. However, the overall percentage of fuzzy matches across other ranges suggests contrary.

The primary performance metric of a fuzzy matching algorithm is its capacity to accurately compare and give the proper similarity score to two input strings. The performance of the algorithm was assessed using performance metrics as shown in Table 4. The precision, which measures how accurate the algorithm is in properly identifying the true matches, was at 98.1%. This means that 98.1% of the matches identified by the algorithm are correct. About 1.9% of matches were found to be incorrect. The reason for these incorrect matches was that the input strings were similar but belonged to a distinct participant. This required manual verification to ascertain that the records belong to two different participants. The algorithm achieved the highest percentage of recall at 88.6%, meaning that fewer missed matches were recorded. The missed matches resulted from significant discrepancies in the two input strings caused by typographical mistakes, preventing the records from being matched, although they belonged to the same participant. The F1-score assessed the algorithm's ability to achieve a good balance between the precision and the recall. The result showed that the algorithm achieved a higher F1-score of 93.0%, suggesting that the algorithm maintained a good balance between not missing true matches and avoiding false matches.

Table 4: Metrics validating the performance of the proposed algorithm

Precision	**Recall**	**F1 Score**
98.1%	88.6%	93.0%

The selection of the most suitable threshold is necessary in striking a balance between not missing true matches and avoiding false matches. Figure 3 presents a precision-recall curve to

analyze the effect of the threshold on precision and recall. When the threshold is set at 70%, the algorithm incorrectly matched 14.8% and mismatched about 11.0% of records. In this case, the precision was lowest compared to another threshold. This is because of moderate to high variations in the input sequences. However, increasing the threshold to 80% resulted in decreased incorrect matches by 1.9% while missed matches slightly increased to 11.4%. The matches became more accurate as the threshold increased, with thresholds of 80% and 90% accurately matching about 98% of the records. However, increasing thresholds negatively affected the algorithm's ability to avoid missed matches. This is particularly because these two thresholds filtered out those records with moderate variations, but have the potential to match. So increasing the threshold increases the accuracy while reducing the recall. The threshold set at 80% gives an optimal F1-score, suggesting that it strikes a good balance between false negatives and avoiding false positives compared to other thresholds.

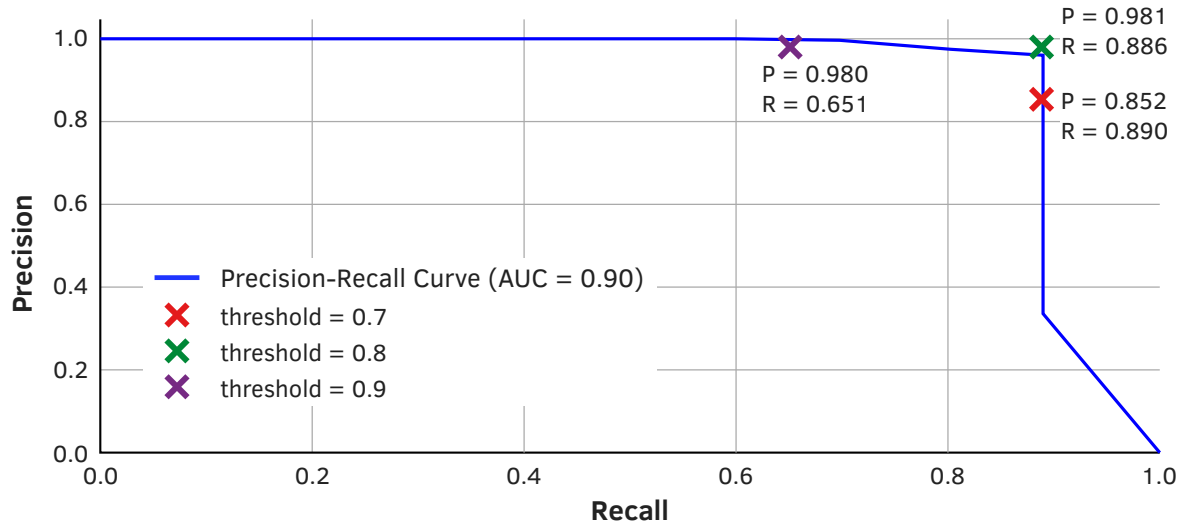

Figure 3: Trade-off between precision and recall curve

5 DISCUSSION

The discrepancies in the schema were a substantial barrier to the successful linking and integration of the datasets. The task of reconciling records from multiple data sources was difficult as a result of data format and structural variations. Data can be stored in a variety of formats, including CSV, XML, JSON, databases, and proprietary formats, making seamless integration problematic. For example, one source may use a date and time stamp in a specific format (e.g., *YYYY-MM-DD*), whereas another may use a different notation (*DD/MM/YYYY*). Data format incompatibilities make it difficult to compare and consolidate records, necessitating preprocessing steps to reconcile them (Rahm & Do, 2000). This adversely affected the linking process and sparked concerns over the precision and comprehensiveness of the data.

The above challenges or issues suggest that there is a need to perform a complex schema analysis before linking the data. Early detection and corrections of schema discordance can result in saving time and resources, hence facilitating a more streamlined integration procedure. To resolve these challenges, it is recommended that solutions such as transforming data into a common format and the use of schema-matching algorithms be used to align different data structures (Doan et al., 2012; Rahm & Bernstein, 2001). In this present study, schema incompatibilities were addressed via a two-step process: first, a data mapping exercise standardized the data into a single schema, including the creation of a mapping table for data subject identifiers (DSIDs) and column restructuring; second, data cleansing was performed using the PDI tool to address inconsistencies and inaccuracies in the transformed data. A crucial element influencing data matching, aside from schema-related problems, was data quality.

Beyond schema-related issues, data quality was a very crucial indicator affecting data linkage. Typographic mistakes, such as misspellings or variations of names and surnames, usually lead to false negatives, where data that should have been linked is missed. This indicates a vital role when it comes to data capturing and storing, since even the most complex algorithm will perform poorly on records that are not accurately captured or stored. In agreement with this, similar records belonging to different participants may produce false positives (e.g., ”Mamabolo Maria” appearing in both datasets but referring to two distinct individuals). To reduce such errors, a hybrid approach that incorporates multiple variables, such as name, date of birth, gender, ID number, and phone number, into the matching algorithm is essential for more accurate classification.

Finney et al. (2011) mentioned that many important real-world circumstances, including healthcare, provide an identifier(s) for each record that is anticipated to be unique to an individual. These would contain identities, such as hospital numbers, as well as identifiers made up of high-cardinality combinations of personal data (surname, forename, phone number), or (surname, forename, date of birth). However, in this study, we found that matching only by phone numbers poses some challenges, resulting in incorrect record linking. For example, a cell phone number may be reallocated to new individuals once discontinued or if it stays active without being used, linking one cell phone number to two different people residing within the same place. Even though a cell phone number may be linked with one individual across multiple datasets can be a useful linkage variable; its reallocation introduces risks of false positives.

The balance between recall and precision was also impacted by the linkage threshold selection. While a looser criterion may result in higher recall at the expense of decreased precision, a stricter threshold decreases false positives but may raise false negatives. According to best practices, match quality can be raised by adding more identifiers, like national ID numbers, or by combining several identifiers in different analyses. A study by Harron et al. (2016) reported the significance of using multiple identifiers to enhance both recall and precision, further supporting the need for a hybrid approach that balances reliability and robustness in the linkage process.

https://doi.org/10.18489/sacjv38i1/20408

6 CONCLUSION

This study hypothesized that the Levenshtein distance algorithm can effectively and precisely link records from semantically and syntactically different datasets into a single source of truth from which a comprehensive data analysis on multimorbidity can be performed. Based on the results, the hypothesis is accepted. The method demonstrated its ability to handle typographical errors and slight differences in identifiers (such as name spellings), which are common issues in record linking. Levenshtein distance increases the recall of true matches across datasets while reducing false negatives.

The issue of whether an algorithm could make it possible to combine disparate datasets into a single, trustworthy analytical source was directly addressed. Despite typographical errors, inconsistent data capturing and storing, and schema conflicts, the findings indicated that the Levenshtein distance algorithm can be used to link disparate data sets. Moreover, by carefully determining matching thresholds and integrating several identifiers, the system was able to balance recall and precision. This hybrid approach strengthened the reliability of the final linked dataset, ensuring that the integrated data could be meaningfully used for subsequent multimorbidity analyses.

7 LIMITATIONS OF THE STUDY

The findings of the present study should be interpreted with the following limitations. Firstly, inconsistencies in data, like missing IDs, unequal cell phone number allocations, and spelling or formatting errors, hinder the data linkage accuracy. Furthermore, connecting outcomes might have been impacted by potential biases if some identifiers (such as ID numbers) were more complete in one dataset than another. Lastly, when linking very large population-level datasets, scalability may be limited due to the computational difficulty of pairwise comparisons using the Levenshtein distance, which increases with dataset size.

8 IMPLICATIONS AND APPLICATIONS

Data linkage and integration have significant ramifications not only for HDSS but other organisations that rely heavily on data, impacting their capacity to extract important insights, render well-informed decisions, and sustain a competitive advantage. With linked and integrated data, organisations can make well-informed decisions based on facts, which results in enhanced operational efficiency, resource allocation, and strategic planning. A single, uniform source of truth enables organizations to reduce the risk of errors and guarantee accurate and consistent data. Furthermore, data linking and integration enable organizations to generate more profound understandings, detect trends, and formulate forecasts that aid in the advancement of innovation, product creation, and a competitive edge in the marketplace. This approach provides a cohesive perspective on the data operations and management within the

organisation.

Integrated and linked data are of paramount importance in augmenting the functionalities of business intelligence (BI) systems through the provision of a more holistic and interconnected data perspective. The comprehensive nature of business intelligence enables entities to detect patterns that transcend functional boundaries, streamline operations, and more efficiently synchronize objectives with organisational operations. The data linking and integration establish a fundamental basis for a business intelligence environment that is more interconnected, precise, and insightful, hence enabling organisations to execute data-driven decisions with greater efficacy.

9 FUTURE DIRECTION

Machine learning and artificial intelligence have the capability to surmount a number of the obstacles encountered by traditional fuzzy matching techniques. Overcoming the curse of dimensionality, enhancing generalizability, optimizing parameters, and striking a balance between cooperation and competition among rules are some of these obstacles (Wu et al., 2020). Through the strategic utilization of machine learning and artificial intelligence capabilities, it may be possible to efficiently tackle these challenges, hence resulting in enhanced functionality and performance of fuzzy matching algorithms. These capabilities will be explored in future to intelligently link the data across several sources.

DECLARATION OF CONFLICT OF INTEREST

Authors declares no conflict of interest regarding this manuscript.

ACKNOWLEDGEMENT

Research reported in this publication was supported by Fogarty International Center of the National Institutes of Health under Award Number U54TW012077. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Alsmadi, M. K., Alzaqebah, M., Jawarneh, S., ALmarashdeh, I., Al-Betar, M. A., Alwohaibi, M., Al-Mulla, N. A., Ahmed, E. A., & AL Smadi, A. (2024). Hybrid topic modeling method based on Dirichlet multinomial mixture and fuzzy match algorithm for short text clustering. *Journal of Big Data*, *11*(1). https://doi.org/10.1186/s40537-024-00930-9

Alsubhi, A., & Munson, E. (2016). Design and usability testing of a user interface for three-way document merging. *Proceedings of the 4th International Workshop on Document Changes: Modeling, Detection, Storage and Visualization*, 1–6. https://doi.org/10.1145/2993585.2993590

Amek, N., Vounatsou, P., Obonyo, B., Hamel, M., Odhiambo, F., Slutsker, L., & Laserson, K. (2015). Using health and demographic surveillance system (HDSS) data to analyze geographical distribution of socio-economic status; an experience from KEMRI/CDC HDSS. *Acta Tropica*, *144*, 24–30. https://doi.org/10.1016/j.actatropica.2015.01.006

Barahama, A. D., & Wardani, R. (2021). Data analysis and data warehouse design based on Pentaho data integration (kettle) to support the determination of student learning achievement. *IOP Conference Series: Materials Science and Engineering*, *1098*(5), 052089. https://doi.org/10.1088/1757-899x/1098/5/052089

Canalle, G. K., Salgado, A. C., & Loscio, B. F. (2020). A survey on data fusion: What for? In what form? What is next? *Journal of Intelligent Information Systems*, *57*(1), 25–50. https://doi.org/10.1007/s10844-020-00627-4

Che Fauzi, A. A., Wan Abdul Rahman, W. F., Fauzi, A., & Weigelt, F. (2021). Managing fragmented database in distributed database environment. *Journal of Mathematics and Computing Science (JMCS)*, *7*(1), 8–14. https://ir.uitm.edu.my/id/eprint/49083

Comber, S., & Arribas-Bel, D. (2019). Machine learning innovations in address matching: A practical comparison of word2vec and CRFs. *Transactions in GIS*, *23*(2), 334–348. https://doi.org/10.1111/tgis.12522

Dasararaju, H. K., & Taori, P. (2019). Data management – Relational database systems (RDBMS). In *Essentials of business analytics* (pp. 41–69). Springer International Publishing. https://doi.org/10.1007/978-3-319-68837-4_3

Dasari, V., & Suresh, S. (2013). HDSS Data Cleaning and Integration Using SAS Business Intelligence Tool. *IBMRD's Journal of Management & Research*, *2*(1), 427–436. https://ibmrdjournal.in/index.php/ibmrd/article/view/52215

Doan, A., Halevy, A., & Ives, Z. (2012). *Principles of data integration*. Elsevier. https://www.sciencedirect.com/book/monograph/9780124160446/principles-of-data-integration

Doğan, O. (2019). Data linkage methods for big data management in Industry 4.0. In *Optimizing big data management and industrial systems with intelligent techniques* (pp. 108–127). IGI Global. https://doi.org/10.4018/978-1-5225-5137-9.ch005

Doukari, O., Naudet, B., & Teulier, R. (2017). Merging IFC-based BIM models: A new paradigm and co-design support tool. *International Journal of 3-D Information Modeling*, *6*(1), 51–64. https://doi.org/10.4018/ij3dim.2017010104

Dubois, D., Prade, H., & Testemale, C. (1993). Weighted fuzzy pattern matching. In *Readings in fuzzy sets for intelligent systems* (pp. 676–685). Elsevier. https://doi.org/10.1016/b978-1-4832-1450-4.50073-0

Dusetzina, S. B., Tyree, S., Meyer, A., Meyer, A., Green, L., & Carpenter, W. R. (2014). *Linking data for health services research: A framework and instructional guide* (tech. rep. No. 14-

EHC033-EF). Agency for Healthcare Research and Quality (US). https://www.ncbi.nlm.nih.gov/books/NBK253313/

Finney, J. M., Walker, A. S., Peto, T. E., & Wyllie, D. H. (2011). An efficient record linkage scheme using graphical analysis for identifier error detection. *BMC Medical Informatics and Decision Making*, *11*(1). https://doi.org/10.1186/1472-6947-11-7

Han, Y., & Lahiri, P. (2018). Statistical analysis with linked data. *International Statistical Review*, *87*(S1), S139–S157. https://doi.org/10.1111/insr.12295

Harron, K. (2022). Data linkage in medical research. *BMJ Medicine*, *1*(1), e000087. https://doi.org/10.1136/bmjmed-2021-000087

Harron, K., Goldstein, H., & Dibben, C. (2016). *Methodological developments in data linkage*. Wiley Online Library. https://doi.org/10.1002/9781119072454

Karabekiroglu, K., Doğangün, B., Hergüner, S., von Salis, T., & Rothenberger, A. (2006). Child and adolescent psychiatry training in Europe: Differences and challenges in harmonization. *European Child & Adolescent Psychiatry*, *15*(8), 467–475. https://doi.org/10.1007/s00787-006-0599-3

Kaur, P., Kotru, S., Singh, S., & Munshi, A. (2022). MiRNA signatures in diabetic retinopathy and nephropathy: Delineating underlying mechanisms. *Journal of Physiology and Biochemistry*, *78*(1), 19–37. https://doi.org/10.1007/s13105-021-00867-0

Khalili, M., Borzooei, R. A., & Deldar, M. (2021). Matching numbers in fuzzy graphs. *Journal of Applied Mathematics and Computing*, *67*(1-2), 1–22. https://doi.org/10.1007/s12190-020-01463-z

Kruse, F., Hassan, A. P., Awick, J., & Marx Gómez, J. (2020). A qualitative literature review on linkage techniques for data integration. *Proceedings of the 53rd Hawaii International Conference on System Sciences*. https://doi.org/10.24251/hicss.2020.132

Lauzanne, O., Frenel, J., Baziz, M., Campone, M., Raimbourg, J., & Bocquet, F. (2022). Optimizing the retrieval of the vital status of cancer patients for health data warehouses by using open government data in France. *International Journal of Environmental Research and Public Health*, *19*(7), 4272. https://doi.org/10.3390/ijerph19074272

Li, G., Yan, L., & Ma, Z. (2019). Pattern match query over fuzzy RDF graph. *Knowledge-Based Systems*, *165*, 460–473. https://doi.org/10.1016/j.knosys.2018.12.014

Lin, J. J., Mochida, T., O'Neill, R. C. W., Yoshida, A., Yamazaki, M., & Sasada, A. (2024). Two online map matching algorithms based on analytic hierarchy process and fuzzy logic. *arXiv*. https://doi.org/10.48550/ARXIV.2402.11866

McCormack, K., & Smyth, M. (2017). A mathematical solution to string matching for big data linking. *Journal of Statistical Science and Application*, *5*(2). https://doi.org/10.17265/2328-224x/2017.0304.001

Mountantonakis, M. (2021). *Services for connecting and integrating big number of linked datasets* [PhD]. University of Crete. http://dx.doi.org/10.12681/eadd/47885

Murray, G. A., & Rivas, L. E. (2020). Techniques for facilitating the joining of datasets [US Patent 10,565,222]. https://patents.google.com/patent/US10565222B2/en

Pikies, M., & Ali, J. (2021). Analysis and safety engineering of fuzzy string matching algorithms. *ISA Transactions*, *113*, 1–8. https://doi.org/10.1016/j.isatra.2020.10.014

Presutti, V., Aroyo, L., Adamou, A., Schopman, B., Gangemi, A., & Schreiber, G. (2011). Extracting core knowledge from linked data. *Proceedings of the Second International Workshop on Consuming Linked Data (COLD2011)*. https://ceur-ws.org/Vol-782/PresuttiEtAl_COLD2011.pdf

Rahm, E., & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *The VLDB Journal*, *10*(4), 334–350. https://doi.org/10.1007/s007780100057

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *Bulletin of the Technical Committee on Data Engineering*, *23*(4), 3–13. https://cs.brown.edu/courses/cs227/archives/2017/papers/data-cleaning-IEEE.pdf

Roldán, M. C. (2013). *Pentaho data integration beginner's guide*. Packt Publishing Ltd.

Rudwan, M. S. M., & Fonou-Dombeu, J. V. (2023). Hybridizing fuzzy string matching and machine learning for improved ontology alignment. *Future Internet*, *15*(7), 229. https://doi.org/10.3390/fi15070229

Rula, A., Maurino, A., & Batini, C. (2016). Data quality issues in linked open data. In *Data and information quality* (pp. 87–112). Springer International Publishing. https://doi.org/10.1007/978-3-319-24106-7_4

Salaki, R. J., Waworuntu, J., & Tangkawarow, I. R. H. T. (2016). Extract transformation loading from OLTP to OLAP data using Pentaho data integration. *IOP Conference Series: Materials Science and Engineering*, *128*, 012020. https://doi.org/10.1088/1757-899x/128/1/012020

Salvadori, I., Huf, A., & Siqueira, F. (2020). Data linking as a service: An infrastructure for generating and publishing linked data on the Web. *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, 262–271. https://doi.org/10.1109/COMPSAC48688.2020.00042

Schildhauer, M. (2017). Data integration: Principles and practice. In *Ecological Informatics* (pp. 129–157). Springer International Publishing. https://doi.org/10.1007/978-3-319-59928-1_8

Shlomo, N. (2018). Overview of data linkage methods for policy design and evaluation. In *Data-driven policy impact evaluation* (pp. 47–65). Springer International Publishing. https://doi.org/10.1007/978-3-319-78461-8_4

Taki, M., & Ardakani, M. M. (2021). A survey on fragmentation in distributed database systems. *Research Square Preprint*. https://doi.org/10.21203/rs.3.rs-395198/v1

Thomas, C., Suresh, B. S., & Vijayalakshmi, M. N. (2020). Data merger computing for an intelligent report generation. *SN Computer Science*, *1*, 167. https://doi.org/10.1007/s42979-020-00157-4

Ubdaidillah, S. H. S. A., & Ahmad, N. (2020). Fragmentation techniques for ideal performance in distributed database –- A survey. *International Journal of Software Engineering and Computer Systems*, *6*(1), 18–24. https://doi.org/10.15282/ijsecs.6.1.2020.3.0066

Wang, H., & Wang, Y. (2022). Research on fragmentation problems of big data analytics in environmental management. *Proceedings of the 3rd International Conference on Green Energy, Environment and Sustainable Development (GEESD2022)*, 522–525. https://doi.org/10.3233/ATDE220322

Wang, X., Brownlee, A. E. I., Weiszer, M., Woodward, J. R., Mahfouf, M., & Chen, J. (2023). An interval type-2 fuzzy logic-based map matching algorithm for airport ground movements. *IEEE Transactions on Fuzzy Systems*, *31*(2), 582–595. https://doi.org/10.1109/tfuzz.2022.3221793

Wesley, D. (2000). Relational database design. *Journal of Insurance Medicine*, *32*(2). https://pubmed.ncbi.nlm.nih.gov/15912904/

Wu, D., Lin, C., Huang, J., & Zeng, Z. (2020). On the Functional Equivalence of TSK Fuzzy Systems to Neural Networks, Mixture of Experts, CART, and Stacking Ensemble Regression. *IEEE Transactions on Fuzzy Systems*, *28*(10), 2570–2580. https://doi.org/10.1109/TFUZZ.2019.2941697

Zhang, Y., Yang, P., Li, C., Zhang, G., Wang, C., He, H., Hu, X., & Guan, Z. (2018). A multi-feature based automatic approach to geospatial record linking. *International Journal on Semantic Web and Information Systems*, *14*(4), 73–91. https://doi.org/10.4018/IJSWIS.2018100104