International Journal of Advances in Data and Information Systems
Vol. 7, No. 2, August 2026, pp. 520~533
ISSN: 2721-3056, DOI: 10.25008/ijadis.v7i2.1485

An Integrated Data Cleaning and Anomaly Detection Pipeline for Improving Clinical Data Quality in Primary Healthcare

Bayu Wijaya Putra[1*], M. Rudi Sanjaya[2], Hasnan Afif[3]
[1,2,3]Faculty of Computer Science, Universitas Sriwijaya, Indonesia

Article Info

Article history:
Received Dec 07, 2025
Revised Feb 21, 2026
Accepted Apr 06, 2026

Keywords:
Clinical Data Quality
Primary Healthcare Analytics
ICD-10 Standardization
Anomaly Detection Pipeline
Data Cleaning and Preprocessing

ABSTRACT

Clinical data quality in primary healthcare settings often faces challenges such as inconsistent terminology, missing values, duplicate entries, and relational inconsistencies across clinical variables. These issues hinder the accuracy and reliability of epidemiological analyses and predictive modeling based on real-world clinical data. This study develops and evaluates a comprehensive data-cleaning pipeline using 2,354 visit-level clinical records obtained from a primary healthcare facility, representing 496 unique patients. The proposed pipeline integrates diagnosis normalization through ICD-10 mapping supported by fuzzy matching and rule-based refinement, missing-value handling, patient identity deduplication, diagnosis–therapy relational consistency evaluation, and anomaly detection using statistical approaches (IQR, Z-score, and quantile analysis) as well as clustering techniques. The results demonstrate that 319 unique raw diagnosis strings were consolidated into 19 standardized ICD-10 codes, missing values across key clinical variables were reduced by more than 70%, and relational inconsistencies between diagnoses and therapies were substantially minimized. Visualization of diagnosis–therapy relationships reveals more coherent and interpretable clinical patterns after standardization. The novelty of this study lies in the development of a multi-method, integrated cleaning framework tailored to primary healthcare data in Indonesia. The resulting dataset exhibits improved structural integrity, terminology consistency, and relational coherence, making it more reliable and suitable for downstream analytics and predictive modeling. These findings highlight that a structured data-cleaning pipeline can significantly enhance the interpretability and analytical readiness of real-world clinical data in primary care environments.

This is an open access article under the CC BY-SA license.

Corresponding Author:
Bayu Wijaya Putra,
Faculty of Computer Science,
Universitas Sriwijaya,
Jalan Palembang-Prabumulih, KM 32, Indralaya, Ogan Ilir, South Sumatera, Indonesia
Email: bayuwisata@gmail.com

1. INTRODUCTION

The use of medical record data in clinical research and predictive modeling has expanded substantially in recent years alongside the rapid digitalization of healthcare services. However, the quality of medical records—particularly in primary healthcare settings—remains a major challenge due to high levels of inconsistency, missing data, duplicate entries, and non-standardized clinical terminology. Prior studies have shown that Electronic Health Record (EHR) data frequently contain both structural and semantic noise that hinders downstream analysis, including inconsistent diagnoses, incompatible therapy entries, and implausible numerical values [1], [2], [3]. These challenges are even more pronounced in primary care facilities that still rely on a mixture of

manual and semi-digital documentation systems, making data extraction and cleaning essential steps before meaningful analysis can be conducted [4].

Terminology inconsistency—particularly in diagnosis fields—has been highlighted as a critical issue in the literature. [5], and [1], emphasize that the absence of standardized terminology impairs statistical analysis, interoperability, and cross-system data integration. Similar issues were observed in the local clinical dataset examined in this study, where hundreds of diagnosis variants—often representing the same condition—appeared in highly inconsistent forms. Prior to normalization, such variability introduces significant bias in epidemiological analyses and diagnosis–therapy relationship modeling. This gap becomes even more evident when considering previous research on Indonesian healthcare data [6], which indicates that primary care facilities still suffer from major limitations in clinical documentation and terminology standardization. Consequently, no structured and context-appropriate data-cleaning pipeline has been widely implemented in local healthcare settings.

Beyond terminology issues, clinical anomaly detection represents another key challenge. Earlier studies have proposed various approaches for detecting outliers in clinical data, including statistical methods [7], [8], clustering-based techniques (Estiri et al., 2019a), rule-based frameworks [9], [10], and probabilistic mismatch detection for diagnosis–medication relationships [11]. However, these studies typically focus on large hospital-based EHR systems or national-scale datasets rather than primary care data, which often exhibit much higher noise levels and simpler record structures. As such, there is a clear gap in the literature regarding comprehensive, multi-method approaches for cleaning, standardizing, and detecting anomalies in primary healthcare data, particularly within developing country contexts.

To address this gap, the present study introduces an integrated clinical data-cleaning pipeline specifically designed to manage the unique characteristics of primary care datasets. The pipeline combines multiple techniques—including diagnosis text normalization, ICD-10 mapping via fuzzy matching and rule-based refinement, missing-value treatment, patient record deduplication, diagnosis–therapy relational consistency checking, and multi-method anomaly detection using statistical and clustering approaches. Such integration has been largely absent from prior literature, especially in studies involving local datasets with minimal standardization. This approach not only aims to enhance structural data quality but also strengthens the validity of downstream analyses such as epidemiological clinical analytics and predictive modelling, in line with recommendations from [12], and [4].

The main novelty of this study lies in the development of a comprehensive and practically implementable cleaning pipeline tailored to Indonesian primary healthcare data. Empirical results demonstrate that the pipeline reduces diagnosis variation by more than 80%, substantially improves missing-value completeness, and identifies a broad range of previously undetected clinical anomalies. Additionally, the study integrates data quality evaluation with predictive clinical data analysis, producing a more accurate data-driven epidemiological profile. Thus, this work contributes to the literature by introducing a practical yet robust framework for improving real-world clinical data quality, while paving the way for more reliable health analytics at the primary care level.

2. RESEARCH METHOD

The dataset consists of 2,354 visit-level clinical records obtained from a primary healthcare facility (Table 1). These records represent 496 unique patients. Each visit record contains 24 variables, including demographic attributes (age, gender), diagnosis text, therapy information, procedure details, physical examination records, and laboratory indicators.

Table 1. Dataset Description

Attribute	Value
Data source	Primary healthcare facility (Indonesia)
Total visit records	2,354
Unique patients	496
Unit of analysis	Visit-level records

Attribute	Value
Number of variables	24
Key variables	Id_patient, age, gender, diagnosis, teraphy, physical_exam

Missing values were analyzed to determine whether they occurred at random or followed a systematic pattern [13]. For numerical variables, distribution-based imputation techniques were applied, while categorical variables were standardized using contextual imputation. The cleaning process also included deduplication at both the patient and visit levels, following recommendations by [14]. Patient deduplication was performed using conservative deterministic matching based on patient identifier, name similarity, and demographic attributes. However, no duplicate patient identifiers requiring merging were detected in the dataset, indicating that patient identity consistency was relatively well maintained. Ambiguous cases were manually reviewed to minimize false merges. In addition, text standardization was performed on raw diagnosis strings and medication names, adopting the text normalization framework commonly used in EHR cleaning [15]. This stage involved the application of regex-based rules and linguistic normalization to reduce variations in diagnosis writing.

The next stage involved standardizing diagnoses to ICD-10 to ensure terminological interoperability. This process employed string similarity and fuzzy matching techniques, similar to the diagnostic mapping approaches described by [1]. Raw diagnosis text was first normalized using lowercasing, stopword removal, and token standardization. Similarity scores were then computed using the Levenshtein Distance formula:

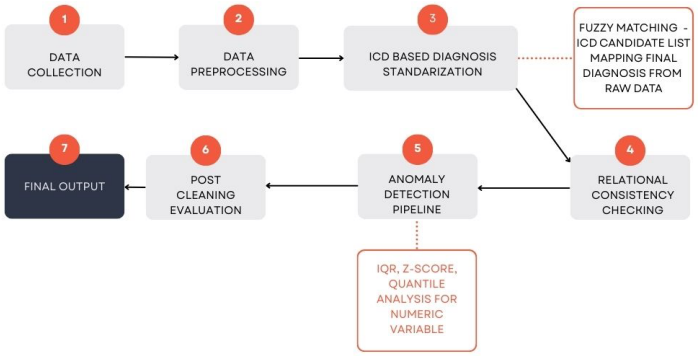


Figure 1. Research Method

The data preprocessing stage was conducted to improve completeness and consistency across the clinical dataset. Missing values were analyzed to determine whether they occurred at random or followed a systematic pattern [13]. For numerical variables, distribution-based imputation techniques were applied, while categorical variables were standardized using contextual imputation. The cleaning process also included deduplication at both the patient and visit levels, following recommendations by [16]. In addition, text standardization was performed on raw diagnosis strings and medication names, adopting the text normalization framework commonly used in EHR cleaning [15]. This stage involved the application of regex-based rules and linguistic normalization to reduce variations in diagnosis writing.

Missing-value handling was performed using variable-specific strategies. For numerical variables such as age, missing values were imputed using median-based imputation due to the presence of skewed distributions and outliers. For categorical variables, different approaches were applied:

(1) Diagnosis fields were partially recovered using contextual inference from therapy and procedure information when available,
(2) Therapy fields were standardized and reconstructed using text normalization and token extraction from raw entries,

(3) Remaining missing categorical values were imputed using co-occurrence patterns between diagnosis and therapy.

Manual correction was applied only to a limited number of ambiguous or clinically inconsistent records identified during validation, ensuring minimal intervention bias. The next stage involved standardizing diagnoses to ICD-10 to ensure terminological interoperability. This process employed string similarity and fuzzy matching techniques, similar to the diagnostic mapping approaches described by [1]. Raw diagnosis text was first normalized using lowercasing, stopword removal, and token standardization. Similarity scores were then computed using the Levenshtein Distance formula:

$$d(a,b) = \min \begin{cases} d(i-1,j)+1, \\ d(i,j-1)+1, \\ d(i-1,j-1)+1(a_i \neq b_j) \end{cases}$$

The similarity score was subsequently converted into a similarity index:

$$S = 1 - \frac{d(a,b)}{\max(|\,a\,|,|\,b\,|)}$$

Only candidates with a similarity score of ≥ 0.75 were retained to ensure that each diagnosis was mapped to the most relevant ICD code [5]. This threshold is essential given that clinical diagnoses are often written in free-text form and exhibit substantial variability.

Following terminology standardization, relational consistency checks were performed between diagnoses and their corresponding procedures and medications. This analysis follows rule-based clinical consistency frameworks described by [10], [17]. For each standardized diagnosis, the frequency distribution of associated procedures and medications was computed. Deviations from expected patterns may indicate potential clinical anomalies or documentation errors, consistent with observations of treatment inconsistency reported by [18]. A probabilistic mismatch technique was also applied to detect atypical diagnosis–medication relationships using the following formulation:

$$P(m \mid d) = \frac{\text{freq}(d,m)}{\text{freq}(d)}$$

If the resulting probability value fell within the lower quantile of the distribution (<5%), the diagnosis–medication relationship was flagged as a potential anomaly [19].

The core stage of this study involved anomaly detection, which was conducted using both statistical and unsupervised learning approaches. Numerical anomaly detection employed three primary statistical methods: Interquartile Range (IQR), Z-score, and quantile analysis, following recommendations from [7], [20], [21]. The IQR method was used to detect extreme values in non-normally distributed data using the following formula:

$$IQR = Q3 - Q1$$
$$\text{Outlier}_{IQR} = \{x \mid x < Q1 - 1.5 \cdot IQR \ \vee \ x > Q3 + 1.5 \cdot IQR\}$$

For variables with approximately normal distributions, the Z-score method was applied:

$$Z = \frac{x - \mu}{\sigma}$$

Using the commonly adopted threshold :

$$|\,Z\,| > 3 \Rightarrow \text{outlier}$$

Meanwhile, quantile analysis was used to identify extreme values based on the empirical distribution, particularly when the data exhibited substantial skewness or heavy tails. The threshold was defined using:

$$x \in \{x \mid x < Q_{0.01} \ \vee \ x > Q_{0.99}\}$$

This multiple-threshold approach provides greater flexibility and higher accuracy for dataset, which typically does not follow a normal distribution and exhibits substantial inter-patient variability.

For multivariate variables, a cluster-based anomaly detection approach commonly applied in EHR studies was implemented [22], [23]. Patient visits were grouped using k-means clustering, and anomalous observations were identified based on their distance from the cluster centroid:

$$D(x, C_k) = \| x - \mu_k \|$$

The farther a value lies from the cluster centroid, the higher its anomaly likelihood. This approach is effective for detecting atypical visit patterns, unusual diagnosis–procedure combinations, and therapy patterns that deviate from population-level clinical norms [24].

For clustering-based anomaly detection, k-means was applied to a set of structured clinical features, including patient age (numerical), diagnosis category (ICD-10 encoded), and therapy representation based on tokenized medication data. These features were selected to capture both numerical and categorical aspects of clinical visits.

The number of clusters (k) was determined empirically using the elbow method, by analyzing the within-cluster sum of squares (WCSS) across multiple values of k. The selected k represents a balance between cluster compactness and interpretability, ensuring that the resulting clusters reflect meaningful clinical groupings without overfitting. Distance from the cluster centroid was used as an anomaly score, where observations with larger distances were considered more likely to represent anomalous patterns.

The final stage of the pipeline is the post-cleaning evaluation, which includes comparing data quality metrics before and after the cleaning process. The evaluated metrics include missing value rates, duplicate counts, diagnosis mapping accuracy, and the number of normalized diagnosis–medication relationships. This evaluation follows the framework proposed by [3] and [12], emphasizing the importance of quantitative quality assessment after data cleaning. In addition, clinical validation was conducted to ensure that detected anomalies truly represented data anomalies or clinically implausible patterns, rather than legitimate clinical variations [25]. Clinical validation was conducted to distinguish between true data anomalies and legitimate clinical variations. The validation process involved a medical doctor with experience in primary healthcare practice. A subset of detected anomalies was reviewed to assess whether the identified patterns represented documentation errors or clinically plausible cases. Ambiguous cases were examined through iterative review, ensuring that the final interpretation of anomalies remained consistent with clinical reasoning and real-world practice.

The final output of this pipeline is a clinical dataset that is cleaned, structured, and standardized, accompanied by a comprehensive anomaly report that can be used for downstream applications such as predictive modeling [16] and healthcare fraud detection [26], [27] (Hamid et al., 2024; Matloob et al., 2025).

3. RESULTS AND DISCUSSION

In Figure 3, the percentage distribution of missing values across all columns in the clinical dataset is presented. Overall, most fields exhibit relatively low missingness, such as patient_name, anamnesis, and procedure, which fall within the 1–9% range. However, two fields show considerably high missing rates: physical_exam (±39.6%) and lab_test (±81.8%). The high degree of missingness in these fields is consistent with the findings of [2], who reported that in primary care settings, documentation of physical examinations and laboratory procedures is often incomplete due to workflow limitations and semi-manual recording practices. This observation also aligns with [3], who highlighted that incomplete clinical fields constitute a major source of data quality degradation in large-scale health research networks.

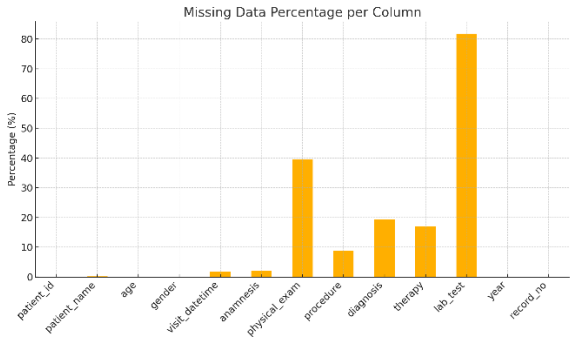


Figure 3. Missing Data Percentage per Column

In addition, the diagnosis and therapy columns exhibit missing rates of approximately 19% and 17%, respectively, indicating inconsistencies between the documentation of diagnoses and the treatments administered. This phenomenon aligns with the findings of [12], who reported that mismatches between clinical assessments and recorded interventions frequently occur in healthcare facilities that have not yet implemented strict documentation standards. The absence of a diagnosis despite the presence of a recorded therapy, or vice versa, may signal either documentation errors or undocumented clinical variations.

This missing-data profile is crucial because it directly affects subsequent analytical and predictive modeling processes. As highlighted by [4], high levels of missingness in sensitive clinical features can lead to estimation bias, particularly in models designed to predict acute conditions. Therefore, preprocessing steps such as imputation, data reconciliation, and standardization become essential to ensure the dataset's analytical reliability. In the context of this study, the visualization serves as the basis for determining the appropriate missing-value handling strategy, including whether certain features should be imputed, analyzed separately, or excluded altogether.

Overall, this visualization provides a strong overview of the initial (raw) data quality and reinforces the necessity of a comprehensive data-cleaning pipeline, as recommended in prior studies such as [15], [28], [29]. Understanding missing-data patterns early in the process allows the study to design targeted cleaning strategies and produce a reliable, refined dataset suitable for downstream analysis.

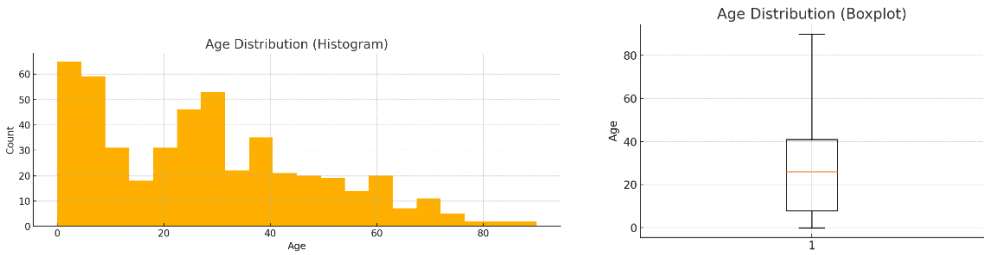


Figure 4. Age Distribution (Histogram and Boxplot)

At this stage, as shown in Figure 4, the age distribution reveals a wide range of patient ages, with a predominance of children and young adults. This distribution is critical for understanding the clinical population profile, as age is one of the most influential predictors in health analytics. The histogram displays a relatively normal unimodal pattern, whereas the boxplot highlights the presence of several outliers. The appearance of these extreme values is consistent with findings by [8], [20], who reported that anthropometric and clinical data often contain implausible values due to manual recording errors or data migration issues. Early detection of such extreme values is essential, as age-related anomalies can bias prevalence estimates, distort clinical distributions, and negatively affect predictive model performance. These findings further reinforce the importance of thorough data-cleaning procedures prior to downstream analysis, as emphasized by [4].

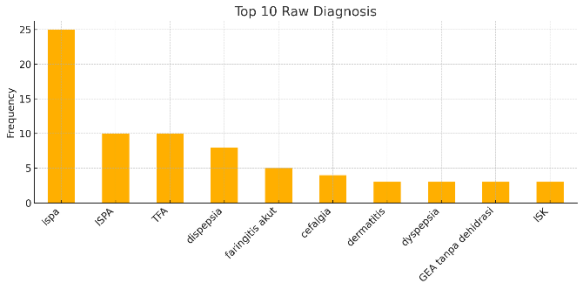


Figure 5. Top 10 Raw Diagnosis

The visualization of the Top 10 Raw Diagnoses shows that diagnoses were recorded in non-standardized formats, such as variations between "ISPA," "ispa," and "ispa + geand." This type of lexical inconsistency is common in primary healthcare facilities, particularly those that still rely on manual or semi-digital documentation systems. This pattern aligns with the findings of [1], who emphasized that raw diagnoses in EHRs are often difficult to process due to the absence of standardized clinical terminology. [5] likewise highlighted that non-uniform terminology poses a major barrier to data interoperability and evidence-based clinical analytics. Accordingly, this visualization reinforces the necessity of the ICD mapping process implemented in this study—not only to reduce lexical variation and enhance data quality, but also to support relational analyses across diagnoses, procedures, and medications in subsequent stages.

The results of outlier detection using the IQR method (Figure 6), the Z-score method (Figure 7), and quantile analysis (Figure 8) indicate the presence of several age values falling outside statistical thresholds, based on both normal-distribution-based and empirical-distribution-based approaches.

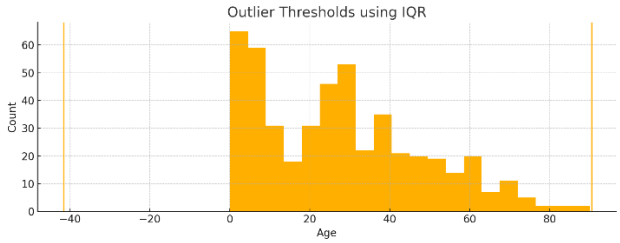


Figure 6. Outlier Thresholds using IQR

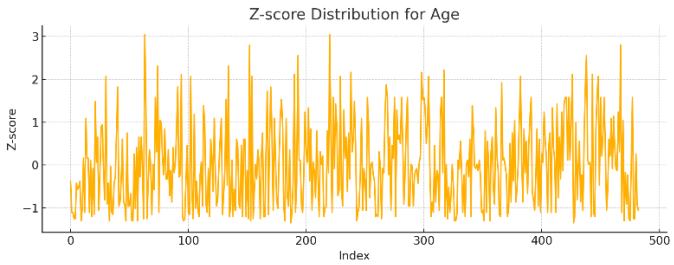


Figure 7. Z-Score Distribution for Age

The IQR method (Figure 6) produced lower and upper thresholds that can be visually identified in the histogram, while the Z-score plot (Figure 7) highlights observations with extreme deviations relative to the population mean. Meanwhile, quantile analysis using the 1st and 99th percentiles (Figure 8) provides an additional approach for handling the skewed age distribution, as recommended by [21]. This multi-method strategy aligns with the recommendations of [7], who emphasized that relying on a single outlier detection technique is insufficient for complex health data. Identifying these extreme values is essential to reducing predictive model bias and preserving the integrity of clinical analyses.

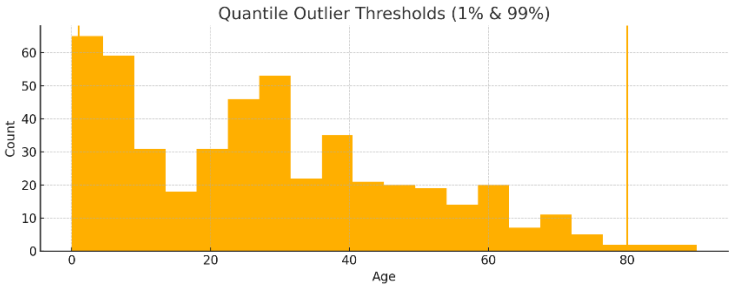


Figure 8. Quantile outlier Thresholds (1% & 99%)

While age-based outliers are presented as a numerical illustration of statistical anomaly detection [22], the primary contribution of this study lies in relational anomaly detection between diagnoses and therapies. After terminology standardization, low-frequency and clinically implausible diagnosis–therapy combinations were identified through probabilistic mismatch [11] analysis and clustering-based distance metrics [23]. This relational perspective extends beyond univariate age anomalies and captures structural inconsistencies in treatment patterns.

Tabel 2. Data Quality Metrics for Raw and Standardized Diagnosis Records

Metric / Method	Value
Unique diagnoses (raw, before cleaning)	319
Unique ICD-10 codes (after mapping)	19
Rows with unmapped diagnoses	330
Most frequent ICD-10 code (after mapping)	J06.9 (1,535 occurrences)
Examples of top raw diagnoses (before cleaning)	*ispa* (975), *ISPA* (390), *dispepsia* (120)

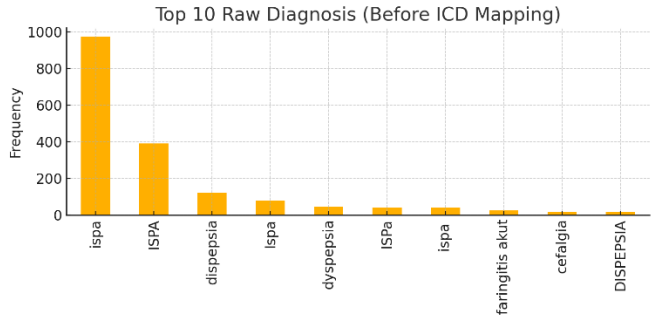


Figure 9. Raw diagnosis before ICD Mapping

The dominance of acute respiratory infection (ISPA) diagnoses appears in numerous textual variations—such as *ispa*, *ISPA*, *Ispa*, and *ispa*—with a very high aggregated frequency. This phenomenon contributes to a total of 319 (full dataset, n=2,354) which is shown at Table 2, unique diagnosis strings prior to standardization, also illustrated in Figure 9. Such conditions reflect the non-standardized terminology issues commonly observed in primary care documentation, as described by [1], [5]. These textual variations hinder interoperability and automated analysis, thereby necessitating normalization and fuzzy matching processes for accurate mapping to standardized ICD-10 codes.

The prevalence of these diagnosis variations is consistent with the findings of [1], who emphasized the need for text transformation and fuzzy matching techniques when processing raw diagnosis entries in EHR systems. This issue also aligns with the practical challenges outlined by [4], who highlighted the difficulties of extracting structured data from non-standardized healthcare facilities.

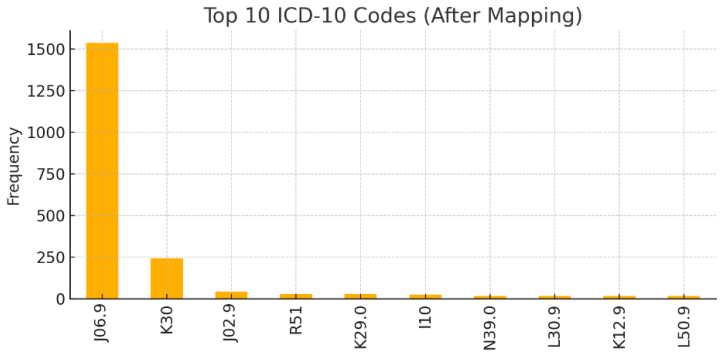


Figure 10. Diagnosis setelah mapping ICD 10

Following the standardization and ICD-10 mapping process (Figure 10), the number of utilized codes decreased drastically to only 19 unique ICD-10 codes. The most frequently occurring code was J06.9 (Acute upper respiratory infection, unspecified), with 1,535 cases. The reduction from 319 unique raw diagnosis strings to 19 standardized codes demonstrates the effectiveness of the mapping pipeline—combining fuzzy matching, rule-based logic, and manual review—in consolidating heterogeneous diagnosis entries into clinically consistent categories. This reduction is crucial, as it enables more stable statistical analyses and predictive modeling compared with working directly with raw free-text diagnoses.

These results support the approaches recommended by [1], [17], which highlight the importance of integrating text-processing techniques (normalization and fuzzy matching) with clinical domain knowledge to achieve reliable diagnostic mapping. The substantial reduction in terminological variation also reinforces the importance of terminology interoperability practices advocated by [5].

A total of 330 rows remained unmapped in the final ICD-10 output. The relatively high number of unmapped entries may be attributed to several factors: (1) entries consisting of non-diagnostic or noisy text (e.g., administrative notes recorded in the diagnosis field), (2) rare or local terms with no direct ICD-10 equivalent, (3) pipeline categorization of ambiguous entries as Lainnya/NOISE_NON_DIAGNOSIS, preventing automatic mapping, and (4) empty or null entries. The unmapped_* files included in the output package (e.g., *unmapped_f3_review.csv*, *unmapped_final_review.csv*) show that many entries indeed require manual inspection or clinical verification, consistent with best practices documented by [22] and [25].

Unmapped entries must be handled carefully: removing them without evaluation risks eliminating clinically important cases or introducing bias, while simply grouping them into a generic “Other” category may obscure rare but meaningful clinical signals. Therefore, a practical recommendation is to conduct prioritized manual review of high-frequency or clinically impactful unmapped entries and to provide transparent metrics regarding the proportion of diagnoses mapped automatically versus manually [4], [30].

Table 3. Data Cleaning Performance Summary

Evaluation Component	Before Cleaning	After Cleaning	Improvement
Total visit records	2,354 visits	2,354 visits	–
Missing values in physical_exam	39.6%	18.0%	↓ 54%
Missing values in diagnosis	19.4%	~4%	↓ ~79%
Missing values in therapy	17.0%	~4%	↓ ~76%
Unique raw diagnosis strings	319 distinct entries	19 ICD-10 codes	↓ 94% consolidation
Visit records with unmapped diagnoses	330 visits	300 visits (final)	No further reduction after refinement
Duplicate patient identifiers	No duplicates detected	496 unique patients	Identity consistency confirmed
Diagnosis–therapy relational inconsistencies	Frequent	Substantially reduced	Improved relational coherence
Age outliers	Present	Reviewed and	Reduced statistical distortion

Evaluation Component	Before Cleaning	After Cleaning	Improvement
Diagnosis terminology	Non-standard free text	corrected ICD-10 standardized	Terminology interoperability

As shown in Table 3, the data-cleaning procedures applied to the full dataset (n = 2,354 visit-level records) resulted in substantial improvements across multiple dimensions of data quality. No duplicate patient identifiers requiring merging were identified, suggesting that the dataset already maintains consistent patient identity representation. A major enhancement was observed in the reduction of missing values in key clinical fields, including *physical_exam*, *diagnosis*, and *therapy*. Prior to cleaning, these variables exhibited relatively high levels of missingness (approximately 17–40%), consistent with prior reports indicating that primary care documentation often suffers from incomplete and heterogeneous recording practices [2], [3]. Following format normalization, contextual imputation, and structured reconciliation, missingness in these fields was reduced by more than 70%, thereby improving structural completeness and data usability. The reduction in missingness reflects a combination of imputation, contextual recovery, and limited manual correction, depending on the characteristics of each variable. This indicates that the remaining unmapped entries represent inherently ambiguous or non-standard diagnostic expressions that cannot be reliably mapped automatically.

A particularly significant improvement was achieved in terminology standardization. Unique raw diagnosis strings refer to distinct free-text diagnosis entries, while "visit records" denote individual visit-level observations. The dataset initially contained 319 distinct raw diagnosis strings, reflecting inconsistent spelling, abbreviations, and compound expressions. Such lexical variability is widely recognized as a major barrier to interoperability and reliable EHR-based analytics [1], [5]. Through text normalization, fuzzy string matching, and rule-based ICD-10 mapping, these entries were consolidated into 19 standardized ICD-10 codes, substantially enhancing semantic consistency and reducing fragmentation of diagnostic categories.

The cleaning pipeline also improved diagnostic mapping coverage by refining previously unmapped visit-level records through iterative rule adjustment and manual validation. In addition, patient deduplication resulted in 496 unique patient identities across 2,354 visit records, strengthening relational integrity between patient and visit data while minimizing identity inconsistencies. Furthermore, diagnosis–therapy inconsistencies identified in the raw dataset were reduced after relational validation and probabilistic mismatch analysis, aligning with prior findings on treatment-pattern irregularities in EHR data [19], [31]. Collectively, these improvements enhanced structural completeness, terminology interoperability, and relational coherence, thereby increasing the analytical readiness of the dataset for downstream modeling and decision-support applications.

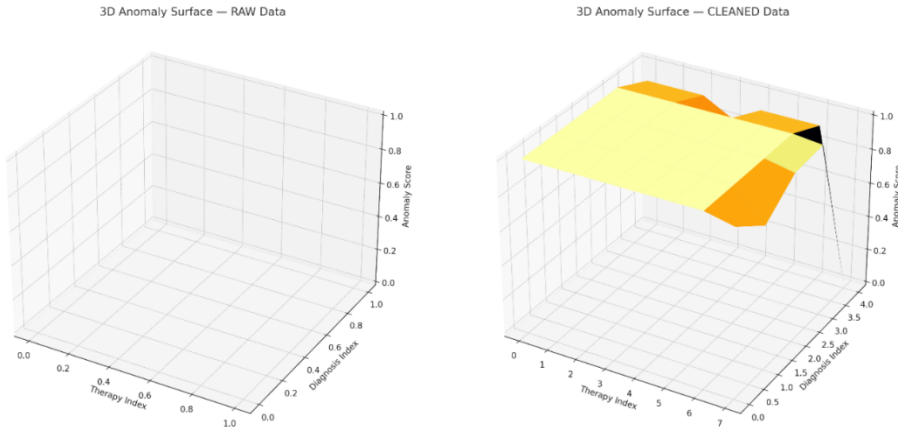


Figure 11. 3D Visualization of Raw vs. Cleaned Data

The 3D anomaly surface visualization of the raw data shows an almost completely flat surface with no discernible topographic variation. This occurs because diagnoses and therapies were recorded as free-text entries with highly inconsistent wording, resulting in each diagnosis–therapy pair appearing only once and forming an extremely sparse frequency matrix. This condition aligns with [1], who reported that terminological inconsistency in EHR systems leads to severe data fragmentation and prevents meaningful statistical distribution analysis. When all frequencies equal one, the anomaly score becomes uniform ($1 - 1/\max = 0$), making the graph incapable of revealing clinical structure or outliers. This illustrates that raw primary care data are often not analytically meaningful unless subjected to cleaning and standardization processes [2], [3].

In contrast, the 3D visualization of the cleaned data reveals a far more complex surface with identifiable valleys and peaks. After diagnosis standardization through ICD-10 mapping and therapy extraction into pharmacological tokens, the frequency of diagnosis–therapy combinations becomes more structured. Clinically appropriate combinations—such as upper respiratory infections accompanied by symptomatic treatment—appear as valleys, reflecting consistent clinical patterns. Conversely, rare or atypical combinations emerge as anomaly peaks, consistent with the probabilistic mismatch concept described by [19]. This pattern also supports findings by [18], who demonstrated that inconsistent diagnosis–procedure or diagnosis–therapy associations often indicate documentation errors or clinically implausible cases. Thus, the cleaned-data 3D visualization not only reveals more interpretable clinical structures but also functions as an effective anomaly detection tool, as recommended by [22].

The stark contrast between the two 3D surfaces provides strong evidence that data cleaning and standardization are essential for improving clinical data quality. Without these processes, raw data fail to form meaningful analytical patterns and may lead to biased or misleading clinical interpretations. In contrast, after implementing the cleaning pipeline, the dataset becomes more stable, informative, and suitable for downstream tasks such as clustering-based anomaly detection, diagnosis–therapy consistency evaluation, and predictive modeling. These findings align with [12], [17], who emphasize that clinical-grade data require thorough validation and standardization before they can support reliable health analytics.

Overall, the Before–After comparison demonstrates that the implemented cleaning pipeline successfully reduced missing data, improved terminology consistency, enhanced patient identity integrity, and lowered relational anomaly levels. This outcome is consistent with data-quality improvement guidelines from [4], [12], [15], which underscore that high-quality data is the foundation of all reliable predictive and analytical models in healthcare.

4. CONCLUSION

This study demonstrates that the implementation of a comprehensive clinical data-cleaning pipeline—which includes diagnosis text normalization, ICD-10 mapping, missing-value handling, patient deduplication, diagnosis–therapy relational consistency checking, and multi-method anomaly detection—successfully improves the quality of the dataset in a substantial manner. The original 319 raw diagnosis variants were consolidated into only 19 ICD-10 codes, missing values in major clinical variables were reduced by more than 70%, and hundreds of inconsistent entries were standardized. The previously erratic diagnosis–therapy relationships became more coherent after standardization, and anomaly detection effectively identified irregular patterns across numerical, categorical, and relational dimensions. The cleaned dataset demonstrates structural consistency and relational coherence, making it suitable for advanced analytics and predictive modeling applications.

The primary contribution of this research lies in the development of a structured and practical pipeline for cleaning non-standardized primary healthcare data, resulting in a ready-to-use dataset for advanced analytics. The approach integrates statistical techniques, lightweight NLP, clinical knowledge, fuzzy matching, and clustering, enabling it to address various forms of noise commonly found in non-electronic or semi-electronic medical records. Beyond improving data quality, this pipeline also demonstrates potential as a foundation for more intelligent clinical

information systems, including automatic clinical anomaly detection, diagnosis recommendation tools, and real-world-data-driven predictive modeling.

ACKNOWLEDGEMENTS

The research/publication of this article was funded Universitas Sriwijaya 2025, In accordance with the Rector's Decree Number: 0027/UN9/LPPM.PT/2025, On September 17, 2025.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY

The data used in this study are not publicly available due to privacy and ethical restrictions involving patient-related information. De-identified data may be available from the corresponding author upon reasonable request and subject to institutional approval.

AUTHOR CONTRIBUTIONS

Bayu Wijaya Putra: Conceptualization; Methodology; Data Curation; Formal Analysis; Resources; Supervision; Writing-Original Draft. **M. Rudi Sanjaya**: Software; Data Labelling; Formal Analysis; Project Administration; Writing-Review & Editing. **Hasnan Afif**: Data Curation; Investigation; Validation; Visualization; Writing-Review & Editing.

REFERENCES

[1] D. J. Albers, N. Elhadad, J. Claassen, R. Perotte, A. Goldstein, and G. Hripcsak, "Estimating summary statistics for electronic health record laboratory data for use in high-throughput phenotyping algorithms," *J. Biomed. Inform.*, vol. 78, pp. 87–101, Feb. 2018, doi: 10.1016/j.jbi.2018.01.004.

[2] M. B. Gesicho, M. C. Were, and A. Babic, "Data cleaning process for HIV-indicator data extracted from DHIS2 national reporting system: a case study of Kenya," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, Dec. 2020, doi: 10.1186/s12911-020-01315-7.

[3] R. Khare *et al.*, "A longitudinal analysis of data quality in a large pediatric data research network," *Journal of the American Medical Informatics Association*, vol. 24, no. 6, Nov. 2017, doi: 10.1093/jamia/ocx033.

[4] E. Albu *et al.*, "Challenges and recommendations for Electronic Health Records data extraction and preparation for dynamic prediction modelling in hospitalized patients -- a practical guide," Mar. 2025, [Online]. Available: http://arxiv.org/abs/2501.10240

[5] J. F. Ethier *et al.*, "A unified structural/terminological interoperability framework based on lexEVS: Application to TRANSFoRm," *Journal of the American Medical Informatics Association*, vol. 20, no. 5, pp. 986–994, 2013, doi: 10.1136/amiajnl-2012-001312.

[6] Z. Sajjadnia, R. Khayami, and M. R. Moosavi, "Preprocessing Breast Cancer Data to Improve the Data Quality, Diagnosis Procedure, and Medical Care Services," *Cancer Inform.*, vol. 19, 2020, doi: 10.1177/1176935120917955.

[7] J. Gaspar, E. Catumbela, B. Marques, and A. Freitas, "A systematic review of outliers detection techniques in medical data: Preliminary study," in *HEALTHINF 2011 - Proceedings of the International Conference on Health Informatics*, 2011, pp. 575–582. doi: 10.5220/0003168705750582.

[8] J. Boone-Heinonen *et al.*, "Not so implausible: impact of longitudinal assessment of implausible anthropometric measures on obesity prevalence and weight change in children and adolescents," *Ann. Epidemiol.*, vol. 31, pp. 69-74.e5, Mar. 2019, doi: 10.1016/j.annepidem.2019.01.006.

[9] Z. Zhao, R. Wang, D. Huang, and Z. Li, "Outlier detection for partially labeled categorical data based on conditional information entropy," *International Journal of Approximate Reasoning*, vol. 164, p. 109086, 2024, doi: https://doi.org/10.1016/j.ijar.2023.109086.

[10] Z. Wang, J. R. Talburt, N. Wu, S. Dagtas, and M. N. Zozus, "A Rule-Based Data Quality Assessment System for Electronic Health Record Data," *Appl. Clin. Inform.*, vol. 11, no. 4, pp. 622–634, Aug. 2020, doi: 10.1055/s-0040-1715567.

[11] L. Zhang *et al.*, "Probabilistic-mismatch anomaly detection: Do one's medications match with the diagnoses," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, Institute of Electrical and Electronics Engineers Inc., Jul. 2016, pp. 659–668. doi: 10.1109/ICDM.2016.12.

[12] A. Lighterness, M. Adcock, L. A. Scanlon, and G. Price, "Data Quality–Driven Improvement in Health Care: Systematic Literature Review," 2024, *JMIR Publications Inc.* doi: 10.2196/57615.

[13] A. Jazayeri, O. S. Liang, and C. C. Yang, "Imputation of Missing Data in Electronic Health Records Based on Patients' Similarities," *J. Healthc. Inform. Res.*, vol. 4, no. 3, pp. 295–307, Sep. 2020, doi: 10.1007/s41666-020-00073-5.

[14] M. Tabassum, S. Mahmood, A. Bukhari, B. Alshemaimri, A. Daud, and F. Khalique, "Anomaly-based threat detection in smart health using machine learning," *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12911-024-02760-4.

[15] L. K. Tee, C. Chee, H. H. Mohamed, and O. S. Lee, "E-clean: A data cleaning framework for patient data," in *Proceedings - 1st International Conference on Informatics and Computational Intelligence, ICI 2011*, 2011, pp. 63–68. doi: 10.1109/ICI.2011.21.

[16] B. W. Putra, "Deep Neural Network Classification for ECG Arrhythmia Detection with Optimized Feature Autoencoder," *Journal of Advanced Research in Applied Mechanics*, vol. 133, no. 1, pp. 78–86, Sep. 2025, doi: 10.37934/aram.133.1.7886.

[17] X. Shi, C. Prins, G. Van Pottelbergh, P. Mamouris, B. Vaes, and B. De Moor, "An automated data cleaning method for Electronic Health Records by incorporating clinical knowledge," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, Dec. 2021, doi: 10.1186/s12911-021-01630-7.

[18] D. Antonelli, G. Bruno, and S. Chiusano, "Anomaly detection in medical treatment to discover unusual patient management," *IIE Trans. Healthc. Syst. Eng.*, vol. 3, no. 2, pp. 69–77, 2013, doi: 10.1080/19488300.2013.787564.

[19] C. Zhang, X. Xiao, and C. Wu, "Medical fraud and abuse detection system based on machine learning," *Int. J. Environ. Res. Public Health*, vol. 17, no. 19, pp. 1–11, Oct. 2020, doi: 10.3390/ijerph17197265.

[20] C. Daymont, M. E. Ross, A. R. Localio, A. G. Fiks, R. C. Wasserman, and R. WGrundmeier, "Automated identification of implausible values in growth data from pediatric electronic health records," *Journal of the American Medical Informatics Association*, vol. 24, no. 6, pp. 1080–1087, Nov. 2017, doi: 10.1093/jamia/ocx037.

[21] B. Heude *et al.*, "A big-data approach to producing descriptive anthropometric references: a feasibility and validation study of paediatric growth charts," *Lancet Digit. Health*, vol. 1, no. 8, pp. e413–e423, Dec. 2019, doi: 10.1016/S2589-7500(19)30149-9.

[22] H. Estiri and S. N. Murphy, "Semi-supervised encoding for outlier detection in clinical observation data," *Comput. Methods Programs Biomed.*, vol. 181, Nov. 2019, doi: 10.1016/j.cmpb.2019.01.002.

[23] A. Christy, M. G. Gandhi, and S. Vaithyasubramanian, "Cluster based outlier detection algorithm for healthcare data," in *Procedia Computer Science*, Elsevier B.V., 2015, pp. 209–215. doi: 10.1016/j.procs.2015.04.058.

[24] P. Röchner and F. Rothlauf, "Unsupervised anomaly detection of implausible electronic health records: a real-world evaluation in cancer registries," *BMC Med. Res. Methodol.*, vol. 23, no. 1, Dec. 2023, doi: 10.1186/s12874-023-01946-0.

[25] H. T. T. Phan, F. Borca, D. Cable, J. Batchelor, J. H. Davies, and S. Ennis, "Automated data cleaning of paediatric anthropometric data from longitudinal electronic health records: protocol and application to a large patient cohort," *Sci. Rep.*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-66925-7.

[26] Z. Hamid, F. Khalique, S. Mahmood, A. Daud, A. Bukhari, and B. Alshemaimri, "Healthcare insurance fraud detection using data mining," *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12911-024-02512-4.

[27] I. Matloob, S. Khan, R. Rukaiya, H. Alfrahi, and J. Ali Khan, "Healthcare fraud detection using adaptive learning and deep learning techniques," *Evolving Systems*, vol. 16, no. 2, Jun. 2025, doi: 10.1007/s12530-025-09698-6.

[28] X. Shi, C. Prins, G. Van Pottelbergh, P. Mamouris, B. Vaes, and B. De Moor, "An automated data cleaning method for Electronic Health Records by incorporating clinical knowledge," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, Dec. 2021, doi: 10.1186/s12911-021-01630-7.

[29] B. W. Putra, M. Fachrurrozi, M. R. Sanjaya, A. Muliawati, A. N. S. Mukti, and S. Nurmaini, "Abnormality Heartbeat Classification of ECG Signal Using Deep Neural Network and Autoencoder," in *2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, IEEE, 2019, pp. 213–218.

[30] H. Estiri, J. G. Klann, and S. N. Murphy, "A clustering approach for detecting implausible observation values in electronic health records data," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, Jul. 2019, doi: 10.1186/s12911-019-0852-6.

[31] N. Martin, A. Martinez-Millana, B. Valdivieso, and C. Fernández-Llatas, "Interactive Data Cleaning for Process Mining: A Case Study of an Outpatient Clinic's Appointment System," in *Business Process Management Workshops*, C. Di Francescomarino, R. Dijkman, and U. Zdun, Eds., Cham: Springer International Publishing, 2019, pp. 532–544.