Research article

A novel approach for standardizing clinical laboratory categorical test results using machine learning and string distance similarity

Syed Ahmmed [a], M. Rubaiyat Hossain Mondal [a,*], Md Raihan Mia [b,d], Mohammad Adibuzzaman [c], Abu Sayed Md. Latiful Hoque [b], Sheikh Iqbal Ahamed [d]

[a] *Institute of Information and Communication Technology, Bangladesh University of Engineering And Technology, Dhaka, Bangladesh*
[b] *Department of Computer Science And Engineering, Bangladesh University of Engineering And Technology, Dhaka, Bangladesh*
[c] *Oregon Clinical and Translational Research Institute, Oregon Health and Science University, Portland, OR, USA*
[d] *Department of Computer Science, Marquette University, Milwaukee, WI, USA*

ARTICLE INFO

ABSTRACT

Standardizing clinical laboratory test results is critical for conducting clinical data science research and analysis. However, standardized data processing tools and guidelines are inadequate. In this paper, a novel approach for standardizing categorical test results based on supervised machine learning and the Jaro-Winkler similarity algorithm is proposed. A supervised machine learning model is used in this approach for scalable categorization of the test results into predefined groups or clusters, while Jaro-Winkler similarity is used to map text terms into standard clinical terms within these corresponding groups. The proposed method is applied to 75062 test results from two private hospitals in Bangladesh. The Support Vector Classification algorithm with a linear kernel has a classification accuracy of 98%, which is better than the Random Forest algorithm when categorizing test results. The experiment results show that Jaro-Winkler similarity achieves a remarkable 99.93% success rate in the test result standardization for the majority of groups with manual validation. The proposed method outperforms previous studies that concentrated on standardizing test results using rule-based classifiers on a smaller number of groups and distance similarities such as Cosine similarity or Levenshtein distance. Furthermore, when applied to the publicly available MIMIC-III dataset, our approach also performs excellently. All these findings show that the proposed standardization technique can be very beneficial for clinical big data research, particularly for national clinical research data hubs in low- and middle-income countries.

1. Introduction

The use of electronic health records (EHRs) has increased significantly over the past few years. More and more evidence points out that generated data may make it easier to identify new medical evidence [1–3]. Additionally, it will improve patient outcomes and medical decision- making. By combining patient data from heterogeneous clinical data sources, EHRs also aid national or multi-institutional health research initiatives. Although EHRs have many benefits, including the generation of large amounts of digitalized

---

Table 1
Example of categorical clinical laboratory test results.

Test Name	Attribute Name	Specimen	Test Result
HIV Test	Sensitivity	Blood	Non-reactive
Blood Type	ABO Group	Serum	"AB"
Hepatitis B Surface Antigen Panel	HBsAg	Blood	Positive
Hepatitis B Surface Antigen Panel	HBsAg	Blood	"Positive (+ve)"
Hepatitis B Surface Antigen Panel	HBsAg	Urine	Positive
Liver Function Test	Bilirubin	Blood	Yellowish
Liver Function Test	Bilirubin	Blood	YELLOW
Liver Function Test	Bilirubin	Blood	L Yellow
Stool Analysis	Blood in the stool	Stool	Present
Stool Analysis	Blood in the stool	Stool	Absent
Stool Analysis	Blood in the stool	Stool	Absent
Blood Type	Rh Factor	Serum	O
Blood Type	Rh Factor	Serum	"0"
Blood Type	Rh Factor	serum	"O"

medical data, the vast majority of this data is still not standardized. Several studies have pointed out that the lack of standardization in EHR data presents a challenge when using the data for research [4–6]. Standardization is especially important regarding categorical test results, which can be prone to variability and inconsistencies in data entry. The categorical test results are broken down into a few exemplary cases and presented in Table 1.

Categorical test results can have more noise and variation than numeric results. Medical laboratories have their own digital system for selecting the names of tests and results that appear on their menus, as well as the reference ranges for normal and abnormal results; sometimes, results are entered by typing manually. In particular, a system may employ local terminology which sometimes leads to incorrect synonyms and abbreviations. Many words were found to describe a positive result, such as "positive," "positive +ve," "Postive," and "Posi". Table 1 shows an instance where a test name "HBsAg" can have the same results denoted as "Positive", "+", "Positive (+ve)". The Liver Function Bilirubin finding of the urine can be "yellowish," "L yellow," "YELLOW," or "yellow". A standardized dictionary can help with the problem of having to pay attention to how laboratory results are presented. Standard formats and identifiers for laboratory tests are provided by systems like the Logical Observation Identifiers Names and Codes (LOINC) [7] and Systematized Nomenclature of Medicine (SNOMED CT) [8].

Prior to initiating the exchange of clinical information across systems, it is essential to establish a standardized framework for clinical data in Bangladesh to ensure the successful integration of health research. With support from the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ), the government adopted the District Health Information Software 2 (DHIS2), an open-source program, in 2009 [9]. The implementation of DHIS2 in Bangladesh accelerated the prioritization of data standardization and interoperability in developing eHealth software and databases. Hospitals in the country typically use the Open Medical Record System (OpenMRS) for electronic medical record keeping. DHIS2 plays the role of aggregator of clinical data from different hospitals' OpenMRS. However, as pointed out by Khan and Hoque [10], the clinical reports generated in these systems still lack international standardization and often make use of incorrect synonyms and abbreviations. Moreover, Mia et al. [11] compared the characteristics of existing health informatics systems in Bangladesh and discovered that these systems do not conform to international standards for interoperability. Even though a lot of EHRs are generated every day; public, private, and non-governmental organization (NGO) healthcare programs are still not connected to this infrastructure. This will likely lead to reports or results that are inconsistent.

In the literature, different ways and tools for standardizing clinical laboratory test results, particularly categorical test results were described. The Hauser et al. [12] team made and tested a tool to make sure that the results of lab tests in a large, multi-center clinical data warehouse were all the same. As part of the standardization process, typos were fixed, categorical results were normalized, inequalities were separated from numbers, and numbers written as words (like "million") were turned into numbers. However, the study included fewer instances of categorical test results being reported in text. To automatically standardize categorical laboratory data, Kim et al. [13] attempted to filter out irrelevant information, organize the remaining data into meaningful groups, and map the results to commonly used concepts. To classify the findings, the researchers developed a rule-based algorithm. The methods proposed by the researchers aimed to establish a gold standard by performing well on their dataset. However, when applied to our new Bangladeshi dataset described later in Section 2.1, the same methods did not yield satisfactory results. The rule-based algorithm developed by the researchers could not classify many groups. Additionally, string similarity measurements, such as the Cosine similarity, showed insufficient outcomes when dealing with test data obtained from the ABO Blood test. The insufficiency became notably apparent while handling outcomes that contained positive signs (+) and a multitude of inconsistent variations, such as "(++)", "++," "(++++)", and even variations like "+++" and "++-." As a consequence, the performance was not satisfactory.

Despite these efforts, a comprehensive solution to the standardization problem is still elusive, especially when the process needs to be scaled up to involve many groups while maintaining perfect standardization. As a result, the issue of standardizing test results requires a solid data standardization protocol. In this study, we propose several techniques for transforming noisy, institution-specific test results into a standardized format. The main contributions of this paper are as follows:

- Application of a supervised machine learning (ML) model for scalable classification of categorical test results into predefined groups

- Implementation and evaluation of Jaro-Winkler similarity between the group's categorical term with noisy result data and standard clinical terms from its predefined group. Jaro-Winkler similarity was evaluated as more accurate than Cosine Similarity or Levenshtein Distance for standardizing noisy clinical categorical data
- Execution of the indicated method to standardize categorical test results for a novel dataset from two private hospitals in Bangladesh

This paper is organized as follows: In the "Introduction" section, we examine a few research papers that discuss the significance of standardizing clinical laboratory test results for interoperability. We also discuss current research demonstrating how clinical laboratory test results are currently standardized, how they are designed, and what issues they have. We discuss and provide a workflow for standardizing the test result in the "Materials and Methods" section using insights from prior research and fresh methodologies to improve the robustness of the standardization process. The experimental results, as well as the model evaluation, analysis, and accuracy, are presented in the "Result" section. The main findings, limitations, and future suggestions are laid out in the "Discussion" section at the end of the paper. Finally, the paper provides concluding remarks.

2. Materials and methods

2.1. Dataset

This section provides an overview of the datasets used in the study. The source dataset was obtained from two of Bangladesh's largest private hospitals, and this dataset collection is part of the development being done for the National Clinical Data Warehouse (NCDW) [11] of Bangladesh. However, the process of data integration raises new technical, semantic, and ethical concerns [14]. To address these issues, we obtained ethical clearance from the Bangladesh Medical Research Council (BMRC) to collect anonymous secondary data (Ref: BMRC/NREC/2019-2022/342). This enabled the integration of the data in a controlled manner and avoided any ethical problems. The study utilized a sizable dataset of health records, which included information from a large number of individuals and a significant number of records. From this dataset, we extracted 75062 categorical test results in text format.

To verify our algorithm's generalizability, we extracted 5657 categorical test results from the publicly available MIMIC-III dataset [15] and employed the procedures to validate the research findings.

2.2. Procedures

The procedure we followed to standardize categorical test results can be broken down into the following steps. The first step was extraction and analysis, in which laboratory data was collected and analyzed for grouping. Grouping test results before comparing them with string distance similarity measures can improve performance by reducing irrelevant comparisons. Thus, identifying the test results in some groups and predefining the vocabulary for the groups can contribute to efficiency by avoiding comparing all test results to all standard terms. In the next step, prepossessing, the text is prepared and refined to be in an appropriate format for analysis. The manual labeling steps consider classifying their test results in order to create an automatic ML model for classifying the terms. Here, under every label name, the group has noisy terms from different institutions and standard terms from standard vocabulary from LOINC and SNOMED CT. Finally, standardization with the string distance similarity step eliminates noise and standardizes the results. Fig. 1 provides an overview of the entire standardization procedure.

2.3. Extraction and preprocess

This section describes the essential steps of data processing, in which unprocessed data from sources is prepared to be compatible with subsequent analysis using various prepossessing techniques. First, clinical laboratory data or reports, including the lab test identifier and the lab test result, were extracted from a database. The study gathered information from two institutions, each with its own sample size. The dataset was split into two categories: categorical data and numerical data. The focus of standardization efforts was mostly on categorical data, assuring consistent representation of outcomes such as "Positive", "Negative", or "Not Specified" as indicated in Table 2. Data cleansing and error detection techniques were used to find and rectify categorical anomalies, such as misspelled categories or non-standard abbreviations. Although the research did not consider numerical data, common examples include blood pressure readings (e.g., 120 80 mmHg), cholesterol levels (e.g., 200 mg/dL), or cell counts (e.g., 5.0 x 10^9 cells/L), which require standardized units and reference ranges for accurate interpretation.

Rows that did not qualify for standardization were removed, as were rows that contained implausible input errors. Regular expression (regex) patterns played a pivotal role in this process. To ensure precise categorization and mitigate potential classification challenges, specific symbols and signs were excluded from the test results through the utilization of regex patterns, which can effectively identify and isolate these unwanted characters. Certain symbols and signs were omitted from the test results to ensure precise categorization and to avoid potential classification difficulties. For reports related to LDL cholesterol and MPV, symbols such as the star sign ("*," "***"), arbitrary dash ("-," "–"), and superscript sign ("," "»>") were removed. Additionally, quotation marks ("") and brackets ("()") were replaced with spaces. These symbols and signs are usually the result of typographical or input errors and have no bearing on the test results. This step is critical for following processes like vectorizing the results for ML since it minimizes the likelihood of inaccurate categorization due to such errors.
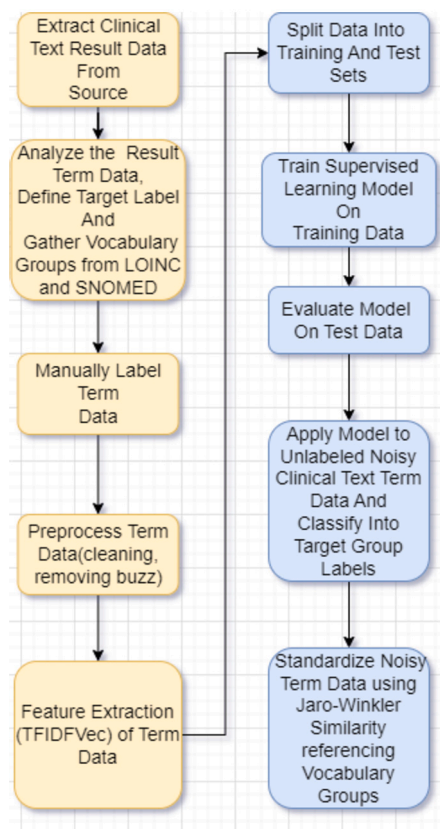
Fig. 1. Workflow of standardization of clinical laboratory categorical test result.

2.4. Defining target labels with vocabulary groups

This section explains how the gold standard groups are defined from the standard vocabularies for the collected test results. Before developing the automated ML model for classification, a table consisting of eight vocabulary groups was established. According to the reference to the Kim et al. [13], in which the test results were divided into five groups, we add three additional groups and divide the test results into eight groups. This grouping aims to divide the test results into distinct groups so that string distance algorithms do not map themselves to other words when counting distances. Such grouping can improve the performance by reducing irrelevant comparisons and identifying similar strings. For instance, the map of "Positive" should not be "Present," and the map of "Nil" should not be "not done." Furthermore, grouping reduces the error of false standardization. Laboratory tests were grouped for this vocabulary group based on their "Test Name" and "Attribute Name". A text mining technique was employed for extracting the specific rows. The dataset was imported into a data frame. Next, the rows were selected according to the presence of "test name" or matching string in result terms from the corresponding group lists such as "Physical Finding", "Present Finding", "Gender Finding", etc. These rows allocated a group name as shown in Table 2. The selected rows were then labeled based on the name of the corresponding group. The data frames of selected rows with labels were combined for each group to generate a single, consolidated data frame of test results. The results terms from the source were queried against the LOINC and SNOMED CT vocabularies, which serve as the gold standard or international standard for standardizing clinical laboratory test results. LOINC and SNOMED are widely recognized as authoritative sources for defining standardized codes and terminology for medical concepts, including laboratory test results. By aligning the test results with these gold standard vocabularies, the selection of appropriate and consistent vocabulary terms for each group was ensured. This adherence to established standards enhances the accuracy, reliability, and interoperability of clinical laboratory data, facilitating effective communication and analysis in healthcare settings. In the second column of Table 2, the predefined vocabulary terms are displayed. In the third column, the corresponding "test name" is listed. This method facilitates the efficient categorization of a vast quantity of medical data and has the potential to be applied to other text analysis disciplines.

2.5. Assigning labels to terms using machine learning algorithms

In the last section, different target groups were identified manually by analyzing the whole dataset. This allowed the data to be put into eight groups. This section introduces a new method for labeling terms using ML algorithms that can considerably reduce the time and effort required to classify large datasets. The automatic classification procedure was then initiated, whereby the raw

Table 2
Target labels with vocabulary groups.

Label	Vocabulary Terms	Example Test Name
Presence Finding	Positive, Negative, Moderately Positive, Weakly Positive, Reactive, Non-Reactive	HIV Test, Dengue Test Syphylis Test,Herpes Test etc. from Different Virus test
Physical Finding	Amber, Blackish, Brown, Clear, Crystal Clear, Dark Brown, Dark Straw, Deep Straw, Deep straw, Greenish, Greyish White, Hazy, Greenish, Yellow, Orange, Reddish, Straw, Turbid, Watery, Whitish, Yellowish	Urine Analysis, Stool Analysis etc.
Blood Group Finding And Another Measure	A, B, C, AB, O, +, ++, +++	Blood Group, Rh Type etc.
Mixed Finding	A Few, Few, Very Few, A Few, Normal, Present, Absent, Occasional, Plenty, Trace, Nil	Urine Analysis etc.
Blood Component Finding	Fresh Frozen Plasma, Packed Red Blood Cells, Platelet Apheresis, Platelet Concentrate, Whole Blood,	Blood Component Test
Compatibility Finding	Compatible, Incompatible	Blood Compatibility Test
Gender Finding	Female, Male	Gender Status*
Non Specific Finding	Not Done, Not Found, Not Specified	Scrapping All Over The Dataset

*Note: The section titled "Example Test Name" provides examples of test names associated with each vocabulary group, but the list is not exhaustive. This column illustrates the test result that may be associated with each vocabulary group, rather than providing an exhaustive list of all tests.
*Gender status label is not a test name, but it is a way to figure out a person's gender. In the same way, the Non-Specific Finding category does not refer to a specific test. Instead, it is a group for labels that do not fit into any other group.

data from the source is sent to the procedure and assigned a label (shown in Table 2) automatically. Fig. 2 shows the structure of the proposed paradigm for using ML to classify test result terms. A pandas DataFrame was populated with a labeled dataset consisting of test result data and their respective labels. In the DataFrame, the "result" column contains unclassified test results, while the "label" column contains the corresponding labels. TfidfVectorizer was employed to convert the test results' data into numerical vectors. TfidfVectorizer is a popular text vectorization technique in ML that transforms a collection of unprocessed documents into a matrix of TF-IDF features, where TF stands for term frequency and IDF stands for inverse document frequency. The resulting matrix can then be utilized as input for ML models. Using the train test split function from scikit-learn [16], the dataset was then divided into training and testing sets. The performance of two popular machine learning algorithms, the Random Forest Classifier (RF) [17] and Support Vector Machine (SVM) Classifier [18], was evaluated. RF is a predictive technique that incorporates decision trees. It effectively manages noisy and high-dimensional data. It randomly selects features and trains each tree using a unique data sample. The trees independently predict the class label or group, and the mode of the predictions is taken as the final prediction. SVM looks for the optimal hyperplane for dividing groups. The method involves projecting the data into a higher dimensional space and then searching for the border between groups that produces the largest margin. For each model, the model was adapted to the training data, predictions were made using the testing data, and the accuracy score was calculated. The accuracy score served as the performance metric for evaluating the models, with a higher accuracy score indicating superior performance in accurately classifying text data into their respective categories. For evaluating the performance of the models, we also computed weighted F1 score, micro-averaged F1 score, and macro-average precision for every model in the labeled dataset [19]. To make it easier to understand how accurate the results were, the weighted F1 score and the micro-averaged F1 score were looked at. This is because accuracy with class imbalance can sometimes lead to too optimistic results. Jupyter Notebook was used to do all of the coding and analysis for this study [20].

2.6. Standardization with string distance similarity

This section describes the method for standardizing test results based on string distance similarity. After classifying all of the test results into different labels, a similarity score was calculated between test result from a group and each of the standard values of the group's reference vocabulary. The measure of lexical and semantic similarity between two pieces of text determines how similar the texts are. Gomaa and Fahmy [21] suggested three methods for measuring similarity: string, corpus, and knowledge-based. The corpus-based and knowledge-based approaches determine the similarity between two terms using information from large corpora or semantic networks. String-based approaches compare string sequences or character compositions for similarity. String-based approaches evaluate the similarity of lexical features, so they cannot detect semantic similarities beyond trivial levels [22]. The text length of the lab test results used in this study was insufficient for semantic similarity extraction. As a consequence, string-based lexical similarity was better for standardizing test results. Ristad and Yianilos [23] suggested a common approach to determine lexical similarity between two short strings. This study utilized the Jaro–Winkler distance, a well-known string distance similarity, to match two strings and calculate their similarity [24–26]. Comparing the applicability of Jaro-winkler distance, Cosine similarity, and Levenshtein distance [27], which are employed in [13], was another focus of this research. The Cosine of the angle between two vectors of two texts is used to calculate Cosine similarity. Levenshtein distance, on the other hand, determines the minimum number
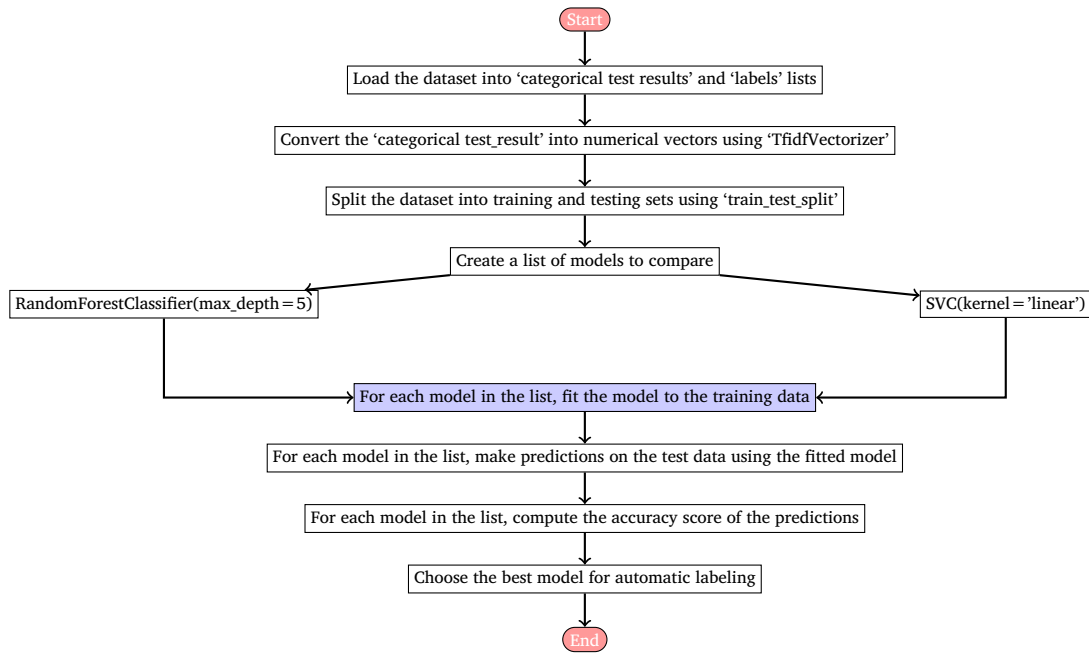
Fig. 2. Flowchart of assigning labels to categorical test result using machine learning algorithms.

of operations required to transform one string into another. Equation (1) and (2) show how Jaro Winkler Similarity is utilized to measure the similarity between two strings [28].

$$sim_j = \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) \tag{1}$$

where m is the number of matching characters. Two characters from s_1 and s_2 are considered matching if they are the same and not farther than $\left\lfloor \frac{max(|s_1|,|s_2|)}{2} \right\rfloor - 1$ characters apart. $|s_1|$ and $|s_2|$ are the lengths of the first and second strings, respectively, and t is the number of transpositions. t is calculated as the number of matching (but different sequence order) characters divided by 2.

$$sim_w = sim_j + \ell p(1 - sim_j) \tag{2}$$

where sim_j is the Jaro similarity for strings s_1 and s_2, ℓ is the length of common prefix at the start of the string up to a maximum of 4 characters, and p is a constant scaling factor for how much the score is adjusted upwards for having common prefixes. p should not exceed 0.25, otherwise the similarity could become larger than 1. The standard value for this constant in Winkler's work is $p = 0.1$.

The pseudocode in Algorithm 1 describes the similarity-scoring algorithm used to measure the lexical similarity between two strings within each data set. The algorithm accepts as input a sample list consisting of test results and a reference list of standard results from the vocabulary group. It then computes the Jaro-Winkler similarity score between each sample and each reference and stores the scores in a list. For each sample, the reference with the highest Jaro-Winkler similarity score is identified and added to a list. This precisely assigns each sample to the most comparable reference based on their Jaro-Winkler scores.

After calculating the similarity score between the terms and comparing the non-standardized term from the classified label to the most similar term from the standard data set, the algorithm returns the optimal match with the highest similarity score. The non-standardized term is then substituted for the closest term from the standard reference.

3. Results

3.1. Evaluation of machine learning models

Prior to standardization with string distance similarity, the test results were classified into eight predefined groups using supervised ML algorithms. This section describes the effectiveness of the models used to organize the test results. Two distinct algorithms were evaluated: RF with a maximum depth of 5 and SVM with a linear kernel.

The RF achieved an accuracy of 0.88 and a weighted average F1 score of 0.83, as shown in Table 3. It was found that the overall average F1 score was 0.47. The macro-average F1 score was 0.47. The macro-average precision and weighted-average precision were 0.90 and 0.97, respectively. The weighted-average recall and macro-average recall, respectively, were 0.88 and 0.48. The SVM, on the other hand, obtained an accuracy of 0.98 and a weighted-average F1 score of 0.98, which is higher than that of RF. The macro-average F1 score was 0.97. The weighted average precision and macro-average precision were, respectively, 0.99 and 0.96. The weighted-average recall and macro-average recall were, respectively, 0.98 and 0.99. The SVM with a linear kernel outperformed the


Algorithm 1: Standardize to vocabulary.

Input : Sample list and vocabulary list
Output: Dataframe with standardized samples and corresponding Jaro-Winkler similarity scores

1   standardize_to_vocabulary($sample\_list, vocabulary\_list$);
2   **Function** standardize_to_vocabulary($sample\_list, vocabulary\_list$):
3     $jw\_scores \leftarrow$ **JaroWinklerMeasures**($sample\_list, vocabulary\_list$);
4     $result \leftarrow []$;
5   **for** $s\_idx, sample$ **in** *enumerate($sample\_list$)* **do**
6        $max\_jw\_score, max\_jw\_idx \leftarrow \max(enumerate(jw\_scores[s\_idx]), key = lambda x : x[1])$; $result$.append([$sample\_list[s\_idx]$, $vocabulary\_list[max\_jw\_idx]$, max_jw_score]);
7   **return DataFrame**($result$, columns = ['Sample', 'Standardized Sample', 'Jaro-Winkler Score']);

8   **Function** *JaroWinklerMeasures*:
9     $sample\_list, vocabulary\_list$

10 **initialize** $jw\_measures$ **as an empty list**
11 **for** *each sample in sample_list* **do**
12   **initialize** jw_row **as an empty list**
13   **for** *each vocabulary in vocabulary_list* **do**
14        $jw\_score \leftarrow$ jaro_Winkler ($sample, vocabulary$) **append** jw_score to jw_row

15   **append**
16   jw_row to $jw\_measures$

17 **return** $jw\_measures$

Table 3
Comparison of classification performance of machine learning models for test results categorization.

Model Name	Parameter	Accuracy	Weighted Avg. F1 score	Macro Avg. F1 score	Weighted Avg. Precision	Macro Avg. Precision	Weighted Avg. Recall	Macro Avg. Recall
RF	max depth = 5	0.88	0.83	0.47	0.90	0.97	0.88	0.48
SVM	kernel = 'linear'	0.98	0.98	0.97	0.99	0.96	0.98	0.99

*Note: RF stands for Random Forest and SVM stands for Support Vector Machine.
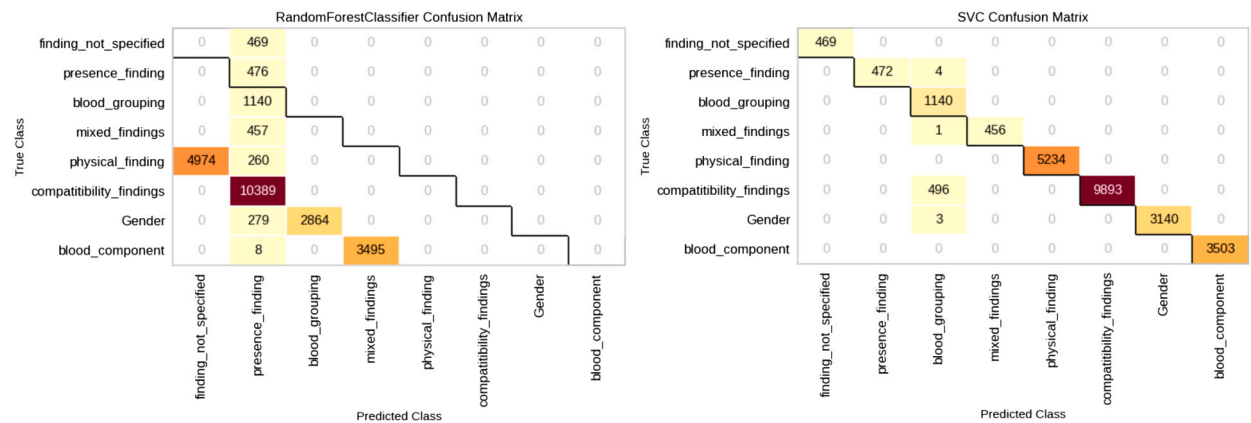
Fig. 3. Comparison of confusion matrices for RF and SVM models for categorizing test results into different labels.

RF algorithm in terms of accuracy and all other performance metrics, according to our findings. Table 3 shows that SVM outperforms RF in this assignment.

Comparison of confusion matrices for RF and SVM models for categorizing categorical test results into various labels is depicted in Fig. 3. The confusion matrix for the RF model reveals a greater number of erroneous predictions than the confusion matrix for the SVM model. Specifically, the RF model made erroneous predictions across multiple categories, whereas the SVM model made errors only in the blood group finding category.

3.2. Overall accuracy of standardized textual test results in clinical laboratory findings

Comparing various string matching algorithms, Table 4 displays the standardized results of the test across multiple groups in the collected dataset. The table displays the number of instances and percentage of correctly standardized textual test results for each group. The second column of the table contrasts the ability of three distinct string matching algorithms, Jaro Winkler similarity, Cosine similarity, and Levenshtein distance, to standardize unique terms for each group. The results indicate that the "Presence

Table 4

Overall standardization results of test results across multiple groups of collected dataset by comparing different string distance similarity.

Group Label	Standardization Status of Unique Term(%)			Supported Data
	Jaro Wink.	Cos. Sim.	Lev. Dist.	
Presence Finding	21 of 21 (100%)	21 of 21(100%)	21 of 21(100%)	10728 (100%)
Physical Finding	29 of 29 (100%)	29 of 29 (100%)	29 of 29 (100%)	9493 (100%)
Blood Group Finding And Another Measure	23 of 35(65.71%)	15 of 35(42.85%)	23 of 35(65.71%)	4961 of 4961 (99.65%)
Mixed Finding	18 of 19 (94.73%)	18 of 19(94.73%)	18 of 19(94.73%)	29865 of 29867 (99.99%)
Blood Component Finding	14 of 14 (100%)	13 of 14(92.85%)	11 of 14 (78.57%)	1312 (100%)
Compatibility Finding	5 of 6 (83.33%)	6 of 6 (100%)	6 of 6 (100%)	1336 of 1337 (99.92%)
Gender Finding	4 of 4 (100%)	4 of 4 (100%)	4 of 4 (100%)	1466 (100%)
Non Specific Finding	9 of 9 (100%)	8 of 9(88.88%)	8 of 9(88.88%)	15900 (100%)
Total Correctly Standardize	92.97(%)	89.91%	90.98%	99.93%

*Note: Jaro Wink. stands for Jaro Winkler Similarity, Cos. Sim. Stands for Cosine Similarity and Lev. Dist. stands for Levenshtein distance.

Finding", "Physical Finding", "Blood Component Finding", "Gender Finding", and "Non-Specific Finding" group are standardized successfully by Jaro Winkler similarity at a rate of 100%. Cosine similarity performs inadequately for "Blood Group Finding and Another Measure" because it misinterprets the "+" sign as an operator, rather than as part of the term, resulting in inaccurate similarity measurements, the standardization accuracy for this method is 42.85%, where another two have 65.71%. In contrast, the Levenshtein distance metric performs poorly for "Blood Component Finding" which contains lengthier terms than usual. This may be because the Levenshtein distance measure is based on the number of editing required to transform one term into another, and longer terms require more edits, which may not reflect the similarity between the terms accurately. So it showed 98.57% where Jaro-Winkler showed 100% and Cosine similarity showed 92.85%. For the standardizing of terms in unique counts, the Jaro Winkler showed a 92.97% accuracy compared with Cosine Similarity (89.91%) and Levenshtein Distance (90.64%). In standardizing categorical test results, Jaro Winkler Similarity obtained higher success rates than Cosine Similarity and Levenshtein Distance. The last column of the table named "Supported Data" concludes by displaying the total number of categorical test results terms data standardized by each group with the percentage of data points when only Jaro-Winkler similarity is considered. The aggregate success rate of the standardization for overall categorical test results is 99.93%, which is remarkably high.

3.3. Dataset analysis and standardization status

The distribution of test results for each of the eight groups is provided in the current section; one group is shown in Table 3 (for more detailed information of the remaining 7 groups, see Table S1 to S7 in Supplementary Material - Additional Performance Results for the Collected Dataset). The source terms and their associated standardized terms, along with their proportion and count, are summarized in these tables. However, some rows have gray markings suggesting that these test results are not standardized properly.

It can be observed that most source terms from "Blood Group Finding And Another Measures" group have been appropriately standardized in Table 5, which comprises 4961 total terms together with their standardized terms from this approach. The greatest percentage is for "B" at 24.53% and the lowest for "Srock" at 0.02%. For instance, the ML model misgrouped the numerous source phrases such as "M","Q.N.S", "Absent" "Incompatible (Minor)", "Packed cell" etc., and produced false results. 17 text terms out of 4961 in the group are not standardized.

In the "Mix Finding And Another Measure" group, the vast majority of the result terms (61.62%) were labeled as "Nil," which is standardized itself. "Absent" (24.35%) and "Normal" (12.41%) are the next two most frequent terms. The remainder category contains less than 2% of all terms, such as "Trace", "Few", "Present", "A Few", and "Occasional". Each one of them perfectly standardized to standard terms. Yet, it is incorrect to standardize "No microorganism found" as "Normal". As can be seen, the entries "Present (++)" and "Present (+)" have been standardized to "Present."

In the "Presence Finding" group, out of a total of 10,728 counts, 50.39% (5,406) are "Negative", and 25.69% (2,756) are "Positive". This technique standardizes a variety of modifications of negative and positive test results, such as "Negative (-ve)" to "Negative" and "Positive (+ve)" to "Positive". Generally, the standardization procedure was successful in resolving the majority of source terms' anomalies.

The "Physical Finding" group has a total of 9,493 terms. "Clear" and "Straw" are the two most frequently used standardized terms itself, making up 45.34% and 45.6% of the overall count, respectively. Once more, several anomalies can be found in the source text results. For example, the names "STRAW" and "L YELLOW" are standardized to "Straw" and "Yellow" respectively, although with significantly lower percentages of 0.25 and 0.11, respectively.

Totaling 1332 results, the "Blood Component Finding" group has 13 unique result terms. "Whole blood" comprises 44.34% (634) of all results, followed by "Packed Red Blood Cells (PRBC)" at 25.73% (368). CCP, STOCK Blood, and Pure Plasma are the least common types of results, accounting for only 0.35% (5) of all results. In this example, the Jaro-Winkler approach is well-suited for standardizing terms such as "FFP" and "PRBC" into more standard terms such as "Fresh Frozen Plasma" and "Packed Blood Cell".

There are 6 distinct result terms in "Compatible Finding" group and they add up to 1337. "Compatible" terms make up 97.01% (1297) of all results, while "Incompatible" results make up 2.24% of all results (30). For this group, all terms from the source were standardized to the proper term.

Table 5
Summary table of source terms and standardized terms for the **Blood Group Finding And Another Measures**.

Source Term	Std. Term	Percentage (%)	Count
"B"	B	24.53	1217
"O"	O	23.22	1152
(+)	+	18.63	924
"A"	A	12.5	620
(++)	++	7.4	367
"AB"	AB	7.16	355
(+++)	+++	3.45	171
"B"	B	0.6	30
"O"	O	0.6	30
(++++)	+++	0.44	22
++	++	0.22	11
"A"	A	0.22	11
+++	+++	0.16	8
+	+	0.14	7
"AB"	AB	0.12	6
++++	+++	0.1	5
M	A	0.06	3
Q.N.S	A	0.06	3
Packed cell	A	0.04	2
B	B	0.04	2
Redish	A	0.02	1
RED	A	0.02	1
RCC	A	0.02	1
O	O	0.02	1
M/F	A	0.02	1
Cloudy	A	0.02	1
InCompatible(Minor)	A	0.02	1
F	A	0.02	1
Ansent	A	0.02	1
AB	AB	0.02	1
"A"	A	0.02	1
"B":	B	0.02	1
"O'"	O	0.02	1
"A"	A	0.02	1
Srock	A	0.02	1
Total Count			4961

The terms in the "Gender Finding" group are male, female, and variants of male's spelling. The breakdown is 85.81% "Male", 13.98% "Female", and 0.14% "male" and 0.07% "MAle" variants. All terms are standardized with perfection.

3.4. Error analysis

To understand the capability of this standardization, if we look attentively at the table, we will find that some errors are seen in the three groups such as 'Blood Group Finding And Another Measure', 'Mixed Finding', and 'Compatibility Finding', as shown in Table 5. Analysis of the category "Blood Group Finding And Other Measures" in Table 4 reveals a mixture of successful standardization and errors. Despite the fact that the vast majority of terms are successfully standardized, there are a few terms that are not perfectly standardized. These terms include "M", "Q.N.S.", "Packed cell", "Reddish", "RED", "RCC", "O", "M/F", "Cloudy", "InCompatible(Minor)," "F", "Ansent", and "Srock" described in Section 3.2. It can be seen that these terms do not belong in this group, it is possible to conclude that the ML system incorrectly classified them. The plot in Fig. 4 depicts the error rate of standardization for each of the groups. We see that three groups: "Blood Group Finding And Another Measure," "Mixed Finding," "Compatibility Finding" and "Blood Group Finding And Another Measure" have error in standardization. The "Blood Group Finding And Another Measure" have 34.29% error rate in standardization for the incorrect classification of ML. This indicates that accurate standardization of terms within this group presents challenges. In contrast, the "Mixed Finding" and "Compatibility Finding" groups exhibit reduced error rates of 5.26% and 16.21%, respectively.

3.5. Validation on MIMIC-III dataset

To validate and evaluate the robustness of the proposed standardization module, the entire procedure was evaluated using the MIMIC-III data set. By analyzing the dataset, five groups were identified, and the terms for each group were extracted with text mining techniques and labeled with their corresponding group names manually. Then, the application of ML models - RF and SVM for automatic labeling or classification perfectly classified text reports of the dataset into the five predefined groups. The RF model obtained an accuracy of 0.80 and a weighted average F1 score of 0.76, while the SVM model with a linear kernel achieved a faultless score across all performance metrics mean 100% score over all the metrics.
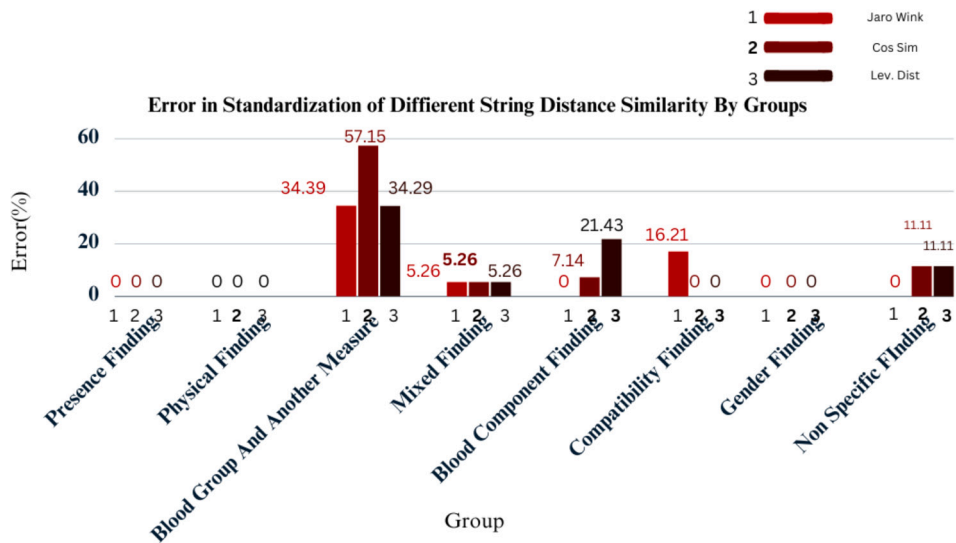
Fig. 4. Error rate of standardization for each string distance similarity by groups.

Table 6
Overall standardization results of categorical test results with Jaro Winkler similarity across multiple groups Of MIMIC-III dataset.

Group Level Label	Standardize Status of Unique Value (%)	Supported Data
Presence Finding	4 of 4 (100%)	2002 (100%)
Physical Finding	8 of 8 (100%)	556(100%)
Blood Group Finding And Another Measure	3 of 3 (100%)	786(100%)
Mixed Finding	13 of 13(100%)	1775(100%)
Non Specific Finding	5 of 5 (100%)	538 (100%)
Total Correctly Standardize		100%

Only Jaro-Winkler similarity is utilized to standardize the categorical test results for the MIMIC-III dataset, as this method in Section 3.3 demonstrated the maximum accuracy among the three distance similarity methods. For more detailed information on the status of the standardization of the MIMIC-III dataset's categorical test results from various groups, ML model's evaluation, and confusion matrix, please refer to Supplementary Material-Additional Performance Results for the MIMIC-III Dataset.

All of the categorical test results from the MIMIC-III dataset have been accurately and flawlessly standardized. The distribution of terms in this dataset and its standardization status is described below. In the "Presence Finding" group, 95.45% of terms are "neg" standardized as "Negative", 2.4% are "negative" standardized as same, and 2.5% are "pos" standardized as "Positive". The majority of terms in the "Physical Finding" group were "clear" (36.87%), followed by "yellow" (32.01%) and "straw" (12.23%). The rest of the terms make less than 10%. All of the terms are precisely standardized. We find that "slhazy" standardize to "Hazy" and "slcloudy" to "Cloudy". In the "Mix Finding" group, the most common terms are "usual" (31.66%), "hold" (17.58%), and "occasional" (14.25%). This group's standardization of "occ" to "Occasional" is perfect. Perfect standardization can be shown in all terms of the "Blood Group Finding and Another Measure" and "Non-specific Finding" from the MIMIC-III dataset. Using Table 6 to evaluate the methodology for MIMIC-III dataset, we can confidently conclude that the standardization of categorical test results across multiple groups in the dataset is highly effective. Standardizing textual test results is a complete success across all groups,

The success of the categorical test results in the standardization procedure relies heavily on the accuracy of the data's manual labeling. For the standardization of the results to be successful, the dataset must be appropriately labeled.

3.6. Comparison with previous studies

This section provides an analysis of our study with existing studies. We applied our technique and an existing method [13] to the publicly accessible MIMIC-III [1] dataset for fair comparison. For standardization of the groups of Presence Finding, Physical Finding, Mix Finding (Normal, Occasional), and Non-Specific Finding, the existing SALT-C method [13] obtained accuracy values of 100%, 100%, 90%, and 80%, respectively. The SALT-C method cannot perform well for the group "Blood Group and other Measures (+, ++, +++)", hence modifications are required in their rule-based approach. On the other hand, our proposed method shows 100% accuracy in standardization in all five groups. Hence, our approach is better than the existing rule-based method.

4. Discussion

4.1. Principal findings

This research paper's key finding is a new method for standardizing clinical textual results using standardized terminologies such as SNOMED-CT and LOINC that combine ML and string distance similarity. It can significantly improve the accuracy and consistency of clinical data interpretation. We used international clinical vocabularies to create predefined groups that covered over 95% of categorical data from the source. By creating an algorithm that grouped source clinical terms and compared distance similarity between source text and standard text from the group, we were able to accurately standardize clinical textual results.

In contrast to previous research [13] that focused on standardizing the test result using rule-based classifiers on a smaller number of groups and distance similarities such as Cosine similarity or Levenshtein distance, we used the supervised ML model SVM achieving a 98% accuracy rate in classifying textual terms and accurately grouping them into more predefined groups. For standardizing, we used Jaro-Winkler distance similarity to standardize the test results. However, to make a fair comparison, different approaches should be applied to the same dataset. Previous studies did not utilize our collected dataset, and as a result, their results cannot be directly applied to the dataset, as they only utilized a smaller number of groups. Existing approaches based on Cosine or Levenshtein distance achieved approximately 90% accuracy when applied to our dataset, whereas our approach achieved 92.97% accuracy. The findings of this paper demonstrate that the use of standardized terminologies can effectively improve the accuracy and consistency of clinical data interpretation.

This research has significant implications for healthcare professionals, as it streamlines the process of interpreting clinical results and ensures consistency in the interpretation of results across healthcare service providers. The use of standardized terminologies also reduces errors and misinterpretations, which can have a significant impact on patient outcomes. By creating predefined groups using international clinical vocabularies and developing a guideline to standardize clinical textual results, this research has demonstrated the effectiveness of using standardized terminologies in clinical data interpretation. This methodology can be implemented in diverse healthcare environments, resulting in enhanced patient outcomes.

4.2. Future work

The limitations identified in this study provide opportunities for enhancing the proposed procedures for standardizing categorical test results across a wider range. Firstly, future research can validate the application's accuracy using datasets from a broader range of healthcare systems to ensure its applicability in a wide range of settings. This may enhance the generalizability of the findings and provide more convincing evidence of the method's effectiveness. Secondly, future research can expand the vocabulary groups used in the study to increase the accuracy and confidence in the results. This can be achieved by adding additional terms and groups to encompass a wider range of laboratory tests and test attributes. In addition, modifications can be made to the preprocessing and labeling steps to increase its adaptability to various datasets. Thirdly, the preprocessing and labeling steps were tailored to the particular datasets used in the study; therefore, modifications may be required to apply the tools to a different dataset. Importantly, the study excluded certain data, such as bacterial and fungal information obtained through the Gram stain test and categorical data exceeding twenty characters. This was done to simplify the process, but it may have limited the study's findings. Future research can investigate the impact of including bacterial and fungal information obtained through the Gram Stain Test and categorical data exceeding twenty characters in length on the guidelines' accuracy. This can provide a more comprehensive understanding of the guidelines' potential to manage a wider range of laboratory data. Furthermore, even though the terminology used in this study is taken from LOINC or SNOMED CT, there are several terms in the groups "Physical Finding" and "Blood Component Finding" that are not found in those sources. As a consequence, we depend on the most frequently used resulted text from the datasets, which is free of anomalies. Even though the study provides practical guidelines for managing categorical test results and demonstrates good accuracy, additional analysis and testing are necessary to determine its effectiveness and applicability to a broader range of datasets.

5. Conclusion

In our study, we presented a clinical data standardization process that facilitates the mapping of test result data to LOINC and SNOMED CT standard texts. This process was applied to a dataset obtained from the two largest private hospitals in Bangladesh and the publicly available MIMIC-III dataset. Unlike previous studies that focused on standardizing test results using rule-based classifiers on a smaller number of groups and distance similarity such as Cosine similarity or Levenshtein distance, we used supervised ML algorithms to classify the term into groups and Jaro-Winkler distance similarity to standardize the test results, also known as categorical clinical data. Using supervised ML algorithms enabled us to improve the accuracy and efficiency of our data categorization process. For the cases considered, the Jaro-Winkler similarity method performed better than other distance similarity methods. Overall, our study offers a novel method for standardizing test results data, which has significant implications for enhancing the interoperability and integration of healthcare data across diverse systems and research platforms.

6. ACRONYM

EHR   Electronic Health Record
OpenMRS Open Medical Record System

NGO   Non-Governmental Organisation
LOINC   Logical Observation Identifiers Names and Codes
SNOMED CT   Systematized Nomenclature of Medicine Clinical Terms
DHIS2   District Health Information System 2
RF   Random Forrest
SVM   Support Vector Machine

CRediT authorship contribution statement

Syed Ahmmed: Designed the study, Developed the algorithm, Designed numerical studies, Performed the experiments and obtained results, Performed data analysis, Designed the content and structure of this article, Drafted the manuscript. **M. Rubaiyat Hossain Mondal:** Defined the scope, Designed the study and interpreted the data, Coordinated the research project, Designed the content of this article, Critically revised the important intellectual content of the manuscript. **Md Raihan Mia:** Data curation, Dataset Collection, Designed the study and interpreted the data, critically revised the important intellectual content of the manuscript. **Mohammad Adibuzzaman:** Interpreted the data, Critically revised the important intellectual content of the manuscript. **Abu Sayed Md. Latiful Hoque:** Initiator of this work, Dataset collection, Designed the study and interpreted the data, Critically revised the important intellectual content of the manuscript. **Sheikh Iqbal Ahamed:** Interpreted the data, Critically revised the important intellectual content of the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e21523.

References

[1] S.I. Khan, A.S.M.L. Hoque, Towards Development of National Health Data Warehouse for Knowledge Discovery, Intelligent Systems Technologies and Applications, vol. 2, Springer, 2016, pp. 413–421.
[2] M.H. Lopez, E. Holve, I.N. Sarkar, C. Segal, Building the informatics infrastructure for comparative effectiveness research (cer): a review of the literature, Med. Care (2012) S38–S48.
[3] C. Safran, M. Bloomrosen, W.E. Hammond, S. Labkoff, S. Markel-Fox, P.C. Tang, D.E. Detmer, Toward a national framework for the secondary use of health data: an American medical informatics association white paper, J. Am. Med. Inform. Assoc. 14 (2007) 1–9.
[4] M.G. Kahn, M.A. Raebel, J.M. Glanz, K. Riedlinger, J.F. Steiner, A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research, Med. Care 50 (2012).
[5] J.F. Burnum, The misinformation era: the fall of the medical record, Ann. Intern. Med. 110 (1989) 482–484.
[6] T. Botsis, G. Hartvigsen, F. Chen, C. Weng, Secondary use of ehr: data quality issues and informatics opportunities, Summit Transl. Bioinform. 2010 (2010) 1.
[7] Logical Observation Identifiers Names and Codes (LOINC), https://loinc.org/. (Accessed 18 April 2023).
[8] SNOMED CT, https://www.snomed.org/snomed-ct. (Accessed 18 April 2023).
[9] M.A.H. Khan, A.K. Azad, V. de Oliveira Cruz, Bangladesh's digital health journey: reflections on a decade of quiet revolution, WHO Southeast Asia J. Public Health 8 (2019) 71–76.
[10] S.I. Khan, A.S.L. Hoque, Privacy and security problems of national health data warehouse: a convenient solution for developing countries, in: 2016 International Conference on Networking Systems and Security (NSysS), IEEE, 2016, pp. 1–6.
[11] M.R. Mia, A.S.M.L. Hoque, S.I. Khan, S.I. Ahamed, A privacy-preserving national clinical data warehouse: architecture and analysis, Smart Health 23 (2022) 100238.
[12] R.G. Hauser, D.B. Quine, A. Ryder, Labrs: a Rosetta stone for retrospective standardization of clinical laboratory test results, J. Am. Med. Inform. Assoc. 25 (2018) 121–126.
[13] M. Kim, S.-Y. Shin, M. Kang, B.-K. Yi, D.K. Chang, et al., Developing a standardization algorithm for categorical laboratory tests for clinical big data research: retrospective study, JMIR Med. Inform. 7 (2019) e14083.
[14] F. Alhazmi, The ethical challenge of conflicts of interest in healthcare, 2019.
[15] A.E.W. Johnson, T.J. Pollard, R.G. Mark, Mimic-iii clinical database (version 1.4), https://doi.org/10.13026/C2XW26, 2016.
[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[17] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[18] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[19] D.M. Powers, Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation, J. Mach. Learn. Technol. 2 (2011) 37–63.

[20] T. Kluyver, B. Ragan-Kelley, F. Pérez, B.E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, C. Willing, Jupyter notebooks – a publishing format for reproducible computational workflows, in: Positioning and Power in Academic Publishing: Players, Agents and Agendas, 2016, pp. 87–90.

[21] W.H. Gomaa, A.A. Fahmy, Simall: a flexible tool for text similarity, in: The Seventeenth Conference on Language Engineering ESOLEC, volume 17, 2017, pp. 122–127.

[22] T. Kenter, M. De Rijke, Short text similarity with word embeddings, in: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, 2015, pp. 1411–1420.

[23] E.S. Ristad, P.N. Yianilos, Learning String Edit Distance, Proceedings of the Fifteenth International Conference on Machine Learning, vol. 1, Morgan Kaufmann Publishers Inc., 1998, pp. 412–420.

[24] M.A. Jaro, Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida, J. Am. Stat. Assoc. 84 (1989) 414–420.

[25] W.E. Winkler, String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage, J. Am. Stat. Assoc. 85 (1990) 274–284.

[26] W.W. Cohen, P. Ravikumar, S.E. Fienberg, A comparison of string distance metrics for name-matching tasks, in: Proceedings of the IJCAI-03 Workshop on Information Integration, volume 3, 2003, pp. 73–78.

[27] V.I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, Sov. Phys. Dokl. 10 (1966) 707–710.

[28] J. Euzenat, P. Shvaiko, et al., Ontology Matching, vol. 18, Springer, 2007.