**Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control**
**Journal homepage: http://kinetik.umm.ac.id**
**ISSN: 2503-2267**
**Vol. 8, No. 2, May, Pp. 581-590**


# Leveraging text-mining techniques on electronic medical records to analyze national drug-insured medication use

**Adhi Dharma Wibawa*[1], Prio Adi Ramadhani[2,4], Ghulam Asrofi Buntoro[2], Ridho Rahman Hariadi[3], Putri Alief Siswanto[1], Shoffi Izza Sabilla[1]**
Department of Medical Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia[1]
Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia[2]
Department of Information Technology, Institut Teknologi Sepuluh Nopember, Surabaya [3]
National Research and Innovation Agency, Jakarta, Indonesia[4]

**Article Info**
**Keywords:**
Text-Mining, Electronic Medical Records, National Drug-Insured, Regular Expression, Levenshtein Distance

**Article history:**
Received: March 26, 2023
Accepted: May 08, 2023
Published: May 31, 2023

**Cite:**
A. D. Wibawa, P. A. Ramadhani, G. A. Buntoro, R. R. Hariadi, P. A. Siswanto, and S. I. Sabilla, "Leveraging Text-Mining Techniques On Electronic Medical Records to Analyze National Drug-insured Medication Use", KINETIK, vol. 8, no. 2, May 2023. https://doi.org/10.22219/kinetik.v8i2.1695

*Corresponding author.
Adhi Dharma Wibawa
E-mail address:
adhiosa@te.its.ac.id

**Abstract**
Processing electronic medical record (EMR) data has become a common practice among scientists for extracting valuable insights and studying diseases. Given the large volumes of text data in EMRs, efficient computerized text-mining techniques are necessary. As academics, we recognize that drug-used analysis from EMR data in Indonesia is currently limited. This study focuses on obtaining meaningful insights from EMR data to make positive recommendations for hospitals. The proposed method uses pattern-based Regular Expressions (regex) to extract drug names and a Levenshtein distance algorithm to check their compatibility. We developed the pattern based on analyzing Indonesia EMR data. The extracted drug names were compared to a list of selected drugs (National Drug-Insured/Fornas) that are required and must be provided at healthcare facilities in Indonesia. The Levenshtein distance threshold was set to two to decide whether the extracted drug names belonged to nationally drug-insured or not. Only about 11.09 – 16.11% of medications given by doctors are listed in the Fornas drug list. Between 2019 and 2021, there was an inaccuracy in the writing of prescriptions for Fornas drugs, with as many as 57.53% to 63.21% of drug names being written incorrectly. The results of this study indicate that the Levenshtein distance algorithm has promising potential for implementation in the Ministry of Health of Indonesia, with a precision rate of 97.07%.

## 1. Introduction

Among the goals of electronic medical records (EMR) is to improve the accuracy and accessibility of information about the health of patients in hospitals [1]. Ideally, successful implementations of EMR will improve patient care and safety [2]. If carried out with compliance and discipline, EMR in a hospital will become a reliable source of big data that can be further processed. A comprehensive overview of healthcare information can be captured from electronic medical record (EMR) data [3], [4]. This study focuses on obtaining meaningful insights from EMR data to make positive recommendations for hospitals.

It is almost impossible to manually process hundreds of thousands per year of EMR data records as it would take a very long time and be prone to human error [5], therefore text processing technology is a necessity. To the best of our knowledge, there have only been three papers devoted to the topic of EMR data processing in Indonesia, namely the study by [6], [7], and [8], indicating a limited number of studies on this subject. The study by [6] used natural language to map patient complaints based on EMR data to support clinical decisions. The study by [7] used machine learning methods to classify diseases from EMR. While the study by [8] used text mining and machine learning methods to classify patient diagnoses based on EMR text data. All previous studies that have processed EMR data in Indonesia have not investigated the accuracy of prescription writing, exploration of new insights through prescriptions, or analysis of national drug-insured usage. The application of EMR itself in Indonesia is still limited [9]. To fill this gap, conducting a real study of the EMR data of hospitals in Indonesia using text-mining techniques is necessary. Based on the study by [10], text mining has shown promising results in revealing new insights in the medical domain.

Some important information that can be explored from EMR data is diagnosis, complaints, and medications. Text-mining based on patient diagnoses and complaints in Hospital EMR data in Indonesia has been carried out by [11] and [12] in 2021 and 2022, but no one has yet explored insights into drug prescription data. Linking diagnostic data to prescription data has the potential to yield unprecedented insights. Moreover, there is Indonesia's national health insurance system which requires a list of selected drugs that are needed and must be available at health service facilities. This list of drugs is regulated in a National Formulary (National Drug-Insured), abbreviated as Fornas in Indonesian [13]. The statistics and distribution of Fornas drug use in a hospital is still a question that cannot be answered

immediately. Therefore, this study will focus more on processing prescription data contained in the EMR to get meaningful insights regarding Fornas drug medication use in Indonesia.

The writing of drug names in the EMR should ideally be following the newest edition of the Indonesian Pharmacopeia or International Non-proprietary Names (INN) or generic names issued by WHO. What percentage of prescription writing follows the rules is still a question, whether the number is only a few or actually quite high. To ensure this, it is necessary to do text processing that can automatically detect the suitability of the drug name. Previously it was common knowledge that the writing of drug names in Indonesia was taken or absorbed from the original name of the drug, such as the drug Gliclazide which in the Indonesian pharmacopeia is written as Gliklazid.

This paper addresses the challenge of extracting drug names from prescription writing in Electronic Medical Records (EMR), which is often unstructured. To tackle this problem, we propose using the Regular Expressions (regex) technique to identify drug names based on prescription writing patterns. By identifying the appropriate pattern, drug name extraction can be accomplished quickly and accurately. The other contribution of this study is to detect and quantify incorrect drug writings in the EMR data. The detection is performed using the Levenshtein distance calculation by comparing drug names in the EMR with those in the Fornas list. Levenshtein Distance will calculate the difference between words or between phrases as sequences by counting the difference in the characters of the letters [14].

This study aims to mine valuable insights related to national drug-insured medication use in Indonesia based on EMR data. We hypothesize that by conducting text-mining of electronic medical records data from Indonesian hospitals, a more complete picture of the state of drug-used and healthcare in Indonesia can be generated and several constructive inputs will be made and discussed in this paper.

## 2. Research Method

The EMR data in this study were collected through informed consent procedures and official permission from the hospital. We process the EMR data from two hospitals in Indonesia. The proposed methodology to process the EMR data in this study can be seen in Figure 1. The EMR data is preprocessed before further processing steps to get normalized text [15]. After preprocessing, the next step is the extraction of the drug's name based on the prescription and the extraction of the International Classification of Diseases (ICD) based on the diagnosis. Then the next step is counting the drug names' frequency by each ICD. In this study, the ICD used is ICD-10 which consists of tens of thousands of internationally recognized diagnosis codes in total [16]. The next step is identifying Fornas drugs using Levenshtein distance and then we map the drugs along with the ICD.

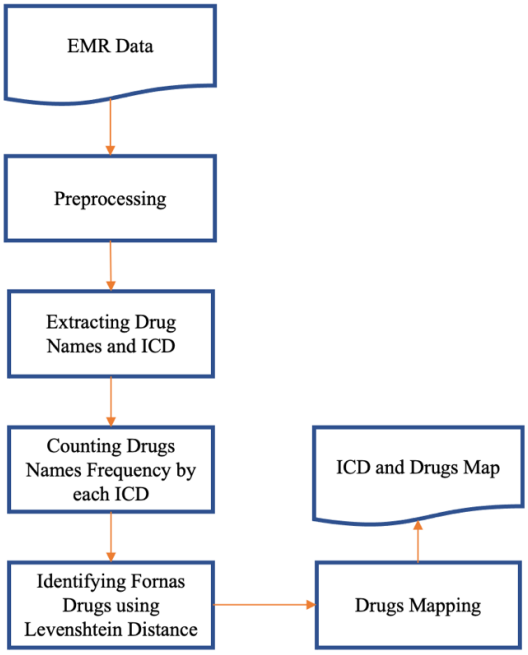


*Figure 1. Proposed Method*

### 2.1 Preprocessing

Preprocessing steps in this study was done on the diagnosis and prescription. Figure 2 shows the flow of preprocessing steps. The preprocessing steps are very important in data-mining, as they make processing easier, shorten processing time, and avoid errors [17]. Figure 2 shows the flow of preprocessing step.

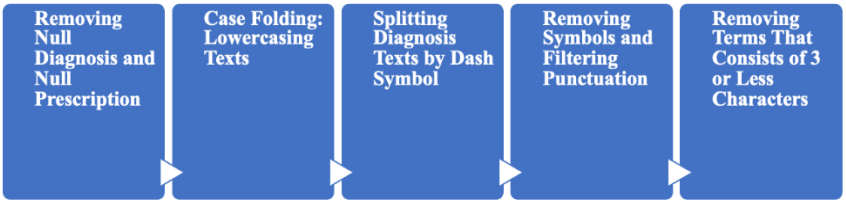


*Figure 2. Preprocessing Steps*

The first step of preprocessing is removing the null diagnosis and null prescription. The second step is case folding, namely lowercasing all the text in diagnosis and prescription to get uniform texts. After that, the diagnosis texts were split by the dash "-" character to get the ICD codes only. Many of the diagnosis texts in this study contain more than one ICD including the description of the ICD. In this study, we assumed that the first written ICD is the primary diagnosis. We use the first written ICD as a single ICD in this study. After some data exploration, there is an exception for ICD Z54.8 - convalescence following other treatments. Z54.8 does not give meaningful information about the disease, because it only tells that the patient is on hospital stay [18]. This ICD is generally followed by other ICDs which are more meaningful. Then we apply a condition, when the ICD is Z54.8 we use the second ICD in the diagnosis text. Some of the ICD in Indonesia EMR data were written without following the ICD-10 rules. For example, the Z54.8 was written Z548. For uniformity purposes, in this study, we remove all the dots in the ICD. The next step is removing symbols and filtering punctuation marks to remove the colon mark and still keep the semicolon mark. Assuming that terms with 3 or fewer characters are non-drug names, the final step in preprocessing involves removing such terms.

### 2.2 Extraction of Drug Names

The extraction of the drug names in this study was done by applying the regex technique. This research uses regex to remove all text or characters that are not drug names. We identified and randomly sampled EMR data to find patterns of drug writing in prescriptions. Based on the identification we conducted, we found a pattern that the drug name is written before the punctuation mark or drug strength. The punctuation mark itself is already removed in a preprocessing step so that the drug name will be always before the writing of drug strength which is always started by a number. The other pattern is the separator among drugs information is always a semicolon or return carriage.

The flow of the extraction of drug names can be seen in Figure 3. The first step is to separate the text of the prescription based on semicolons to separate the drugs written on each prescription. The drug names still contain non-drug name information and other information such as drug strength and drug form. The second step is to find and remove the non-drug names using the regex technique. The third step is to find the drug's strength using regex. Drug strength is detected through the presence of digits in the drug name. The fourth step is to remove the drug strength and all characters after it. Through these four steps, the drug names that need to be extracted will be obtained.

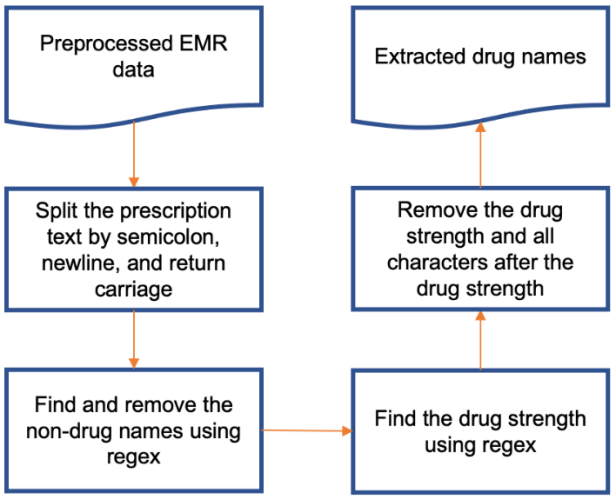


*Figure 3. Extraction of the Drug Names*

### 2.3 Counting the Drug Names Frequency by each ICD and Identifying the Fornas Drugs

The counting of the drug names' frequency was done after the extraction process. The calculation of the frequency of drug names per ICD was carried out to analyze the use of certain drugs in certain diseases. The drugs included in the Fornas list can also be analyzed in terms of the percentage of use compared to all the drugs on the

prescription. A comparative analysis of the number of Fornas drugs can produce a picture of whether the use of Fornas drugs dominates or not. The Fornas drugs were identified by comparing all the extracted drug names with the official list of Fornas drugs issued by the Ministry of Health, Indonesia. However, based on data exploration, we found that the writing of Fornas drug names in the EMR data we processed did not match the writing of the official Fornas drug names. Therefore, in this study, we applied the Levenshtein distance calculation to identify the Fornas drug present in our EMR data. For each drug name that is compared with the official Fornas drug name, the Levenshtein distance value is calculated. If the Levenshtein distance value is less than or equal to two, the extracted drug being compared is labeled as Fornas drug.

By using Levenshtein distance, the non-standard writing of drug names like *gliklazid* in Indonesia can be identified as the same drug as gliclazide. The distance between those two names using Levenshtein calculation is two. The other example is isosorbide dinitrate, which in Indonesia is written as *isosorbid dinitrat*, with a Levenshtein distance equal to two. In this study, the Levenshtein distance threshold was set to two to decide whether the extracted drug names belonged to national drug-insured or not. Levenshtein distance was used to calculate how many different characters there are among the sequence of letters that make up the names of the compared medicines.

### 2.4 Drugs Mapping based on ICD and Fornas Drugs List

The mapping of drugs based on ICD and the Fornas drugs list is intended to obtain an overview of the distribution of national drugs insured for certain diseases. Through this drug mapping, it can be determined how many types of diseases are treated with Fornas drugs. The mapping is carried out by taking the names of drugs that are identified as Fornas drugs in the EMR data and grouping these drugs into each ICD. Each ICD represents a specific patient with a particular disease according to the doctor's diagnosis and prescription in the EMR data record.

## 3. Results and Discussion

The results of this study will be presented and explained in this section. We conducted text-mining on EMR data in this study with a total of 277,043 records from two hospitals in Indonesia. After preprocessing, we obtained 171,619 clean data records. A sample of the data before and after processing can be seen in Table 1. For a more easily understandable and verifiable explanation of the data processing results in this study, please refer to Figure 4. By following the flow outlined in the pseudocode in Figure 4, the processing results can be obtained as shown in Table 1.

Some interesting insights that will be discussed include the low percentage of Fornas drug usage in hospitals and the low number of diseases prescribed using Fornas drugs. The other discussion is the use of psychotropic Fornas drugs, an overview of diseases in Indonesia based on the number of diagnoses, and an evaluation of Levenshtein distance results.

**Algorithm 1** Extracting (*extr_drug*) Drug Names from Prescription (*presc*)

```
non_drug_names = array of non-drug names
function preprocessing presc :
    convert presc to lowercase
    remove symbols and punctuation except semicolons in presc
    remove terms consisting of 2 or fewer characters in presc
    return preprocessed_presc
function processing:
    split the preprocessed_presc text by semicolons, newlines, or return carriages
    for drug_text in the split_presc do
        remove non_drug_names in drug_text
        remove digits and all data after the digits in drug_text
        if drug_text is not empty do
            append drug_text in extr_drug
    return extr_drug
```

*Figure 4. Pseudocode for Extracting Drug Names from Prescription*

*Table 1. Data Capture Before and After Processing*

| Before Processing | | After Processing | | |
|---|---|---|---|---|
| ICD | Prescription | ICD | Prescription | Extracted Drug Names |
| I10 - Essential (primary) hypertension | AMITRIPTYLINE 25 MG TABLET :15.00:0-0-1/2; ANS NEURODEX TABLET :30.00:1-0-0; ASPILET 80 MG TABLET :30.00:0-1-0 | I10 | ['amitriptyline 25 mg tablet 15.000-0-1/2', 'neurodex tablet 30.001-0-0', 'aspilet 80 mg tablet 30.000-1-0'] | ['amitriptyline', 'neurodex tablet', 'aspilet'] |
| Z548 - Convalescence following other treatment; M4306 - Spondylolysis, lumbar region | AMITRIPTYLINE 25 MG TABLET :4.00:1/2-0-1/2; ANS PREGABALIN 75 MG CAPSUL :7.00:0-0-1; ANS MECOBALAMIN 500 MCG TABLET :14.00:2 dd tab 1 | M4306 | ['amitriptyline 25 mg tablet 4.001/2-0-1/2', 'pregabalin 75 mg capsul 7.000-0-1', 'mecobalamin 500 mcg tablet 14.002 dd tab 1'] | ['amitriptyline', 'pregabalin', 'mecobalamin'] |
| I10 - Essential (primary) hypertension | ANS AMLODIPIN 10 MG TAB **:30.00:1-0-0; asam mefenamat 300mg Tramadol 25mg Diazepam 1mg Amitriptilin 5mg Caffeine 20mg mfla pulv da in caps:15.00:3 dd tab 1; NATRIUM DICLOFENAK 50MG TABLET :10.00:2 dd tab 1 | I10 | ['amlodipin 10 mg tab **30.001-0-0', 'asam mefenamat 300mg', 'tramadol 25mg', 'diazepam 1mg', 'amitriptilin 5mg', 'caffeine 20mg', 'mfla pulv da in caps15.003 dd tab 1', 'natrium diclofenak 50mg tablet 10.002 dd tab 1'] | ['amlodipin', 'asam mefenamat', 'tramadol', 'diazepam', 'amitriptilin', 'caffeine', 'mfla pulv', 'natrium diclofenak'] |

### 3.1 Percentage of National Drug-Insured (Fornas) Drugs

The total number of official Fornas drugs is 658 based on data from the Ministry of Health, Indonesia. We calculated the percentage of national drug use in Hospital A and Hospital B compared to the total available Fornas drugs. It was found that the use of Fornas drugs in Hospital A was only around 11.09 - 16.11%. While in Hospital B the use of Fornas drugs is 23.56%. In terms of ICDs, there are 700 ICDs in Hospital A and 16 ICDs in Hospital B that do not contain Fornas drugs in drug prescriptions.

There are several possible reasons for the small percentage of Fornas drug use in these two hospitals. The first possibility is following the findings in previous studies that the procurement of national medicine in Indonesian hospitals has many problems so its availability cannot be ensured in hospitals [19]. The next possibility is that the number of patients registered with the Indonesian National Health Insurance is less than the number of regular patients who come to the two hospitals discussed in this study. But on the other hand, there is the fact that the number of Indonesians registered with the National Health Insurance system reaches 88% [20][21]. In this way, the possible reason for the small percentage of Fornas drug use is closer to the first possibility, namely because of the problem of the availability of Fornas drugs in hospitals.

We use the Levenshtein distance to identify the Fornas drugs list contained in the prescription. If a drug name on the prescription is misspelled with the official Fornas drug name by a Levenshtein distance value of 1 or 2, it is considered incorrect. As a result, there are as many as 57.53% - 63.21% inaccuracies in the writing of drug names between 2019 and 2021. This inaccuracy causes checking Fornas drug names on the website will not show the desired results. Therefore, we recommend that the EMR system of hospitals in Indonesia be connected to the Fornas website for checking. In addition, the Hospital Information System for recording the EMR should also decrease the typing methods by the hospital staff that tend to create errors in inputting the drug's name.

### 3.2 The Use of Psychotropic Fornas Drugs

The use of psychotropic drugs should receive special attention. One of the high risks due to the consumption of psychotropic drugs for a long time is the increased risk of falls [22], [23]. One of the most misused psychotropic drug is Alprazolam [24]. In this study, by mapping the drug names and the ICD, we analyze the situation of the use of psychotropic fornas drugs in Indonesia. We found that there was some use of psychotropic drugs. The psychotropic drugs used belong to groups 3 and 4. Group 3 psychotropics have a moderate level of addiction, while group 4 has a low level of addiction. The percentage of psychotropics drugs use in 2019 was 2%, in 2020 it remained 2%, and in 2021 it was 1%. The use of psychotropic drugs in groups 3 and 4 does not violate regulations in Indonesia, but their use is still very strict and must be according to a doctor's prescription. Consumption of psychotropic drugs, even with moderate and low levels of addiction, is still at risk of death if consumed excessively [25].

According to the study by [26], the administration of Alprazolam to patients other than anxiety disorders can be performed on patients with heart disease and hypertension. Our study found that Alprazolam, apart from being prescribed for anxiety disorders, was also prescribed at a high frequency to patients with atherosclerotic heart disease, essential (primary) hypertension, and type-2 diabetes mellitus. According to [26], Alprazolam appears to be a promising treatment for alleviating anxiety in patients suffering from heart failure or myocardial infarction. Therefore, the decision of doctors in Indonesia prescribes Alprazolam to heart disease patients is correct in terms of professional healthcare services. Giving Alprazolam to hypertensive patients is also a correct decision because according to the findings by [27], in hypertensive individuals with a systolic blood pressure of more than 160 mmHg, Alprazolam is equally effective as captopril in decreasing blood pressure.

### 3.3 Overview of Drugs and Diseases in Indonesia

The total number of unique drug names from Hospital A is 111,687 and from Hospital B is 503. We merge all the drugs names and rank them based on the frequency of occurrence. The top 10 drugs names are Furosemide, Spironolactone, Amlodipine, Clopidogrel, Candesartan, Nitrokaf Retard, Piracetam, Aptor, V-Bloc, and Bisoprolol. Table 2 shows more detailed information about the extracted drug names.

*Table 2. Top 10 Extracted Drug Names*

| Drug Name | ICD | Frequency of Occurences |
|---|---|---|
| Candesartan | I25.1 & I10 | 13,805 |
| Furosemide | I25.1 | 11,347 |
| Spironolactone | I25.1 | 9,004 |
| Amlodipine | Z54.8 | 8,537 |
| Clopidogrel | I25.1 | 8,248 |
| Nitrokaf Retard | I25.1 | 7,489 |
| Piracetam | Z54.8 | 6,805 |
| Aptor | Z54.8 | 6,087 |
| V-bloc | I25.1 | 5,939 |
| Bisoprolol | I25.1 | 5,374 |

As stated in the preprocessing section, ICD Z54.8 does not give meaningful information about the disease, and the other ICD written in the EMR should be extracted. But in some cases, ICD Z54.8 was not followed by other ICDs, so ICD Z54.8 is still used in this study. We found that most of the drugs in our EMR data were used to treat heart disease and hypertension. Heart disease and hypertension were indeed the most diseases detected in the EMR data in this study. The top 10 diseases found in this study are shown in Table 3.

The advantage that can be taken by calculating the frequency of occurrence of this drug is to assist in the preparation of the Hospital Drug Needs Plan. Drugs that are used very much should be the focus of management in planning drug purchases in the following years because the process of buying drugs in Indonesia is not simple, especially drugs that are classified as national drugs-insured / Fornas [19]. As stated by the Pharmacy Study Program of the Islamic University of Indonesia in a special report, drugs classified as Fornas drugs are difficult to meet in hospitals [19]. The vacancy of Fornas drugs has resulted in pharmaceutical installations delaying drug purchases.

*Table 3. Top 10 Extracted ICDs*

| Disease | ICD | Frequency of Occurrences |
|---|---|---|
| Convalescence following other treatment | Z54.8 | 31,368 |
| Atherosclerotic heart disease | I25.1 | 23,023 |
| Essential (primary) hypertension | I10 | 15,177 |
| Type 2 diabetes mellitus | E11 | 5,180 |
| Cerebral infarction due to thrombosis of cerebral arteries | I633 | 5,101 |
| Cerebral infarction | I63 | 5,058 |
| Type 2 diabetes mellitus without complications | E119 | 4,767 |
| Hypertensive heart disease without heart failure | I119 | 4,539 |
| Type 2 diabetes mellitus with circulatory complications | E115 | 2,649 |
| Hypertensive heart disease | I11 | 2,626 |

As seen in Table 3, the dominant disease ranging from 2019 – 2021 are heart disease, hypertension, diabetes, and stroke. If it is associated with the level of salt and sugar consumption in Indonesia, it is natural that heart disease, hypertension, and diabetes are the most common diseases. According to the Director General of Disease Control and Environmental Health of the Ministry of Health in 2013, daily salt intake in Indonesia can reach 15 grams, higher than the amount recommended by WHO [28]. In addition, based on Indonesia Basic Health Research in 2007 and 2010 the Indonesian population who often consumes sweet foods and sweet drinks is 65.2 percent [28]. So, with the findings in this study that three of the four most common diseases found in the EMR data are heart disease, hypertension, and diabetes, it shows that it is likely that the level of salt and sugar consumption by people in Indonesia is still high. Cerebral infarction disease is also suspected to be caused by high salt intake based on studies by [29] and [30].

The government must immediately take concrete steps to prevent more Indonesians from experiencing heart disease. ICD I.251 is one of the ICDs included in ischemic heart disease (IHD)[31], where IHD is a leading cause of death worldwide [32]. Prevention of IHD will certainly have an impact on increasing the level of public health in Indonesia and reducing the costs that must be incurred by the government in treating patients with IHD. Reducing salt consumption is one of the initiatives that can be carried out by the Indonesian government as well as by Indonesian society, as recommended by [33] and [34], in addition to reducing alcohol consumption, refined carbohydrates, and eliminating tobacco and trans-fats. Unfortunately, the percentage of smokers in Indonesia is 25.13% compared to the total population of Indonesia [35].

In this study, based on the data in Table 2, it was found that one of the most widely used drugs is Furosemide to treat heart disease patients. Furosemide is a diuretic that functions to reduce sodium [36]. Furosemide is a first-line treatment of hypertension [37] and a second-line agent to treat heart disease [38]. Based on the data obtained on the EMR in this study, it can be said that most heart disease patients in Indonesia also have hypertension. It seems that prevention is better than cure because if Indonesia's National Health Insurance System must cover patients with hypertension and hypertensive diseases, there will be a lot of drugs that have to be used. In addition, the total number of ICD I10 was increasing 24.6 times in 2020 and 31.7 times in 2021 compared to 2019.

### 3.4 Levenshtein Distance Evaluation

To evaluate the Levenshtein distance we count the total number of identified Fornas drugs with a Levenshtein distance less than or equal to 2 and the total number of Fornas drugs that should be identified with that condition. Table 4 shows the total number of uniquely identified Fornas drugs using the Levenshtein distance metric.

*Table 4. Identified Fornas Drugs*

| Identified Fornas Drugs with Levenshtein Distance ≤ 2 (a) | Fornas Drugs that should be Identified with Levenshtein Distance ≤ 2 (b) |
|---|---|
| 239 | 232 |

We calculate the error value (ev) by the percentage of the difference between (a) and (b) using the Equation 1.

$$ev = \frac{|\, a - b \,|}{a}\ x\ 100\% \tag{1}$$

By calculating the above formula, the error percentage is 2.93%. So, the precision is 97.07% for the Levenshtein distance threshold less than or equal to 2.

## 4. Conclusion

We successfully leveraged text-mining techniques on 171,619 EMR data records to produce new and useful insights. In this study, a text-mining method using regex and Levenshtein distance calculation techniques was presented. We successfully extracted the names of drugs from unstructured prescription writing. The study found that there is still a low percentage of national drug-insured prescriptions, ranging from only 11.09% to 16.11%. This condition may be caused by several factors, including the limited availability of national drugs-insured in hospitals.

Additionally, many inaccuracies were found in the writing of the names of national drugs-insured or Fornas drugs in hospitals. There were inaccuracies ranging from 57.53% to 63.21% between 2019 and 2021. This highlights the need for integration of the EMR system in hospitals with the Fornas database of the Indonesian Ministry of Health. The inaccuracies in the drug name writing could be efficiently detected using the Levenshtein distance calculation. The use of Levenshtein distance appears promising with an accuracy rate of 97.07% in this study. The accuracy can be further improved by changing the threshold value in the Levenshtein distance calculation.

In the future, the dataset in this study needs to be labeled with drug names as the ground-truth result of the extraction process. Manual labeling by doctors or pharmacists can be done, but it takes a very long time. Automatic labeling using pre-trained pipelines can also be an option, but the accuracy of the labels may not be perfect.

## Acknowledgment

This work was supported by Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia.

## References

[1] S. Honavar, "Electronic medical records – The good, the bad and the ugly," *Indian Journal of Ophthalmology*, Vol. 68, No. 3, P. 417, 2020. https://doi.org/10.4103/ijo.IJO_278_20

[2] A. de Benedictis, E. Lettieri, L. Gastaldi, C. Masella, A. Urgu, and D. Tartaglini, "Electronic Medical Records implementation in hospital: An empirical investigation of individual and organizational determinants," *PLOS ONE*, Vol. 15, No. 6, P. e0234108, 2020. https://doi.org/10.1371/journal.pone.0234108

[3] K. Adane, M. Gizachew, and S. Kendie, "The role of medical data in efficient patient care delivery: a review," *Risk Management and Healthcare Policy*, Vol. Volume 12, Pp. 67–73, 2019. https://doi.org/10.2147/RMHP.S179259

[4] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, "Data processing and text mining technologies on electronic medical records: A review," *Journal of Healthcare Engineering*, Vol. 2018. Hindawi Limited, 2018. https://doi.org/10.1155/2018/4302425

[5] F. Ridzuan and W. M. N. Wan Zainon, "A Review on Data Cleansing Methods for Big Data," *Procedia Computer Science*, Vol. 161, Pp. 731–738, 2019. https://doi.org/10.1016/j.procs.2019.11.177

[6] S. Kusumadewi, C. I. Ratnasari, and L. Rosita, "Natural language parsing of patient complaints in Indonesian language," in *2015 International Conference on Science and Technology (TICST)*, Nov. 2015, Pp. 292–297. https://doi.org/10.1109/TICST.2015.7369373

[7] A. H. Sangaji, Y. Pamungkas, S. M. S. Nugroho, and A. D. Wibawa, "Rule-based Disease Classification using Text Mining on Symptoms Extraction from Electronic Medical Records in Indonesian," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, 2022. https://doi.org/10.22219/kinetik.v7i1.1377

[8] M. Jamaluddin and A. D. Wibawa, "Patient Diagnosis Classification based on Electronic Medical Record using Text Mining and Support Vector Machine," in *2021 International Seminar on Application for Technology of Information and Communication (iSemantic)*, Sep. 2021, Pp. 243–248. https://doi.org/10.1109/iSemantic52711.2021.9573178

[9] Y. Maryati and A. Nurwahyuni, "Evaluasi Penggunaan Electronic Medical Record Rawat Jalan di Rumah Sakit Husada dengan Technology Acceptance Model," *Jurnal Manajemen Informasi Kesehatan Indonesia*, Vol. 9, No. 2, Pp. 2337–585, 2021. https://doi.org/10.33560/jmiki.v9i2.374

[10] O. Metsker, E. Bolgova, A. Yakovlev, A. Funkner, and S. Kovalchuk, "Pattern-based Mining in Electronic Health Records for Complex Clinical Process Analysis," *Procedia Computer Science*, Vol. 119, Pp. 197–206, 2017. https://doi.org/10.1016/j.procs.2017.11.177

[11] M. Jamaluddin and A. D. Wibawa, "Patient Diagnosis Classification based on Electronic Medical Record using Text Mining and Support Vector Machine," in *2021 International Seminar on Application for Technology of Information and Communication (iSemantic)*, 2021, Pp. 243–248. https://doi.org/10.1109/iSemantic52711.2021.9573178

[12] A. H. Sangaji, Y. Pamungkas, S. M. S. Nugroho, and A. D. Wibawa, "Rule-based Disease Classification using Text Mining on Symptoms Extraction from Electronic Medical Records in Indonesian," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, 2022. https://doi.org/10.22219/kinetik.v7i1.1377

[13] S. Winda, "National Formulary (FORNAS) and e-Catalogue of Medicines as Efforts to Prevent Corruption in Drug Administration of National Health Insurance (JKN)," *Jurnal Integritas*, Vol. 4, No. 2, Pp. 177–206, 2018. https://doi.org/10.32697/integritas.v4i2.328

[14] J. Beernaerts, E. Debever, M. Lenoir, B. de Baets, and N. Van de Weghe, "A method based on the Levenshtein distance metric for the comparison of multiple movement patterns described by matrix sequences of different length," *Expert Systems with Applications*, Vol. 115, Pp. 373–385, 2019. https://doi.org/10.1016/j.eswa.2018.07.076

[15] M. Kashina, I. D. Lenivtceva, and G. D. Kopanitsa, "Preprocessing of unstructured medical data: the impact of each preprocessing stage on classification," *Procedia Computer Science*, Vol. 178, Pp. 284–290, 2020. https://doi.org/10.1016/j.procs.2020.11.030

[16] A. Blanco, S. Remmer, A. Pérez, H. Dalianis, and A. Casillas, "Implementation of specialised attention mechanisms: ICD-10 classification of Gastrointestinal discharge summaries in English, Spanish and Swedish," *Journal of Biomedical Informatics*, Vol. 130, P. 104050, 2022. https://doi.org/10.1016/j.jbi.2022.104050

[17] J. Santos-Pereira, L. Gruenwald, and J. Bernardino, "Top data mining tools for the healthcare industry," *Journal of King Saud University - Computer and Information Sciences*, Vol. 34, No. 8, Pp. 4968–4982, 2022. https://doi.org/10.1016/j.jksuci.2021.06.002

[18] C. Chojenta, J. Byles, and B. K. Nair, "Rehabilitation and convalescent hospital stay in New South Wales: an analysis of 3,979 women aged 75+," *Australian and New Zealand Journal of Public Health*, Vol. 42, No. 2, Pp. 195–199, 2018. https://doi.org/10.1111/1753-6405.12731

[19] I. U. of I. Pharmacy Study Program, "Availability of Medicine in the Era of National Health Insurance," Yogyakarta, 2018.

[20] H. Humas, "BPJS Hears 2022 Nets of Feedback on JKN Management in the Future," Indonesia Health Social Security Administering Agency Official Website, Jul. 24, 2022.

[21] O. Z, "273 Million Indonesian Population Updated Version of the Ministry of Home Affairs," Ministry of Home Affairs Official Website, Feb. 24, 2022.

[22] L. J. Seppala *et al.*, "Fall-Risk-Increasing Drugs: A Systematic Review and Meta-Analysis: II. Psychotropics," *Journal of the American Medical Directors Association*, Vol. 19, No. 4, Pp. 371.e11-371.e17, 2018. https://doi.org/10.1016/j.jamda.2017.12.098

[23] S. Laberge and A. M. Crizzle, "A Literature Review of Psychotropic Medications and Alcohol as Risk Factors for Falls in Community Dwelling Older Adults," *Clinical Drug Investigation*, Vol. 39, No. 2, Pp. 117–139, 2019. https://doi.org/10.1007/s40261-018-0721-6

[24] N. Ait-Daoud, A. S. Hamby, S. Sharma, and D. Blevins, "A Review of Alprazolam Use, Misuse, and Withdrawal," *Journal of Addiction Medicine*, Vol. 12, No. 1, Pp. 4–10, 2018. https://doi.org/10.1097/ADM.0000000000000350

[25] BNN Public Relations, "What is Psychotropic and its Dangers?," BNN Official Website, Jan. 02, 2019.

[26] M. J. Anwar, K. K. Pillai, R. Khanam, M. Akhtar, and D. Vohora, "Effect of alprazolam on anxiety and cardiomyopathy induced by doxorubicin in mice," *Fundamental & Clinical Pharmacology*, Vol. 26, No. 3, Pp. 356–362, 2012. https://doi.org/10.1111/j.1472-8206.2011.00925.x

[27] S. Yilmaz, M. Pekdemir, Ü. Tural, and M. Uygun, "Comparison of alprazolam versus captopril in high blood pressure: A randomized controlled trial," *Blood Pressure*, Vol. 20, No. 4, Pp. 239–243, 2011. https://doi.org/10.3109/08037051.2011.553934

[28] N. Ridarineni and D. Muhammad, "Daily Salt Consumption in Indonesia 15 Grams Higher," Republika News - Leasure, Oct. 06, 2013.

[29] Y. Li *et al.*, "Longitudinal Change of Perceived Salt Intake and Stroke Risk in a Chinese Population," *Stroke*, Vol. 49, No. 6, Pp. 1332–1339, 2018. https://doi.org/10.1161/STROKEAHA.117.020277

[30] M. Hu *et al.*, "High-salt diet downregulates TREM2 expression and blunts efferocytosis of macrophages after acute ischemic stroke," *Journal of Neuroinflammation*, Vol. 18, No. 1, P. 90, 2021. https://doi.org/10.1186/s12974-021-02144-9
[31] S. K. Park *et al.*, "The risk for incident ischemic heart disease according to estimated glomerular filtration rate in a korean population," *Journal of Atherosclerosis and Thrombosis*, Vol. 27, No. 5, Pp. 461–470, 2020. https://doi.org/10.5551/jat.50757
[32] M. A. Khan *et al.*, "Global Epidemiology of Ischemic Heart Disease: Results from the Global Burden of Disease Study," *Cureus*, 2020. https://doi.org/10.7759/cureus.9349
[33] R. Gupta and D. A. Wood, "Primary prevention of ischaemic heart disease: populations, individuals, and health professionals.," *Lancet (London, England)*, Vol. 394, No. 10199, Pp. 685–696, 2019. https://doi.org/10.1016/S0140-6736(19)31893-8
[34] A. Grillo, L. Salvi, P. Coruzzi, P. Salvi, and G. Parati, "Sodium intake and hypertension," *Nutrients*, Vol. 11, No. 9, 2019. https://doi.org/10.3390/nu11091970
[35] H. Litbangkes, "Adult smokers in Indonesia have increased in the last ten years," Health Research and Development Agency, Ministry of Health of Indonesia Official Website.
[36] S. W. Oh and S. Y. Han, "Loop diuretics in clinical practice," *Electrolyte and Blood Pressure*, Vol. 13, No. 1. Korean Society of Electrolyte and Blood Pressure Research, Pp. 17–21, Jun. 01, 2015. https://doi.org/10.5049/EBP.2015.13.1.17
[37] G. C. Roush, R. Kaur, and M. E. Ernst, "Diuretics: a review and update.," *Journal of cardiovascular pharmacology and therapeutics*, Vol. 19, No. 1, Pp. 5–13, 2014. https://doi.org/10.1177/1074248413497257
[38] T. M. Khan, R. Patel, and A. H. Siddiqui, Furosemide. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing.