Jurnal Matematika Vol. 13, No.2,Desember 2023, pp. 95-104 ISSN: 1693-1394
Article DOI: 10.24843/JMAT.2023.v13.i02.p163

ICD-10 WHO Search With RAKE Algorithm

I Gusti Ngurah Lanang Wijayakusuma
Matematika, Universitas Udayana Bali
e-mail: lanang_wijaya@unud.ac.id

I Putu Winada Gautama
Matematika, Universitas Udayana Bali
e-mail: winadagautama@unud.ac.id

***Abstract**: In most cases, clinicians do not use the ICD-10 standard established by the WHO for diagnosing diseases. These issues may result in outcomes that are undesirable from the standpoints of patient safety and the law. The WHO ICD-10 standard collection of diagnoses that can't be effectively searched using MySQL's native search mechanism. Therefore, in order to automatically produce several keywords for each ICD-10 code, academics are interested in analyzing the natural language analysis of WHO ICD-10 data. When diagnosing illnesses, it is envisaged that the availability of numerous types of keywords can lead to more fruitful search results. Natural language analysis, a technique for removing stop words from sentences and simultaneously assessing the semantics of the language from which the keywords will be extracted, makes it possible to do this.*

***Keywords**: Natural Language Analysis, ICD-10, Stop Words, Semantics.*

***Abstrak**: Dalam kebanyakan kasus, dokter tidak menggunakan standar ICD-10 yang ditetapkan oleh WHO untuk mendiagnosis penyakit. Masalah ini dapat mengakibatkan hasil yang tidak diinginkan dari sudut pandang keselamatan pasien dan hukum. Standar ICD-10 WHO adalah kumpulan diagnosis yang tidak dapat dicari secara efektif menggunakan mekanisme pencarian asli MySQL. Oleh karena itu, untuk menghasilkan beberapa kata kunci secara otomatis untuk setiap kode ICD-10, para akademisi tertarik untuk menganalisis analisa bahasa alami dari data ICD-10 WHO. Saat mendiagnosis penyakit, diperkirakan bahwa ketersediaan berbagai jenis kata kunci dapat menghasilkan hasil yang lebih bermanfaat. Analisa bahasa alami, teknik untuk menghilangkan kata-kata berhenti dari kalimat, dan secara bersamaan menilai semantik bahasa dari mana kata kunci akan diekstraksi, memungkinkan dilakukannya hal ini.*

***Kata Kunci**: Analisis Bahasa Alami, ICD-10, Hentikan Kata, Semantik*

1. Introduction

The Covid-19 epidemic, which started in 2019, has taught us that people's health will have a significant influence on other elements of their life. On January 30,2020, the World Health Organization (WHO) declared 2019-n COV to be a Public Health Emergency of International Concern (PHEIC), due to significant increase in confirmed new cases in various countries (Susanna, 2020). Politics, education, and the economy are all significantly impacted by health sector. World Health Organization suggests a safe physical distance of at least one meter from other surrounding people (physical distancing)(World Health Organitation, 2020). In Indonesia, physical distancing has been implemented, although not very successfully. It its intended to avoid direct contact with infected people and possible virus transmission from those who are asymptomatic (Orang Tanpa Gejala/OTG). The policy physical distancing was followed by social distancing, which banned people from gathering close to schools, or in workplaces wet markets, malls, public transport, and religious and wedding ceremonies, amongst other. Peduli Lindungi application was introduced by the government as one of the numerous measures it has taken to combat the Covid-19 pandemic, including the rapid and accurate detection of health issues in specific regions. As a result, the Minister of Health of the Republic of Indonesia enacted Regulation Number 24 of 2022 concerning Medical Records in September 2022 (Sadikin & Laoly, 2022; Satria, 2022). The Need for timely, accurate and representative health care data has become increasingly evident since the first cases of the coronavirus disease 2019 (Covid-19) appeared.

When we discuss medial records, we are essentially discussing how health services are delivered from patient registration to the doctor diagnosis, but the diagnostic given by doctors to the patient's in indonesia is written in a highly non-standard manner. As one of the most commonly used nosologies, the international Classification of Disease (ICD diagnosis codes) are an attractive tool for identifying and tracking cases to spport healthcare surveillance efforts and facilitate epidemiological research (Lynch et al., 2021). Pre-ICD-10, some scholars track the origin of ICD to 1763. The French physician and botanist Dr Franc ois Bossier de Sauvages deLacroix developed a categorization of 10 distinct classes of diseases, which were further divided into 2400 unique diseases (Jetté et al., 2010). Recognizing the importance of disease classification, the first International Statistical Congress helding Brussels in 1853 appointed Jacob Marc d'Espine and William Farr to develop a system of classifying causes of mortality that could be used across borders and languages (Helling et al., 2019; Henderi et al., 2022; Kusuma et al., 2019; Purba & Sondang, 2022; World Health Organitation, 2015).

Although the WHO's ICD-10 document contains the standard for patient diagnosis, indonesia has not full accepted it due to a number of issues that clinicians must deal with, one of which is the challenge is locating the appropriate ICD-10 code for the illness of the

patient being evaluated, furthermore the doctor's knowledge of ICD-10 is inadequate. The finding also demonstrate that the hospital administration has not planned socialization for doctors on disease coding necessitating the development of a system by the hospital to support the use of ICD-10 by medical professionals, because there is also a risk of miscoding when assigning an ICD-10 code to the patient medical record (Kamal et al., 2020; Noor et al., 2014; Nordgaard et al., 2016; Wijayakusuma & Yowani, 2022).

The development of a diagnosis search engine based on the ICD-10 diagnostic keywords is one technological advancement that can help with the coding of this illness. The concept is that clinicians just need to enter a small number of diagnostic keywords, after that, a list of diagnose that have already been assigned an ICD-10 code will appear. Naturally, we anticipate that this search engine will be useful in assisting physicians in providing patient diagnoses in accordance with the WHO's ICD-10 coding standard.

The development of several applications and technologies in the health industry has reached a turning point with the appearance of ministerial regulation. The author predicts that the ICD-10 diagnostic search engine will advance among other technologies at very fast pace.

2. Research Method

The International Statistical Classification of Disease and Related Health Problems, or ICD-10, is in its tenth revision. The World Health Organization (WHO) coded disease and their signs, symptoms , abnormal findings, complaints, social contexts, and environmental factors that have contributed to an accident or illness in ICD-10 (Noor et al., 2014; Nordgaard et al., 2016).

The automatic (or semi-automatic) processing of human language is known as natural language processing or NLP. NLP is the widely used technique to extract key phrases from large chunk of data. Natural language processing (NLP) is ability of a computer program to understand human language as it is spoken (Beltagy et al., 2019). NLP is a component of artificial intelligence (AI). Natural language refers to the way we humans communicate with each other namely, speech and text (Baruni & Sathiaseelan, 2020). It deals with formal language theory, construction method, theorem proof, machine learning, and human-computer interaction in computer science (Armentano et al., 2014). Spelling and grammar checkers, screen readers for blind and low vision users, augmentative and alternative communication, information-seeking, document classification , and document grouping are some of the applications of NLP (Khader et al., 2018; Shenoi et al., 2020)

RAKE is one of the information retrieval industry's methodologies for keyword extraction. RAKE was created due to the discovery that keywords frequently include compound

words but lack conjunctions and stop words. RAKE bases its automated keyword generation process on a database of conjunctions (Rose et al., 2012; Shih et al., 2021).

The Gianyar District Health Service Facility in Bali, specifically, was the site of this study. The intended time frame for this study is from February 2022 to October 2022. And natural language analysis is used to generate keywords automatically, and it is integrated into the MySQL DBMS. The steps involved in the research process include the following:

1. Design of the WHO ICD-10 Database Schema
2. Mysql DBMS implementation of the WHO ICD-10 Database Schema
3. Import of the ICD-10 Database into the created database
4. Design of the RAKE Algorithm to build the index
5. Implementation of Keyword Extractor
6. Designing a diagnostic search engine algorithm
7. Developing a diagnostic search engine algorithm
8. Analyzing keyword and search engine accuracy

The development comes first in this research's sequence, followed by design and then implementation. Figure 1 is the focus of the research flow.

3. Result and Discussion

The *icd10_chapter* table and the *icd10_keywords* table are the two interconnected tables that make up the WHO ICH-10 database schema architecture. There are 12 columns in the *icd10_chapter* table, which will subsequently be used to contain WHO ICD-10 code property is divided into 4 attributes. The *icd10_keywords* table, which has five attribute columns and is related to the *icd10_chapter* table, will serve as the foundation for the diagnostic search engine because it is based on automatically produces keywords. The database structure that can be created based on the two tables above is as follows :

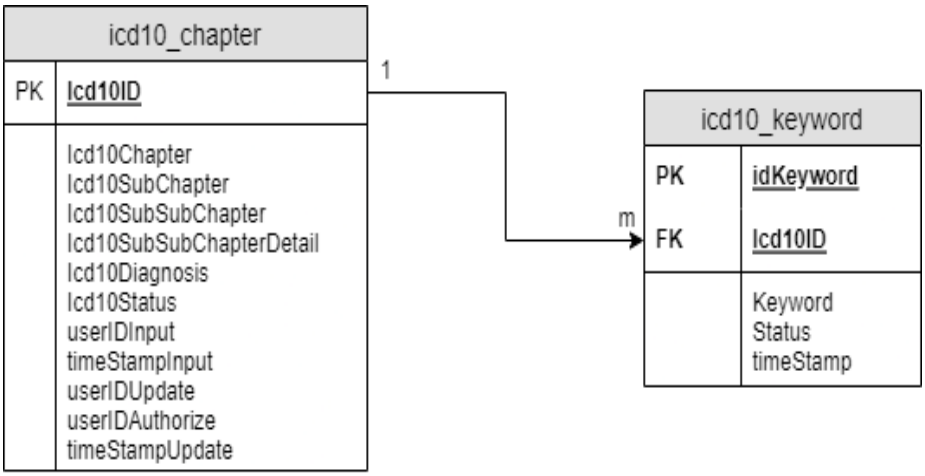

Figure 2. Shemiatic for a diagnostic search engine

Jurnal Matematika Vol. 13, No.2,Desember 2023, pp. 95-104 ISSN: 1693-1394
Article DOI: 10.24843/JMAT.2023.v13.i02.p163

The outcome of integrating the aforementioned database structure into the MySQL DBMS is as follows:

#	Name	Type	Collation	Attributes	Null	Default	Comments	Extra	Action
1	icd10ID	bigint(20)		UNSIGNED	No	None		AUTO_INCREMENT	Change Drop More
2	icd10Chapter	varchar(100)	latin1_swedish_ci		No	None			Change Drop More
3	icd10SubChapter	text	latin1_swedish_ci		No	None			Change Drop More
4	icd10SubSubChapter	text	latin1_swedish_ci		No	None			Change Drop More
5	icd10SubSubChapterDetail	varchar(100)	latin1_swedish_ci		No	-			Change Drop More
6	icd10Diagnosis	text	latin1_swedish_ci		No	None			Change Drop More
7	icd10Status	varchar(50)	latin1_swedish_ci		No	ok			Change Drop More
8	userIDInput	bigint(20)		UNSIGNED	No	None			Change Drop More
9	timeStampInput	timestamp			No	current_timestamp()			Change Drop More
10	userIDUpdate	bigint(20)		UNSIGNED	No	0			Change Drop More
11	userIDAuthorize	bigint(20)		UNSIGNED	No	0			Change Drop More
12	timeStampUpdate	timestamp			No	current_timestamp()		ON UPDATE CURRENT_TIMESTAMP()	Change Drop More

#	Name	Type	Collation	Attributes	Null	Default	Comments	Extra	Action
1	idKeyword	bigint(20)		UNSIGNED	No	None		AUTO_INCREMENT	Change Drop More
2	icd10ID	bigint(20)		UNSIGNED	No	None			Change Drop More
3	keyword	varchar(255)	utf8mb4_general_ci		No	None			Change Drop More
4	status	varchar(100)	utf8mb4_general_ci		No	None			Change Drop More
5	timeStamp	timestamp			No	current_timestamp()		ON UPDATE CURRENT_TIMESTAMP()	Change Drop More

Figure 3. Implimentation database into MySQL DBMS

3.1. Rapid Authomatic Keyword Extraction (RAKE)

Based on the algorithm that (Rose et al., 2012) had previously created, the RAKE algorithm used in this study was created. The candidate keywords were extracted using RAKE. RAKE is an automatic domain-independent method for extracting single document keyword (Anjali et al., 2019; Baruni & Sathiaseelan, 2020; Benita & Baizal, 2022). RAKE (Rapid Automatic Keyword Extraction) utilizes a stop-list to locate candidate keyword. Any sequence of words that appear between two stop-list words and/or punctuation marks are marked as candidate keywords. Then the frequency and the degree values of each word in the list of candidate keywords are calculated. The frequency of a word is the total number of its occurrences within the list of candidate keywords. The degree of a word is the total number of words that it appears with, within the list of candidate keywords. Then each word is computed by summing up the scores of the words that it contains. The top third scoring candidate keywords are extracted as keywords (Pay et al., 2019). The RAKE algorithm's flowchart is shown below :

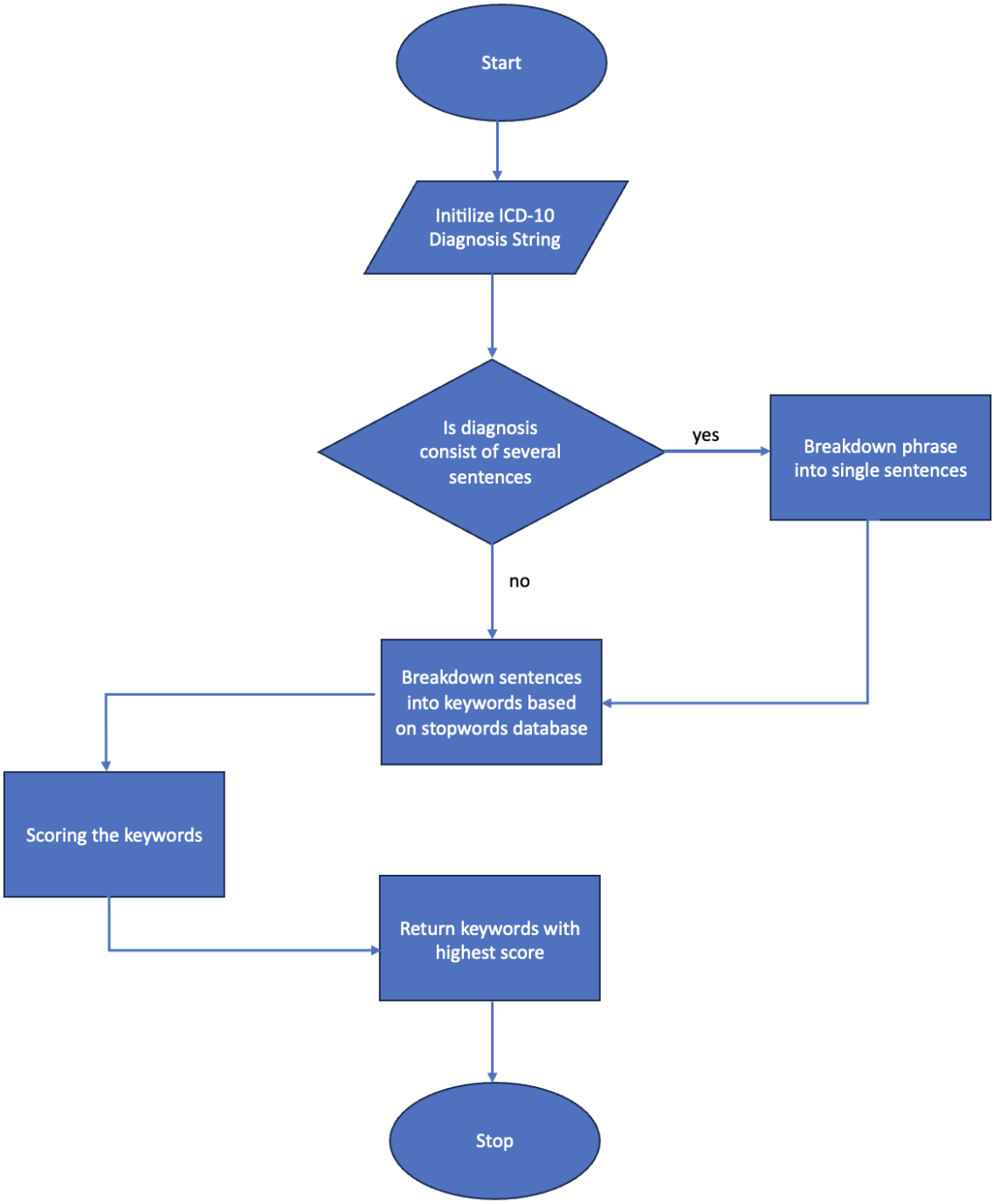

Figure 4. Rake algorithm flowchart

3.2. Implementation Rake into Diagnostic Search

The keyword list produced by the RAKE algorithm is shown as an example in the figure below. In order to validate the keyword results that were produced automatically, the author of this study approached an expert (in this case, a doctor) for assistance.

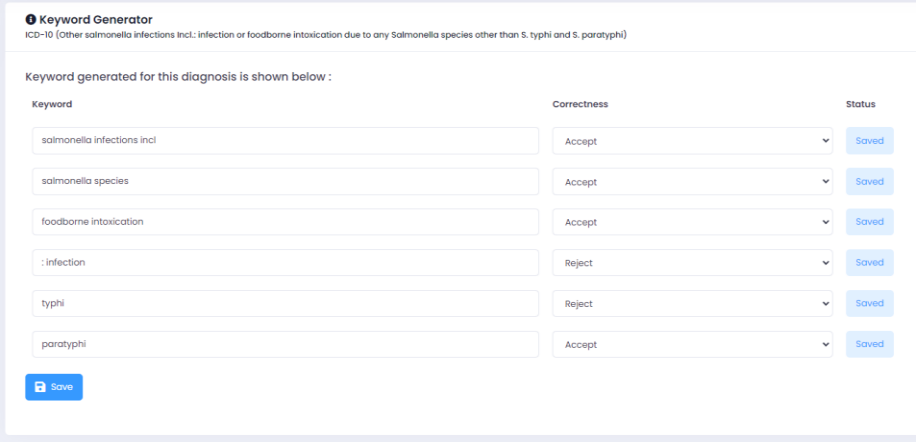

Figure 5. Rake result keyword

The accuracy of the implemented algorithm will be ascertained using the validation findings. Experts have the authority to accept or reject created keywords. so that you may later determine how accurate the keywords RAKE created are.

3.3. Analysis of Rake Algorithm Implementation

One hundred WHO ICD-10 diagnoses were examined. There were 100 samples, and 284 keywords were taken out. Following that, the expert chose 139 of these 284 keywords to be accepted and 145 tove rejected. Using the formula below, we can determine precision.

$$Precision = \frac{Totalaproved}{TotalExtracted} \quad (1)$$

The following outcomes are obtained by using the data we currently have in the formula above :

$$Precision = \frac{139}{284} = 0.489 \approx 0.491 \quad (2)$$

The precision results are below 50%, which is unquestionably unacceptable. The stopword list that was employed was not carefully curated to extract keywords from the health industry, which may be the source of the poor precision result.

3.4. Design and Development of Diagnostic Search Engines

The creation of new keywords and their storage in the database will serve as the foundation for the diagnostic search engine's architecture. The languages used for the search are HTML, PHP, and SQL. After testing, the term database utilized to create a diagnostic search engin can yield 100% precision and 100% recall. This is because each keywords has a distinct relationship with each WHO ICD-10 diagnostic.

4. Conclusion

A good keyword extraction algorithm is Rapid Automatic Keyword Extraction (RAKE). This is so that RAKE may extract keywords using stopwords in addition to delimiters. RAKE then offers a maximum score for the created keywords in addition to extracting them. The ranking of the keywords will be determined afterward using the score.

The choice of stopwords has a significant impact on the level of precision or precision of RAKE. Because words that cannot serve as keywords will also be favorable if the stopwords utilized have been carefully chosen. Conversely, poor keyword extract results stem from poorly curated stopwords, as seen in this research the precision of RAKE algorithm is 0,491 which is not very good.

Following the implementation of RAKE in the doctor's appointment ICD-10 diagnosis search program. It turns out that doctors are now quicker and more accurate in their search for a WHO ICD-10 diagnosis. This is undoubtedly highly beneficial for clinicians as it increases the precision with which patient diagnoses are chosen following the WHO ICD-10 code. Additionally, doctors will be more helpful in general when conducting patient examinations because of this.

To develop a data search software or even a search engine, RAKE might be selected as the keyword extractor algorithm. Even if the essential stopwords have been carefully selected by professionals, this must of course be supplemented by a decent selection of stopwords.

Acknowledgments

Thanks to **DIPA BLU Universitas Udayana Tahun Anggaran 2022 sesuai dengan Surat Perjanjian Penugasan Pelaksanaan Penelitian Unggulan Program Studi Nomor : B/78.157/UN14.4.A/PT.01.03/2022**

References

Anjali, S., Meera, N. M., & Thushara, M. G. (2019). A graph based approach for keyword extraction from documents. *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, 1–4.

Armentano, M. G., Godoy, D., Campo, M., & Amandi, A. (2014). NLP-based faceted search: Experience in the development of a science and technology search engine. *Expert Systems with Applications*, *41*(6), 2886–2896.

Baruni, J., & Sathiaseelan, J. (2020). Keyphrase Extraction from Document Using RAKE and TextRank Algorithms. *International Journal of Computer Science and Mobile Computing*, *9*(9), 83–93.

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *ArXiv Preprint ArXiv:1903.10676*.

Benita, I. R., & Baizal, Z. K. A. (2022). News Recommender System Based on User Log History Using Rapid Automatic Keyword Extraction. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *6*(4), 2341–2345.

Helling, L. S., Wahyudi, E., & Hasanudin, H. (2019). Siremis: Sistem Informasi Rekam Medis Puskesmas Kecamatan Matraman Jakarta. *INTENSIF: Jurnal Ilmiah Penelitian Dan Penerapan Teknologi Sistem Informasi*, *3*(2), 116–129.

Henderi, H., Al Khudhorie, F., Maulani, G., Millah, S., & Devana, V. T. (2022). A proposed model expert system for disease diagnosis in children to make decisions in first aid. *INTENSIF: Jurnal Ilmiah Penelitian Dan Penerapan Teknologi Sistem Informasi*, *6*(2), 139–149.

Jetté, N., Quan, H., Hemmelgarn, B., Drosler, S., Maass, C., Oec, D.-G., Moskal, L., Paoin, W., Sundararajan, V., & Gao, S. (2010). The development, evolution, and modifications of ICD-10: challenges to the international comparability of morbidity data. *Medical Care*, 1105–1110.

Kamal, W., Björnsdottir, S., Kämpe, O., & Trolle Lagerros, Y. (2020). Concordance between ICD-10 codes and clinical diagnosis of hypoparathyroidism in Sweden. *Clinical Epidemiology*, 327–331.

Khader, M., Awajan, A., & Al-Naymat, G. (2018). The effects of natural language processing on big data analysis: Sentiment analysis case study. *2018 International Arab Conference on Information Technology (ACIT)*, 1–7.

Kusuma, D. H., Shodiq, M. N., Yusuf, D., & Saadah, L. (2019). Si-Bidan: Sistem Informasi Kesehatan Ibu dan Anak. *INTENSIF: Jurnal Ilmiah Penelitian Dan Penerapan Teknologi Sistem Informasi*, *3*(1), 43–53.

Lynch, K. E., Viernes, B., Gatsby, E., DuVall, S. L., Jones, B. E., Box, T. L., Kreisler, C., & Jones, M. (2021). Positive predictive value of COVID-19 ICD-10 diagnosis codes across calendar time and clinical setting. *Clinical Epidemiology*, 1011–1018.

Noor, V. M. M., Ansyori, A., & Hariyanto, T. (2014). Peran Pengetahuan dan Sikap Dokter dalam Ketepatan Koding Diagnosis berdasar ICD 10. *Jurnal Kedokteran Brawijaya*, *28*(1), 65–67.

Nordgaard, J., Jessen, K., Sæbye, D., & Parnas, J. (2016). Variability in clinical diagnoses during the ICD-8 and ICD-10 era. *Social Psychiatry and Psychiatric Epidemiology*, *51*, 1293–1299.

Pay, T., Lucci, S., & Cox, J. L. (2019). An ensemble of automatic keyword extractors: TextRank, RAKE and TAKE. *Computación y Sistemas*, *23*(3), 703–710.

Purba, R. A., & Sondang, S. (2022). Design and Build Monitoring System for Pregnant Mothers and Newborns using the Waterfall Model. *INTENSIF: Jurnal Ilmiah Penelitian Dan Penerapan Teknologi Sistem Informasi*, *6*(1), 29–42.

Rose, S. J., Cowley, W. E., Crow, V. L., & Cramer, N. O. (2012). *Rapid automatic keyword extraction for information retrieval and analysis*. Google Patents.

Sadikin, B. G., & Laoly, H. Y. (2022). Peraturan Menteri Kesehatan Republik Indonesia Nomor 24 Tahun 2022 Tentang Rekam Medis. *Berita Negara Republik Indonesia Tahun 2022*, 1–20.

Satria, S. B. (2022). *MEMAHAMI Perbedaan Peraturan Menteri Kesehatan Nomor 24 Tahun 2022 Tentang REKAM MEDIS dengan Permenkes No 269 Tahun 2008 Tentang REKAM MEDIS*.

Shenoi, S. J., Ly, V., Soni, S., & Roberts, K. (2020). Developing a search engine for precision medicine. *AMIA Summits on Translational Science Proceedings*, *2020*, 579.

Shih, C.-H., Lin, C.-J., & Jeng, S.-Y. (2021). Improved rapid automatic keyword extraction for voice-based mechanical arm control. *Sensors and Materials*, *33*(8), 2897–2909.

Susanna, D. (2020). When will the COVID-19 pandemic in indonesia end? *Kesmas: Jurnal Kesehatan Masyarakat Nasional (National Public Health Journal)*, *15*(4).

Wijayakusuma, I. G. N. L., & Yowani, S. C. (2022). WHO ICD-10 BASED ONLINE DISEASE DIAGNOSIS FOR GENERATING DIGITAL MEDICAL RECORD APPLICATION DEVELOPMENT. *SINTECH (Science and Information Technology) Journal*, *5*(1), 24–30.

World Health Organitation. (2015). History of the Development of the ICD. *World Health Organitation*.

World Health Organitation. (2020). *Considerations for school-related public health measures in the context of COVID-19*.