Medical Named Entity Recognition for Enhanced Electronic Health Record Maintenance

Muralikrishna S. N[1], Raghavendra Ganiga[2]*, Raghurama Holla[3], Ruppikha Sree Shankar[4]
Department of Computer Science and Engineering, Manipal Institute of Technology,
Manipal Academy of Higher Education, Manipal, Karnataka, India-576104[1, 4]
Department of Information and Communication Technology, Manipal Institute of Technology,
Manipal Academy of Higher Education, Manipal, Karnataka, India-576104[2]
Department of Data Science and Computer Applications, Manipal Institute of Technology,
Manipal Academy of Higher Education, Manipal, Karnataka, India-576104[3]
Centre of Indian Language Data Lab, Manipal Institute of Technology,
Manipal Academy of Higher Education, Manipal, Karnataka, India-576104[1, 2, 4]

***Abstract*—The increasing use of electronic health records (EHRs) has led to a surge in unstructured data, making it challenging to extract valuable insights. This study proposes Natural Language Processing (NLP) based techniques to standardize Electronic Health Record (EHR) data. Conducted in a healthcare setting, the research focuses on transforming unstructured EHR text into structured data using Part-of-Speech tagging and Named Entity Recognition (NER). NER techniques are applied to extract and categorize medical terms, enhancing data accuracy and consistency. The framework's performance is evaluated using precision and recall rates. Experimental results demonstrate that NER effectively identifies and organizes medical entities, facilitating improved data analysis and decision-making in healthcare. This approach promises to enhance interoperability and the overall utility of EHR systems.**

***Keywords*—Electronic health records; named entity recognition; natural language processing; part-of-speech**

I. Introduction

In recent years, EHR systems have been widely adopted in hospitals to effectively manage patient information, including diagnoses, lab results, medications etc. in a digital format. However, this increased adoption has also led to the generation of unstructured data in the form of raw clinical notes [1], [2]. Therefore, standardizing this data is essential for decision making and interoperability across different healthcare systems. One way to tackle this challenge is by using NER to standardize and structure EHR data. NER is a technique in NLP used for the identification and extraction of specific entities from unstructured text. In EHR systems, structured patient information, such as diagnoses and medications, can be extracted using NER. This helps create a standardized and well-organized database that multiple healthcare organizations can access for efficient data comparison and analysis [3], [4].

NER has several applications in clinical decision support systems (CDSS) and population health management. In CDSS, NER is used to standardize EHR data, enabling healthcare providers to identify critical patients and recommend appropriate treatments. Additionally, NER helps healthcare providers analyze large patient populations and supports more effective public health interventions [5] [6] [7]. As NER can be used to convert unstructured clinical data into structured data, it has a significant role in medical research for achieving better outcomes [8]. The main advantage of NER in EHR standardization is its ability to reduce errors and inconsistencies in clinical data.

EHRs must be maintained by various healthcare providers, which can lead to discrepancies due to variations in terminology and documentation practices. NER addresses this issue by eliminating inconsistencies, ensuring improved accuracy and uniformity across different healthcare systems [9], [10]. In addition, it helps reduce manual effort by automating the process, allowing healthcare providers to focus primarily on patient care rather than relevant data searching [11], [12], [13], [14].

Implementing Named Entity Recognition (NER) involves several challenges, including the extensive training required to achieve high accuracy in extracting clinical entities. Additionally, integrating NER into existing healthcare systems demands careful consideration of patient privacy and data security. Further complications arise from significant variations in clinical notes across healthcare providers, making it difficult to develop universal extraction approaches. The contextual ambiguity of medical terms adds another layer of complexity, while inconsistencies in local coding practices hinder interoperability. To address these challenges, we propose a standardization technique utilizing an NLP pipeline. Our approach incorporates metadata generation, XML representation, Part-of-Speech tagging, chunking, and Named Entity Recognition to achieve standardized representation [15].

In conclusion, applying Named Entity Recognition to generate standardized EHR data is an effective approach for enhancing the accuracy, consistency, and usability of healthcare data. As NER transforms unstructured text into structured information, it enhances decision-making, interoperability, and overall healthcare outcomes.

In Section II, we provide literature related to EHR standardization, followed by the methodology in Section III. In Section IV, we present the experimental setup and results, followed by conclusions in Section V.

*Corresponding Author.

II. BACKGROUND

India's healthcare system operates at three levels: primary, secondary, and tertiary care. Each level generates vast amounts of data daily, including structured, unstructured, and semi-structured formats. This includes clinical notes, patient records, and medical narratives. The challenge for healthcare professionals is making sense of this data. By extracting and converting unstructured data into a structured format, we can enable better analysis and improve decision-making for patient care [16].

The implementation of EHR in India is still evolving. While large corporate hospitals have adopted EHR systems to some extent, small and medium-sized hospitals continue to rely on a hybrid record-keeping approach. Improving data accuracy and reducing errors are critical challenges, and NER plays a significant role in advancing these efforts [17].

As highlighted by Durango et al. [18] NER could standardize free-text notes across various healthcare providers, minimizing variations and errors, and contributing to more reliable EHRs. Additionally, NER automates the extraction of relevant information from clinical notes, saving time and reducing the manual effort required by healthcare providers. Pinheiro et al. [19] highlighted that automation improves efficiency, enabling healthcare providers to focus more on clinical decision-making rather than data processing.

NER plays a key role in improving the accuracy and efficiency of EHR systems. However, its implementation faces significant challenges, as it requires a large dataset and specialized training tailored to medical terminology. General language datasets often fail to capture the nuances of medical records, making domain-specific data essential for effective NER in healthcare. Mishra et al. [20] emphasized that without these domain-specific datasets, the performance of NER systems can be compromised. Another challenge is integrating NER into existing EHR systems, which often have diverse data structures. Variations in formats can complicate data mapping and interoperability, hindering seamless integration [21], [22], [23], [24].

In summary, while challenges remain in the use of NER for standardizing EHR data, its benefits—such as improved accuracy, time efficiency, and support for decision-making—highlight its potential to transform healthcare systems and contribute to better patient care and research outcomes.

III. METHODOLOGY AND RESEARCH DESIGN

Generating standardized EHR from semi-structured or unstructured data is vital in the healthcare industry as a globally acceptable standardization protocol. Most health records are written by hand or are in a semi-structured digital format. Using deep learning techniques to solve computer vision problems has made it possible for handwritten documents to be automatically turned into digital files. However, in the second phase, where the semi-structured data needs to be brought to a standardized format for effective and seamless exchange of information in an application-independent environment, we address this major issue using a novel methodology. The proposed method uses a natural language processing backbone in the framework. We achieve the following objectives with the proposed framework as shown in Fig. 1:

- Read semi-structured data from .xls file and text files.
- Convert the semi-structured data to a well-defined XML format.
- The well-defined XML format automatically generates the meta-data including disease classes, drug information, with relevant ICD10 codes as a standardization method.

A. Read Semi-Structured Data from .xls File

In this step, data is read from a semi-structured data source, which is an Excel (.xls) file in this case. Semi-structured data refers to data that does not have a formal structure but has some organization. For example, the data in an Excel file may have a header row and be organized in columns, but there may be cells that contain multiple pieces of information. To read this data, a program could use a library or tool that is capable of reading and parsing Excel files, such as Pandas or OpenPyXL.

B. Convert the Semi-Structured Data to a Well-Defined XML Format

In this step, the semi-structured data is transformed into a well-defined XML format. This involves defining a schema or template for the XML document that specifies the structure of the data and how it should be organized. The program could use a library or tool to perform this transformation, such as lxml or ElementTree. The resulting XML file should be structured in a way that makes it easy to process and extract information from.

C. Generate Meta-Data Including Disease Classes and Drug Information with Relevant ICD10 Codes

In this step, the well-defined XML format is used to automatically generate meta-data, including disease classes, drug information, and relevant ICD10 codes. This process involves extracting relevant information from the XML document and using it to populate metadata fields. The metadata could be generated using tools such as NER using fine-tuned clinical BERT model. Once the metadata is generated, it can be used as a standardization method for the data, making it easier to analyze and compare with other datasets.

Overall, these steps involve reading semi-structured data from an Excel file, transforming it into a well-defined XML format, and generating metadata from the XML document using NLP and NER algorithms. The resulting XML file and metadata can then be used for analysis and standardization of the data.

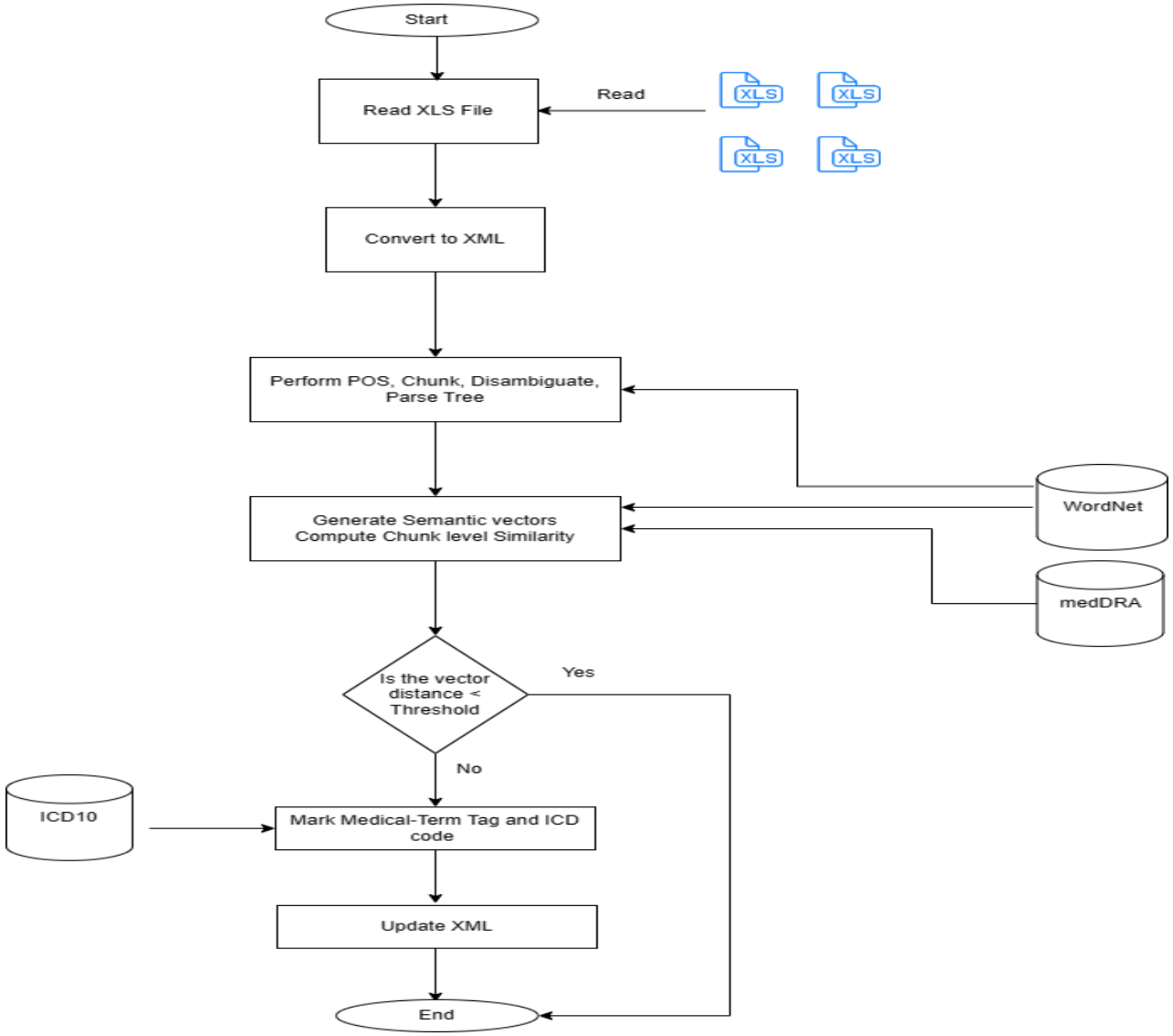

Fig. 1. The proposed framework for standardizing EHR.

D. System Architecture

The diagram shows five main components in architecture:

Cloud Infrastructure: This component provides the underlying infrastructure for the cloud-based healthcare system, including the computing resources, storage, and networking infrastructure required to support the system.

1) Healthcare data storage: This component provides a secure, scalable, and accessible storage solution for healthcare data. This may include electronic health record (EHR) data, medical images, and other types of healthcare data.

2) Data analytics and decision support: This component provides tools for data analytics and decision support, including machine learning algorithms and other data analysis techniques. This component can help healthcare providers to identify patterns and trends in patient data, make more informed decisions, and provide more personalized care.

3) Mobile and web applications: This component provides a user-friendly interface for healthcare providers and patients to access the system. This may include mobile and web-based applications that allow patients to view their medical records, communicate with healthcare providers, and manage their healthcare needs.

4) Security and compliance: This component provides a security and compliance framework for the cloud-based healthcare system. This may include access control, data encryption, and other security measures to protect patient data and comply with relevant regulations.

Overall, this architecture provides a flexible and scalable solution for managing healthcare data in the cloud. It can help healthcare providers to improve the quality of care, reduce costs, and provide better patient experience. Fig. 2 illustrates the process flow for standardizing EHR.

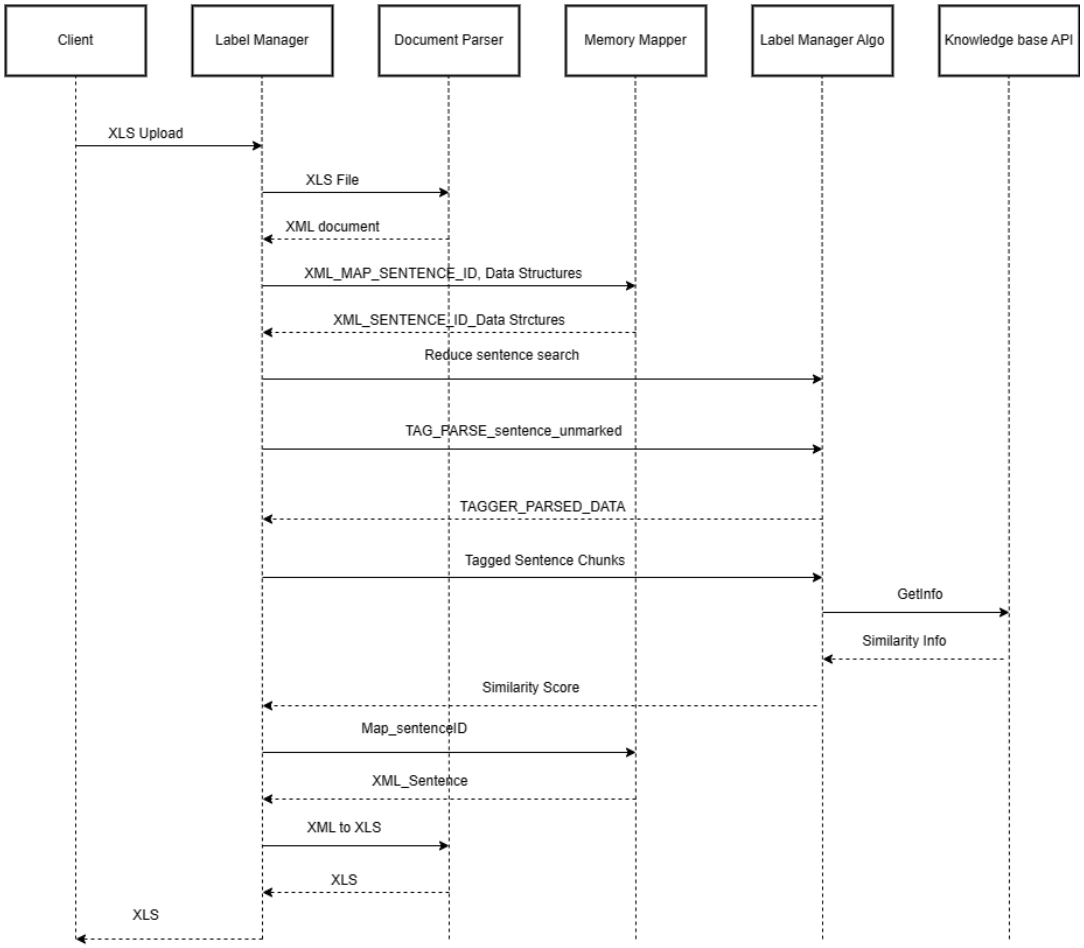

Fig. 2. Sequence diagram illustring the process flow for standardizing HER.

E. Process Steps

The diagram shows four main steps in the process:

1) Data acquisition: In this step, EHR data is acquired from various sources, including electronic health record systems, clinical notes, laboratory results, and radiology reports. The data is typically in a semi-structured or unstructured format, making it difficult to extract and standardize.

2) Data preprocessing: In this step, the EHR data is preprocessed to extract relevant information and prepare it for standardization. This may involve cleaning the data, removing duplicates, and structuring the data into a suitable format for further processing.

3) Standardization using NER: This step uses named entity recognition (NER) to identify and extract specific entities from the EHR data, such as disease classes, drug information, and relevant ICD10 codes. NER is a technique in natural language processing (NLP) that can automatically identify and extract named entities from text data.

4) Output: In this final step, the standardized data is output in a well-defined XML format, which includes the meta-data generated from the NER process. The output data can be used for various applications, including clinical decision support, patient risk analysis, and epidemiological studies.

Overall, this process provides a standardized method for extracting and organizing EHR data using NER techniques. This can improve the accuracy and efficiency of healthcare data analysis, making it easier to identify patterns and trends in patient data.

IV. Experimental Setup and Results

The primary objective of this study was to transform unstructured electronic health record text into a structured format using natural language processing techniques, including NER, Part-of-Speech (POS) tagging, and medical coding integration (MedDRA & ICD-10). To achieve this, we synthesized 50 patient records containing detailed medical histories, symptoms, diagnoses, medications, and follow-up instructions.

The unstructured text was preprocessed and converted into an Excel format, where each sentence was assigned a unique Sentence-ID for easy tracking. Through NER and chunking, key medical entities—such as symptoms, diseases, medications, and healthcare providers—were extracted and categorized as shown in Table I, Table II and Table III. Additionally, POS tagging helped identify grammatical structures within the clinical text, improving the accuracy of entity recognition as shown in Table

IV. We used pretrained clinical BERT model for the identification of named entities.

To ensure medical standardization, extracted terms were mapped to ICD-10 and MedDRA codes, allowing for systematic classification of symptoms and diagnoses as reported in Table V. The results demonstrate that NLP-based automation significantly improves the efficiency and accuracy of data extraction from unstructured patient records. The following sections provide a detailed breakdown of the key findings from our analysis.

A. Dataset Generation

We synthesized 50 patient records using OpenAI's Language Model (LLM) in raw text format using carefully designed prompt engineering. The generated records contained unstructured text, including patient demographics, medical history, symptoms, medications, and follow-up details. The dataset was augmented with clinical terms to generate random and less frequent words.

Example input text:

"John Doe, a male patient born on March 15, 1985, presents with a persistent cough, shortness of breath, wheezing, and chest tightness, indicating an asthma exacerbation likely triggered by seasonal allergies and recent cold weather exposure. He is prescribed Albuterol Inhaler, Fluticasone Propionate, Montelukast, and Loratadine."

B. Data Preprocessing and Standardization

To facilitate structured analysis, we converted raw text into an Excel (xls) format, assigning each sentence a unique Sentence-ID. Data standardization was achieved by categorizing key medical information into structured fields:

TABLE I. PATIENT CATEGORIES AND EXTRACTED INFORMATION

Category	Extracted Information
Patient Demographics	Name, age, gender, date of birth
Symptoms	Persistent cough, shortness of breath, wheezing

TABLE II. PATIENT TREATMENT SUMMARY

Diagnoses	Asthma exacerbation, seasonal allergies
Medications	Albuterol, Fluticasone, Montelukast, Loratadine
Prescribing Physician	~~Dr. xxxxxxxxxx, Pulmonologist~~
Follow-up Instructions	Follow-up in 4 weeks at Springfield Medical Center

C. Named Entity Recognition (NER) and Chunking

To extract meaningful medical entities, we applied NER and chunking. This process identified symptoms, diseases, and prescribed medications.

D. Part-of-Speech (POS) Tagging and Analysis

POS tagging based on Stanford CoreNLP was applied to medical terms to enhance entity recognition.

Example POS tagging output:

"John Doe, a male patient born on March 15, 1985, presents with a persistent cough, shortness of breath, wheezing, and chest tightness, indicating an asthma exacerbation likely triggered by seasonal allergies."

TABLE III. HEALTHCARE DATA CATEGORIES

Category	Example
Symptoms	Shortness of breath, wheezing
Diseases	Asthma exacerbation
Medications	Albuterol, Fluticasone
Doctors & Facilities	~~Dr. xxxxxxxx, Springfield Medical Center~~

TABLE IV. MAPPING OF WORDS TO POS TAGS

Word	POS Tag
John	NNP (Proper Noun)
patient	NN (Noun)
presents	VBZ (Verb)
persistent	JJ (Adjective)
cough	NN (Noun)
asthma	NN (Noun)

E. Medical Coding: MedDRA and ICD-10 Mapping

To enhance standardization, we mapped extracted terms to ICD-10 and MedDRA codes using rule-based method with a lookup table.

TABLE V. MEDICAL TREMS WITH ICD-10 CODES AND MEDDRA CATEGORIES

Medical Term	ICD-10 Code	MedDRA Category
Persistent cough	R05	Respiratory Symptoms
Asthma exacerbation	J45.901	Respiratory Diseases
Shortness of breath	R06.02	Breathing Abnormalities

The implementation of natural language processing (NLP) techniques significantly improved the efficiency of extracting key medical information from unstructured patient records. By automating the identification of medications, symptoms, and diagnoses, manual effort was substantially reduced, allowing for faster and more accurate data processing. One of the primary advantages of this approach was error reduction, as automated entity recognition minimized inconsistencies commonly found in manual data entry. Additionally, time efficiency was greatly enhanced, enabling rapid extraction of critical medical details and facilitating structured data collection. To ensure standardization, the extracted terms were mapped to MedDRA and ICD-10 classifications, improving interoperability across different healthcare systems.

In total, 50 synthetic patient records were successfully processed using NLP-based techniques. The NER and Part-of-Speech (POS) tagging played a crucial role in accurately identifying and extracting medical terms. Furthermore, the integration of ICD-10 and MedDRA mapping ensured that symptoms, diagnoses, and treatments were classified according to standardized medical codes. The result of each task is shown

in Table VI. The transformation of unstructured text into a structured data format improved both readability and consistency, making the data more suitable for further analysis and clinical decision support.

TABLE VI. PERFORMANCE MEASURE POS TAGGING, NER AND MEDDRA/ICD10 RULE BASED INTEGRATION

Metric\Task	POS	NER	MedDRA/ICD10 Rule based Integration
Precision	0.82	0.77	0.68
Recall	0.80	0.74	0.65
F1-accuracy	0.81	0.75	0.67

V. CONCLUSION

NER is a powerful tool for standardizing EHR data by extracting structured information from unstructured text. This standardization enables healthcare providers to efficiently compare and analyze patient data across different systems, improving interoperability and data consistency. Despite certain challenges, the benefits of NER—such as enhanced accuracy, automation, and efficiency—make it an asset in healthcare data management. The test was conducted on synthetic data due to the lack of publicly available real-world EHR datasets. This limitation could impact the generalizability of the results. To enhance reliability, future research should validate the proposed method using real-world clinical data. Additionally, exploring the effects of different POS tagging and entity recognition approaches would help optimize accuracy and robustness. As healthcare technology advances, the role of NER in optimizing EHR utilization is expected to grow, further enhancing clinical decision-making and research capabilities.

ACKNOWLEDGMENT

Authors would like to acknowledge the support of Manipal Academy of Higher Education.

REFERENCES

[1] Meystre, S. M., Savova, G. K., & Kipper-Schuler, K. C. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. Yearbook of medical informatics, 17(01), 128-144.

[2] Mahbub M, Srinivasan S, Danciu I, Peluso A, Begoli E, Tamang S, Peterson GD. Unstructured clinical notes within the 24 hours since admission predict short, mid & long-term mortality in adult ICU patients. PLoS One. 2022 Jan 6;17(1):e0262182.

[3] Liu, S., Yang, L., Zhang, C., Luan, H., Chute, C. G., & Zhu, Q. (2017). Extraction of medication information from clinical text via a jointly trained deep neural network. Journal of biomedical informatics, 76, 41-48.

[4] Murphy, S. N., Weber, G., Mendis, M., Gainer, V., Chueh, H. C., Churchill, S., & Kohane, I. (2010). Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). Journal of the American Medical Informatics Association, 17(2), 124-130.

[5] Jehangir, Basra, Saravanan Radhakrishnan, and Rahul Agarwal. "A survey on named entity recognition—datasets, tools, and methodologies." Natural Language Processing Journal, 3 (2023).

[6] Navarro, D. F., Ijaz, K., Rezazadegan, D., Rahimi-Ardabili, H., Dras, M., Coiera, E., & Berkovsky, S. (2023). Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. *International Journal of Medical Informatics*, *177*, 105122.

[7] Narzary, S., Brahma, A., Nandi, S., & Som, B. (2024). Deep Learning based Named Entity Recognition for the Bodo Language. *Procedia Computer Science*, *235*, 2405-2421.

[8] Sherman, R. E., Anderson, S. A., Dal Pan, G. J., Gray, G. W., Gross, T., Hunter, N. L.,& Califf, R. M. (2016). Real-world evidence—what is it and what can it tell us. *N Engl J Med*, *375*(23), 2293-2297.

[9] Kong, H. J. (2019). Managing unstructured big data in healthcare system. *Healthcare informatics research*, *25*(1), 1-2.

[10] Shao, M., Fan, J., Huang, Z., & Chen, M. (2022). The Impact of Information and Communication Technologies (ICTs) on Health Outcomes: A Mediating Effect Analysis Based on Cross-National Panel Data. *Journal of environmental and public health*, *2022*(1), 2225723.

[11] Murdoch, T. B. & Detsky, A. S. The inevitable application of big data to health care. J. Am. Med. Assoc. 309, 1351–1352 (2013).

[12] Zhang, D., Yin, C., Zeng, J., Yuan, X., & Zhang, P. (2020). Combining structured and unstructured data for predictive models: a deep learning approach. *BMC medical informatics and decision making*, *20*, 1-11.

[13] Vest, J. R., Grannis, S. J., Haut, D. P., Halverson, P. K. & Menachemi, N. Using structured and unstructured data to identify patients' need for services that address the social determinants of health. Int. J. Med. Inform. 107, 101–106 (2017).

[14] Kharrazi, H., Anzaldi, L. J., Hernandez, L., Davison, A., Boyd, C. M., Leff, B., & Weiner, J. P. (2018). The value of unstructured electronic health record data in geriatric syndrome case identification. *Journal of the American Geriatrics Society*, *66*(8), 1499-1507.

[15] Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F.& Botsis, T. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *Journal of biomedical informatics*, *73*, 14-29.

[16] Koleck, T. A., Dreisbach, C., Bourne, P. E., & Bakken, S. (2019). Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *Journal of the American Medical Informatics Association*, *26*(4), 364-379.

[17] Sheikhalishahi, S., Miotto, R., Dudley, J. T., Lavelli, A., Rinaldi, F., & Osmani, V. (2019). Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR medical informatics*, *7*(2), e12239.

[18] Durango MC, Torres-Silva EA, Orozco-Duque A. Named Entity Recognition in Electronic Health Records: A Methodological Review. Healthcare Informatics Research. 2023;29:286–300.

[19] Da Silva, D. P., da Rosa Fröhlich, W., de Mello, B. H., Vieira, R., & Rigo, S. J. (2023). Exploring named entity recognition and relation extraction for ontology and medical records integration. *Informatics in medicine unlocked*, *43*, 101381.

[20] Ahmad, Pir Noman, Adnan Muhammad Shah, and KangYoon Lee. "A review on electronic health record text-mining for biomedical name entity recognition in healthcare domain." Healthcare. Vol. 11. No. 9. MDPI, 2023.

[21] Reisman, M. (2017). EHRs: the challenge of making electronic data usable and interoperable. *Pharmacy and Therapeutics*, *42*(9), 572.

[22] Raza, S., Reji, D. J., Shajan, F., & Bashir, S. R. (2022). Large-scale application of named entity recognition to biomedicine and epidemiology. *PLOS Digital Health*, *1*(12), e0000152.

[23] Navarro, D. F., Ijaz, K., Rezazadegan, D., Rahimi-Ardabili, H., Dras, M., Coiera, E., & Berkovsky, S. (2023). Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. *International Journal of Medical Informatics*, *177*, 105122.

[24] Ahmad, P. N., Shah, A. M., & Lee, K. (2023, April). A review on electronic health record text-mining for biomedical name entity recognition in healthcare domain. In *Healthcare* (Vol. 11, No. 9, p. 1268). MDPI.