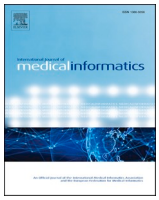
International Journal of Medical Informatics 204 (2025) 106049



Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Review article



Natural language processing in medical text processing: A scoping literature review

Luis B. Elvas [a,b,c,d,*], Ana Almeida [d], João C. Ferreira [a,b,d]

[a] Department of Logistics, Molde University College, Molde 6410, Norway
[b] Inov Inesc Inovação – Instituto de Novas Tecnologias, 1000-029 Lisbon, Portugal
[c] Breast Cancer Research Program, Champalimaud Foundation, Lisbon, Portugal
[d] ISTAR, Instituto Universitário de Lisboa (ISCTE-IUL), 1649-026 Lisbon, Portugal

ARTICLE INFO

Keywords:
NLP
NER
BERT
Medical records
Healthcare

ABSTRACT

Background: The exponential growth of digitized medical data has created significant challenges for healthcare professionals, as medical documentation transitions from simple text records to complex, multi-dimensional data structures. Natural Language Processing (NLP), particularly Named Entity Recognition (NER), has emerged as a crucial tool for extracting and categorizing critical information from clinical texts. The development of transformer-based models like BERT and the ability to fine-tune pre-trained AI models have revolutionized the field, offering unprecedented opportunities to enhance the efficient and precise interpretation of medical data across diverse languages and healthcare contexts.

Objective: This literature review aimed to analyze recent NLP approaches for medical text processing, examining techniques, performance metrics, and advancements across different languages and healthcare contexts.

Method: Following the Preferred Reporting Items for Systematic Reviews and Meta Analyses (PRISMA) methodology, a scoping search was conducted in Scopus and PubMed databases, focusing on studies published between 2019–2024. The review included studies on language model fine-tuning and information extraction in healthcare, with a specific search query designed to capture relevant NLP techniques.

Results: Of 67 initial records, 31 studies were ultimately included. Bidirectional Encoder Representations from Transformers (BERT)-based approaches, neural networks, and CRF/LSTM techniques dominated, consistently achieving F1-scores above 85 %. The studies covered multiple languages, with 51.5 % in English, 27.3 % in Chinese, and smaller representations in Italian, German, and Spanish. Hybrid approaches and techniques addressing data privacy and limited labeled data were notably prevalent.

Conclusions: The review revealed that modern NLP techniques, particularly BERT-based models and hybrid approaches, show significant promise in medical text processing across different languages. While challenges remain in cross-lingual adaptation and data availability, these technologies demonstrate potential to enhance medical data interpretation and analysis.

1. Introduction

The volume of digitized medical data has been steadily increasing, offering significant opportunities for developing effective, modern solutions. As medical documentation evolves from simple text records to complex, multi-dimensional data structures, the workload of healthcare professionals becomes increasingly challenging, making the efficient and precise interpretation of this data a daunting task [1].

In this context, NLP has emerged as a crucial tool for extracting pertinent information from voluminous medical data, with NER techniques playing a pivotal role in automatically identifying and categorizing relevant information in clinical texts, including diagnoses, medications, and symptoms, thereby facilitating a more detailed and structured analysis of medical reports [2].

Fine-tuning enables the customization of pre-trained AI models for specific applications, thereby enhancing their accuracy and adaptability to the nuances of medical data [2]. This is particularly relevant given the emergence of transformer-based models like BERT [3], which have revolutionized the field of NLP since their introduction in late 2018.

Before the widespread adoption of transformer-based architectures,

* Corresponding author at: Department of Logistics, Molde University College, Molde 6410, Norway.
E-mail address: luis.m.elvas@himolde.no (L.B. Elvas).

https://doi.org/10.1016/j.ijmedinf.2025.106049
Received 29 March 2025; Received in revised form 4 July 2025; Accepted 15 July 2025
Available online 17 July 2025
1386-5056/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

clinical NLP research primarily relied on rule-based systems [4], statistical methods such as Conditional Random Fields (CRFs) [5] and Support Vector Machines (SVMs) [6], and early neural models including RNNs and LSTMs [7]. These approaches laid important groundwork for tasks such as named entity recognition and de-identification but often required extensive feature engineering and lacked contextual flexibility. The advent of pretrained transformer models significantly enhanced performance and generalizability, especially in multilingual and low-resource settings.

The interest and need to adopt these solutions are driving research into the steps needed to effectively integrate these innovations into clinical practice, with a view to a more efficient and accurate approach to medical data analysis. Understanding how these technologies perform across different languages and healthcare contexts is crucial for their successful implementation.

While recent advancements in transformer-based models have significantly improved NLP performance in general domains [8], their application to medical texts, particularly across different languages and healthcare systems, remains fragmented and underexplored [9]. To date, there is limited consolidated evidence assessing how fine-tuned NLP models perform in diverse clinical settings and languages. This review aims to address this gap by examining studies that apply NER, to clinical texts, with attention to cross-linguistic performance, domain adaptation, and practical implementation in clinical workflows. The scope of this review encompasses both monolingual and multilingual approaches, with a focus on methods leveraging modern transformer architectures.

Several literature reviews have addressed the use NLP approaches in healthcare settings [10], focusing on clinical applications or focusing in English corpora [11]. However, few have systematically reviewed specifically transformer-based NER applications in multilingual clinical settings.

Given the increasing requirement for sophisticated analysis methods in medical reporting, this chapter aims to provide a review of the current state of the art in NLP approaches for medical text processing. The review is divided into three sections. Firstly, it outlines the research methodology, explaining the scoping review approach based on the PRISMA methodology. The following sub-section presents the results of the research, elaborating on the findings across different languages and technical approaches. The final sub-section discusses the evidence from the articles under study, summarizing the principal findings, identifying future research directions and conclusions.

2. Review methodology

In this chapter, we specify the Methodology followed for the search of articles to be reviewed, how were they selected and reviewed. The literature related to the proposed topic was screened and critically evaluated based on the PRISMA methodology [12]. PRISMA, initially created for the medical research field, offers evidence-based protocols to guide authors in transparently documenting their review methodology, procedures, and results.

This approach facilitates a thorough examination of existing research, offering a comprehensive survey of the current landscape regarding the application of NLP techniques, to automate the analysis of medical reports.

While we did not perform a formal quality appraisal of the included studies, we relied on the assumption that publications indexed in Scopus and PubMed have undergone peer review and meet a baseline standard of methodological rigor. This approach is consistent with prior reviews in the field, particularly given the technical nature of NLP studies and the focus on recent work from reputable sources.

2.1. Search strategy

The search was conducted on October 2024, using the Scopus [13] and PubMed [14] databases, two comprehensive and authoritative databases in scientific and medical literature, ensure a wide-ranging and reliable search for relevant studies in the field. The preliminary research aimed to identify relevant articles published within the last six years (2019–2024) that were written in English. Limiting the search to the last five 5 years ensures that the review includes the most recent and relevant studies, reflecting the latest advancements, technologies, and methodologies in the field. Since BERT [3], one of the most popular Transformer architectures, and language models (LMs) were launched at the end of 2018, we searched for studies from 2019 onwards.

2.2. Inclusion criteria

The criteria were defined to include studies focused on continuing the training of LMs to achieve healthcare-domain LMs or studies focused on creating Information Extraction (IE) models within the healthcare field. We formulated a query that encompasses the training or fine-tuning of language models or information extraction models within the context of healthcare, using medical reports or records as data. The query specifically includes terms related to "Language Model," "NER," and "Information Extraction" to capture relevant AI technologies. We also included terms like "Medical Report" and "Medical Record" to focus on studies using healthcare data.

To ensure relevance to our study's aim, we included terms related to model adaptation such as "Fine-tuning" and "Training." The healthcare context was maintained by including terms like "Healthcare," "Health Care," "Clinical," and "Medical.".

We decided to exclude large language models (LLMs) from our search. While there is significant semantic similarity between LMs and LLMs, the latter typically have billions of parameters and are primarily suited for text generation, which differs from the purpose of our study. LMs, in contrast, often require fine-tuning with labeled data for specific downstream NLP tasks, aligning more closely with our research focus.

The refined search query was as follows:

("Language Model" OR "NER" OR "Information Extraction") AND ("Medical Report" OR "Medical Record") AND ("Fine-tuning" OR "Training") AND ("Healthcare" OR "Health Care" OR "Clinical" OR "Medical") AND NOT ("Large Language Models" OR "LLM").

Although this review focuses specifically on NER as a foundational task in clinical NLP, related information extraction tasks such as relation extraction, entity linking, and question answering are indeed integral to downstream applications, such as including clinical decision support and EHR summarization, they typically build upon accurate entity recognition. The decision to range the review around NER was intentional, aimed at exploring model performance and challenges specific to entity detection. Nevertheless, we acknowledge that many recent works, including those leveraging BERT and large language models, increasingly integrate NER with these complementary tasks to enable end-to-end clinical text understanding.

The exclusion of LLMs was due to their generative nature and significantly higher computational requirements, but it's still important to recognize their emerging role in information extraction tasks. Recent models such as BioGPT [15] and GPT-based NER [16] systems have demonstrated promising results in extracting structured data from clinical narratives. However, the implementation and fine-tuning of such models demand substantial computational resources and infrastructure, which are often beyond the capabilities of typical deployment environments in many healthcare settings. For this reason, we focus on lighter, task-specific language models that are more practical for real-world integration.

The screening and evaluation of the studies were conducted independently by two researchers. In the event of disagreement, a third researcher was consulted to make the final decision. Although inter-rater reliability was not formally calculated using statistical measures, this consensus-based approach was applied to minimize bias and ensure consistent selection criteria.

3. Results

This chapter presents the outcomes of our scoping literature review, detailing both the selection process and the subsequent analysis of the selected studies. Through careful application of our search strategy and inclusion criteria, we identified and analyzed relevant research in the field of medical text processing using NLP techniques. Our review encompasses studies across multiple languages and various technical approaches, providing an overview of current methodologies and their effectiveness in processing medical texts.

Structuring the study selection process was crucial for efficiently managing and evaluating the relevant information. The creation of a flowchart maps out the steps that were followed from the initial identification of potentially relevant articles to the final selection of studies included in our research.

Fig. 1 presents the PRISMA Flow Diagram, which divides the selection process into four main stages. The progression through these stages involves systematically narrowing down the pool of articles based on specific criteria. At each stage, we assessed the relevance of articles by examining key elements such as titles, abstracts, keywords, and full-text content. This structured approach ensures a comprehensive and transparent selection of studies for our review.

The search process, as illustrated in Fig. 1, began with the identification of potential studies from the databases mentioned in the previous chapter. Initially, 67 records were identified, however, 5 of those records were duplicates and were removed before the screening.

During the screening phase, 62 records were assessed based on their titles and abstracts. This initial screening resulted in the exclusion of 18 records that did not meet the basic criteria or were clearly out of scope for our research objectives.

In the next screening phase, 44 reports were identified for retrieval. However, 6 of these were inaccessible, resulting in their exclusion.

In the eligibility assessment stage, the full texts of the remaining 38 reports were carefully examined. This review led to the exclusion of 7 reports based on their applicability, meaning that the objectives of the studies didn't align with our goal.

Ultimately, the rigorous selection process resulted in 31 studies being included in the final review. These studies met all the inclusion criteria and were deemed highly relevant to our research objectives.

The reviewed studies utilized models across various languages, a diversity we deliberately maintained in our analysis. This decision was motivated by the fact that our own model was applied to a language other than English. By including studies with different linguistic contexts, we aimed to capture a broader range of insights that could be relevant to our work. This approach allowed us to examine how language-specific factors might influence model performance and outcomes, potentially providing valuable perspectives for our research.

While English-language studies dominate (51.5 %) (see Table 1), there's significant representation from other languages, particularly Chinese (27.3 %). The inclusion of Italian, German, and Spanish studies,

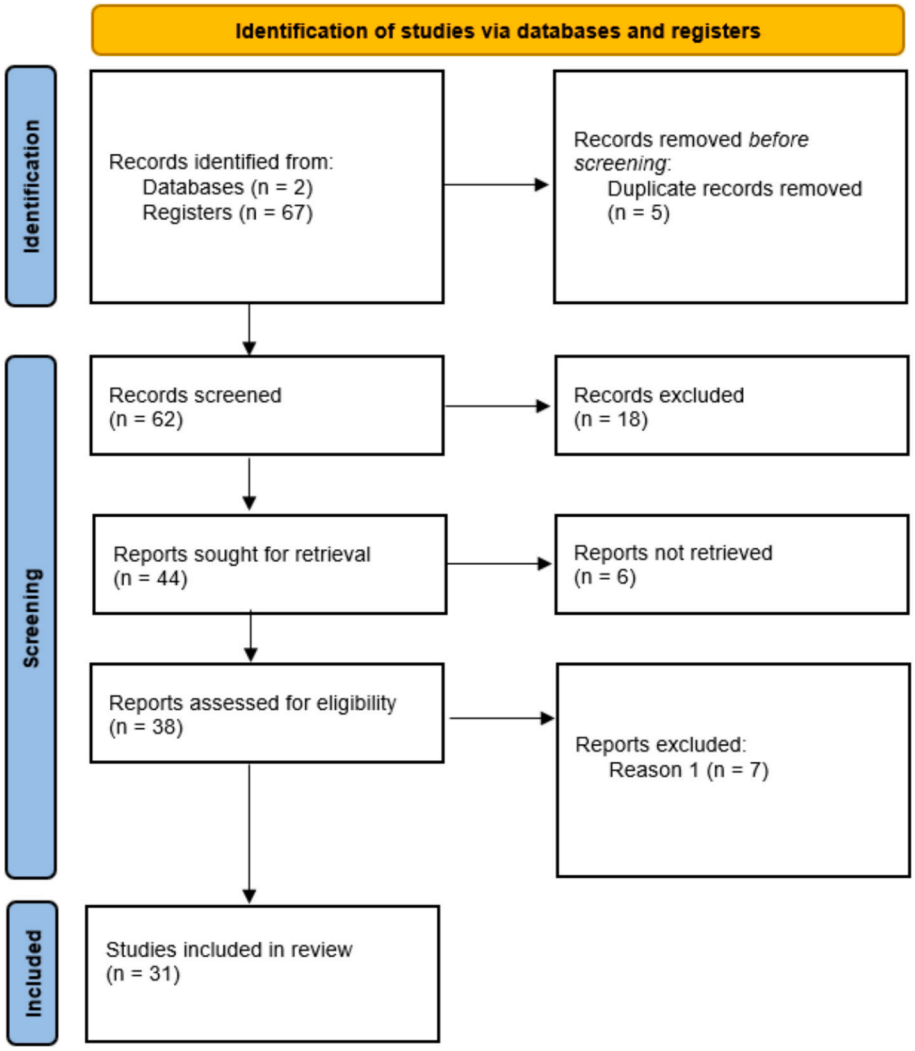


Fig. 1. PRISMA Flow Diagram with the state-of-the-art results.

Table 1
Reviewed articles by their data languages.

	References	Count	Percentage
Language			
English	[17–33]	17	51.5 %
Chinese	[34–42]	9	27.3 %
Italian	[43,44]	2	6.1 %
German	[45,46]	2	6.1 %
Spanish	[47]	1	3.0 %

though in smaller proportions, adds valuable cross-linguistic perspectives. This linguistic diversity highlights the potential for NLP techniques to be adaptable across different medical language contexts.

According to the keywords selected in the search query, the studies chosen for review focus on Information extraction or NER, but the techniques vary. We identified the following techniques:

The most used technique (see Table 2) was BERT or some variation/combination of it, with a representation of 36.4 % in the total articles reviewed. In second and third place come CRF or LSTM and then NN, with a representation of 33.3 % and 30.3 %, respectively. We decided to group CRF and LSTM because they appeared together in 7 out of the 10 articles where they make an appearance. The rest of the techniques appear with much lower frequency, which might suggest that BERT, NN, CRF, and LSTM are currently the most established and effective methods in the field. This prevalence could indicate their robustness and adaptability across different languages and tasks, while also highlighting potential areas for innovation using less common techniques like Federated Learning or Knowledge-Enhanced Frameworks.

3.1. Results from English-language clinical corpora

The following sub-sub-section examines studies that utilized English-language medical texts, representing 51.5 % of the reviewed literature.

3.2. BERT-based approaches

Recent studies have demonstrated the exceptional capabilities of BERT-based approaches in medical text processing, particularly in clinical NER and Electronic Health Records (HER) analysis. A study [17] utilizing deep neural network (DNN) and a biomedical corpus statistics achieved an F1-Score of 88.34 %, successfully addressing out-of-vocabulary challenges through improved handling of medical term morphology. Similarly, entity-BERT [21] further exemplified this success in medical electronic records processing, achieving remarkable precision and recall scores above 97 %, with accuracy scores exceeding 92.5 % on ACE 2004 and 96.5 % on ACE 2005.

Table 2
Overview of model architectures, input representations, and datasets used across the reviewed studies.

	References	Count	Percentage
Techniques			
BERT or variation	[18,21,27,28,30,32,34,37–430,47]	12	36.4 %
CRF (Conditional Random Field) or LSTM (Long Short-Term Memory)	[23–26,31,36–40,43]	11	33.3 %
NN (Neural Networks)	[17,23,24,26,28,29,34,39,41,44]	10	30.3 %
Rule-Based	[22,24,42,45,46]	5	15.2 %
GA (Genetic Algorithm) with Machine Learning	[20,34]	2	6.1 %
FL (Federated Learning)	[19]	1	3.0 %
KTGF (Knowledge-Enhanced Two-Stage Generative Framework)	[35]	1	3.0 %
SNER (Semi-Supervised Named Entity Recognition)	[33]	1	3.0 %

In emergency case management, BERT-based systems also showed promising results. A system designed for intracerebral hemorrhage (ICH) cases [27] achieved an accuracy of 88.55 %, with specificity of 85.71 % and sensitivity of 91.83 %. In the realm of clinical data curation, a BERT-based pipeline for identifying immune checkpoint inhibitor (ICI)-colitis from EHR notes [18] achieved 84 % precision while reducing manual review load by 75 %. This extraction pipeline reduces manual note review load and makes EHR more accessible for research.

The application of BERT extended beyond purely medical applications, as demonstrated by the crop disease diagnosis model based on BERT, specific to the crop disease domain and recurrent convolutional neural network (RCNN) (CdsBERT-RCNN) [28], which achieved an F1-score and accuracy of approximately 85.65 %. Integration studies combining BERT for NER and Relation Extraction (RE) tasks in Electronic Medical Records (EMR) [30] demonstrated precision, recall, and F1-scores exceeding 90 %. The research results demonstrate that incorporating semantic information can significantly enhance the understanding and analysis of medical imaging, providing valuable insights for medical intelligence research and application.

A study using a medical dictionary enhanced (Med-BERT) variant [32], showed a strong performance in handling long medical entities. When combined with the proposed a span flat-lattice transformer (Span-FLAT) method, it achieved recall results between 92–95 %, surpassing existing state-of-the-art methods.

3.3. NN-based approaches

Models using NN incorporated with LSTM configurations have shown strong performance across different applications. A character-level neural network using LSTM-CRF with CNN and Bidirectional Long Short-Term Memory (BiLSTM) [23] achieved results with an F1-score of 84.61 and recall of 85.41, demonstrating effective performance without external resources. This success was further exemplified in medical record de-identification, where a hybrid framework combining CNN and BiLSTM + CRF models [26] achieved exceptional results, with F1-scores of 99.16 % on their corpus and 96.24 % on the 2014 i2b2 i2b2/UTHealth de-identification dataset [48]—a benchmark comprising over 1,300 longitudinal clinical records annotated for protected health information.

Recurrent Neural Network (RNN) models, particularly those incorporating gated recurrent unit (GRU) and LSTM [24], have demonstrated superior performance compared to traditional statistical models in information extraction tasks. In parallel, CNN-based approaches [29] have shown effectiveness in specialized applications, such as section detection within medical documents, where they enhanced information extraction through improved contextual understanding.

In the realm of contextual embedding applications, a study [25] examined a bootstrapped neural network approach for Coronary Heart Disease (CAD) records using deep language models pre-trained on unannotated CAD corpus. The study found that embeddings from language models (ELMo) and BERT-based contextual embedding methods significantly outperformed traditional word embedding approaches, particularly when incorporating information from unannotated CAD corpora. This work highlights the potential of contextual embedding in maximizing the utility of limited annotated data while leveraging larger unannotated clinical corpora, though the authors acknowledge room for further methodological improvements.

3.4. Other techniques

Rule-based approaches, while less common in the reviewed

literature, showed effectiveness in specific applications. An example is an NLP algorithm developed for summarizing narrative breast pathology reports [22], which achieved exceptional performance with a micro-F1 score of 99.50 % and a macro-F1 score of 98.97 %. This algorithm assists clinicians in highlighting the key points from narrative reports and serves as a preliminary step in promoting synoptic reporting in the clinical domain.

A recent study evaluated FL's effectiveness across 8 biomedical corpora using 6 language models, including BERT, GPT, and the RNN-based and BiLSTM-CRF [19], while ensuring data privacy. Results showed that FL models achieved strong F1-scores (ranging from 88 % to 99 %), highlighting their comparable performance to centralized models. However, performance diminished as the number of clients increased, indicating a trade-off between client count and model accuracy in FL applications for biomedical NLP. In parallel, a hybrid machine learning model combining GAs [20] achieved accuracies of 84.13 % and 78.88 % for PubMed 20 K RCT and Heart Failure datasets respectively, supporting ensemble learning's effectiveness in text classification.

In the domain of specialized applications, a novel SNER model [33] demonstrated superior performance in processing diabetes data, particularly excelling with limited labeled data. Similarly, a CRF based study for clinical de-identification [31] showed promising results across various Protected Health Information categories, with particularly high precision for hospital (96 %) and ID (99 %) recognition, offering potential solutions for enterprise-wide data warehousing challenges.

These results demonstrate the diversity of approaches available for medical text processing and highlight the importance of selecting appropriate methodologies based on specific use cases and data availability.

3.5. Results from non-English clinical

Now we will analyze the remaining research conducted on non-English medical texts, encompassing data in Chinese (27.3 %), Italian (6.1 %), German (6.1 %), and Spanish (3.0 %).

While the architectural categories remain similar, the performance of techniques across languages varies due to linguistic and resource-specific factors. The following subsections retain the same structure to facilitate comparison.

3.6. BERT-based approaches

As with English-language datasets, BERT-based models dominate performance in non-English corpora. However, language-specific challenges and dataset variability still influence outcomes, as illustrated by the following examples.

In entity recognition tasks, three Chinese-language studies demonstrate progressive improvements in performance. A hybrid model combining bidirectional LSTM and CNN [34] achieved an F1-score of 85.75 %. Building upon this, the Medical Chinese BERT with graph neural network (GCN) and CRF (MCBERT-GCN-CRF) joint training model [38] reached 86.29 % F1 when tested on the CCKS-2019 dataset for medical entity identification. The most advanced approach—Medical Chinese BERT with BiLSTM and CRF (MC-BERT-BiLSTM-CRF) [40] – applied to rheumatoid arthritis EMRs, achieved 92.96 % F1 for entity recognition and 95.29 % for relation extraction, demonstrating excellent performance even with limited annotated data.

Another implementation, the BERT-based improved convolutional neural network (imConvNet) with BiLSTM and CRF (BERT-imConvNet-BiLSTM-CRF) [39], tested on a clinical corpus of impacted wisdom teeth, achieved F1-scores between 91 % and 93 %. This four-layer architecture further validates the effectiveness of combining pretrained language models with deep neural networks in specialized medical domains.

For Spanish-language data, the study by Chizhikova et al. [47] presents a radiologist-annotated corpus and evaluates several transformer-based models, including BioBERT and RoBERTa. The best-performing variant, the RoBERTa biomedical-clinical model, achieved 86.7 % micro and 83.2 % macro F1-score, highlighting the potential of fine-tuned models even in lower-resource linguistic settings.

3.7. NN-based approaches

A Chinese study focused on enhancing NER in EMR by developing a two-stage enhancement of the Transformer Encoder for Named Entity Recognition (TENER) architecture [41]. First, they integrated radical features into the base TENER model, and then augmented it with CNN-extracted features specifically for entity category recognition. This enhanced architecture was tested across three distinct datasets, demonstrating consistent performance with F1-scores ranging from 70 % to 82 %.

Parallel efforts in Italian medical text processing demonstrated the effectiveness of combining traditional and neural approaches. Researchers developed a hybrid system that integrated a RNN with dictionary lookup methods for clinical event extraction from medical reports [44]. This combined approach proved to be effective, achieving 91.7 % recall, 88.6 % precision, and a 90.1 % F1-score, demonstrating that sophisticated neural architectures can be successfully adapted for non-English medical text processing while maintaining high performance levels.

3.8. CRF and LSTM approaches

In the Chinese medical domain, two studies advanced different approaches to medical text processing. The first introduced a multi-standard active learning method [36] that achieved a remarkable 66.67 % reduction in required labeled data compared to traditional supervised learning approaches. Building on this efficiency breakthrough, another study combined Transformer architecture with Soft Term Position Lattice [37], reaching f-scores of 91.6 % for long medical term recognition and 90.36 % for abbreviations and numbers. This approach can capture potential interactions between Chinese characters and words.

In the Italian medical context, researchers developed a COVID-19 de-identification dataset utilizing a BiLSTM + CRF architecture with stacked embedding [43]. This study demonstrated a hierarchical performance pattern: achieving an exceptional 97.94 % F1-score at the binary token level, while maintaining robust performance of 80.17 % and 80.49 % at the entity and category levels, respectively. These results highlighted the benefits of combining contextualized and character-level language models with sub-word embedding.

3.9. Other techniques

One study presents a KTGF for extracting term-status pairs from medical dialogues, addressing a key challenge in automated EMR scribing [35]. KTGF innovates through a two-phase approach: first generating medical terms and then assigning their status, while incorporating prior medical knowledge. Tested on two Chinese datasets, the framework demonstrated significant improvements over existing models, with performance gains of up to 5.55 % in term identification and 5.28 % in full evaluations, particularly excelling in low-resource scenarios.

A rule-based text mining approach is introduced to extract key patient features from unstructured clinical documents in a German memory clinic database [46]. The system, requiring minimal training data, achieved over 95 % F1-scores for classes like memory disturbances, creating a machine-readable dataset from 8,000 patient visits over ten years. This highlights text mining's potential to reveal valuable clinical data from routine documentation.

Researchers developed a specialized NLP pipeline for extracting Tumor Node Metastasis (TNM) classification values from malignant tumor data, using German medical records [45]. Their key innovation

was the implementation of readable, well-structured rules within the text extraction pipeline, particularly using Apache Uima Ruta language (AURL). This approach significantly improved system maintainability, allowing complex Regular expression (REGEX) rules to be simplified and reprogrammed efficiently – reducing the process to just 5 min – while enhancing the system's transferability to different contexts.

Another study using Chinese EMR, focused on extracting clinical information for hepatocellular carcinoma patients through a dual methodological approach [42]. The researchers developed both pure rule-based methods and hybrid solutions that combined rule-based approaches with CRF-segmentation techniques. Both methods achieved robust performance metrics with precision, recall, and F1-scores exceeding 80 %. When applied to the practical task of hepatocellular carcinoma (HCC) staging, the automated extraction showed a 75 % concordance rate with manual expert review, demonstrating its potential for clinical application.

4. Discussion

In this chapter, we will examine the findings from our literature review, synthesizing the key trends, achievements, and challenges identified in the field of medical text processing using NLP techniques. Our analysis focuses on three main areas: principal findings that emerged from the reviewed literature, highlighting patterns and breakthrough achievements; and future research directions that could address current limitations and expand the field's capabilities; and finally, the main conclusions drawn from our analysis.

4.1. Principal findings

The analysis of the reviewed articles demonstrated a clear trend in methodological approaches, with BERT-based models, NN, and CRF/LSTM approaches emerging as the dominant techniques. These methods consistently achieved high performance metrics, with F1-scores frequently exceeding 85 % across various applications.

BERT-based approaches showed particularly strong results, with some implementations achieving precision and recall scores above 97 %. Examples include entity-BERT's performance on ACE datasets (accuracy > 92.5 %) and Med-BERT's handling of long medical entities (recall 92–95 %). These results suggest that BERT and its variants are well-suited for medical text processing tasks.

The review also highlighted the effectiveness of hybrid approaches that combine multiple techniques. For instance, the integration of CNN with BiLSTM + CRF models achieved noteworthy results in medical record de-identification (F1-scores > 99 %), while combinations of BERT with other architectures, such as BERT-imConvNet-BiLSTM-CRF, demonstrated strong performance (F1-scores 91–93 %) in specialized medical contexts.

A significant finding for our research was the successful adaptation of these techniques across different languages. While English-language studies dominated, substantial achievements were documented in Chinese, Italian, German, and Spanish implementations. This linguistic diversity demonstrates that modern NLP techniques can be effectively adapted to non-English contexts. The results across multiple languages suggests that the fundamental architectures and methodologies are language-agnostic, given appropriate training data and language-specific adjustments.

There was also an emphasis on data privacy and protection across the reviewed literature. Some articles addressed privacy concerns through various approaches, including de-identification techniques and secure data handling protocols. This focus on privacy reflects the field's recognition of the sensitive nature of medical data and the importance of maintaining patient confidentiality while advancing NLP capabilities.

The review also revealed developments in handling limited labeled data, with some studies achieving significant reductions in required labeled data (up to 66.67 %) through active learning approaches. This advancement is important for medical applications where labeled data can be scarce or resource-intensive to obtain.

Rule-based approaches, while less prevalent, demonstrated effectiveness in specific applications, achieving F1-scores as high as 99.50 % in tasks such as summarizing narrative pathology reports. This suggests that traditional rule-based methods still have valuable applications, particularly in well-defined, specialized medical tasks.

By looking at Table 3, we can observe in this cross-study comparison that BERT-based models consistently outperform other techniques in medical NER tasks across a variety of languages and contexts. In English, BERT-based models consistently achieve F1-scores above 90 %, with some cases exceeding 97 % precision and recall, as seen in the Entity-BERT model. In Chinese, advanced BERT architectures—such as MC-BERT combined with BiLSTM, CNN, and CRF—reach F1-scores above 92 %, outperforming other approaches like TENER (maximum ~ 83 %) and traditional CRFs. Even in Spanish, where the available corpora are smaller, a BERT variant (RoBERTa Biomedical) still achieves 83.2 % F1, surpassing traditional baselines.

In contrast, non-BERT techniques tend to deliver more variable results. Traditional CRF or LSTM-based models usually range between 80–86 % F1, and while neural networks (e.g., CNN + BiLSTM) can

Table 3
Model performance by technique and language.

Technique	Language	Model/Study	F1-score	Precision	Recall/ Accuracy
BERT-based	English	Entity-BERT	>97	>97	>92.5
BERT-based	English	BERT-ICH	91.83	85.71	88.55
BERT-based	English	BERT-ICI Colitis	84	–	–
BERT-based	English	CdsBERT-RCNN	85.65	–	85.65
BERT-based	English	BERT for NER + RE	>90	>90	>90
BERT-based	English	Med-BERT + Span-FLAT	92–95	–	–
NN	English	LSTM-CRF (Char level)	84.61	85.41	–
NN	English	CNN + BiLSTM + CRF (de-id)	99.16	–	–
CRF	English	CRF (de-id)	–	96	–
Rule-based	English	Narrative breast reports	98.97	–	99.50
BERT-based	Chinese	MC-BERT + BiLSTM + CNN + MHA + CRF	94.22/ 86.47/ 92.28	–	–
BERT-based	Chinese	MC-BERT + BiLSTM + MHA + CRF (variant)	94.22/ 86.27/ 92.82	–	–
BERT-based	Chinese	MC-BERT baseline (BiLSTM + CRF)	91.91	–	–
BERT-based	Chinese	TENER dual-branch model	82.67/ 74.37/ 70.16	–	–
NN	Chinese	TENER model (CMeEE benchmark)	68.39	–	–
CRF	Chinese	Soft Term Lattice	91.6	–	–
Rule-based	Chinese	Rule + CRF (HCC staging)	>80	>80	>80
BERT-based	Spanish	RoBERTa Biomedical	83.2	–	86.7
NN	Italian	Hybrid RNN + Dictionary	90.1	88.6	91.7
CRF	Italian	BiLSTM + CRF (de-id, Flair + FastText)	96.96	–	–
Rule-based	German	Memory Clinic	>95	–	–

occasionally achieve exceptional performance—such as the 99.16 % F1 score reported for a de-identification task—these results are task-specific and not directly comparable to general NER applications. De-identification tasks typically involve clearly defined entities like names, dates, or IDs, which can be learned effectively by traditional or hybrid models when sufficient labeled data is available. However, for broader NER tasks that require complex contextual understanding and domain adaptation, BERT-based models exhibit superior generalizability and robustness, consistently ranking as the most effective and adaptable solution in modern clinical NLP.

Although model performance is well documented across the reviewed studies, our analysis reveals a clear gap in real-world applicability. Only a small subset of papers mention key deployment-related aspects such as *external validation*, *cross-site evaluation*, or *clinical scalability*. For instance, one study [27] references external validation for an automatic system in hemorrhage management, while a few others [19,26] mention generalizability or scalability in theoretical terms. However, most models were tested only on internal datasets, often from a single site, with no assessment of performance stability across institutions, EHR systems, or clinical workflows. This limits the current evidence on how such models would perform in production environments. We emphasize that future research in clinical NLP should report on multi-site validation, deployment in hospital settings, and operational constraints, to bridge the gap between proof-of-concept performance and clinical utility.

Beyond covering multiple languages, our review identified language-specific challenges that impact the performance and generalizability of NLP models in clinical settings. For Chinese, several studies acknowledge the complexity of word segmentation and the choice between character-level vs. word-level tokenization, which significantly affects model design and accuracy [23,30]. In German and Italian, issues such as morphological complexity, compound words, and lack of annotated corpora were frequently cited [43,46]. Spanish studies also noted limited corpus availability and challenges in transferring models trained on high-resource languages [47]. Overall, multilingual clinical NLP remains constrained by language-specific preprocessing needs, data scarcity, and morphological variability, underlining the importance of tailored solutions and cross-lingual transfer methods for under-resourced languages.

Furthermore, although technical performance is a primary focus in most reviewed studies, ethical and regulatory considerations remain underexplored. Only a minority of papers explicitly address concerns such as data privacy, bias in medical data, or compliance with standards like GDPR or HIPAA. For instance, [19] discuss the role of privacy in federated learning, [26] propose a de-identification framework that considers real-world data protection, and [43] highlight anonymization strategies for Italian clinical text. However, these are exceptions. Discussions on bias mitigation, fairness, and external auditability are rare, limiting the practical readiness of current NLP models for deployment in diverse clinical settings. Future research should prioritize ethical-by-design NLP pipelines that integrate data protection principles and address algorithmic fairness to enhance trust and clinical relevance.

Comparative analysis across the reviewed studies suggests that model effectiveness is closely tied to the nature of the task and clinical data. For example, BERT-based models consistently outperform traditional methods in complex entity recognition tasks due to their ability to model long-range dependencies and context. In contrast, hybrid models—such as BiLSTM combined with CRFs or CNNs—perform competitively in structured tasks like de-identification, where entities like dates and names follow predictable patterns. Furthermore, language-specific preprocessing demands (e.g., segmentation in Chinese or morphological variation in German) also affect model design choices and performance. Studies using transformer-based models tailored to biomedical domains (e.g., BioBERT, MC-BERT) report higher F1-scores across languages and domains, especially when fine-tuned on task-specific corpora. These findings reinforce that no single architecture is universally optimal; instead, model selection should consider task complexity, data structure, and linguistic context.

Despite the promising performance of BERT-based models, their scalability for real-time clinical integration remains uncertain. Most reviewed studies focused on offline evaluation using static datasets, without addressing latency, computational requirements, or integration with existing hospital information systems. Transformer-based architectures, while powerful, can be resource-intensive, posing challenges for deployment in time-sensitive clinical settings with limited infrastructure. Additionally, few studies presented user-centered evaluations or workflow integration strategies, which are essential for real-world adoption. Future research should consider lightweight or distilled variants of these models and focus on deployment pipelines that align with clinical decision-making timelines and operational constraints.

4.2. Future work

Future research in healthcare NLP presents several promising directions for advancement. In the technical domain, developing improved embedding techniques and efficient model architectures could enhance performance while reducing computational demands. Comparative research is also needed to guide model selection based on task type, language, and data complexity—reinforcing that no single architecture is universally optimal.

Data quality and availability remain crucial areas for improvement. Expanding and diversifying datasets across different healthcare institutions could enhance model generalizability and reduce biases. Emphasis should be placed on external validation, cross-site evaluation, and benchmarking across clinical settings to assess scalability and robustness in real-world use.

Privacy-preserving approaches and secure data-sharing frameworks—such as federated learning and differential privacy—can facilitate collaborative research while maintaining patient confidentiality and complying with regulations like GDPR or HIPAA. Future work should also address fairness and bias mitigation to ensure equitable clinical NLP applications.

Cross-lingual and multilingual capabilities represent another significant frontier. Developing models capable of processing medical data across different languages and adapting to varied healthcare contexts could broaden the global applicability of NLP solutions. This includes improving language model adaptability, addressing segmentation and morphological challenges, and enhancing cross-domain performance.

Finally, addressing the challenge of limited labeled data through advanced techniques such as weak supervision, data augmentation, and synthetic data generation could reduce annotation bottlenecks. Optimizing methods for handling non-standard terminology and heterogeneous clinical documentation across systems remains a key goal for practical deployment.

5. Conclusion

This review of recent developments in medical text processing highlights the dominance of BERT-based approaches, which consistently achieve state-of-the-art results across a variety of clinical NLP tasks. These models excel in complex scenarios requiring contextual understanding, particularly in named entity recognition. However, hybrid architectures—such as combinations of BiLSTM, CNN, and CRF—continue to perform competitively in structured tasks like de-identification, especially when dealing with well-defined entity types.

Multilingual studies confirm that transformer-based models, when adapted to biomedical and linguistic contexts, retain high performance across diverse healthcare settings. Yet, language-specific preprocessing challenges and the scarcity of annotated corpora in low-resource languages remain significant hurdles.

While technical performance is robust, most studies still lack external validation and deployment insights. Real-world integration of these

models faces challenges related to scalability, interoperability, ethical compliance, and bias mitigation. Addressing these gaps will be essential to move from proof-of-concept results to clinically adopted NLP solutions.

Modern NLP techniques are well-positioned to transform healthcare text analysis, but future progress must align technical innovation with practical deployment, multilingual accessibility, and ethical safeguards to ensure equitable and trustworthy applications in clinical practice.

CRediT authorship contribution statement

Luis B. Elvas: Writing – review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Ana Almeida:** Writing – original draft, Methodology. **João C. Ferreira:** Supervision.

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.

References

[1] H.-P. Chan, R.K. Samala, L.M. Hadjiiski, C. Zhou, Deep learning in medical image analysis, Adv. Exp. Med. Biol. 1213 (2020) 3–21. PMID:32030660.

[2] M.C. Durango, E.A. Torres-Silva, A. Orozco-Duque, Named entity recognition in electronic health records: a methodological review, Healthc. Inform. Res. 29 (4) (2023 Oct) 286–300. PMID:37964451.

[3] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv; 2019. Available from: http://arxiv.org/abs/1810.04805 [accessed Oct 15, 2024].

[4] C. Friedman, L. Shagina, Y. Lussier, G. Hripcsak, Automated encoding of clinical documents based on natural language processing, J. Am. Med. Inform. Assoc. JAMIA 11 (5) (2004) 392–402. PMID:15187068.

[5] Settles B. Biomedical named entity recognition using conditional random fields and rich feature sets. Proc Int Jt Workshop Nat Lang Process Biomed Its Appl USA: Association for Computational Linguistics; 2004. p. 104–107.

[6] O. Uzuner, Y. Luo, P. Szolovits, Evaluating the state-of-the-art in automatic de-identification, J. Am. Med. Inform. Assoc. JAMIA 14 (5) (2007) 550–563. PMID: 17600094.

[7] A.N. Jagannatha, H. Yu, Bidirectional RNN for medical event detection in electronic health records, Proc Conf. Assoc. Comput. Linguist. North Am. Chapter Meet 2016 (2016 Jun) 473–482. PMID:27885364.

[8] A. Gillioz, J. Casas, E. Mugellini, O.A. Khaled, Overview of the Transformer-based Models for NLP Tasks, in: 2020 15th Conf Comput Sci Inf Syst FedCSIS, 2020, pp. 179–183, https://doi.org/10.15439/2020F20.

[9] (PDF) Weakly-Supervised Multilingual Medical NER for Symptom Extraction for Low-Resource Languages. ResearchGate 2025 Jun 10; doi: 10.3390/app15105585.

[10] (PDF) A Systematic Review of Natural Language Processing in Healthcare. ResearchGate. doi: 10.5815/ijitcs.2015.08.07.

[11] H. Eguia, C.L. Sánchez-Bocanegra, F. Vinciarelli, F. Alvarez-Lopez, F. Saigí-Rubió, Clinical decision support and natural language processing in medicine: systematic literature review, J. Med. Internet Res. 26 (1) (2024 Sep 30) e55315, https://doi.org/10.2196/55315.

[12] PRISMA statement. PRISMA Statement. Available from: https://www.prisma-statement.org [accessed Jun 20, 2024].

[13] Scopus – Advanced search | Signed in. Available from: https://www.scopus.com/search/form.uri?display=advanced [accessed Sep 9, 2024].

[14] PubMed. PubMed. Available from: https://pubmed.ncbi.nlm.nih.gov/ [accessed Oct 15, 2024].

[15] Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, Liu T-Y. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform. 2022 Nov 1;23(6):bbac409. doi: 10.1093/bib/bbac409.

[16] Wang S, Sun X, Li X, Ouyang R, Wu F, Zhang T, Li J, Wang G. GPT-NER: Named Entity Recognition via Large Language Models. arXiv; 2023. doi: 10.48550/arXiv.2304.10428.

[17] A. Dash, S. Darshana, D.K. Yadav, V. Gupta, A clinical named entity recognition model using pretrained word embedding and deep neural networks, Decis Anal. J. 10 (2024), https://doi.org/10.1016/j.dajour.2024.100426.

[18] P. Rahman, C. Ye, K.F. Mittendorf, M. Lenoue-Newton, C. Micheel, J. Wolber, T. Osterman, D. Fabbri, Accelerated curation of checkpoint inhibitor-induced colitis cases from electronic health records, JAMIA Open 6 (1) (2023), https://doi.org/10.1093/jamiaopen/ooad017.

[19] L. Peng, G. Luo, S. Zhou, J. Chen, Z. Xu, J. Sun, R. Zhang, An in-depth evaluation of federated learning on biomedical natural language processing for information extraction, Npj Digit Med 7 (1) (2024), https://doi.org/10.1038/s41746-024-01126-4.

[20] G. Ben Abdennour, K. Gasmi, R. Ejbali, An optimal model for medical text classification based on adaptive genetic algorithm, Data Sci. Eng. (2024), https://doi.org/10.1007/s41019-024-00257-8.

[21] W. Lu, J. Jiang, Y. Shi, X. Zhong, J. Gu, L. Huangfu, M. Gong, Application of Entity-BERT model based on neuroscience and brain-like cognition in electronic medical record entity recognition, Front. Neurosci. 17 (2023), https://doi.org/10.3389/fnins.2023.1259652.

[22] W.-M. Tan, K.-H. Teoh, M.D. Ganggayah, N.A. Taib, H.S. Zaini, S.K. Dhillon, Automated generation of synoptic reports from narrative pathology reports in University Malaya Medical Centre using natural language processing, Diagnostics 12 (4) (2022), https://doi.org/10.3390/diagnostics12040879.

[23] L. Li, W. Xu, H. Yu, Character-level neural network model based on Nadam optimization and its application in clinical concept extraction, Neurocomputing 414 (2020) 182–190, https://doi.org/10.1016/j.neucom.2020.07.027.

[24] S. Malmasi, W. Ge, N. Hosomura, A. Turchin, Comparing information extraction techniques for low-prevalence concepts: the case of insulin rejection by patients, J. Biomed. Inform. 99 (2019), https://doi.org/10.1016/j.jbi.2019.103306.

[25] X. Cen, J. Yuan, C. Pan, Q. Tang, Q. Ma, Contextual embedding bootstrapped neural network for medical information extraction of coronary artery disease records, Med. Biol. Eng. Comput. 59 (5) (2021) 1111–1121, https://doi.org/10.1007/s11517-021-02359-1.

[26] L. Liu, O. Perez-Concha, A. Nguyen, V. Bennett, L. Jorm, De-identifying Australian hospital discharge summaries: an end-to-end framework using ensemble of deep learning models, J. Biomed. Inform. 135 (2022), https://doi.org/10.1016/j.jbi.2022.104215.

[27] B. Deng, W. Zhu, X. Sun, Y. Xie, W. Dan, Y. Zhan, Y. Xia, X. Liang, J. Li, Q. Shi, L. Jiang, Development and validation of an automatic system for intracerebral hemorrhage medical text recognition and treatment plan output, Front. Aging Neurosci. 14 (2022), https://doi.org/10.3389/fnagi.2022.798132.

[28] J. Ding, B. Li, C. Xu, Y. Qiao, L. Zhang, Diagnosing crop diseases based on domain-adaptive pre-training BERT of electronic medical records, Appl. Intell. 53 (12) (2023) 15979–15992, https://doi.org/10.1007/s10489-022-04346-x.

[29] M.Y. Landolsi, L. Hlaoua, L.B. Romdhane, Hybrid method to automatically extract medical document tree structure, Eng. Appl. Artif. Intell. 120 (2023), https://doi.org/10.1016/j.engappai.2023.105922.

[30] B. Guo, H. Liu, L. Niu, Integration of natural and deep artificial cognitive models in medical images: BERT-based NER and relation extraction for electronic medical records, Front. Neurosci. 17 (2023), https://doi.org/10.3389/fnins.2023.1266771.

[31] T. Paul, M.K.Z. Rana, P.A. Tautam, T.V.P. Kotapati, Y. Jampani, N. Singh, H. Islam, V. Mandhadi, V. Sharma, M. Barnes, R.D. Hammer, A.S.M. Mosa, Investigation of the utility of features in a clinical de-identification model: a demonstration using EHR Pathology reports for advanced NSCLC patients, Front. Digit Health 4 (2022), https://doi.org/10.3389/fdgth.2022.728922.

[32] N. Liu, Q. Hu, H. Xu, X. Xu, M. Chen, Med-BERT: a pretraining framework for medical records named entity recognition, IEEE Trans. Ind. Inform. 18 (8) (2022) 5600–5608, https://doi.org/10.1109/TII.2021.3131180.

[33] J. Zuo, Q. Qian, Y. Liu, S. Lu, B. Li, Y. Guo, SNER: semi-supervised named entity recognition for large volume of diabetes Data, IEEE J. Biomed. Health Inform. (2024) 1–14, https://doi.org/10.1109/JBHI.2024.3412716.

[34] Y. Wang, Y. Sun, Z. Ma, L. Gao, Y. Xu, A hybrid model for named entity recognition on chinese electronic medical records, ACM Trans. Asian Low-Resour. Lang. Inf. Process. 20 (2) (2021), https://doi.org/10.1145/3436819.

[35] Z. Hu, Z. Ni, J. Shi, S. Xu, B. Xu, A knowledge-enhanced two-stage generative framework for medical dialogue information extraction, Mach. Intell. Res. 21 (1) (2024) 153–168, https://doi.org/10.1007/s11633-023-1461-5.

[36] Q. Pan, C. Huang, D. Chen, A method based on multi-standard active learning to recognize entities in electronic medical record, Math. Biosci. Eng. 18 (2) (2021) 1000–1021, https://doi.org/10.3934/MBE.2021054.

[37] S. Guo, W. Yang, L. Han, X. Song, G. Wang, A multi-layer soft lattice based model for Chinese clinical named entity recognition, BMC Med. Inform. Decis. Mak. 22 (1) (2022) 201, https://doi.org/10.1186/s12911-022-01924-4.

[38] M. Li, C. Gao, K. Zhang, H. Zhou, J. Ying, A weakly supervised method for named entity recognition of Chinese electronic medical records, Med. Biol. Eng. Comput. 61 (10) (2023) 2733–2743, https://doi.org/10.1007/s11517-023-02871-6.

[39] Y Z, Z H, Y C, X D, J S, W Y, H H. An imConvNet-based deep learning model for Chinese medical named entity recognition. BMC Med Inform Decis Mak BMC Med Inform Decis Mak; 2022 Nov 21;22(1). PMID:36411432.

[40] F. Liu, M. Liu, M. Li, Y. Xin, D. Gao, J. Wu, J. Zhu, Automatic knowledge extraction from Chinese electronic medical records and rheumatoid arthritis knowledge graph construction, Quant. Imaging Med. Surg. 13 (6) (2023) 3873–3890, https://doi.org/10.21037/qims-22-1158.

[41] H. Peng, Z. Zhang, D. Liu, X. Qin, Chinese medical entity recognition based on the dual-branch TENER model, BMC Med. Inform. Decis. Mak. 23 (1) (2023), https://doi.org/10.1186/s12911-023-02243-y.

[42] L. Chen, L. Song, Y. Shao, D. Li, K. Ding, Using natural language processing to extract clinically useful information from Chinese electronic medical records, Int. J. Med. Inf. 124 (2019) 6–12, https://doi.org/10.1016/j.ijmedinf.2019.01.004.

[43] R. Catelli, F. Gargiulo, V. Casola, G. De Pietro, H. Fujita, M. Esposito, A Novel COVID-19 data set and an effective deep learning approach for the de-identification of italian medical records, IEEE Access 9 (2021) 19097–19110, https://doi.org/10.1109/ACCESS.2021.3054479.

[44] N. Viani, T.A. Miller, C. Napolitano, S.G. Priori, G.K. Savova, R. Bellazzi, L. Sacchi, Supervised methods to extract clinical events from cardiology reports in Italian, J. Biomed. Inform. 95 (2019), https://doi.org/10.1016/j.jbi.2019.103219.

[45] N. Ladas, F. Borchert, S. Franz, A. Rehberg, N. Strauch, K.K. Sommer, M. Marschollek, M. Gietzelt, Programming techniques for improving rule

readability for rule-based information extraction natural language processing pipelines of unstructured and semi-structured medical texts, Health Informatics J. 29 (1) (2023), https://doi.org/10.1177/14604582231164696.

[46] L. Langnickel, K. Krockauer, M. Uebachs, S. Schaaf, S. Madan, T. Klockgether, J. Fluck, Information extraction from German clinical care documents in context of Alzheimer's disease, Appl. Sci. Switz. 11 (22) (2021), https://doi.org/10.3390/app112210717.

[47] M. Chizhikova, P. López-Úbeda, J. Collado-Montañez, T. Martín-Noguerol, M. C. Díaz-Galiano, A. Luna, L.A. Ureña-López, M.T. Martín-Valdivia, CARES: a corpus for classification of Spanish Radiological reports, Comput. Biol. Med. 154 (2023), https://doi.org/10.1016/j.compbiomed.2023.106581.

[48] A. Stubbs, C. Kotfila, O. Uzuner, Automated systems for the de-identification of longitudinal clinical narratives: overview of 2014 i2b2/UTHealth shared task Track 1, J. Biomed. Inform. 58 (Suppl) (2015 Dec) S11–S19. PMID:26225918.