Automatic patient pre-screening for clinical trials: a narrative review

Anonymous ACL submission

Abstract

We present a narrative review of recent advances in Natural Language Processing for automating patient pre-screening in clinical trials. We review the state-of-the-art across three core tasks: (1) automatic generation of eligibility surveys from trial protocols, (2) extraction of structured patient information from electronic health records (EHRs) and (3) automatic patient-trial matching. We analyze recent trends in using neural architectures, and we highlight current bottlenecks in linguistic variability, data interoperability and hallucination in generative systems. Our survey aims to synthesize a fragmented landscape and provide future directions towards clinical trials improvement.

1 Introduction

The digitalisation of healthcare data transformed the management of patient information, opening opportunities to improve clinical trials and the pre-screening process. However, this process remains a challenge due to the complexities of determining eligibility from the unstructured data contained in EHRs. Identifying eligible patients for clinical trials involves analysing inclusion and exclusion criteria, a task conducted manually by healthcare professionals. This approach is time consuming, error prone, and limited by the resources available at each healthcare institution.

In the matrix Table 2Appendix B, we synthesize works that contributed to the state-of-the-art in biomedical NLP and EHR processing. Rows represent the goals aimed by the studies, columns the NLP tasks. Since the 1960s (Slack et al., 1966), there has been a transition from manual processing to more sophisticated, automated methods for managing clinical data. Prototypes such as COSTAR (Barnett et al., 1979) laid the groundwork for modern EHR systems (Embi et al., 2005). Improvements in computational power, storage, and data management systems have made widespread adoption of EHRs feasible, marking a shift toward their use in research settings, including clinical trials.

NLP has emerged as an enabler in the automation of EHR interpretation. Clinical NLP systems such as cTAKES (Savova et al., 2010) and MedCAT (Kraljevic et al., 2021) have been developed to extract structured information—diagnoses, medications and observations from free-text records. Others like MedEX (Xu et al., 2010) based on semantic taggers were developed to extract medication information from discharge summaries. These capabilities are leveraged to identify trial-eligible patients. However, the task remains difficult due to linguistic variability, inconsistencies in clinical documentation and limited interoperability across EHR systems.

The landscape has been transformed by LLMs and deep learning. Models such as BioBERT (Lee et al., 2020), ClinicalBERT (Huang et al., 2019) and GPT-4 have demonstrated performance in medical text understanding and text generation. They offer promise in automating clinical trial recruitment process, from generating patient-facing eligibility questions to matching structured EHR outputs.

Despite the advances, challenges remain. Linguistic diversity within medical texts, including synonyms, abbreviations... complicate the extraction of relevant information. Furthermore, the interoperability of different healthcare systems remains a barrier to data sharing and integration between institutions. Finally, maintaining data privacy and ensuring compliance with regulations such as Health Insurance Portability and Accountability Act (HIPAA) adds complexity to the implementation of automated systems for clinical trial recruitment (Zhang et al., 2020).

This narrative review synthesizes recent advances in automating patient pre-screening for clinical trial. As a high-level review of existing sys-

tematic and narrative studies, it offers a comprehensive perspective on three key tasks: generating structured eligibility surveys from trial protocols, extracting clinical information from EHRs and matching patients to trials. We examine how NLP approaches address persistent challenges in recruitment, including linguistic variability, data heterogeneity and decision accuracy.

The primary contributions of this review are as follows:

- Comprehensive synthesis: we consolidate a fragmented body of literature by comparing rule-based, statistical and LLM-driven approaches across the clinical trial pre-screening pipeline. The rule-based system is a deterministic automat that takes as input the feature extracted from text mining.
- Critical analysis of challenges: we identify and analyze barriers that hinder the real-world deployment of NLP systems.
- Improvement with LLMs: we check the benefits that language models provide to the task.
- Future research agenda: we outline promising directions for advancing the field

2 Related works

Systematic literature reviews (SLRs), umbrella reviews (UR) and narrative reviews play a vital role in structuring knowledge in areas such as clinical NLP. While SLRs and URs aim to answer defined questions through exhaustive evidence collection, narrative reviews are closer to manual search overviews where the position is based on qualitative literature review.Table 1 in the appendix B summarizes the most cited SLRs and umbrella reviews on NLP in clinical domain.

Some SLRs and umbrella reviews traced the evolution of patient-trial matching. (Meystre et al., 2008; Uzuner, 2008) describe the challenges and opportunities for extracting structured data from clinical narratives. The development of tools like cTAKES (Savova et al., 2010) marked a turning point in domain-specific NLP pipelines becoming a top-level software of Apache [1].

Early systems demonstrated gains in efficiency (Ni et al., 2015) by targeting trial recruitment directly. Efforts toward automation of systematic reviews emerged with (Beller et al., 2018), who proposed coordination frameworks like ICASR and (Ofori-Boateng et al., 2024), who synthesized deep learning applications across literature reviews, although neither focused on clinical trials. (Idnay et al., 2021)conducted the first systematic review dedicated to NLP systems for eligibility pre-screening, categorizing models and evaluating performance and (Panayi et al., 2023) demonstrated how machine learning tools can support semi-automated data extraction, yet their evaluation was limited to literature review workflows. In contrast, (Jin et al., 2024; Hamer et al., 2023) introduced LLMs into trial-patient matching, highlighting the potential of transformer-based models for joint text understanding, though both lacked robustness studies and real-world clinical deployment. (Rahmanian et al., 2023) proposed a prompt-based eligibility classification model with high adaptability, but minimal attention to interoperability or ethical constraints.

(Kuziemsky et al., 2024) propose an UR to describe AI in healthcare and describe a systematic review protocol. Most recently, (Sharif and Rehman, 2025) conducted a systematic review comparing LLM-based and classical approaches to eligibility matching. While these studies provide insights, LLM matching—there is no unified synthesis covering the entire clinical trial recruitment pipeline. This work addresses that gap by synthesizing approaches across clinical trial questionnaire generation and automatic patient pre-screening.In the next section, we describe our methodology to achieve it.

3 Methodology

As it is a narrative review, the knowledge and literature gathering and absorption is unconventional. It is based on a qualitative, non-systematic search of the literature using Google Scholar, ACL Anthology and PubMed, supplemented by iterative queries to ChatGPT for cross-verification and exploratory surfacing of under-indexed works. The use of ChatGPT instead of other commercial models is due to the fact that we are used to it and how to engineer prompts so it does what it's asked.

We first looked in Google Scholar for the keywords *historical, biomedical, NLP* in the same query. We used them to extract the most relevant articles that match the topic. We opened the first 20 results of the search and kept the articles that men-

[1] "An integral part of Mayo's clinical data management infrastructure, processing more than 80 million clinical notes" https://en.wikipedia.org/wiki/Apache_cTAKES

tioned NLP and a medical-related concept in the title or the abstract. Then we combined the keywords with *EHR, criteria extraction, question generation, patient pre-screening, patient-trial matching, LLM, few-shot, zero-shot*, filtering by the dates of 2001-2010 to get an historic overview and from 2019 to LM-related articles. As Google Scholar prioritises cited articles, we used this method to search in the ACL Anthology and PubMed.Whereas the ACL Anthology search engine struggled to provide results, the PubMed provided bibliography that made possible our work.

As for ChatGPT, using the model 4o (Shahriar et al., 2024) we instructed it with prompts, which 3 of them are available in the Appendix A. ChatGPT was useful regarding the search of recent overviews and surveys. We doubled-checked all the references provided by the tool,and forced it to search in NLP conference proceedings as well as medical databases. As we did not conduct a rigorous evaluation of its hallucination rate, around 80% of the references actually existed. The remaining 20% where either invented or did not correspond to the same article as the output claimed. However, the systematic reviews described in the Section 2 have been found by ChatGPT. Besides, the search is easier to customize in natural language than in Google Scholar. We were able to instruct it to look for special venues: LREC for resources and evaluation, TALN and SEPLN for linguistics-related topics.

In the next two sections, we discuss about questionnaire generation state-of-the-art and techniques to realize automatic patient pre-screening.

4 Questionnaire generation from eligibility criteria

According to the literature, the transformation of clinical trial eligibility criteria into patient-directed questionnaires can be implemented through two primary approaches: (1) a modular pipeline in which criteria are first structured into formal representations before being converted into questions and (2) an end-to-end generative approach in which models produce questions from free-text eligibility statements using LLMs. The most famous clinical trials database is clinicaltrial.gov[2], mentioned in Section 5.3.2, where the trial description is written in natural language, with a distinction between inclusion and exclusion criteria (Zarin and Keselman, 2007).

Figure 1 illustrates a question generation pipeline. The upper section represents the criteria extraction phase where, after an manual annotation phase, a named entity recognizer (NER) is trained to detect entities in the criteria (see matrix in Table 2 in Appendix B). Recent studies show that few-shot learning approaches (Naguib et al., 2024) reach similar accuracy than fully supervised NER, as well as zero-shot learning (Wornow et al., 2025) and even prompt-tuning-based systems such as Autocriteria (Datta et al., 2024). The lower section represents the questionnaire generation model training, that takes as input the clinical trials and examples of human generated questionnaires (Lei et al., 2024). Few-shot (Izacard et al., 2022), zero-shot (Zeng et al., 2023) and LLM-based solutions were also implemented. Depending on whether it is strategy (1) or (2), the generation model takes as input the extracted criteria (Dhomse, 2024).

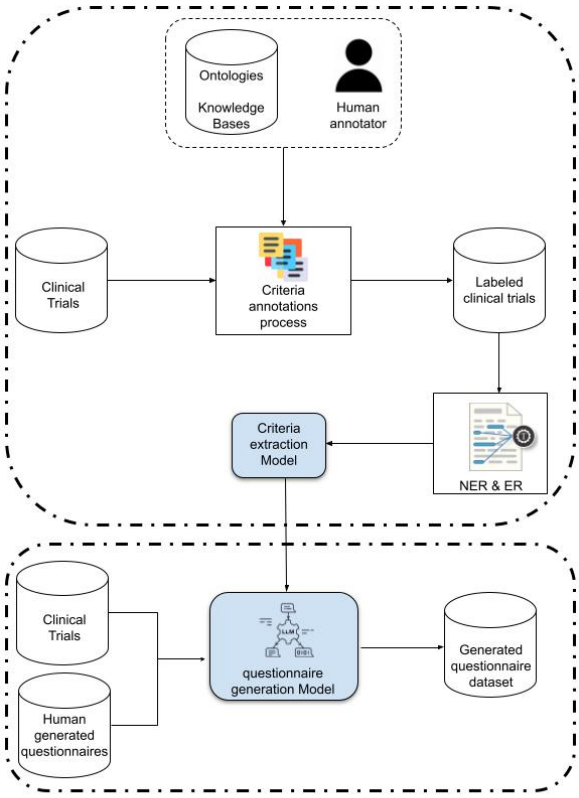

Figure 1: Questionnaire generation scheme: on top is the criteria extraction phase and on the bottom the questionnaire generation.

[2]https://clinicaltrials.gov/

4.1 Criteria extraction and question generation

This initial step involves transforming unstructured eligibility criteria texts into structured, computable representations. Systems such as MedSpaCy (Eyre et al., 2021) implement rule-based or hybrid methods that link clinical entities to controlled vocabularies (UMLS, SNOMED...). These structured outputs facilitate subsequent automated reasoning or question generation tasks.

However, these methods require extensive manual annotation, ontological grounding and predefined rule sets, limiting their adaptability and scalability to heterogeneous criteria texts (Tian et al., 2023). Recent approaches addressing these limitations incorporate neural-based strategies, such as BioBERT and zero-shot methods leveraging pretrained transformer models for entity extraction (Averly and Ning, 2025).

Nevertheless, limitations remain significant, including semantic ambiguity, domain specificity, negation handling and temporal reasoning (Mehrabi et al., 2015). Errors in the extraction stage propagate downstream, potentially compromising questionnaire accuracy, clinical relevance and overall system performance (Olex and McInnes, 2021).

Performance evaluation in two-step approaches involves distinct metrics for each stage. For criteria extraction, metrics include precision, recall and F1-score, benchmarked on datasets such as n2c2 shared tasks (Mahajan et al., 2023).

The second phase converts structured representations derived in the first step into natural language questions. Traditional generation methods rely on template-based approaches, where predefined linguistic templates map structured slots to surface-level question forms (Yuan et al., 2019). Although these approaches are straightforward and interpretable, they lack flexibility when criteria complexity increases or novel expressions appear.

Recent studies employ neural generation models trained on aligned datasets of structured criteria and human-authored questions. Such methods demonstrate greater fluency, adaptability and linguistic variability. These methods are analogous to machine translation tasks, where encoder-decoder architectures have been used to transform input texts into semantically equivalent target languages or formats (Ma et al., 2022). Nevertheless, these neural approaches face issues regarding logical consistency, semantic fidelity, and clinical appropriateness, requiring rigorous human validation and specialized evaluation frameworks.

BLEU (Papineni et al., 2001), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019) are used to evaluate surface fluency and lexical overlap, but they fail to capture semantic equivalence or clinical soundness, prompting recent calls for domain-specific validation metrics such as expert scoring or scenario-based evaluation. Moreover, these metrics assess surface-level linguistic fluency rather than clinical accuracy or logical consistency. Consequently, recent reviews mentioned in Section 2 advocate for more clinically-oriented evaluationsto more accurately measure real-world utility and robustness of generated questions.

4.2 End-to-end questionnaire generation

Unlike modular approaches, end-to-end methods map raw, free-text clinical trial eligibility criteria to patient-facing questions, eliminating explicit intermediate structuring steps. Leveraging advances in LLMs, these approaches formulate the questionnaire generation task as a single sequence-to-sequence problem (Brown et al., 2020). The underlying hypothesis is that LLMs can implicitly learn internal representations capable of capturing linguistic, logical, and semantic complexities without separate structuring modules. These methods improve scalability, adaptability to new medical domains and flexibility for heterogeneous criteria (Frayling et al., 2024).

Recent studies investigate few-shot learning: using a limited set of criteria-question examples, these methods demonstrate generalization to new criteria without substantial domain-specific annotations (Lin et al., 2024; Poon et al., 2024). Zero-shot and prompt-based learning were also experimented: leveraging pretrained LLMs through specialized task prompts eliminating fine-tuning entirely (Elsahar et al., 2018).

End-to-end methods simplify the pipeline by modeling eligibility criteria-to-question mappings, thus avoiding cascading errors inherent to modular approaches (Ferber et al., 2024). Additionally, LLMs capture linguistic complexity, providing scalability across diverse medical contexts without explicit rule engineering or ontology mapping (Bohra et al., 2023; Wong et al., 2023). However, they suffer from logical inconsistencies and semantic inaccuracies, including omissions and hallucinations, compromising clinical reliability (Wang

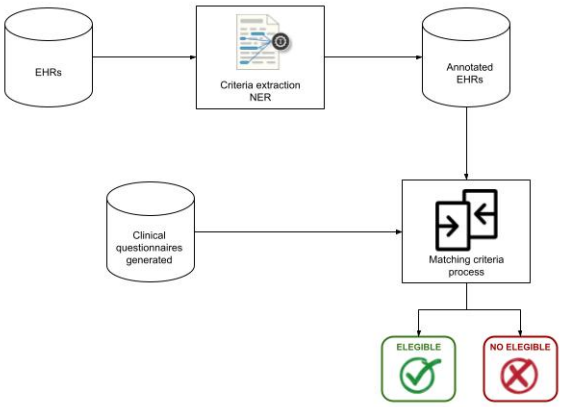

Figure 2: Automatic patient-trial matching scheme

et al., 2024a). Furthermore, their limited interpretability and reliance on high-quality training data pose challenges for clinical validation, trustworthiness and generalization (Singh et al., 2024).

The question that arises now is how this question generation can be used to fulfill the patient-trial matching task. We first discuss the challenges in EHR parsing and the solutions to tackle them as well as for patient pre-screening.

5 Automatic patient-trial matching

This section describes the challenges we face when trying to solve the criteria extraction task from EHR. It also presents methods and tools aimed at achieving the automatic pre-screening of patients. It can be summarized by: if a patient meets all inclusion criteria and none of the exclusion criteria, they are eligible; otherwise, they are not.

A diagram is provided to illustrate a possible workflow pipeline in Figure 2, which distinguishes between the extraction process of criteria from EHR medical reports. It uses a NER model similar to that described in Section 4 and the mapping process between the extracted eligibility criteria and the patients' pathological and molecular characteristics.

5.1 Challenges in extracting inclusion and exclusion criteria

The extraction eligibility criteria presents challenges due to linguistic variability, both diachronic (evolution of word meaning through time) and synchronic (different usage of the same word in the same time period). Additionally, clinical reports may be written in different languages, introducing complexity when aiming to extract features.

One challenge is terminological variability (Cohen and Elhadad, 2013), including synonyms and technical jargonFurthermore, complex grammatical structures (Lonsdale et al., 2008) and the presence of conditional expressions (Ross et al., 2010) demand advanced syntactic analysis techniques to avoid misinterpretations. Added to this is semantic ambiguity (Amosa et al., 2023), since many medical terms are polysemous and require contextual understanding. Expressions like "patients with a history of cardiovascular disease" can be ambiguous if the term "history" is not defined.

Regarding diachronic issues, the evolution of medical terminology over time can make interpreting older texts difficult, as terms may have changed or acquired new meanings. Meanwhile, synchronic problems arise from the coexistence of multiple ways to refer to the same concept in a given period, requiring extraction systems to recognize these variations. Dialectal differences also play a role, like in the USA that possesses a huge hispanic community. To address this, resources like the *Diccionario panhispánico de términos médicos*[3] have been developed to promote a shared medical vocabulary and improve interoperability in Spanish-speaking contexts.

Lastly, in multilingual settings, the translation of eligibility criteria must be not only linguistically accurate but also culturally appropriate. Studies in bilingual regions show that switching languages during medical consultations can enhance doctor-patient communication. Moreover, research has found that the quality of online health information varies by language, reinforcing the need for adaptation when translating these criteria to ensure medical communication across diverse populations (Fefer et al., 2020; Schlicht et al., 2025). Let us discuss some methods to tackle these challenges.

5.2 Manual methods for extracting criteria from patient reports

The extraction of eligibility criteria for clinical trials relies on an analysis of patients' medical records to identify those who meet the requirements. This process involves a review of clinical history to detect evidence matching the established criteria.

A key step is the disambiguation of medical terms (Jonnalagadda et al., 2017), as many concepts may have multiple meanings. To resolve these ambiguities, medical dictionaries and special-

[3] https://dptm.es/

ized ontologies are used, considering the clinical context in which terms appear to ensure interpretation.

Another is information extraction (Adupa et al., 2016), organizing it into categories such as demographic data, medical conditions, treatments and results. The information structuring relies on linguistic norms to interpret relationships between clinical concepts and maintain a representation of the information. Validation of the information (Kraljevic et al., 2021) is essential to ensure its reliability. This involves checking consistency between the extracted data and the patient's medical history and consulting specialists to confirm the findings. This methodological approach, supported by scientific literature, ensures that the extracted data is trustworthy and suitable for determining patient participation. We present now some methods that aim to reproduce this criteria extraction process.

5.3 Automation of criteria extraction from medical reports

5.3.1 Methods and models

The automation of eligibility criteria extraction for clinical trials from EHRs has advanced thanks to a range of methodologies and models:

- **Apache cTAKES** (Savova et al., 2010): an open-source NLP system based on OpenNLP[4] and designed to extract clinical information from unstructured EHR text.
- **MedCAT (Medical Concept Annotation Toolkit)** (Kraljevic et al., 2021): based on the spaCy framework, MedCAT combines rule-based and machine learning methods to process clinical texts.
- **Linguamatics I2E** (Rath et al., 2023): used to extract information from clinical and biomedical texts based on rules.
- **GPT-4** (Datta et al., 2024): a large-scale language model developed by OpenAI. GPT-4 has shown advanced capabilities in understanding and generating text.
- **GatorTron** (Yang et al., 2022): a large-scale clinical language model trained on over 90 billion words to process unstructured EHRs.
- **Text Nailing** (Kartoun, 2017): a hybrid method that combines human input with NLP techniques to extract structured information from unstructured documents.
- **DICE** (Ma et al., 2023): the DICE model automates the extraction of relevant clinical events from medical records, facilitating patient pre-screening for clinical trials. DICE employs a conditional generation strategy and contrastive learning to define medical mentions.

[4] https://opennlp.apache.org/

5.3.2 Datasets

The automation of eligibility criteria extraction for clinical trials relies on datasets used to train and evaluate NLP pipelines:

- **ClinicalTrials.gov** (Stergiopoulos et al., 2019): a public database providing information on registered clinical trials, including eligibility criteria, interventions, and outcomes. It is useful for developing models based on real-world clinical trial data.
- **MIMIC-III** (Johnson et al., 2016): a publicly available clinical dataset containing information from patients admitted to intensive care units. It includes several detailed data about patients' condition.
- **FAERS (FDA Adverse Event Reporting System)** (Polepalli Ramesh et al., 2014): a database containing reports of adverse events and drug reactions. It is valuable for adverse event detection and pharmacovigilance.
- **AIDS Clinical Trials Group Study 175** (Fu, 2024): contains health statistics and categorical information for patients diagnosed with AIDS. Its primary task is to predict whether a patient died within a given time frame.
- **i2b2 Clinical Notes Dataset** (Eguia et al., 2024): a collection of annotated clinical notes used in clinical information extraction challenges.
- **DermatES** (Torre et al., 2024): a dataset consisting of dermatological clinical reports in Spanish, collected from various healthcare centers in Spain.

These datasets are essential in advancing the automation of eligibility criteria extraction in clinical

trials. They support the development of NLP models capable of transforming unstructured clinical text into structured data, improving the efficiency and accuracy of patient identification for clinical trials. Table 3 in Appendix B summarizes the characteristics of each of them.

5.3.3 Methods benchmark

A comparative table of key models used for the automated extraction of eligibility criteria in clinical trials is presented in Table 4 in Appendix B, including their evaluation metrics and datasets used.

To assess the performance of models in the task of automatic patient pre-screening, a variety of metrics are employed. One study explored the use of InstructGPT (Ouyang et al., 2022) to assist physicians in determining patient eligibility based on summarized medical profiles. The study measured the model's ability to identify eligibility criteria and classify patient suitability, finding that the LLM could reduce physicians' workload by filtering out non-applicable criteria (Hamer et al., 2023).

Furthermore, the QUEST framework was proposed for the human evaluation of LLMs in healthcare applications. This framework encompasses five key principles: *information quality, understanding and reasoning, expression and persona, safety and harm, and trust*. While not focused on patient pre-screening, QUEST offers a structured guide to assess the effectiveness and safety of LLMs in clinical contexts (Tam et al., 2024).

5.4 Patient-trial automatic matching

As we mentioned in Section 1, automatic patient–trial matching entails aligning a patient's medical profile with a trial's eligibility criteria through a dedicated matching system. Recent NLP research has introduced LLMs and inference architectures to tackle this task. (Aguiar et al., 2025) built NLI4PR, a natural language inference approach where patients describe their profiles in everyday language. (Jin et al., 2024) propose TrialGPT, a zero-shot LLM framework that retrieves candidate trials, evaluates criterion-level eligibility, and aggregates results into trial-level scores and (Gupta et al., 2024) created OncoLLM, a language model that outperforms GPT 4 in an empirical evaluation of clinical trial matching using real-world EHRs. As for cohort selection, (Dasgupta et al., 2020) implemented a co-training-based model to select patient cohorts automatically, outperforming fully supervised pipelines. (Wornow et al., 2025) report a zero-shot LLM-based matching system reaching state-of-the-art performance on the n2c2 2018 cohort screening benchmark. Additionally, (Shi et al., 2025) introduce MAKAR, a multi-agent knowledge-augmented reasoning system that integrates domain knowledge. Finally, a multimodal LLM-powered pipeline by (Callies et al., 2025) outperforms manual review times with a minimal drop in accuracy.

Despite strong performance, current systems face unresolved challenges: generalizing across medical specialties, handling complex multi-criteria logic, scaling to large trial sets and ensuring explainable, auditable decisions. Research must prioritize multi-site validation, transparent failure analysis, and user-centered design in deployment environments. We detailed them in the next section.

6 Ethical, technical and ecological limitations

6.1 Technical challenges

Clinical notes are unstructured, with inconsistent formats, sections and notation standards (Tang et al., 2019). Semantic interoperability remains a challenge: different systems use distinct terminologies and data models, complicating data integration (Torab-Miandoab et al., 2023; Ademola et al., 2024). Moreover, medical narratives contain domain-specific abbreviations, typographical errors, and hedge language. For example, BERT-based clinical text models still struggle with resolving negations ("no evidence of metastasis"), temporal cues ("6 months prior") and jargon-heavy constructions ("SOB on exertion") (Liu et al., 2023).

LLMs trained on general medical corpora require fine-tuning to handle specialty domains such as oncology or mental health triage. Studies show that domain-specific pretraining improves performance but demands substantial annotated in-domain data (Kerner, 2024). Finally, EHR texts include typos, shorthand, markup remnants and section headers, reducing text mining accuracy.

6.2 Ethical issues

The use of LLMs for clinical trial pre-screening raises ethical and legal concerns in the domain of clinical NLP (Šuster et al., 2017). One issue is privacy and data access: efforts highlight how access to large-scale clinical text is restricted by data protection policies, necessitating the use of privacy-preserving NLP approaches and secure data han-

dling to comply with regulations such as GDPR and HIPAA.

Another concern is algorithmic bias and health equity. Clinical embeddings have been shown to carry gender and race-based biases which can perpetuate disparities in care (Shah et al., 2020; Sogancioglu et al., 2022). Further, the clinical NLP field is moving toward greater interpretability and explainability (Huang et al., 2024). Models that lack traceability present risks if deployed in patient-facing settings, as chatbots have demonstrated concerns over reproducing racially biased output, raising questions about model validation and developer liability[5].

6.3 Ecological impact

LLMs deployed for clinical-trial pre-screening consume substantial computational resources during both training and inference, raising environmental concerns (Hershcovich et al., 2022). The NLP community has begun to acknowledge this impact. (Strubell et al., 2019; Wang et al., 2023) highlight that large transformer models require immense energy to train and deploy.

More recent work benchmarks inference energy within NLP tasks (Morrison et al., 2025). (Liu et al., 2025) introduce "functional unit" analysis for prompt-based LLM deployment, revealing that even a single query to GPT-style models may emit more carbon dioxide than certain web searches. Lightweight transformer variants targeted to reduce model complexity and decoding cost without performance loss, have been demonstrated effective in clinical text-processing tasks, offering a direction toward sustainable deployments (Bannour et al., 2021).

7 Conclusion and future works

In this work, we successfully connected two close fields of NLP for bio-informatics: questionnaire generation for clinical trials and automatic patient pre-screening. By searching literature of both fields, we were able to conduct a narrative review that outlines recent progress for these two tasks and current limitations as well as the links between them. We showed that automation of patient pre-screening for clinical trials through NLP has shown significant progress. The combination of LLMs and specialized tools, such as MedSpaCy, has improved the extraction of eligibility criteria from clinical trial documents and EHRs. We also highlighted that the generation of clinical questionnaires based on extracted trial criteria has demonstrated considerable potential to streamline the patient screening process, enhancing accuracy and reducing recruitment time.

However, our review also reveals several challenges. The heterogeneity of clinical data, the lack of standardization in EHRs, and the limitations of NLP tools in handling unstructured data are major barriers to large-scale implementation. For future work, several areas need to be addressed: data interoperability across platforms, improving the accuracy of NLP models in clinical contexts and increasing the volume of annotated clinical data. Implementing hybrid methods that combine supervised and unsupervised learning could enhance the extraction of clinical events and the classification of eligibility criteria in real-world environments. Likewise, the development of continuous feedback systems (integrating real-time data from clinical practice) is emerging as a direction to optimize model accuracy and adaptability.

Another line of future work concerns issues related to the generation of questions. In particular, the standardization of generated questions is essential. In the medical field, it is common to use acronyms or initialisms, which may appear in either their abbreviated or expanded forms. Standardizing these formats will allow for faster and more reusable patient assessments.

It is worth noting the importance of analyzing the correspondence between eligibility criteria and generated questions. Although the most common scenario is a one-on-one correspondence between criterion and question, there are cases where a single criterion must generate multiple questions. Conversely, multiple criteria may lead to a single question. An example of each case is provided in the Table 5 from Appendix B. Therefore, designing a procedure to determine such correspondences could improve the accuracy of these systems.

The use of current state-of-the-art LLMs in combination with models capable of extracting relevant information from the original text appears to be a promising approach. Supplying this extracted information through the model's prompt provides an additional layer that could help improve the quality of the generated output compared to cases where such information is not included.

[5] https://apnews.com/article/ai-chatbots-racist-medicine-chatgpt-bard-6f2a330086acd0a1f8955ac995bdde4d

Limitations

Although the authors did their best to cover the extensive literature about automatic patient pre-screening, this work presents several limitations, due to it's own goal or to the current state-of-the-art:

- Systematic literature review and text mining: the authors argue that this work is a narrative overview that completes a SLR or an UR. However it lacks the rigor and the statistical work of the former that select all the articles that mention in the title or in the abstract the current subject. Besides, all the linguistics difficulties mentioned over the sections of this work are based on literature and not corroborated by a real experiment, or a text mining pipeline that could demonstrate them.

- Absence of state-of-the-art benchmark: many models, techniques and methods are mentioned all along this article. Despite this extensive list and explanation, there is no real comparison between them, nor a proper benchmark that could justify the strengths and weaknesses of each of them.

- Ethical issues limitation: several ethical, gender-bias, hallucination and explainability issues have been outlined in the work. The authors do not propose any solution other than manual supervision to tackle them, which could either be consider as a future work, or a limitation.

- Ecological impact: most studies mentioned by the authors about the ecological damaged provoked by LLMs are based on carbon footprint. However they do not take into account other issues like production of the material to pre-train the models or the electrical consuming thus the financial cost for any deployment of a LLM-based solution. A more specific work must be made in order to measure the real ecological impact of the automation of patient pre-screening in clinical trials.

- Interoperability and explainable AI: regarding semantic interoperability, some standards are being designed to tackle this challenge. We can cite Fast Healthcare Interoperability Resources (FHIR) that aim to ease the implementation and usage of clinical workflows and the Observational Medical Outcomes Partnership (OMOP) whose purpose is to normalize healthcare data for research and analysis (Wang et al., 2024b; Tabari et al., 2024). The integration of one or both of them into an NLP pipeline could solve the interoperability challenge but we did not focus on it. As for explainable AI and hallucinations, recent approaches leverage grounded generation techniques that condition LLMs on structured inputs (e.g., EHR or knowledge bases), improving factual consistency (Lewis et al., 2020). Additionally, rationale-enhanced models generate explicit natural language explanations to improve transparency and support clinical validation (Rajani et al., 2019). However this work does not get deeper into this exploratory leads.

- Baseline benchmark: the lack of a concise benchmark summary in the article makes it harder to identify which technique is state-of-the-art or under which conditions a method works best for patient-trial matching. However, to our knowledge there is no benchmark that compares the question generation plus patient pre-screening double task.

- AI-assisted literature search: we mention several databases as information source for this narrative review. We do say we use a generative model-based like ChatGPT to conduct the rest of the search and complete it. However, we do not compare it with other commercial generative models such as Gemini (Saab et al., 2024), Claude (Jiang et al., 2025), Le Chat[6] or QWen[7] nor with open-source models such as LLaMa[8] or DeepSeek (DeepSeek-AI et al., 2025) . Moreover, some AI-based tools are built to assist academic research, such as Scispace (Tozuka et al., 2025) or NotebookLM (Jain et al., 2024). One improvement could be comparing the paper search evaluating both the quality of the search (hallucination rate, articles pertinence) and the time spent to gather the literature.

[6] https://chat.mistral.ai/chat

[7] https://chat.qwen.ai/

[8] https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/

References

Adetunji Ademola, Carlisle George, and Glenford Mapp. 2024. Addressing the interoperability of electronic health records: The technical and semantic interoperability, preserving privacy and security framework. *Applied System Innovation*, 7(6).

Abhishek Kalyan Adupa, Ravi Prakash Garg, Jessica Corona-Cox, Sanjiv J Shah, and Siddhartha R Jonnalagadda. 2016. An information extraction approach to prescreen heart failure patients for clinical trials.

Mathilde Aguiar, Pierre Zweigenbaum, and Nona Naderi. 2025. Am I eligible? natural language inference for clinical trial patient recruitment: the patient's point of view. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 243–259, Albuquerque, New Mexico. Association for Computational Linguistics.

Anton M Alekseev, Z Miftahutdinov, Elena Tutubalina, Artem Shelmanov, V Ivanov, V Kokh, Alexander Nesterov, Manvel Avetisian, Andrey Chertok, and S Nikolenko. 2022. Medical crossing: A cross-lingual evaluation of clinical entity linking. *LREC*, pages 4212–4220.

Erick Alphonse, Sophie Aubin, Philippe Bessières, Gilles Bisson, Thierry Hamon, Sandrine Lagarrigue, Adeline Nazarenko, Alain-Pierre Manine, Claire Nédellec, Mohamed Ould Abdel Vetah, Thierry Poibeau, and Davy Weissenbacher. 2006. Event-based information extraction for the biomedical domain: the caderige project.

Temitope Ibrahim Amosa, Lila Iznita Bt Izhar, Patrick Sebastian, Idris B Ismail, Oladimeji Ibrahim, and Shehu Lukman Ayinla. 2023. Clinical errors from acronym use in electronic health record: A review of NLP-based disambiguation techniques. *IEEE Access*, 11:59297–59316.

Emilia Apostolova, Tony Wang, Tim Tschampel, Ioannis Koutroulis, and Tom Velez. 2019. Combining structured and free-text electronic medical record data for real-time clinical decision support. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, Stroudsburg, PA, USA. Association for Computational Linguistics.

E Aramaki, Yoshinobu Kano, Tomoko Ohkuma, and Mizuki Morita. 2016. MedNLPDoc: Japanese shared task for clinical NLP. pages 13–16.

Vahan Arsenyan, Spartak Bughdaryan, Fadi Shaya, Kent Small, and Davit Shahnazaryan. 2023. Large language models for biomedical knowledge graph construction: Information extraction from EMR notes.

Shadnaz Asgari, Fabien Scalzo, and Magdalena Kasprowicz. 2019. Pattern recognition in medical decision support. *Biomed Res. Int.*, 2019:6048748.

Reza Averly and Xia Ning. 2025. Entity decomposition with filtering: A zero-shot clinical named entity recognition framework. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2935–2951, Stroudsburg, PA, USA. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Nesrine Bannour, Sahar Ghannay, Aurélie Névéol, and Anne-Laure Ligozat. 2021. Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 11–21, Virtual. Association for Computational Linguistics.

G O Barnett, N S Justice, M E Somand, J B Adams, B D Waxman, P D Beaman, M S Parent, F R Van Deusen, and J K Greenlie. 1979. COSTAR—A computer-based medical information system for ambulatory care. *Proc. IEEE Inst. Electr. Electron. Eng.*, 67(9):1226–1237.

Elaine Beller, Justin Clark, Guy Tsafnat, Clive Adams, Heinz Diehl, Hans Lund, Mourad Ouzzani, Kristina Thayer, James Thomas, Tari Turner, Jun Xia, Karen Robinson, Paul Glasziou, and founding members of the ICASR group. 2018. Making progress with the automation of systematic reviews: principles of the international collaboration for the automation of systematic reviews (ICASR). *Syst. Rev.*, 7(1):77.

Arth Bohra, Govert Verkes, Artem Harutyunyan, Pascal Weinberger, and Giovanni Campagna. 2023. BYOC: Personalized few-shot classification with co-authored class descriptions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13999–14015, Singapore. Association for Computational Linguistics.

Theodora S Brisimi, Tingting Xu, Taiyao Wang, Wuyang Dai, William G Adams, and Ioannis Ch Paschalidis. 2018. Predicting chronic disease hospitalizations from electronic health records: An interpretable classification approach. *Proc. IEEE Inst. Electr. Electron. Eng.*, 106(4):690–707.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the*

34th International Conference on Neural Information Processing Systems, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Anatole Callies, Quentin Bodinier, Philippe Ravaud, and Kourosh Davarpanah. 2025. Real-world validation of a multimodal llm-powered pipeline for high-accuracy clinical trial patient matching leveraging ehr data. *Preprint*, arXiv:2503.15374.

Leonardo Campillos Llanos, Dhouha Bouamor, Éric Bilinski, Anne-Laure Ligozat, Pierre Zweigenbaum, and Sophie Rosset. 2015. Description of the PatientGenesys dialogue system. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Campos, Sergio Matos, and Jose Luis. 2012. Biomedical named entity recognition: A survey of machine-learning tools. In *Theory and Applications for Advanced Text Mining*. InTech.

Jose Castano, María Laura Gambarte, Hee Joon Park, Maria del Pilar Avila Williams, David Perez, Fernando Campos, Daniel Luna, Sonia Benitez, Hernan Berinsky, and Sofía Zanetti. 2016. A machine learning approach to clinical terms normalization. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Raphael Cohen and Michael Elhadad. 2013. Effect of out of vocabulary terms on inferring eligibility criteria for a retrospective study in hebrew EHR. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing, BioNLP@ACL 2013, Sofia, Bulgaria, August 8, 2013*, pages 116–119. Association for Computational Linguistics.

Tirthankar Dasgupta, Ishani Mondal, Abir Naskar, and Lipika Dey. 2020. Extracting semantic aspects for structured representation of clinical trial eligibility criteria. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 243–248, Online. Association for Computational Linguistics.

Surabhi Datta, Kyeryoung Lee, Hunki Paek, Frank J Manion, Nneka Ofoegbu, Jingcheng Du, Ying Li, Liang-Chin Huang, Jingqi Wang, Bin Lin, Hua Xu, and Xiaoyan Wang. 2024. AutoCriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models. *J. Am. Med. Inform. Assoc.*, 31(2):375–385.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Louise Deléger, Cyril Grouin, and Pierre Zweigenbaum. 2010. Extracting medical information from narrative patient records: the case of medication-related information. *J. Am. Med. Inform. Assoc.*, 17(5):555–558.

Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. 2009. What can natural language processing do for clinical decision support? *J. Biomed. Inform.*, 42(5):760–772.

Yihan Deng, M Stoehr, and K Denecke. 2014. Retrieving attitudes: Sentiment analysis from clinical narratives. pages 12–15.

Santoshi Deshmukh and Utkarsha Pacharaney. 2025. Enhancing healthcare communication: A study on automated speech-to-text conversion and analysis of doctor-patient dialogues for improved clinical documentation and patient care. In *2025 4th International Conference on Sentiment Analysis and Deep Learning (ICSADL)*, pages 229–234. IEEE.

Kanchan Babaji Dhomse. 2024. Dynamic question generation using NER with various feature extraction and NLP techniques. *Adv. Nonlinear Var. Inequal.*, 27(3):639–652.

Hans Eguia, Carlos Luis Sánchez-Bocanegra, Franco Vinciarelli, Fernando Alvarez-Lopez, and Francesc Saigí-Rubió. 2024. Clinical decision support and natural language processing in medicine: Systematic literature review. *J. Med. Internet Res.*, 26:e55315.

Hady Elsahar, Christophe Gravier, and Frederique Laforest. 2018. Zero-shot question generation from knowledge graphs for unseen predicates and entity types. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 218–228, New Orleans, Louisiana. Association for Computational Linguistics.

Peter J Embi, Anil Jain, Jeffrey Clark, and C Martin Harris. 2005. Development of an electronic health record-based clinical trial alert system to enhance recruitment at the point of care. In *AMIA Annual Symposium Proceedings*, volume 2005, page 231.

Hannah Eyre, Alec B Chapman, Kelly S Peterson, Jianlin Shi, Patrick R Alba, Makoto M Jones, Tamára L Box, Scott L DuVall, and Olga V Patterson. 2021. Launching into clinical space with medspacy: a new clinical text processing toolkit in python. *AMIA Annu. Symp. Proc.*, 2021:438–447.

Maia Fefer, Carolyn C Lamb, Abra H Shen, Peter Clardy, Vinayak Muralidhar, Phillip M Devlin, and Edward Christopher Dee. 2020. Multilingual analysis of the quality and readability of online health information on the adverse effects of breast cancer treatments. *JAMA Surg.*, 155(8):781–784.

Dyke Ferber, Lars Hilgers, Isabella C. Wiest, Marie-Elisabeth Leßmann, Jan Clusmann, Peter Neidlinger, Jiefu Zhu, Georg Wölflein, Jacqueline Lammert, Maximilian Tschochohei, Heiko Böhme, Dirk Jäger, Mihaela Aldea, Daniel Truhn, Christiane Höper, and

Jakob Nikolas Kather. 2024. End-to-end clinical trial matching with large language models. *Preprint*, arXiv:2407.13463.

Erlend Frayling, Jake Lever, and Graham McDonald. 2024. Zero-shot and few-shot generation strategies for artificial clinical records.

Xiaowen Fu. 2024. *On the differential privacy of statistical analysis in clinical studies*. Ph.D. thesis.

J J García Adeva, J M Pikatza Atxa, M Ubeda Carrillo, and E Ansuategi Zengotitabengoa. 2014. Automatic text classification to support systematic reviews in medicine. *Expert Syst. Appl.*, 41(4):1498–1508.

Shashi Gupta, Aditya Basu, Mauro Nievas, Jerrin Thomas, Nathan Wolfrath, Adhitya Ramamurthi, Bradley Taylor, Anai N. Kothari, Regina Schwind, Therica M. Miller, Sorena Nadaf-Rahrov, Yanshan Wang, and Hrituraj Singh. 2024. Prism: Patient records interpretation for semantic clinical trial matching system using large language models. *npj Digital Medicine*, 7(1):305.

Danny M den Hamer, Perry Schoor, Tobias B Polak, and Daniel Kapitan. 2023. Improving patient pre-screening for clinical trials: Assisting physicians with large language models.

Daniel Hershcovich, Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. 2022. Towards climate awareness in NLP research. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2480–2494, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Michael Hindelang, Sebastian Sitaru, and Alexander Zink. 2024. Transforming health care through chatbots for medical history-taking and future directions: Comprehensive systematic review. *JMIR Med. Inform.*, 12:e56628.

Lynette Hirschman, Ralph Grishman, and Naomi Sager. 1976. From text to structured information: Automatic processing of medical reports. In *Proceedings of the June 7-10, 1976, national computer conference and exposition on - AFIPS '76*, New York, New York, USA. ACM Press.

Chih-Wei Huang, Bethany C Y Wu, Phung Anh Nguyen, Hsiao-Han Wang, Chih-Chung Kao, Pei-Chen Lee, Annisa Ristya Rahmanti, Jason C Hsu, Hsuan-Chia Yang, and Yu-Chuan Jack Li. 2023. Emotion recognition in doctor-patient interactions from real-world clinical video database: Initial development of artificial empathy. *Comput. Methods Programs Biomed.*, 233(107480):107480.

Guangming Huang, Yingya Li, Shoaib Jameel, Yunfei Long, and Giorgos Papanastasiou. 2024. From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality? *Computational and Structural Biotechnology Journal*, 24:362–373.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv:1904.05342*.

Betina Idnay, Caitlin Dreisbach, Chunhua Weng, and Rebecca Schnall. 2021. A systematic review on natural language processing systems for eligibility prescreening in clinical research. *J. Am. Med. Inform. Assoc.*, 29(1):197–206.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Atlas: Few-shot learning with retrieval augmented language models.

Siddhant Jain, Asheesh Kumar, Trinita Roy, Kartik Shinde, Goutham Vignesh, and Rohan Tondulkar. 2024. Scispace literature review: Harnessing ai for effortless scientific discovery. In *Advances in Information Retrieval*, pages 256–260, Cham. Springer Nature Switzerland.

Qile Jiang, Zhiwei Gao, and George Em Karniadakis. 2025. Deepseek vs. chatgpt vs. claude: A comparative study for scientific computing and scientific machine learning tasks. *Theoretical and Applied Mechanics Letters*, 15(3):100583.

Qiao Jin, Zifeng Wang, Charalampos S Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun, and Zhiyong Lu. 2024. Matching patients to clinical trials with large language models. *Nat. Commun.*, 15(1):9074.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Sci. Data*, 3(1):160035.

Siddhartha R Jonnalagadda, Abhishek K Adupa, Ravi P Garg, Jessica Corona-Cox, and Sanjiv J Shah. 2017. Text mining of the electronic health record: An information extraction approach for automated identification and subphenotyping of HFpEF patients for clinical trials. *J. Cardiovasc. Transl. Res.*, 10(3):313–321.

Corinne Jorgenson, Ali I Ozkes, Jurgen Willems, and Dieter Vanderelst. 2024. Virtual assistants are unlikely to reduce patient non-disclosure. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:659–669.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mahtab Karami and Azin Rahimi. 2019. Semantic web technologies for sharing clinical information in health care systems. *Acta Inform. Med.*, 27(1):4–7.

Uri Kartoun. 2017. Text nailing: An efficient human-in-the-loop text-processing method. *Interactions*, 24(6):44–49.

Tobias Kerner. 2024. Domain-specific pretraining of language models: A comparative study in the medical field. *arXiv preprint arXiv:2407.14076*.

Martijn G Kersloot, Florentien J P van Putten, Ameen Abu-Hanna, Ronald Cornet, and Derk L Arts. 2020. Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies. *J. Biomed. Semantics*, 11(1):14.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, Rebecca Bendayan, Mark P Richardson, Robert Stewart, Anoop D Shah, Wai Keong Wong, Zina Ibrahim, James T Teo, and Richard J B Dobson. 2021. Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit. *Artif. Intell. Med.*, 117(102083):102083.

Craig E Kuziemsky, Dillon Chrimes, Simon Minshall, Michael Mannerow, and Francis Lau. 2024. AI quality standards in health care: Rapid umbrella review. *J. Med. Internet Res.*, 26(1):e54705.

Khai Le-Duc, Khai-Nguyen Nguyen, Long Vo-Dang, and Truong-Son Hy. 2024. Real-time speech summarization for medical conversations. In *Interspeech 2024*, pages 1960–1964, ISCA. ISCA.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yan Lei, Liang Pang, Yuanzhuo Wang, Huawei Shen, and Xueqi Cheng. 2024. Qsnail: A questionnaire dataset for sequential question generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13407–13418, Torino, Italia. ELRA and ICCL.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jennifer Liang, Ching-Huei Tsou, and Ananya Poddar. 2019. A novel system for extractive clinical note summarization using. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 46–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. *Annu Meet Assoc Comput Linguistics*, pages 74–81.

Zefeng Lin, Weidong Chen, Yan Song, and Yongdong Zhang. 2024. Prompting few-shot multi-hop question generation via comprehending type-aware semantics. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3730–3740, Mexico City, Mexico. Association for Computational Linguistics.

Sicen Liu, Xiaolong Wang, Yongshuai Hou, Ge Li, Hui Wang, Hui Xu, Yang Xiang, and Buzhou Tang. 2023. Multimodal data matters: Language model pre-training over structured and unstructured electronic health records. *IEEE Journal of Biomedical and Health Informatics*, 27(1):504–514.

Yifei Liu, Zuo Gan, Zhenghao Gan, Weiye Wang, Chen Chen, Yizhou Shan, Xusheng Chen, Zhenhua Han, Yifei Zhu, Shixuan Sun, and Minyi Guo. 2025. Efficient serving of llm applications with probabilistic demand modeling. *Preprint*, arXiv:2506.14851.

D W Lonsdale, C Tustison, C G Parker, and D W Embley. 2008. Assessing clinical trial eligibility with logic expression queries. *Data Knowl. Eng.*, 66(1):3–17.

Jessica López Espejel. 2019. Automatic summarization of medical conversations, a review. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume III : RECITAL*, pages 487–498, Toulouse, France. ATALA.

P López-Úbeda, Manuel Carlos Díaz Galiano, L A U López, M Martín-Valdivia, T Martín-Noguerol, and A Luna. 2020. Transfer learning applied to text classification in spanish radiological reports. pages 29–32.

Hui Ma, Jian Wang, Hongfei Lin, and Bo Xu. 2022. Graph augmented sequence-to-sequence model for neural question generation. *Appl. Intell.*

Mingyu Derek Ma, Alexander Taylor, Wei Wang, and Nanyun Peng. 2023. DICE: Data-efficient clinical event extraction with generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Diwakar Mahajan, Jennifer J Liang, Ching-Huei Tsou, and Özlem Uzuner. 2023. Overview of the 2022 n2c2 shared task on contextualized medication event extraction in clinical notes. *J. Biomed. Inform.*, 144(104432):104432.

Saeed Mehrabi, Anand Krishnan, Sunghwan Sohn, Alexandra M Roch, Heidi Schmidt, Joe Kesterson, Chris Beesley, Paul Dexter, C Max Schmidt, Hongfang Liu, and Mathew Palakal. 2015. DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx. *J. Biomed. Inform.*, 54:213–219.

S M Meystre, G K Savova, K C Kipper-Schuler, and J F Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb. Med. Inform.*, 17(01):128–144.

Jacob Morrison, Clara Na, Jared Fernandez, Tim Dettmers, Emma Strubell, and Jesse Dodge. 2025. Holistically evaluating the environmental impact of creating language models. *Preprint*, arXiv:2503.05804.

Emmanuel Mutabazi, Jianjun Ni, Guangyi Tang, and Weidong Cao. 2021. A review on medical textual question answering systems based on deep learning approaches. *Appl. Sci. (Basel)*, 11(12):5456.

Marco Naguib, Xavier Tannier, and Aurélie Névéol. 2024. Few-shot clinical entity recognition in english, french and spanish: masked language models outperform generative model prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6829–6852, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yizhao Ni, Jordan Wright, John Perentesis, Todd Lingren, Louise Deleger, Megan Kaiser, Isaac Kohane, and Imre Solti. 2015. Increasing the efficiency of trial-patient matching: automated clinical trial eligibility pre-screening for pediatric oncology patients. *BMC Med. Inform. Decis. Mak.*, 15(1):28.

Regina Ofori-Boateng, Magaly Aceves-Martins, Nirmalie Wiratunga, and Carlos Francisco Moreno-Garcia. 2024. Towards the automation of systematic reviews using natural language processing, machine learning, and deep learning: a comprehensive review. *Artif. Intell. Rev.*, 57(8).

Amy L Olex and Bridget T McInnes. 2021. Review of temporal reasoning in the clinical domain for timeline extraction: Where we are and where we need to be. *J. Biomed. Inform.*, 118(103784):103784.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Antonia Panayi, Katherine Ward, Amir Benhadji-Schaff, A Santiago Ibanez-Lopez, Andrew Xia, and Regina Barzilay. 2023. Evaluation of a prototype machine learning tool to semi-automate data extraction for systematic literature reviews. *Syst. Rev.*, 12(1):187.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Riccardo Pellecchia. 2024. Leveraging ai via speech-to-text and llm integration for improved healthcare decision-making in primary care. Master's thesis, Politecnico di Milano, Milan, Italy. Academic Year 2022/2023.

Claude Pirtle, Harrison Whyte, Edward Goode, Shilo Anders, Christoph Lehmann, and Yaa Kumah-Crystal. 2018. Electronic health record interactions through voice: A review. *Appl. Clin. Inform.*, 09(03):541–552.

Balaji Polepalli Ramesh, Steven M Belknap, Zuofeng Li, Nadya Frid, Dennis P West, and Hong Yu. 2014. Automatically recognizing medication and adverse event information from food and drug administration's adverse event reporting system narratives. *JMIR Med. Inform.*, 2(1):e10.

Yin Poon, John Sie Yuen Lee, Yu Yan Lam, Wing Lam Suen, Elsie Li Chen Ong, and Samuel Kai Wah Chu. 2024. Few-shot question generation for reading comprehension. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 21–27, Bangkok, Thailand. Association for Computational Linguistics.

Sameer Pradhan, Noémie Elhadad, Brett R South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy W Chapman, and Guergana Savova. 2015. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *J. Am. Med. Inform. Assoc.*, 22(1):143–154.

Mojdeh Rahmanian, Seyed Mostafa Fakhrahmad, Zahra Mousavi, and Ashkan Sami. 2023. Towards efficient patient recruitment for clinical trials: Application of a prompt-based learning model.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

K Rajeshkumar, S Dhanasekaran, and V Vasudevan. 2023. Efficient and secure medical big data management system using optimal map-reduce framework and deep learning. *Multimed. Tools Appl.*, 83(16):47111–47138.

N Rath, W Harrison, E Louvet, W Dunlop, and B Liljas. 2023. MSR112 assessing the feasibility of applying natural language processing for systematic literature reviews: A case study in non-small-cell lung cancer. *Value Health*, 26(12):S414.

Ajay Madhavan Ravichandran, Julianna Grune, Nils Feldhus, Aljoscha Burchardt, Roland Roller, and Sebastian Möller. 2024. XAI for better exploitation of text in medical decision support. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 506–513, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jessica Ross, Samson Tu, Simona Carini, and Ida Sim. 2010. Analysis of eligibility criteria complexity in clinical trials. *Summit On Translat. Bioinforma.*, 2010:46–50.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, and 48 others. 2024. Capabilities of gemini models in medicine. *Preprint*, arXiv:2404.18416.

Efsun Sarioglu, Kabir Yadav, and Hyeong-Ah Choi. 2013. Topic modeling based classification of clinical reports. *Proc. Conf. Assoc. Comput. Linguist. Meet.*, 2013:67–73.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *J. Am. Med. Inform. Assoc.*, 17(5):507–513.

Ipek Baris Schlicht, Zhixue Zhao, Burcu Sayin, Lucie Flek, and Paolo Rosso. 2025. Do llms provide consistent answers to health-related questions across languages? In *Advances in Information Retrieval*, pages 314–322, Cham. Springer Nature Switzerland.

Lennart Seitz, Sigrid Bekmeier-Feuerhahn, and Krutika Gohil. 2022. Can we trust a chatbot like a physician? a qualitative study on understanding the emergence of trust toward diagnostic chatbots. *Int. J. Hum. Comput. Stud.*, 165(102848):102848.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Sakib Shahriar, Brady D. Lund, Nishith Reddy Mannuru, Muhammad Arbab Arshad, Kadhim Hayawi, Ravi Varma Kumar Bevara, Aashrith Mannuru, and Laiba Batool. 2024. Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. *Applied Sciences*, 14(17).

Muhammad Talha Sharif and Abdul Rehman. 2025. Systematic literature review on clinical trial eligibility matching.

Hanwen Shi, Jin Zhang, and Kunpeng Zhang. 2025. Enhancing clinical trial patient matching through knowledge augmentation and reasoning with multi-agent. *Preprint*, arXiv:2411.14637.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. Rethinking interpretability in the era of large language models. *Preprint*, arXiv:2402.01761.

Maria Skeppstedt. 2013. Annotating named entities in clinical text by combining pre-annotation and active learning. *Annu Meet Assoc Comput Linguistics*, pages 74–80.

W V Slack, G P Hicks, C E Reed, and L J Van Cura. 1966. A computer-based medical-history system. *N. Engl. J. Med.*, 274(4):194–198.

Gizem Sogancioglu, Fabian Mijsters, Amar van Uden, and Jelle Peperzak. 2022. Gender bias in (non)-contextual clinical word embeddings for stereotypical medical categories. *Preprint*, arXiv:2208.01341.

Stella Stergiopoulos, Kenneth A Getz, and Christine Blazynski. 2019. Evaluating the completeness of ClinicalTrials.Gov. *Ther. Innov. Regul. Sci.*, 53(3):307–317.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Simon Šuster, Stéphan Tulkens, and Walter Daelemans. 2017. A short review of ethical challenges in clinical natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 80–87, Valencia, Spain. Association for Computational Linguistics.

Parinaz Tabari, Gennaro Costagliola, Mattia De Rosa, and Martin Boeker. 2024. State-of-the-art fast healthcare interoperability resources (FHIR)-based data model and structure implementations: Systematic scoping review. *JMIR Med. Inform.*, 12:e58445.

Thomas Yu Chow Tam, Sonish Sivarajkumar, Sumit Kapoor, Alisa V Stolyar, Katelyn Polanska, Karleigh R McCarthy, Hunter Osterhoudt, Xizhi Wu, Shyam Visweswaran, Sunyang Fu, Piyush Mathur, Giovanni E Cacciamani, Cong Sun, Yifan Peng, and Yanshan Wang. 2024. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ Digit. Med.*, 7(1):258.

Matthew Tang, Priyanka Gandhi, Md Ahsanul Kabir, Christopher Zou, Jordyn Blakey, and Xiao Luo. 2019. Progress notes classification and keyword extraction using attention-based deep learning models with bert. *Preprint*, arXiv:1910.05786.

Cui Tao, Harold R Solbrig, and Christopher G Chute. 2011. CNTRO 2.0: A harmonized semantic web ontology for temporal relation inferencing in clinical narratives. *AMIA Summits Transl. Sci. Proc.*, 2011:64–68.

Shubo Tian, Pengfei Yin, Hansi Zhang, Arslan Erdengasileng, Jiang Bian, and Zhe He. 2023. Parsing clinical trial eligibility criteria for cohort query by a multi-input multi-output sequence labeling model. *Proceedings (IEEE Int. Conf. Bioinformatics Biomed.)*, 2023:4426–4430.

Abhisek Tiwari, Manisimha Manthena, Sriparna Saha, Pushpak Bhattacharyya, Minakshi Dhar, and Sarbajeet Tiwari. 2022. Dr. can see: Towards a multi-modal disease diagnosis virtual assistant. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, New York, NY, USA. ACM.

Amir Torab-Miandoab, Taha Samad-Soltani, Ahmadreza Jodati, and Peyman Rezaei-Hachesu. 2023. Interoperability of heterogeneous health information systems: a systematic literature review. *BMC Med. Inform. Decis. Mak.*, 23(1):18.

Leon-Paul Schaub Torre, Pelayo Quiros, and Helena Garcia Mieres. 2024. Automatic detection of diseases in spanish clinical notes combining medical language models and ontologies.

Ryota Tozuka, Hisashi Johno, Akitomo Amakawa, Junichi Sato, Mizuki Muto, Shoichiro Seki, Atsushi Komaba, and Hiroshi Onishi. 2025. Application of notebooklm, a large language model with retrieval-augmented generation, for lung cancer staging. *Japanese Journal of Radiology*, 43(4):706–712.

Ozlem Uzuner. 2008. Second i2b2 workshop on natural language processing challenges for clinical records. *AMIA Annu. Symp. Proc.*, pages 1252–1253.

Ozlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R South. 2012. Evaluating the state of the art in coreference resolution for electronic medical records. *J. Am. Med. Inform. Assoc.*, 19(5):786–791.

Li Wang, Xi Chen, XiangWen Deng, Hao Wen, MingKe You, WeiZhi Liu, Qi Li, and Jian Li. 2024a. Prompt engineering in consistency and reliability with the evidence-based guideline for llms. *npj Digital Medicine*, 7(1):41.

Liwei Wang, Andrew Wen, Sunyang Fu, Xiaoyang Ruan, Ming Huang, Rui Li, Qiuhao Lu, Andrew E Williams, and Hongfang Liu. 2024b. Adoption of the OMOP CDM for cancer research using real-world data: Current status and opportunities.

Xiaorong Wang, Clara Na, Emma Strubell, Sorelle Friedler, and Sasha Luccioni. 2023. Energy and carbon considerations of fine-tuning bert. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, page 9058–9069. Association for Computational Linguistics.

Cliff Wong, Sheng Zhang, Yu Gu, Christine Moung, Jacob Abel, Naoto Usuyama, Roshanthi Weerasinghe, Brian Piening, Tristan Naumann, Carlo Bifulco, and Hoifung Poon. 2023. Scaling clinical trial matching using large language models: A case study in oncology. In *Proceedings of the 8th Machine Learning for Healthcare Conference*, volume 219 of *Proceedings of Machine Learning Research*, pages 846–862. PMLR.

Michael Wornow, Alejandro Lozano, Dev Dash, Jenelle Jindal, Kenneth W Mahaffey, and Nigam H Shah. 2025. Zero-shot clinical trial patient matching with LLMs. *NEJM AI*, 2(1).

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine.

Ying Xiong, Xin Yang, Linjing Liu, Ka-Chun Wong, Qingcai Chen, Yang Xiang, and Buzhou Tang. 2023. EARA: Improving biomedical semantic textual similarity with entity-aligned attention and retrieval augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8760–8771, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hua Xu, Shane P Stenner, Son Doan, Kevin B Johnson, Lemuel R Waitman, and Joshua C Denny. 2010. MedEx: a medication information extraction system for clinical narratives. *J. Am. Med. Inform. Assoc.*, 17(1):19–24.

Xi Yang, Nima Pour Nejatian, Hoo Chang Shin, Kaleb Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona Flores, Ying Zhang, Tanja Magoc, Christopher Harle, Gloria Lipori, Duane Mitchell, William Hogan, Elizabeth Shenkman, Jiang Bian, and Yonghui Wu. 2022. GatorTron: A large clinical language model to unlock patient information from unstructured electronic health records.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Med. Inform. Decis. Mak.*, 19(Suppl 3):71.

Chi Yuan, Patrick B Ryan, Casey Ta, Yixuan Guo, Ziran Li, Jill Hardin, Rupa Makadia, Peng Jin, Ning Shang, Tian Kang, and Chunhua Weng. 2019. Criteria2Query: a natural language interface to clinical databases for cohort definition. *J. Am. Med. Inform. Assoc.*, 26(4):294–305.

Deborah A Zarin and Alla Keselman. 2007. Registering a clinical trial in ClinicalTrials.gov. *Chest*, 131(3):909–912.

Hongwei Zeng, Bifan Wei, Jun Liu, and Weiping Fu. 2023. Synthesize, prompt and transfer: Zero-shot conversational question generation with pre-trained language model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8989–9010,

Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT.

Xingyao Zhang, Cao Xiao, Lucas M Glass, and Jimeng Sun. 2020. DeepEnroll: Patient-trial matching with deep embedding and entailment prediction. In *Proceedings of The Web Conference 2020*, New York, NY, USA. ACM.

Yaoyun Zhang, Min Jiang, Jingqi Wang, and Hua Xu. 2016. Semantic role labeling of clinical text: Comparing syntactic parsers and features. *AMIA Annu. Symp. Proc.*, 2016:1283–1292.

Angelo Ziletti and Leonardo D'Ambrosi. 2024. Retrieval augmented text-to-SQL generation for epidemiological question answering using electronic health records.

A ChatGPT prompts

In this section, the following list gives 3 examples of prompts that were used to instruct ChatGPT in searching for references.

1. When it was about the related works we prompted: *any other related works or umbrellas review to enrich this section ?*
   (a) Reply: *Ai Quality Standards (2025) conducted a rapid umbrella review examining quality benchmarks across healthcare AI, revealing inconsistent reporting and evaluation practices in systematic reviews ChatGPT in Consumer Health (2025) reviewed existing systematic reviews on ChatGPT in healthcare, emphasizing strengths in accessibility but noting weak evaluation frameworks and scarce primary clinical studies . AI in Vaccine Research (2025) deployed an umbrella review to consolidate findings on AI applications in antigen discovery, trial design, and supply chains, underscoring a consistent over-reliance on machine learning lagging behind deep learning and system integration.*
   (b) Reply: *In contrast, existing reviews in patient pre-screening (see Table 1) focus narrowly on specific tasks: Meystre et al. (2008) and Uzuner (2008) defined foundational methods in EHR extraction; cTAKES (Savova et al., 2010) delivered a modular pipeline;Ni et al. (2015) piloted rule-based recruitment tools; Beller et al. (2018) and Ofori-Boateng et al. (2021) broadened scope to systematic reviews; Idnay et al. (2022) systematically reviewed NLP systems for eligibility screening; Panayi et al. (2023) explored semi-automated data extraction; Jin et al. (2023), Hamer et al. (2023), Rahmanian et al. (2024), and Sharif and Rehman (2025) introduced and evaluated LLM-based matching and prompt-based learning systems.*
2. About enhancing the list about criteria extraction: *any NLP ref I haven't mentioned?*
   - ***Criteria2Query (2019)*** *– Introduced a hybrid NLP pipeline to parse free-text eligibility criteria into structured, executable cohort queries. The system achieved F1 scores of approximately 0.80 for entity and relation extraction and supports interactive, human-in-the-loop query refinement.*
   - ***DeepEnroll (2020)*** *– Proposed a cross-modal deep learning model that embeds eligibility criteria and longitudinal EHR data into a shared latent space, incorporating a numerical entailment module — outperforming baseline matching systems by up to 12.4% F1.*
   - ***Criteria2Query 3.0 / TrialGPT (2024)*** *– Upgraded the original Criteria2Query pipeline with GPT-3.5/GPT-4 to enhance concept extraction, SQL generation, and reasoning explanation—marked by a significant leap in usability and accuracy.*
   - ***PRISM (2024)*** *– A deep neural network system that directly embeds patient records and trial criteria to compute match scores, with special handling for unstructured oncology data and real-world note streams.*
   - ***Entity-Extraction + Neural Re-ranking (2023)*** *– A two-stage pipeline combining NER-enabled lexical retrieval with Transformer-based re-rankers, boosting precision by approximately 15% compared to traditional retrieval approaches.*
   - ***LLM Distillation for Matching (2023)*** *– Benchmarked GPT-3.5/4 and open-*

source LLaMA models for patient–trial matching. Through fine-tuning, it showed that lightweight LLMs can rival proprietary models in accuracy.

- ***Zero-Shot LLM Matching (2024)*** *– Demonstrated how zero-shot prompting with GPT variants achieves state-of-the-art results on n2c2 cohort selection benchmarks, also generating coherent, clinician-readable explanations.*

3. *I need a last reference list about patient-trial automatic matching, it must really be patient EHR, clinical trial and a matching system. it must contain new references from NLP conferences, performance limitation and metrics.*

   (a) Reply 1:

   - *Savova et al. (2010) – cTAKES: Clinical Text Analysis and Knowledge Extraction System*
   - *Kraljevic et al. (2021) – MedCAT: Medical Concept Annotation Toolkit*
   - *Neumann et al. (2021) – MedSpacy: A Clinical Text Processing Toolkit*
   - *Luo et al. (2022) – Challenges in NLP for Clinical Trials*
   - *Lee et al. (2020) – BioBERT: Pre-trained Biomedical Language Representation Model for Biomedical Text Mining*
   - *Liu et al. (2021) – Zero-shot Biomedical Entity Recognition with GPT-style Models*
   - *Rahmanian et al. (2024) – Autocriteria: Prompt-tuned Criteria Extraction for Clinical Trials*
   - *Mehrabi et al. (2015) – Handling Negation and Uncertainty in Clinical Text*
   - *Weng et al. (2019) – Criteria2Query: Converting Eligibility Criteria into Computable Queries*
   - *Choi et al. (2022) – Error Propagation in Clinical NLP Pipelines*
   - *Johnson et al. (2021) – Temporal Reasoning and Ambiguity in EHR Criteria Extraction*

   (b) Reply 2 after adding "give me the URL of each of them":

   - Savova et al. (2010) – cTAKES
   - Kraljevic et al. (2021) – MedCAT
   - Neumann et al. (2021) – MedSpacy
   - Lee et al. (2020) – BioBERT
   - Liu et al. (2021) – Zero-shot NER with GPT
   - Rahmanian et al. (2024) – Autocriteria
   - Mehrabi et al. (2015) – DeepEN Negation
   - Weng et al. (2019) – Criteria2Query
   - *Choi et al. (2022) – No public URL available (secondary citation)*
   - *Johnson et al. (2021) – No direct URL found*

B Complementary information

In this section, complementary information is given. Table 1 shows the chronological list of all systematic literature and umbrella reviews about NLP tasks and research applied to biomedical topics. Table 2 provides a mapping between EHR features and associated NLP/AI techniques. Table 3 is a summary of the datasets used for training and evaluating NLP models in eligibility criteria extraction. Table 4 shows performance comparison of models for automated eligibility criteria extraction. Table 5 shows three examples of correspondence between the number of criteria and questions to be generated.

Year	Title
2008	Extracting information from textual documents in EHR (Meystre et al., 2008)
2008	The i2b2 2008 workshop (Uzuner, 2008)
2010	Clinical Text Analysis and Knowledge Extraction System (Savova et al., 2010)
2015	Automated clinical trial eligibility prescreening (Ni et al., 2015)
2018	ICASR (Beller et al., 2018)
2021	Automation of systematic reviews (Ofori-Boateng et al., 2024)
2021	Systematic review of NLP systems for eligibility pre-screening (Idnay et al., 2021)
2023	Evaluation of a prototype ML tool for literature reviews (Panayi et al., 2023)
2023	Improving patient pre-screening with LLMs (Hamer et al., 2023)
2023	Prompt-based learning for efficient clinical trial recruitment (Rahmanian et al., 2023)
2024	Matching patients to clinical trials with LLMs (Jin et al., 2024)
2024	AI Quality Standards in HealthCare: rapid umbrella review (Kuziemsky et al., 2024)
2025	Systematic review on clinical trial eligibility matching (Sharif and Rehman, 2025)

Table 1: Chronological summary of key works on NLP for clinical trial recruitment and eligibility screening

EHR Feature	Document validation	Knowledge representation	Content structuring	NLU	Person	Big Data
"Complete" medical information	Named entity recognition (Aramaki et al., 2016)	Semantic web (Karami and Rahimi, 2019)	Categorization (Yao et al., 2019)	Semantic role labeling (Zhang et al., 2016)	Speech to text (Deshmukh and Pacharaney, 2025)	Information retrieval (Joulin et al., 2017)
Avoid loss of information	Information extraction (Alphonse et al., 2006)	Ontologies (Torre et al., 2024)	Automatic formatting (Hirschman et al., 1976)	Normalization (Castano et al., 2016)	Chatbot (Hindelang et al., 2024)	Text mining (Apostolova et al., 2019)
Registered in an understandable way	Template mapping (Kersloot et al., 2020)	Normalization (Pradhan et al., 2015)	Classification (García Adeva et al., 2014)	Sentiment analysis (Deng et al., 2014)	Speech to text (Pirtle et al., 2018)	Semantic similarity (Xiong et al., 2023)
So that the professional reading the record can understand it	Entity linking (Alekseev et al., 2022)	Ontologies (Tao et al., 2011)	Topic modelling (Sarioglu et al., 2013)	Question answering (Mutabazi et al., 2021)	Task oriented dialog system (Campillos Llanos et al., 2015)	Retrieval augmented generation (Xiong et al., 2024)
Minimize time spent	Active learning (Skeppstedt, 2013)		Code switching (López-Úbeda et al., 2020)	Speech to text (Pellecchia, 2024)	Virtual assistant (Tiwari et al., 2022)	
Maintain the patient-doctor relationship	Emotion detection (Huang et al., 2023)	Chatbot (Seitz et al., 2022)	Automatic summarization (López Espejel, 2019)	Sentiment analysis (Demner-Fushman et al., 2009)	Virtual assistant (Jorgenson et al., 2024)	
Information contained in a large number of medical records	Template mapping (Rajeshkumar et al., 2023)	Knowledge graph (Arsenyan et al., 2023)	Keyword extraction (Deléger et al., 2010)	Pattern recognition (Asgari et al., 2019)	Decision support system (Ravichandran et al., 2024)	Information retrieval / Retrieval augmented generation (Ziletti and D'Ambrosi, 2024)
Accurate and concise recording	Automatic summarization (Liang et al., 2019)	Named entity recognition (Campos et al., 2012)	Classification (Brisimi et al., 2018)	Coreference resolution (Uzuner et al., 2012)	Speech to text (Le-Duc et al., 2024)	Semantic similarity (López Espejel, 2019)

Table 2: Mapping between EHR features and associated NLP/AI techniques.

Dataset	Language	Annotations	Type	Size	Access
ClinicalTrials.gov[9]	English	No	Protocols, trial outcomes	Over 400,000 studies	Public
MIMIC-III[10]	English	Yes	Clinical notes, ICU data	Over 40,000 patients	Requires access
FAERS[11]	English	No	Adverse event reports	Over 10 million reports	Public
AIDS CTG Study 175[12]	English	Yes	Clinical notes	2,139 participants	Public
i2b2[13]	English	Yes	Annotated clinical notes	∼1,500 notes	Requires access
DermatES[14]	Spanish	Yes	Dermatological reports	8,800 dermatology EHRs	Public

Table 3: Summary of datasets used for training and evaluating NLP models in eligibility criteria extraction.

Model	Main Metric	Value	Dataset(s) Used	Reference
AutoCriteria	F1-Score	89.42	ClinicalTrials.gov	(Datta et al., 2024)
DICE	F1-Score	70.46–75.22	MACCROBAT-EE	(Ma et al., 2023)
GPT-4 (C2Q 3.0)	F1-Score	89.1	ClinicalTrials.gov	(Yuan et al., 2019)
GPT-4 (C2Q 3.0)	F1-Score	64.8–72.5	Oncology clinical trials	(Yuan et al., 2019)

Table 4: Comparison of models in automated extraction of eligibility criteria for clinical trials.

1 criterion	**1 question**
Predicted life expectancy > 3 months	Is the patient's predicted life expectancy > 3 months?
1 criterion	**4 questions**
Previous or current malignancies of other histologies within the last 2 years, except for in situ carcinoma of the cervix, and adequately treated basal cell or squamous cell carcinoma of the skin	Has the patient had any other previous or current malignancy?
	Is that malignancy an in situ carcinoma of the cervix?
	Is that malignancy an adequately treated non-melanoma skin cancer?
	Has there been evidence of that malignancy within the last 2 years?
4 criteria	**1 question**
Three cohorts of subjects are defined in this prospective multicenter study:	Has the patient been diagnosed with: (1) triple negative breast cancer? (2) HER 2 positive breast cancer? (3) Non-small cell lung cancer?
Cohort 1: Triple-negative breast cancer (TNBC)	
Cohort 2: HER 2 positive breast cancer (HER2+ BC)	
Cohort 3: Non-small cell lung cancer (NSCLC)	

Table 5: Examples of correspondence between the number of criteria and questions to be generated. The criteria are extracted from the clinical trial NCT05278975. The questions are generated by the authors of this article