Neural language models for clinical trial eligibility criteria

Klaudia Kantor

Under the supervision of Professor Mikolaj Morzy

DOCTORAL THESIS
Poznan University of Technology

August 2024

Except where otherwise indicated, this thesis is my own original work.

Klaudia Kantor
29 August 2024

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my thesis advisor, Prof. Mikołaj Morzy, for his continuous support, guidance, and encouragements thorough this research. His constructive feedback, ideas and suggestions have been invaluable in shaping this work.

I would also like to thank Dr. Marek Dąbrowski, whose experience and knowledge were a great source of inspiration and learning. His assistance in structuring my unstructured thoughts has been particularly important.

I am deeply grateful to my colleagues at Roche, whose support made this research possible. Special thanks to Michael DAndrea and Marek Dąbrowski for their engagement during the hackathon and other key phases of the project. I also appreciate Roche for granting me the opportunity to undertake these studies, providing the necessary time and support without which this research would not have been feasible.

A big thank you to the domain experts, who provided their insights and supported the evaluation process. Their input was critical in shaping the final solution.

My heartfelt thanks go to the members of the CTDN initiative, led by Brian Goodman, which served as a supportive community for my research and a source of inspiration.

I would also like to acknowledge the support of my friends and family, particularly my husband Mateusz Kantor, whose patience and understanding have been a constant source of strength. Their belief in me kept me motivated through the challenging moments.

Finally, I extend my gratitude to the Polish Ministry of Education and Science and the Faculty of Computing and Telecommunications of Poznan University of Technology for the financial support provided during this research. Without their assistance, this project would not have been possible.

To everyone who contributed to this thesis in one way or another, your support has been deeply appreciated. Thank you.

Abstract

Clinical trials are crucial for advancing medical knowledge and introducing new therapies to the market. However, the recruitment of patients is a significant challenge, often leading to delays and trial terminations. A major contributor to this issue is the manual process of patient screening, which is highly inefficient due to the stringent and complex nature of eligibility criteria. These criteria are written in free text, making them difficult to process with existing algorithms. Currently, there is no tool capable of automating this process effectively. This thesis explores the potential of neural language models, particularly recent advances in natural language processing (NLP), to transform clinical trial eligibility criteria into a machine-readable format, and consequently automate and optimize the recruitment process.

The study begins by investigating the impact of trial design on trial execution performance, followed by a comprehensive exploration of available biomedical NLP resources. A critical scoping review reveals that the majority of approaches to parsing eligibility criteria rely on outdated and non-generalizable methods such as regular expressions, rule-based systems, and early word embedding techniques like word2vec and FastText. Surprisingly, only a small portion of recent studies have employed pretrained transformer models, with a few leveraging generative AI, despite its rapid advancements.

The experimental phase of this research involves evaluating text embeddings in the biomedical domain for semantic textual similarity (STS) tasks and comparing the performance of BERT models with GPT-4 in named entity recognition (NER) from eligibility criteria. This work is conducted within the context of an industrial PhD program, where scientific findings are applied in practice. The implementation phase includes organizing a hackathon at Roche to assess the potential of large language models (LLMs) in parsing eligibility criteria and developing a Proof of Concept (PoC) tool using GPT-4o, along with advanced prompt engineering. The tool was developed in collaboration with domain experts and evaluated in a carefully designed experiment involving their participation. The results emphasize its practical utility in the pharmaceutical industry.

The findings indicate that the problem of criteria processing is complex and extends beyond basic STS or NER tasks. Effective parsing requires the extraction of more nuanced information, including relationships between entities and advanced reasoning—a task where generative LLMs demonstrate significant potential, particularly in the context of limited annotated data. Additionally, the support of domain experts is crucial for the successful implementation of such tools. The research also identifies a gap in benchmark datasets for evaluating models that process eligibility criteria and a lack of reliable methods for evaluating LLMs in information retrieval tasks, highlighting an area for future exploration.

The contributions of this thesis include the most updated and extensive scoping review in the field, practical requirements for implementing NLP tools in the pharmaceutical industry, a comparison of BERT models and LLMs in biomedical applications, and the design and development of a PoC tool guided by domain expert input and advanced prompt engineering techniques. The conclusions underscore the need for further refinement of the PoC for broader applicability across different therapeutic areas and criteria types, more rigorous evaluation under diverse conditions, and enhanced collaboration with domain experts. Despite these challenges, the research demonstrates that integrating NLP with domain expertise can significantly improve patient recruitment processes in clinical trials.

Keywords: generative AI, large language models, transformers, named entity recognition, information retrieval, eligibility criteria, clinical trials

Streszczenie

Badania kliniczne są kluczowe dla rozwoju medycyny i wprowadzania nowych metod leczenia na rynek. Rekrutacja pacjentów do badań klinicznych stanowi istotne wyzwanie, które często powoduje opóźnienia lub nawet przerwanie badań. Jedną z głównych przyczyn tego problemu jest ręczny proces selekcji pacjentów, który jest wysoce nieefektywny ze względu na bardzo rygorystyczne i złożone kryteria kwalifikacji. Kryteria te są pisane hermetycznym językiem medycznym, co utrudnia ich bezpośrednie przetwarzanie przez istniejące algorytmy. Obecnie nie istnieje żadne narzędzie, które skutecznie zautomatyzowałoby proces porównywania danych pacjentów z tymi kryteriami. Niniejsza praca doktorska bada potencjał modeli językowych opartych na sieciach neuronowych w przekształcaniu kryteriów kwalifikacji do badań klinicznych w ustrukturyzowaną formę czytelną dla algorytmów, w konsekwencji badając, czy te modele mogą pomóc w automatyzacji i optymalizacji procesu rekrutacji.

Rozprawa rozpoczyna się od analizy wpływu elementów projektu badania klinicznego na jego przebieg, a następnie opisuje dostępne zasoby do przetwarzania tekstu w dziedzinie biomedycyny. Wnikliwy przegląd literatury ujawnia, że większość badań wokół przetwarzania kryteriów kwalifikacji używa przestarzałych metod, takich jak wyrażenia regularne, algorytmy oparte na regułach oraz stare techniki wektoryzacji tekstów, takie jak word2vec i FastText. Co ciekawe, tylko niewielka część publikacji wykorzystuje modele oparte o architekturę Transformer, a zaledwie kilka bazuje na generatywnych modelach językowych, które bardzo szybko rozwijają się w ostatnich latach.

Faza eksperymentalna rozprawy obejmuje ocenę metod wektoryzacji tekstów biomedycznych w zadaniu podobieństwa semantycznegoc (STS) oraz porównanie modeli opartych o BERT i modelu GPT-4 w zadaniu rozpoznawania nazwanych jednostek (NER) w tekstach kryteriów kwalifikacji. Ze względu na to, że praca ta jest doktoratem wdrożeniowym, jej wyniki naukowe są wdrażane w przemyśle. Faza implementacyjna obejmuje organizację hackathonu w firmie Roche w celu oceny potencjału dużych modeli językowych (LLM) w przetwarzaniu kryteriów kwalifikacji oraz opracowanie prototypu narzędzia (PoC) parsującego te kryteria. Narzędzie to wykorzystuje model GPT-4o oraz zaawansowane technki inżynierii zachęt, uwzględniające wyspecjalizowaną wiedzę domenową. Narzędzie zostało zbudowane we współpracy z ekspertami dziedzinowymi, a jego efektywność oceniono w starannie zaprojektowanym eksperymencie ewaluacyjnym z udziałem tychże ekspertów. Wyniki ewaluacji podkreślają jego praktyczną użyteczność w branży farmaceutycznej.

Rozprawa dowiodła, że problem przetwarzania kryteriów kwalifikacji jest bardzo złożony i wykracza poza podstawowe zadania NLP, takie jak STS, czy NER. Przetworzenie kryteriów do formatu użytecznego w rekrutacji pacjentów wymaga wydoby-

cia bardziej złożonych informacji, na przykład relacji między kryteriami, i zaawansowanego rozumowania. Jest to zadanie, w którym generatywne duże modele językowe osiągają wysokie wyniki, szczególnie w przypadku braku dużych zbiorów uczących. Ponadto wsparcie ekspertów domenowych jest kluczowe dla skutecznego rozwoju i wdrożenia takich narzędzi. W rozprawie rozpoznano deficyt dostępnych referencyjnych zbiorów danych, które mogłyby być użyte do oceny modeli procesujących kryteria kwalifikacji, a także brak rzetelnych metod mierzenia dokładności modeli generatywnych w zadaniach ekstrakcji informacji z tekstu. Zidentyfikowano tutaj potancjalny obszar dalszych badań naukowych.

Kontrybucja tej pracy obejmuje najbardziej aktualny i rozległy przegląd literaturowy w tej dziedzinie, zebranie praktycznych wymagań dotyczących wdrożenia narzędzia NLP w przemyśle farmaceutycznym, porównanie modeli BERT i LLM w dziedzinie biomedycznej oraz zaprojektowanie i zbudowanie narzędzia PoC, które integruje wiedzę domenową z najnowszą technologią. Dalsze badania w tym obszarze powinny skupić się na poszerzeniu zakresu zastosowania opracowanego narzędzia o nowe obszary terapeutyczne i typy kryteriów, opracowaniu bardziej rygorystycznego sposobu ewaluacji tego narzędzia oraz włączenie większej liczby ekspertów domenowych. Pomimo ograniczeń, wyniki rozprawy pokazują, że integracja NLP z wiedzą ekspercką może znacząco poprawić efektywność rekrutacji pacjentów do badań klinicznych.

Słowa kluczowe: generatywna sztuczna inteligencja, duże modele językowe, transformer, BERT, GPT, NER, ekstrakcja informacji, kryteria kwalifikacji, badania kliniczne

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Clinical trials are essential for the advancement of medicine, because they validate the efficacy and safety of new treatments before they reach the market. They also play a crucial role for patients, offering early access to innovative therapies that may be their only treatment options, sometimes serving as critical, life-saving interventions. A fundamental component of those trials are the eligibility criteria, which define the characteristics of participants suitable for a study. Efficient parsing and interpretation of these criteria are vital for successful patient recruitment, which, in turn, influences the success of the trial. Patient recruitment has been identified as the biggest obstacle to the successful completion of trials, causing delays and even terminations, as finding eligible patients who meet the inclusion and exclusion criteria of clinical trials is challenging and time-consuming.

In recent years, advancements in Natural Language Processing (NLP) have created new opportunities for automating the parsing of eligibility criteria. Neural language models like BERT and its biomedical variants, and particularly large language models (LLMs) like GPT-3 and GPT-4, have demonstrated significant performance in human language understanding and generation. This thesis explores the applications of these advanced models in eligibility criteria parsing, aiming to increase the efficacy of patient recruitment for clinical trials. Current manual and semi-manual methods are very labour-intensive and not scalable. Developing an automated recruitment system requires a machine-readable format of eligibility criteria. Therefore, this research seeks to develop a robust and reliable AI-powered system that can assist in structuring eligibility criteria, which can be further utilized in automated patient eligibility screening.

This problem is both interesting and significant due to the high complexity of the eligibility criteria expression, which makes it a non-trivial task to structure the criteria. It requires not only the knowledge of recent advances in NLP, but also medical expertise to design the solution and structure the criteria correctly. This subject combines the disciplines of artificial intelligence, natural language processing, medical informatics, and medical sciences. Developing a solution that can handle this complexity is a challenging task with significant influence in the biomedical domain.

This subject is critically important for Roche, a pharmaceutical company that launches over 50 new trials each year, seeking for participants in numerous countries.

Currently there are 136[1] Roche clinical trials recruiting patients in over 30 countries. All of them have planned enrollment numbers that need to be met in order to deliver significant evidence on treatment efficacy and safety. The company relies on clinical representatives to recruit patients for trials, but this process is highly inefficient. Each day of delay incurs significant costs. Numerous trials at Roche have ended in a failure due to an insufficient number of patients, resulting in wasted money and resources. Efficiently parsing eligibility criteria can lead to a more automated recruitment process, reduce trial delays, and accelerate the introduction of new therapies to the market.

The research hypothesis guiding this thesis is that **neural language models can significantly enhance the efficacy of parsing clinical trial eligibility criteria, outperforming traditional methods and consequently enhancing patient recruitment in clinical trials.**

The primary objective is to investigate the potential of neural language models to effectively handle the complexity and variability of eligibility criteria in order to produce a robust and generalize parsing solution that can be integrated into clinical trial management systems. This research aims to:

- analyze the complexity of eligibility criteria and challenges of parsing them,
- review previous works related to this context,
- evaluate the performance of various NLP models and techniques in different NLP tasks related to eligibility criteria, including traditional methods and cutting-edge LLMs,
- gather practical requirements for a parsing tool implementation at Roche,
- develop and test a prototype tool to automate the extraction and structuring of eligibility criteria,
- assess the business impact of the proposed tool.

The key elements of the new proposal include:

- **Utilizing large language models:** applying state-of-the-art neural language models such as GPT-4 to parse the eligibility criteria;
- **Few-shot prompt engineering:** utilizing in-context learning capabilities of LLMs to improve parsing accuracy with limited annotated data;
- **Chain-of-thought approach:** leveraging chain-of-thought prompting to guide the model through the reasoning process;
- **Explainability:** ensuring that the model's decisions are transparent and understandable by providing the source criterion for each of the parsed outputs;

[1] Counted based on ClinicalTrials.gov records accessed on July 13, 2024.

- **Domain expertise:** engaging domain experts in the development process and incorporating their insights to refine and evaluate the solution.

The original contributions of this thesis are:

- **Scoping review:** the most comprehensive and up-to-date scoping review on NLP and ML solutions for parsing eligibility criteria;
- **NLP resource overview:** a broad overview of available resources for biomedical NLP;
- **Sentence embeddings evaluation:** the most extensive evaluation of various sentence embeddings for biomedical domain;
- **Comprehensive evaluation of GPT vs. BERT:** conducting extensive experiments comparing LLMs with BERT-based models in low-resource setting;
- **Innovative prompt engineering techniques**: introducing and evaluating complex prompts, including few-shot templates, chain of thought, and domain knowledge instructions;
- **Practical insights:** providing practical guidelines and insights on the development and implementation of such a tool in a real-world pharmaceutical context;
- **Development of a prototype tool:** creating a working prototype that demonstrates the feasibility of using LLMs in eligibility criteria parsing.

The thesis is structured as follows:

- **Chapter 2: Clinical Trial Eligibility Criteria** - provides an overview of clinical trials and their main pain points, the importance of eligibility criteria , and the challenges related to parsing these criteria. It lays the foundation for understanding the context and significance of the research problem.
- **Chapter 3: Machine Learning Prediction of Clinical Trial Operational Efficiency** - studies the impact of trial features, including eligibility criteria, on trial recruitment and overall operational efficiency.
- **Chapter 4: NLP Resources for Biomedical Domain** - explores various resources available for biomedical NLP, including ontologies, knowledge bases, and pretrained language models. It offers an overview of the tools applicable for this research.
- **Chapter 5: Natural Language Processing in Clinical Trial Eligibility Criteria Parsing** - provides a comprehensive scoping review of studies on NLP applications for clinical trial eligibility criteria. It discusses the findings and identifies research gaps and potential directions of future research.

- **Chapter 6: Biomedical Semantic Textual Similarity** - explores different techniques of sentence embedding in the biomedical domain and evaluates them in a semantic textual similarity task.
- **Chapter 7: Named Entity Recognition in Eligibility Criteria** - examines the ability of the GPT-4-turbo model to recognize named entities in the eligibility criteria and compares this model with the state-of-the-art pretrained Transformer models based on the BERT architecture in a scenario of limited annotated data.
- **Chapter 8: Prompt Engineering Hackathon at Roche** - describes the hackathon conducted to explore prompt engineering feasibility in eligibility criteria parsing and provides the key findings from the event.
- **Chapter 9: Requirements for an Eligibility Criteria Parsing Tool Using LLMs** - defines the requirements for developing and implementing a robust and compliant AI-driven tool for criteria parsing at Roche, covering aspects such as data standardization, MLOps practices, and security.
- **Chapter 10: Implementation of the Eligibility Criteria Parsing Tool: A Proof of Concept** - presents the design and implementation of the parsing tool, detailing the methodology, tool architecture, user interface, and evaluation results.
- **Chapter 11: Conclusions** - summarizes the research findings, discusses the limitations, and suggests future research directions.

The research was conducted as part of the Polish Ministry of Education and Science's Industrial Doctorate Program implemented from 2020 to 2024 (Contract No. DWD/4/24/2020).

Chapter 2

Clinical Trial Eligibility Criteria

This chapter examines the crucial role of eligibility criteria in the context of clinical trials, which are essential for the development and approval of new treatments. It focuses on the critical obstacles and opportunities in optimizing clinical trial execution.

Section 2.1 outlines the key aspects of clinical trials, emphasizing their significance in advancing medical science and offering critical treatment options to patients. Section 2.2 describes the challenges of trial recruitment, a critical factor that affects the success and efficiency of a clinical trial. Section 2.3 explores technical difficulties associated with eligibility criteria parsing.

2.1 Clinical trials

A clinical trial is a research study that tests new drugs, treatments and medical devices on selected groups of volunteers or patients. Its goal is to determine efficacy and safety of the studied product. Clinical trials are crucial to the progress of medicine and serve important roles for both pharmaceutical companies and patients. Each drug must be evaluated in a clinical trial before it can be approved for market introduction, which motivates drug manufacturers to conduct trials as efficiently and swiftly as possible. For patients, clinical trials offer access to innovative therapies before they are publicly available. These trials often provide critical treatment options that may be pivotal for their health and life. Efficient execution of the trial is also very important to patients not involved in the studies, because the sooner the drug is approved, the sooner it is available as an additional - and sometimes the only - treatment option on the market.

Clinical trials consist of four phases, which differ in objectives and patient numbers:

- **Phase 1**: Evaluates the safety of the therapy, testing it on a small group of people (tens), most often healthy volunteers.
- **Phase 2**: Validates the efficacy of the therapy and continues safety monitoring. The treatment is given to a larger group of people (hundreds) with a specified disease.

- **Phase 3**: Assesses whether the therapy is more effective or safer than existing treatments. Large group of people (thousands) is involved in the study.
- **Phase 4**: Evaluates long-term effects after the therapy has been introduced to the market.

An example of clinical trial metadata is illustrated in Figure 2.1. The figure presents a Phase 3 clinical trial on breast cancer, designed for HER2-positive patients.

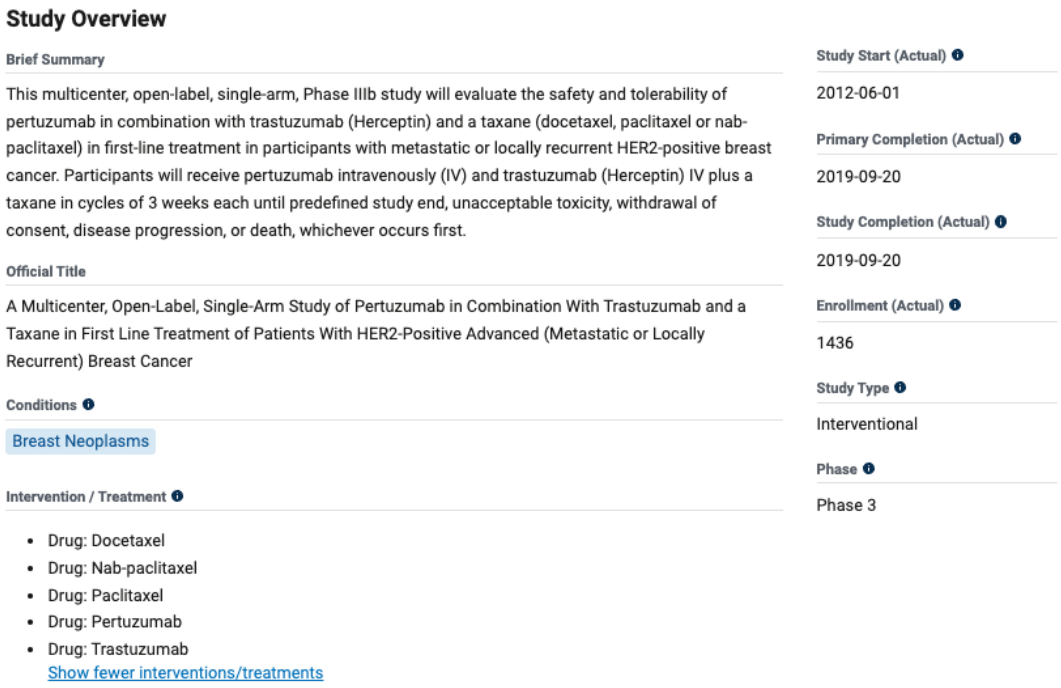


Figure 2.1: An example of a breast cancer trial published on ClinicalTrials.gov.

The complexity of conducting clinical trials results in lengthy execution times, typically around 90 months [1], and substantial financial burdens for pharmaceutical companies. Recent studies indicate that bringing a new drug to market costs an average of $1.3 billion [2] and clinical trials are the most expensive part of that process [3]. The cost of a clinical trial varies by therapeutic area, but estimates suggest that conducting all three phases can range from $50 millions to about $100 millions [1]. Phase 3 is notably the most expensive stage of a trial, as it involves the largest group of patients.

ClinicalTrials.gov is the primary registry for clinical trials, which serves as a database for ongoing and past trials, and as a repository for their results. As of April 2024, the database contains over 377 500 trials conducted in 223 countries, including more than 64 000 active studies. Figure 2.2 shows that the number of newly registered trials has been increasing annually. Among the registered studies, there

are over 36 000 trials that have been terminated or withdrawn, indicating a trial failure. Such outcomes result in significant financial losses for pharmaceutical companies, with costs rising in later study phases due to larger participant groups and the completion of prior phases.

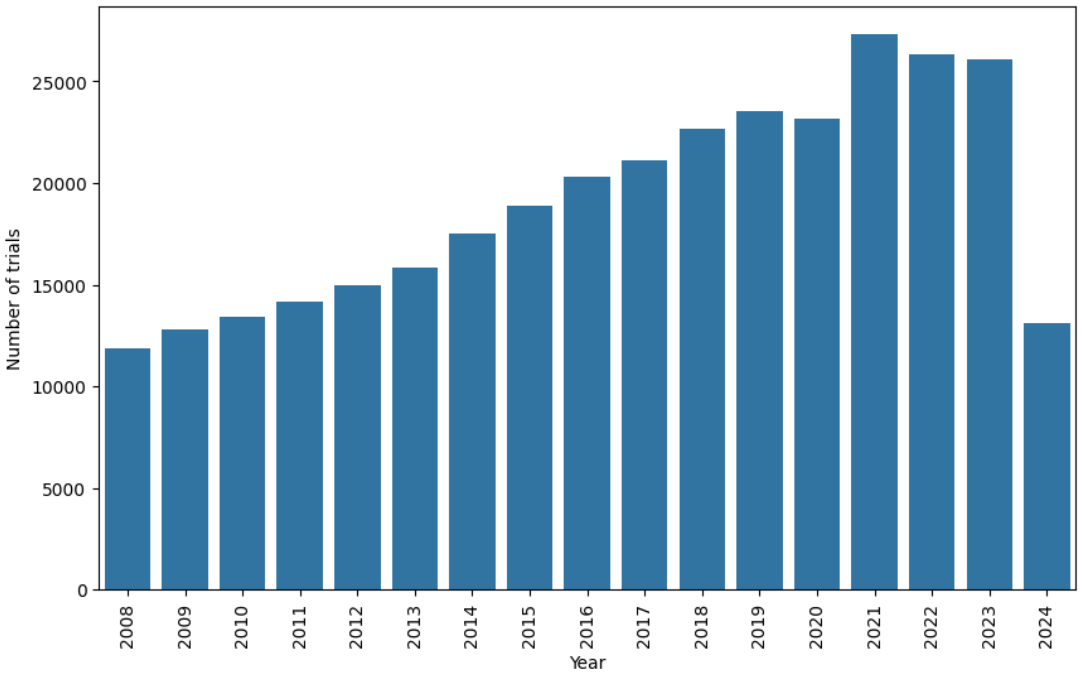


Figure 2.2: Number of newly registered clinical trials per year.

2.2 Patient recruitment

One of the key aspects of clinical trials is patient recruitment. Clinical trials require an appropriate sample size to validate the efficacy and safety of a new treatment. Patients must meet specific requirements, which are defined in eligibility criteria. These criteria specify the characteristics of the trial population, such as required age, gender, medical history, and current health status. They not only ensure the safety of the patients but also guarantee that the data collected during the study provide the answers to the research questions stated at the beginning of the trial. The eligibility criteria are divided into two sections: inclusion criteria and exclusion criteria, which define the conditions that must or must not be met, respectively, to include patients in a trial. Trial screening, which is the process of comparing eligibility criteria against patient records, is currently performed manually by the research personnel. This method is time-consuming, labor-intensive and prone to human errors [4]. It limits the number of patient that can be screened. On average, screening takes 15-20 minutes per patient per trial, but for more complex studies, it can take more than 2 hours to assess patient eligibility [5]. The cost of screening a single patient has been

estimated to range from approximately $130 to $340 [6]. However, not all screened patients are eligible and enrolled in the study. Many candidates are evaluated more than once, for various trials, and often, evaluations are repeated over time due to disease progression. For a pharmaceutical company, this means a waste of resources and time, for patients, it extends the waiting time for potentially life-saving treatment.

Patient recruitment, also known as patient accrual, is a crucial determinant of a trial's success [7]. Low accrual is the most frequent reason for trial failure, with over 40% of terminated trials failing because they did not manage to recruit enough patients [8]. Additionally, up to 85% of all trials experience delays due to low patient enrolment [9], and each day of delay can cost the pharmaceutical company between $600 000 and $8 millions [10]. However, this does not imply a lack of eligible individuals. Studies have shown that even 60% of eligible patients are not considered in the recruitment process [11], often due to a lack of awareness of available studies among both patient and clinician [12]. This not only slows down the advancement of medical science but also deprives patients of the opportunity to participate in a therapy, which sometimes is the only option for their cure. It has been observed that only 3-5% of cancer patients participate in clinical trials, although 20% are eligible [13]. These issues surrounding trial recruitment affect not only the pharmaceutical companies, but also the quality of healthcare and patients' lives.

Eligibility screening is considered a trial bottleneck and one of the largest barriers to the successful completion of clinical trials. Therefore, the National Institutes of Health has prioritized enrolment improvements in recent years [14]. The screening issue has two facets: patients are unaware of trials for which they are eligible, and trial staff is unable to identify eligible patients within the planned timeframe. The volume of available trials and potential participants overwhelms the capacity of clinical research staff and patients. Currently, there is no centralized clinical trial database with an effective search engine. While ClinicalTrials.gov allows searches for trials recruiting patients with a specific condition, such as *"breast cancer"*, it does not enable filtering based on more detailed eligibility criteria, such as *"breast cancer AND ER+ AND no surgery in the past AND HIV allowed"*. A search for *"breast cancer"* on ClinicalTrials.gov returns over 1800 currently recruiting clinical trials.[1] Each trial has its own set of inclusion and exclusion criteria that need to be analysed along with patient data, which is a task demanding medical expertise due to the complex and often implicit language used in these criteria. This information overload results in enrolment difficulties. Automating the patient-trial matching process could significantly benefit the pharmaceutical industry, accelerate recruitment and increase the number of successful trials.

[1] State as of April 2024.

2.3 Complexity of eligibility criteria parsing

There are numerous challenges associated with patient screening automation, one of which is parsing the eligibility criteria. These criteria are included in the trial protocol and written in free text form, which must be converted into a computer-readable format for processing by a patient-trial matching algorithm. As presented in Figure 2.3, the language used in the criteria is complex, including medical terminology, acronyms, numerical values, and units. Additionally, the same measurements might be expressed in different units, and there are many synonyms which refer to the same medical conditions. The syntax of the criteria also introduces significant challenges, including:

- **Negations**: The inclusion criteria section contains exclusions, eg. *"Inclusion Criteria: [...] patients must not have had a blood transfusion within 28 days prior to registration."*[2]

- **Exceptions**: An inclusion or exclusion criterion may not apply in specific situations, eg. *"Exclusion Criteria: [...] Systemic corticosteroids (oral or injectable) within 7 days of first dose of 852A (topical or inhaled steroids are allowed)."*[3]

- **Logical dependencies**: Not all criteria need to be met to include or exclude a patient from the trial, eg. *"Stage I to III breast cancer with surgical resection of the primary tumor that is confirmed to be either: TNBC, irrespective of BRCA status or HR+/HER2- breast cancer."*[4]

- **Temporal restrictions**: Some criteria are only applicable if a condition was met within a defined period, eg. *"Subjects must have ended hormonal replacement therapy (HRT) at least 1 month (30 days) prior to receiving the first dose of randomized therapy."*[5]

- **Subpopulations**: Certain criteria apply only to specific groups of people, eg. *"For participants with bilateral BC, HER2-positive status must be demonstrated in both locations or in a metastatic site."*[6]

Furthermore, not all conditions are explicitly stated in the text. Experienced clinical staff can deduce additional required or excluded conditions based on other terms mentioned. For example, the criterion *"metastatic breast cancer"* implies that the participant must have stage IV breast cancer. Similarly, *"triple-negative breast cancer"* translates to *"no expression of HER2, ER and PR biomarkers"*. Also, the phrase: *"no prior systemic anti-cancer therapy"* excludes patients who have undergone chemotherapy, hormonal therapy, immunotherapy, targeted therapy, or biological therapy.

[2]NCT02595905 trial
[3]NCT00319748 trial
[4]NCT04915755 trial
[5]NCT00073528 trial
[6]NCT03153163 trial

All these factors make parsing criteria into a reliable form a difficult task that requires medical expertise and a sophisticated system capable of handling the complex dependencies mentioned above. Simple regular expressions, rule-based approaches, or basic named entity recognition models may not be sufficient for this task.

Inclusion Criteria:

- Stage I to III breast cancer with surgical resection of the primary tumor that is confirmed to be either: TNBC, irrespective of BRCA status or HR+/HER2- breast cancer with a known and documented deleterious or suspected deleterious tBRCA mutation.
- Estrogen receptor (ER) and/or progesterone receptor (PgR) negativity is defined as immunohistochemistry (IHC) nuclear staining less than (<) 1 percentage (%), or by Allred scoring system where TNBC is defined to be 0 out of 8 or 2 out of 8, or staining in <1 % of cancer cells.
- Completed prior standard therapy for curative intent.
- Participants with HR+ breast cancer must be on a stable regimen of endocrine therapy.
- Detectable ctDNA as measured by central testing.
- An archival tumor tissue specimen of the primary tumor sufficient in quality and quantity for ctDNA assay design and tBRCA and Homologous recombination deficiency (HRD) testing is required.
- An Eastern Cooperative Oncology Group (ECOG) performance status of 0 or 1.

Exclusion Criteria:

- Prior treatment with a Poly Adenosine-diphosphate Ribose Polymerase (PARP) inhibitor.
- Current treatment with a Cyclin-dependent kinase (CDK)4/6 inhibitor or endocrine therapy other than anastrozole, letrozole, exemestane, and tamoxifen with or without ovarian suppression.
- Participants have any sign of metastasis or local recurrence after comprehensive assessment conducted per protocol.
- Participants have shown no definitive response to preoperative chemotherapy by pathologic, radiographic or clinical evaluation, in cases where preoperative chemotherapy was administered.
- Participants have inadequately treated or controlled hypertension.
- Participants have received live vaccine within 30 days of planned start of study randomization.
- Participants have a second primary malignancy.
- Exceptions are the following: (a) Adequately treated non-melanoma skin cancer, curatively treated in situ cancer of the cervix, Ductal carcinoma in situ (DCIS) of the breast, Stage I Grade 1 endometrial carcinoma. (b) Other solid tumors and lymphomas (without bone marrow involvement) diagnosed >=5 years prior to randomization and treated with no evidence of disease recurrence and for whom no more than 1 line of chemotherapy was applied.
- Participant is pregnant, breastfeeding, or expecting to conceive children while receiving study treatment and/or for up to 180 days after the last dose of study treatment (except France).
- Participant is immunocompromised. Participants with splenectomy are allowed. Participants with known human immunodeficiency virus (HIV) are allowed if they meet protocol-defined criteria.
- Participants have a known history of myelodysplastic syndrome (MDS) or acute myeloid leukemia (AML).

Figure 2.3: An example of clinical trial eligibility criteria.

Chapter 3

Machine Learning Prediction of Clinical Trial Operational Efficiency

As outlined in Chapter 2, many clinical trials struggle with low efficiency which results in high costs, trial failures, and fewer drug approvals. In recent years, trials have become significantly more expensive due to their increased complexity [15], involving more endpoints, procedures, eligibility criteria, countries, sites, and patients than in the previous decade [16]. Therefore, there is a great need to improve the operational efficiency of trials in order to reduce costs and shorten the lag in improving patient access to novel and innovative treatments. The expertise of the trial designer has been shown to be a significant factor in determining trial success [17].

This chapter describes a study on predicting multiple operational efficiency metrics using machine learning models and identifying the the trial design features that influence these metrics. It is a condensed version of a research article published in The AAPS Journal [18], in which I was a co-author. The other authors of this paper are: Kevin Wu,[1] Eric Wu,[2] Michael DAndrea,[3] Nandini Chitale,[3] Melody Lim,[3] Marek Dabrowski,[4] Hanoor Rangi,[5] Ruishan Liu,[2] Marius Garmhausen,[6] Navdeep Pal,[3] Chris Harbron,[7] Shemra Rizzo,[3] Ryan Copping,[3] James Zou.[1,2] My contributions to this study included preparing the dataset, supervising the experimentation process, providing feedback on the content, and reviewing the publication. This was the first research related to clinical trials that I was involved in, which revealed the importance of trial design features (including eligibility criteria) on efficiency and sparked my interest in the topic.

Section 3.1 defines the metrics used to measure the trial operational efficiency. These are the target variables of the fitted ML models. Section 3.2 presents other works investigating the impact of clinical trial complexity on operational efficiency and determines how this study differ from them. Section 3.3 describes the dataset

[1]Department of Biomedical Data Science, Stanford University, Stanford, California, USA
[2]Department of Electrical Engineering, Stanford University, Stanford, California, USA
[3]Genentech, South San Francisco, San Francisco, California, USA
[4]Roche Pharmaceuticals, Warsaw, Poland
[5]Roche Pharmaceuticals, Mississauga, Canada
[6]Roche Pharmaceuticals, Basel, Switzerland
[7]Roche Pharmaceuticals, Welwyn Garden City, UK

used in the experiments, while Section 3.4 outlines the data preprocessing method, selected model, and evaluation metrics. Section 3.5 presents the results, and Section 3.6 discusses the findings.

3.1 Measuring Operational Efficiency

Complex trials often include extensive patient recruitment requirements and protocol-related delays, leading to significant operational inefficiencies. The rise in protocol procedures and amendments greatly increases site work burden and operational delays. In this study, trial operational efficiency is assessed through various metrics associated with patient recruitment and trial duration.

Patient recruitment is assessed via two distinct metrics:

- **Screen failure ratio:** The fraction of screened patients that do not end up enrolled in a trial. This metric is commonly used to measure patient recruitment efficiency [19, 20, 21]. A high screen failure ratio indicates that a trial requires more money and time to acquire its patients.

- **Dropout ratio:** The fraction of enrolled patients that do not complete the trial. This is an important metric to estimate in the study design phase [22]. Patients can be withdrawn from a trial for various reasons, such as adverse events, noncompliance, protocol deviations, and safety issues. Excessive dropout can lead to costly protocol amendments or under-powered studies [23], affecting the quality of data that can be used to improve patient outcomes.

Trial duration is measured with the use of three metrics, capturing different components of trial execution:

- **Pre-enrollment duration:** The median number of days per site between site selection and the enrollment of the first patient. This measures the time required to complete organizational prerequisites (e.g., contract negotiation and site training). A lengthy pre-enrollment period can imply a high regulatory and organizational burden.

- **Enrollment duration:** The median number of days per site between enrolling the first patient and the last patient across sites. Extended site enrollment delays can be due to unnecessarily stringent eligibility criteria and screening protocols, rare patient populations, and competing clinical trials.

- **Study duration:** The median number of days per site between site activation and the last patient visit, capturing the end-to-end time required for a study to complete across sites.

All five metrics are visualised in Figure 3.1.

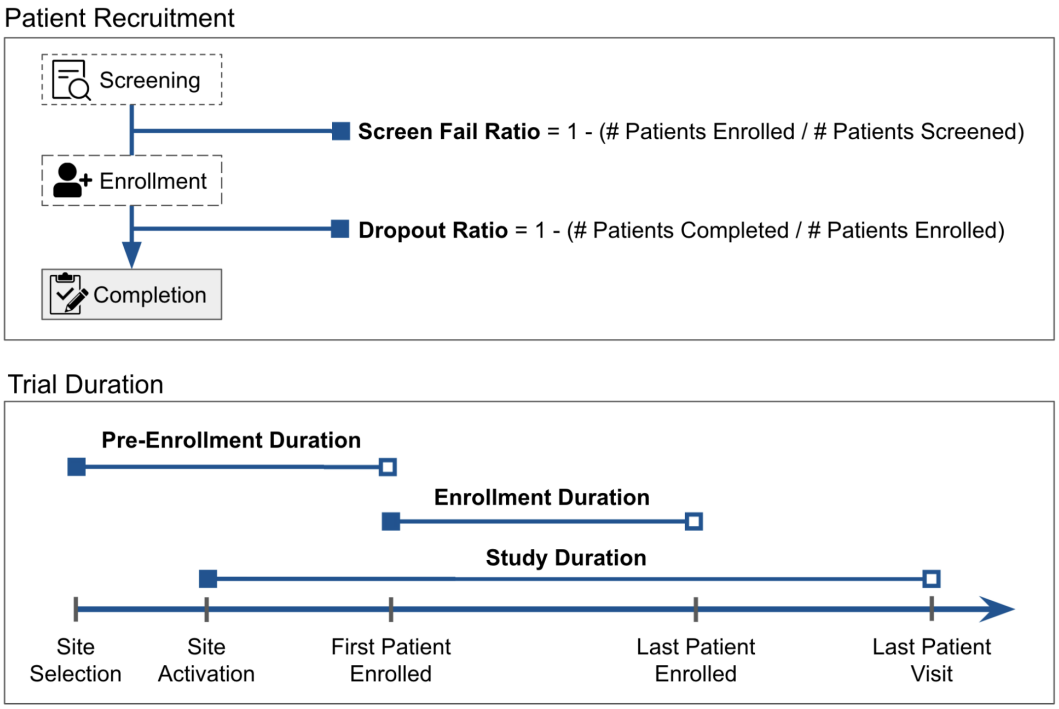


Figure 3.1: Patient recruitment metrics displayed across the patient funnel from screening to completion and trial duration metrics across an abridged timeline of clinical trials (the timeline presented applies to a single site and these events can be asynchronous between sites).

3.2 Related works

There exists extensive literature studying the growing complexity of clinical trials [24, 25, 26, 27, 28], including patient recruitment and trial duration [15, 29, 23, 19, 20, 21, 30, 31, 22, 32, 33, 34, 35, 36]. Fogel [23] provides a systematic overview of how operational inefficiencies impact the likelihood of overall trial success.

Getz *et al.* [37] show that more complex trials with more procedures perform worse in patient recruitment and retention compared to low complexity trials. Ross *et al.* [38] find that complex trials discourage trial participants, while Boericke and Gwinn [39] identify that trials with more eligibility criteria tend to be more prone to delays. Additionally, Andersen *et al.* [40] find that patient dropout is much higher in more complex trials.

This study aims to unify these analyses by modeling a variety of trial features in relation to patient recruitment and trial duration. By doing so, it is possible to estimate the collective impact of multiple trial features on trial efficiency.

Prior work in applying machine learning methods to improve clinical trial efficiency includes natural language processing methods for patient recruitment and extracting structured data from eligibility criteria [41, 42]. Machine learning has been applied to clinical trial data for the purpose of predicting the overall likelihood of

approvals [43, 44, 45]. This study differs from prior research by focusing on specific efficiency metrics, a level of granularity that is of particular interest to trial designers.

3.3 Data

The dataset comprises 2051 completed clinical trials conducted by Roche, with starting dates ranging from 2009 to 2020. Due to varying levels of missing data across different efficiency metrics, the analyses include 1922 trials for enrollment duration, 1395 for screen fail ratio, 932 for pre-enrollment duration, 526 for study duration, and 361 for withdrawal ratio. These discrepancies are caused by differences in the data collection process; for instance, enrollment duration is more easily calculated with start and end dates, while withdrawal ratios require follow-up reports for each patient. The study incorporates a total of 23 operational features, detailed in Table 3.1. These features include study phase, therapeutic area, experimental design, number of endpoints, number of eligibility criteria, and specifics about planned procedures. The data represent 288 unique drugs and 219 unique indications, with an average of 11 inclusion criteria, 15 exclusion criteria, and 4 countries per trial.

3.4 Methods

3.4.1 Data Preprocessing

Categorical features, including drug names and indications, are encoded using a one-hot encoding scheme. To manage the rare feature values (appearing in less than 1% of the data), we grouped them into a single category labeled *"other"*. Missing values are imputed using the mean. Trials with target variables falling outside two standard deviations are considered outliers and are excluded from the dataset.

3.4.2 Model

In the study, we chose the LightGBM algorithm due to its high performance on tabular data. Separate models were trained for each of the five target variables. Although AutoSklearn was also executed on the data, LightGBM remained the top-performing model, confirming it as a suitable choice for this purpose.

To support study design decisions, both specific point estimates and the uncertainty around them are provided. This uncertainty is quantified by producing predictive intervals using a quantile loss function, trained at quantiles 0.05 and 0.95 to achieve a 90% predictive interval. Point estimates are derived from the 0.5 quantile, representing the median.

The dataset is split into training, test, and validation sets in a 60/30/10% ratio. Model performance is optimized through a grid search of hyperparameters on the validation set and evaluated using the test set.

Table 3.1: Descriptions of trial features used in the model predicting trial operational efficiency.

Feature Name	**Description**
Study Phase	Trial phase (ie. I, II, III, IV)
Therapeutic Area	Oncology, I2O (Immunology, Infectious Diseases, and Ophthalmology), Neuroscience, and Other
Distribution	Single-site, Single-country, Multi-country
Comparison	Active controlled, Non-controlled, etc.
Randomization	Randomized, Non-randomized
Intervention	Interventional study, Non-interventional study
Blinding	Open label, Double blind, Single blind
Num Primary Endpoints	Number of distinct primary endpoints
Num Secondary Endpoints	Number of distinct secondary endpoints
Num Inclusion Criteria	Number of inclusion criteria
Num Exclusion Criteria	Number of exclusion criteria
Num Countries	Number of countries planned for the study
Min Age	Minimum eligible age
Max Age	Maximum eligible age
Patient Gender	Eligible gender
Num Planned Examination Procedures	Number of planned examination procedures (physicals, observations, questionnaires, and other measurements)
Num Planned Diagnostic Procedures	Number of planned diagnostic procedures (lab work such as blood tests and imaging work/biopsies that are usually performed by technicians).
Num Planned Core Procedures	Number of planned core procedures (mandatory and common procedures such as adverse events and informed consent)
Num Planned Non-Core Procedures	Number of planned non-core procedures (procedures not counted as core procedures)
Num Planned Drug Treatment Procedures	Number of planned drug treatment procedures (administering treatment to patients)
Num Planned Visits	Number of unique planned visits
Num Planned Enrollment	Target number of patients to enroll
Drug Name	Name of drug being studied

3.4.3 Evaluation

To evaluate the models consistently across different regression tasks, the c-index is used. It measures the proportion of correctly ordered pairs in the test set, indicating the proportion of concordant pairs among all evaluation pairs in the test set. A c-index of 1 indicates perfect prediction accuracy, while a c-index of 0.5 suggests performance no better than chance. For example, if the actual screen failure ratios for two trials are 0.75 and 0.90, predicted values of 0.60 and 0.80 would count as a concordant pair. The c-index thus reflects the model's ability to predict the direction of change in operational efficiency based on trial features. Additionally, the R-squared score and mean absolute error (MAE) are reported.

3.5 Results

For each target metric, we evaluate model performance on all trials from the test set as well as subsets selected by therapeutic area and study phase. The c-index results are presented in Table 3.2. Models predicting patient recruitment perform excellently with c-index values around 0.80, while those predicting trial duration metrics have c-index values around 0.70, indicating that predicting study duration is more challenging than predicting patient recruitment. The R-squared scores and MAE are reported in Table 3.3.

Table 3.2: The c-index results for the LightGBM models evaluated across all trials and stratified by therapeutic area and study phases.

Efficiency Metric	Overall C-Index	Therapeutic Area (C-Index)				Study Phase (C-Index)			
		I2O	Neuroscience	Oncology	Other	I	II	III	IV
Screen Failure Ratio	0.801	0.795	0.765	0.789	0.808	0.622	0.788	0.802	0.771
Dropout Ratio	0.791	0.750	0.651	0.715	1.000	0.784	0.801	0.804	0.771
Pre-Enrollment Duration	0.705	0.724	0.635	0.611	0.687	0.675	0.565	0.587	0.597
Enrollment Duration	0.706	0.680	0.709	0.683	0.672	0.764	0.692	0.647	0.609
Trial Duration	0.728	0.644	0.766	0.624	0.756	0.808	0.656	0.610	0.666
Average	0.746	0.719	0.705	0.684	0.784	0.731	0.700	0.690	0.683

Table 3.3: R-squared score and mean absolute error from the LightGBM models across five operational efficiency metrics.

Efficiency metric	R-squared	Mean absolute error
Screen failure ratio	0.463	0.097
Dropout ratio	0.513	0.179
Pre-enrollment duration	0.319	60.0
Enrollment duration	0.26	245
Study duration	0.32	405
Average	0.375	-

Two additional validations are conducted:

- **Evaluation on unseen drugs:** To assess the potential overfitting due to trials on the same drug appearing in both the training and test sets, the data is split based on randomly selected sets of drugs. The model is trained on 209 unique drugs and tested on 79 different unique drugs. The results, included in Table 3.4, show that performance slightly decreases without prior knowledge of the drug, indicating that such knowledge can support the prediction of trial efficiency, though the impact is not large.
- **Evaluation on newest trials:** The models are evaluated for potential time-specific biases by splitting the data into two periods: 2009-2011 (training and validation set) and 2012-2020 (test set). The c-index for each metric is reported in Table 3.5. The overall performance is slightly lower due to a smaller training set size, but no significant differences in performance are observed.

Table 3.4: Validation on unseen drugs across five efficiency metrics.

Validation on Unseen Roche Drugs (C-index)	**Training Drug Set (N=339)**	**Testing Drug Set (N=359)**
Screen Failure Ratio	0.781	0.712
Dropout Ratio	0.757	0.738
Pre-Enrollment Delay	0.674	0.634
Enrollment Duration	0.673	0.665
Trial Duration	0.699	0.679
Average Across Metrics	0.717	0.686

Table 3.5: Validation of trials from two time periods across five efficiency metrics.

Validation Across Time (C-index)	**Trials completed 2009-2011 (N=439)**	**Trials completed 2012-2020 (N=376)**
Screen Failure Ratio	0.742	0.726
Dropout Ratio	0.630	0.682
Pre-Enrollment Delay	0.673	0.680
Enrollment Duration	0.711	0.669
Study Duration	0.704	0.717
Average	0.692	0.695

To quantify how actionable features of trial design correlate with operational efficiency metrics, a separate multivariate regression model is fitted. Features that are fixed aspects of a trial are used as covariates but excluded from the table (eg. trial phase, therapeutic area, randomization), while the features that can be changed during trial design (e.g., number of eligibility criteria, endpoints, countries, and procedures) are assessed and presented in the Table 3.6.

Table 3.6: Coefficients of a multivariate linear model fitted on trial features (next to the coefficients, the p-values for each feature in each model are reported for each cell: * <= 0.05, ** <= 0.01, *** <= 0.001, ns = no significance).

Trial Operational Feature	**Screen Failure Ratio**	**Dropout Ratio**	**Pre-Enrollment Duration**	**Enrollment Duration**	**Study Duration**
Num Primary Endpoints	0.0064 **	ns	ns	ns	ns
Num Secondary Endpoints	0.0046 ***	ns	ns	-7.2121 **	ns
Number Planned Countries	0.0036 ***	ns	1.2799 **	-7.9442 ***	10.2753 **
Num Eligibility Criteria	ns	ns	ns	1.5514 *	ns
Num Planned Examination Procedures	ns	0.0114 **	ns	ns	ns
Num Planned Diagnostic Procedures	ns	ns	ns	ns	ns
Num Planned Non-Core Procedures	-0.0029 *	ns	ns	ns	ns
Num Planned Drug Treatment Procedures	ns	ns	ns	ns	ns
Num Planned Core Procedures	ns	ns	ns	ns	ns
Num Unique Planned Visits	ns	0.0024 ***	-0.2941 *	0.9518 *	3.7001 ***
Planned Patient Enrollment	ns	ns	ns	0.0164 **	ns

Key findings include:

- A higher number of countries is associated with a longer pre-enrollment and study duration, and a higher screening failure ratio, but a shorter site-specific enrollment duration.
- A higher number of primary and secondary endpoints is associated with a higher screen fail ratio.
- A higher number of planned patient visits correlates with an increased dropout ratio and a longer study duration.
- A higher planned patient enrollment is correlated with a longer enrollment duration.

Additionally, the feature importance for the LightGBM, defined as the information gained from each feature with respect to the loss function, is calculated and presented in Figure 3.2. For visual clarity, the importance scores are normalized to sum to one for each metric. Moreover, for interpretability, only the importance of a subset of actionable features is reported, rather than the whole set of features used by the model. The analysis shows that planned patient enrollment and the number of eligibility criteria are among the most important features, which were not identified by the linear model. This indicates a strong nonlinear relationship between those features and operational efficiency. The analysis also reveals that the number of planned visits has a significant impact on the dropout ratio.

3.6 Conclusion

The results of this study indicate that trial features can be robust predictors of the trial operational efficiency. We observe consistent outcomes across drug names, time periods, sponsor companies, therapeutic areas, and study phases. Notably, complex trials generally achieve worse results in patient recruitment and trial duration. The analysis demonstrates that models predicting patient recruitment perform better than those forecasting trial duration when using the selected features.

Operational efficiency is multifaceted and influenced by the interaction of multiple trial features. Trial investigators, however, must balance operational and scientific efficiency - a task that sometimes involves retaining complex trial designs due to their scientific value. This study also highlights that the relationships between trial features and operational efficiency are not strictly linear. Furthermore, it is critical to note that some features, such as the number of eligibility criteria and the numbers of procedures, are considered at a high level, without analyzing their specific content. For instance, criteria related to gender and age are usually less restrictive than those concerning specific cancer biomarkers and stages, which could significantly narrow down the eligible patient population and complicate recruitment. Nonetheless, the feature importance analysis reveals that even a general count of criteria significantly influences model prediction.

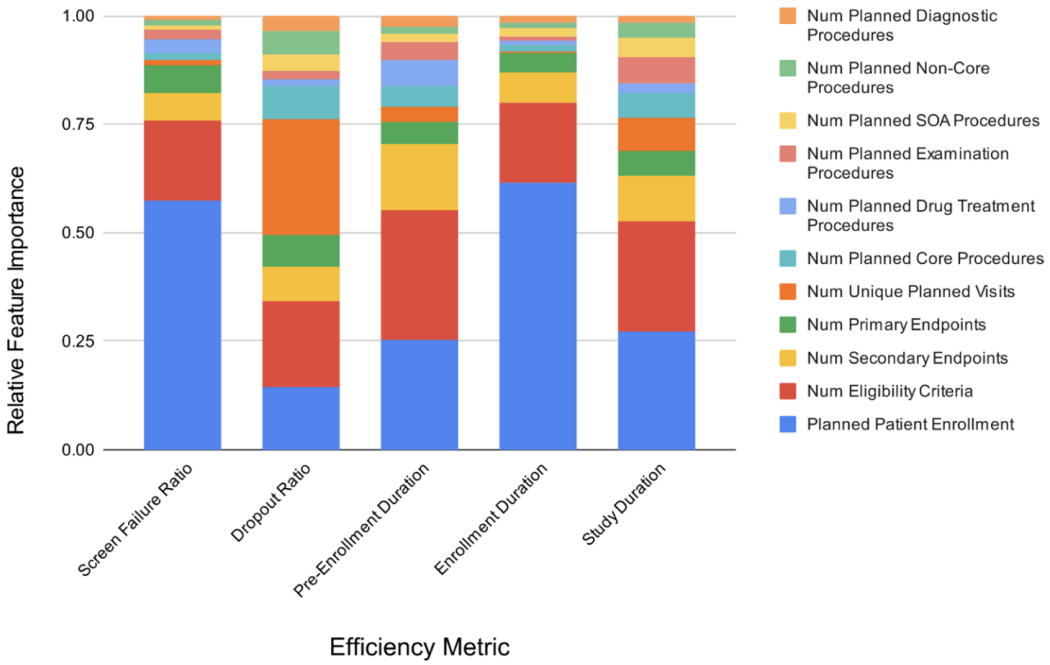


Figure 3.2: The importance of features used in the LightGBM models predicting trial operational efficiency metrics.

This experiment demonstrates the potential of machine learning to predict operational metrics of clinical trials using design features. It explores the impact of modifiable features on efficiency and provides insights valuable for planning and design future trials. However, the current study does not examine the impact of the specific content of eligibility criteria on trial efficiency, due to the complexities involved in parsing these criteria. This gap in research, along with the absence of reliable methods for parsing eligibility criteria - which limits various analyses and the automation of trial processes - has directly motivated my doctoral project. There is a need for future research to focus on developing methods that parse and analyze the content of eligibility criteria. This advancement is crucial for further enhancing trial design and operational efficiency.

Chapter 4

NLP Resources for Biomedical Domain

The increasing availability of biomedical text data has led to rapid developments in the field of biomedical natural language processing (BioNLP). This field supports the optimization of healthcare-related processes by integrating natural language processing, bioinformatics, medical informatics, and computational linguistics [46]. The effectiveness of BioNLP is supported by the availability of numerous reusable resources, which are crucial in overcoming common challenges, such as the lack of large annotated corpora and the complexity of semantics and syntax in biomedical texts.

This chapter reviews widely used BioNLP resources, including knowledge bases (Section 4.1), ontologies (Section 4.2), and pretrained language models (Section 4.3). Section 4.4 discusses large language models and the Generative AI revolution.

4.1 Knowledge bases

Knowledge bases are essential components of the Biomedical NLP field, acting as repositories of both structured and unstructured biomedical information. They contain extensive knowledge on various healthcare topics, such as medical conditions, treatments, drugs, and procedures. This section examines the most commonly used biomedical knowledge bases.

4.1.1 PubMed

One of the fundamental biomedical resources is PubMed, a literature database created by the National Center of Biotechnology Information at the U.S National Library of Medicine and available since 1996 [47]. It provides access to more than 37 million citations and abstracts from biomedical literature, coming from three sources [48]:

- **MEDLINE** - a primary bibliographic database maintained by the U.S. National Library of Medicine and available since 1971. It contains over 31 million references to publications on life sciences, with main focus on biomedicine. It is the largest component of PubMed.

- **PubMed Central (PMC)** - an archive of full-text journal articles, author manuscripts and preprints on biomedical and life sciences, which is a part of the U.S. National Library of Medicine collection. It contains over 8 million records.
- **Bookshelf** - a repository of books and documents on life sciences and healthcare. It includes over 9000 records.

4.1.2 UMLS

Another resource is the Unified Medical Language System (UMLS), which is a collection of biomedical vocabularies created by the U.S. National Library of Medicine [49]. It includes 2.5 million names mapped to about 900 000 concepts from over 60 families of biomedical vocabularies and more than 12 million relations between those concepts. A concept in the UMLS is a meaning that can have different names. It is a cluster of synonymous terms, linked to other concepts in the database, forming a rich graph. There are three main components of the UMLS [50]:

- **Metathesaurus:** A large multi-lingual vocabulary database containing information about biomedical concepts and their relations. It includes terms and codes from various vocabularies, e.g. RxNorm, SNOMED-CT, MeSH, and ICD-10. This component maintains the hierarchy of terms, their definitions, and attributes. It is the central element of the UMLS.
- **Semantic Network:** Provides categorization for the Metatheasurus concepts. It includes a set of 127 categories (semantic types) and a set of 54 relationships between these types (semantic relations). This network simplifies the Metathesaurus by grouping concepts into semantic types, which are the nodes of the network, connected by links (relations).
- **SPECIALIST Lexicon and Tools:** A large syntactic lexicon of biomedical and general English terms, including morphological, orthographic, and syntactic information for each term form, along with a collection of processing tools for term normalization and creating lexical variants.

Additionally, the UMLS is integrated with MetaMap, an application developed by National Library of Medicine, which maps biomedical terms to concepts in the UMLS Metathesaurus. It is very useful for information extraction or term normalization purposes.

4.1.3 DrugBank

DrugBank is a database of drugs and their targets, created by the University of Alberta in 2006 [51]. It is considered a gold standard knowledge resource for drug data, including chemical, pharmacological and pharmaceutical properties. For instance, it provides information on drug indications, mechanism of action, metabolism, toxicity, adverse events, and interactions with other drugs and food. The database also

includes a list of various brand names for each drug and provides data on related proteins such as targets and transporters.

The latest release of DrugBank in March 2024 includes 16 602 drugs and 5293 distinct proteins, such as drug targets, enzymes, transporters, and carriers, which are linked to these drugs [52].

4.2 Ontologies

Biomedical ontologies are crucial for structuring and standardizing domain knowledge. An ontology serves as a catalog of entities, outlining the relationships between them. Each entity is associated with a unique concept identifier, which connects to all entities that share the same semantic meaning. Typically, this includes at least one entity but often extends to several due to the presence of many synonyms [53]. For instance, in an ontology, the entity *"liver cancer"* should share the same identifier as *"hepatocellular carcinoma"*, and this identifier should also be linked to other natural language terms, such as *"liver neoplasm"*, *"HCC"*, *"liver carcinoma"*, etc.

A biomedical ontology is an invaluable resource in numerous BioNLP tasks, such as text normalization and named entity recognition. The biomedical language contains a vast number of medical terms and abbreviations that require specialized medical knowledge for accurate processing by AI systems. An ontology can incorporate this knowledge into the system, consequently enhancing the effectiveness of the solution. This section provides an overview of the main biomedical ontologies and ontology-like vocabularies.

4.2.1 SNOMED CT

SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) is a multilingual clinical terminology released in 2002, used by over 80 countries and considered to be the most comprehensive clinical vocabulary in the world [54]. It allows for a standardized representation of clinical terms in healthcare systems. The terminology consists of 3 components [55]:

- **Concepts:** These are the basic elements of the terminology. Each concept represents a unique clinical meaning and is associated with a unique numeric identifier.
- **Descriptions:** These are the unique names of the concepts and their synonyms, for example, *"liver cancer"* is the unique name, and the synonyms include *"liver neoplasm"* and *"hepatocellular carcinoma"*.
- **Relationships:** These are links between two concepts, which are machine-readable and maintain semantic connections.

SNOMED CT includes over 350 000 concepts and 1 million relationships [55]. The types of concepts range from clinical findings and procedures to substances, pharmaceutical products, and body structures. All concepts are arranged into hierarchical

structures with an *"IS A"* relationship, such as *"viral pneumonia" IS A "infective pneumonia"*. There are 19 separate hierarchies in SNOMED CT. Concepts from different hierarchies are linked by attribute relationships (e.g. *"finding site"*, *"causative agent"*). For example, the *"finding site"* relationship connects the *"viral pneumonia"* concept with the *"lung"* concept, whereas these concepts originate from the different hierarchies of clinical findings and body structures, respectively. Fig. 4.1 shows the hierarchy of SNOMED CT concepts for *"Diabetes mellitus type 2"*, which stem from *"Clinical finding"* hierarchy, and a list of relations to other hierarchies - *"body structures"* in that example.

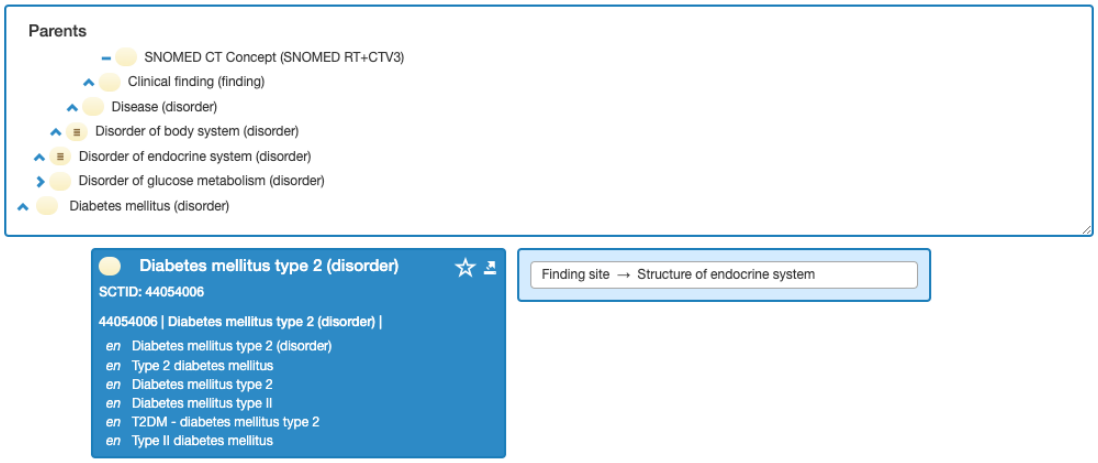


Figure 4.1: Hierarchy of Diabetes mellitus type 2 in SNOMED CT (screenshot from: https://browser.ihtsdotools.org/).

4.2.2 MeSH

Medical Subject Heading (MeSH) is a thesaurus for the biomedical domain, created by the U.S. National Library of Medicine (NLM) in 1960 [56]. It was developed to index life sciences publications on PubMed. Each record in PubMed is tagged with a set of MeSH terms, which are subject descriptors that facilitate the searching of publications. This vocabulary allows users to enter various terms in a search query in PubMed and retrieve a consistent set of records, even if the terms are different but related to the same concept. For example, searching for *"myocardial infraction"* will also return articles on *"heart attack"*, *"acute myocardial injury"*, and their plural variants.

MeSH serves as an ontology by providing a hierarchical structure of medical terms. This hierarchy is particularly useful for searching in PubMed, where using broader terms can yield papers indexed under narrower related terms. For instance, a search for *"neoplasms"* publications will not only retrieve records tagged with *"neoplasms"*, but also those under *"breast neoplasms"*, *"lung neoplasms"*, etc. A fragment of the *"Lung Neoplasms"* hierarchy is presented in Fig. 4.2.

MeSH consists of three types of records:

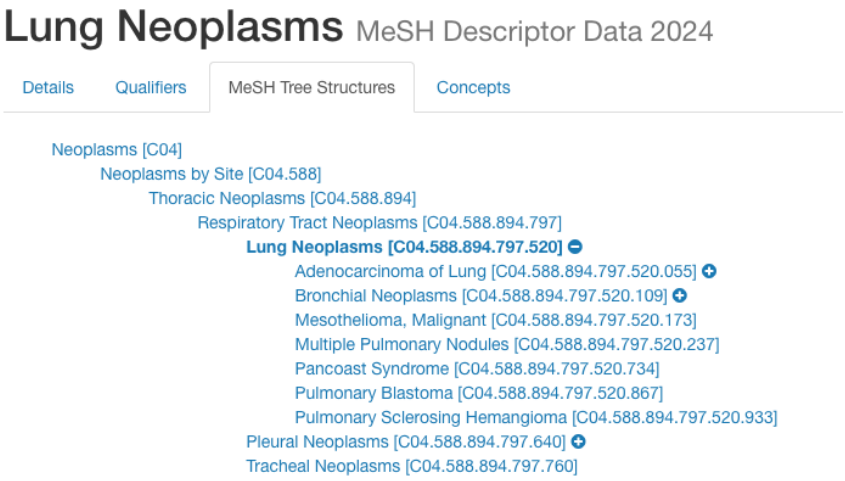


Figure 4.2: Hierarchy of Lung Neoplasms in MeSH (screenshot from https://meshb.nlm.nih.gov/).

- **Descriptors:** These are terms that describe the subject of publications in PubMed and other NLM databases.
- **Qualifiers:** These are narrower terms that complement descriptors. They specify details of the discussed subject. For example, in a publication about drug treatment for lung neoplasms, *"Lung Neoplasms"* is a descriptor, while *"drug therapy"* is a qualifier.
- **Entry Terms:** These are synonyms and closely related terms to descriptors, which can be used interchangeably in searching publications. For instance, *"lung cancer"* and "*pulmonary cancer"* are entry terms for *"lung neoplasms"*.

An illustration of entry terms and qualifiers for the *"Lung Neoplasms"* descriptor is shown in Fig. 4.3

4.2.3 ICD

The International Classification of Diseases (ICD) is a global standard for classifying and coding health information, such as diseases, symptoms, injuries, causes of death, and other medical conditions [57]. Since 1948, the World Health Organization (WHO) has been responsible for publishing and maintaining this system. ICD provides a common language for the healthcare industry, enabling the sharing and interpretation of medical data at various medical sites and by different professionals across countries and regions. It has been translated into 43 languages and is utilized by a diverse group of stakeholders, including physicians, nurses, policymakers, insurers, and IT professionals working with medical data.

As a medical terminology and classification framework, ICD maps medical terms to unique codes organized into a hierarchical structure, thus serving as a medical

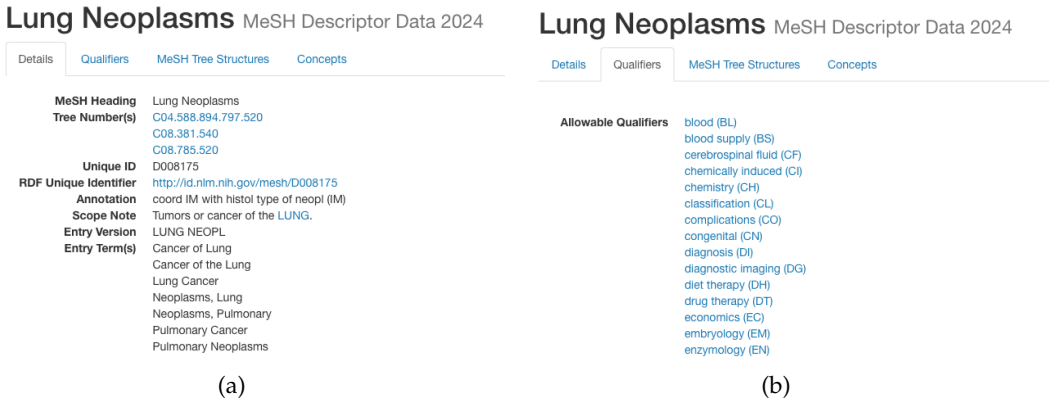


Figure 4.3: Records for Lung Neoplasms in MeSH. (a) Entry terms. (b) Qualifiers. (screenshots from: https://meshb.nlm.nih.gov/)

ontology that supports the development of NLP systems. ICD includes all diseases, each associated with their diagnostic characteristics and unique identifiers, composed of alphanumeric characters. The latest version, ICD-11, contains over 120 000 medical terms linked to 17 000 codes ranging from 1A00.00 to ZZ9Z.ZZ [58] [59]. The initial character of a code indicates a chapter, such as *"Diseases of the visual system"*, or *"Sleep-wake disorders"*, and the subsequent characters detail specific conditions. For example, the code 2A00.00 corresponds to *"Glioblastoma of brain"*, which is the most detailed level in the hierarchy, as illustrated in Fig. 4.4. Higher levels include *"Gliomas of brain"* (2A00.0), *"Primary neoplasms of brain"* (2A00), and *"Neoplasms of brain or central nervous system"* (2A0), up to the broader *"Neoplasms"* chapter indicated by the code 2.

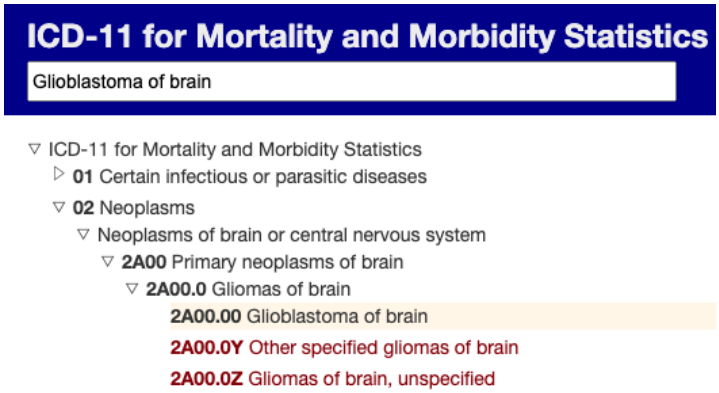


Figure 4.4: ICD-11 hierarchy for Glioblastoma of brain (screenshot from: https://icd.who.int/browse/2024-01/mms/en).

4.2.4 LOINC

Logical Observation Identifiers Names and Codes (LOINC) is another biomedical standardized vocabulary created by the Regenstrief Institute and released in 1995 [60]. It supports the electronic exchange of clinical results, such as laboratory tests and clinical observations, by providing universal identifiers for them. The current version of LOINC contains over 103 000 records. Unlike ICD or SNOMED CT, LOINC does not incorporate hierarchical information within the observation codes. Codes are assigned sequentially as observations are added to the system. LOINC includes codes not only for laboratory tests like microbiology, chemistry, or serology, but also for clinical observations, such as EKG measurements, vital signs, or radiology report findings [61].

Each LOINC record is assigned a unique code and name, composed of five or six parts, which include [61]:

- **Component:** The substance or element being measured, e.g. hemoglobin, hepatitis C antigen.
- **Property:** The type of measurement property, e.g. mass, length, volume.
- **Timing:** The time interval over which the measurement is taken, e.g. point in time, 24-hour urine collection.
- **System:** The type of sample or organ examined, e.g. urine, chest, blood.
- **Scale:** The measurement scale type, e.g. quantitative, nominal, ordinal.
- **Method:** The method used to obtain the measurement, e.g. manual count, immuno blot; this is the only optional part of the name.

Fig. 4.5 illustrates the LOINC hierarchy and fully-specified name structure for a laboratory test, BAL, which detects bacteria in the lungs, with the specific code 95074-1. Additionally, LOINC is one of the source vocabularies for the Unified Medical Language System (UMLS), discussed in Section 4.1.2 [60].

4.2.5 RxNorm

RxNorm is a terminology for all medications available in the United States, containing both generic and branded names. It was developed by the U.S. National Library of Medicine in 2001 [62]. The primary stakeholders of RxNorm include hospitals, pharmacies, and other organizations that use electronic systems to process and store drug information. Each brand name of a drug in RxNorm is linked to a normalized name, which consists of three components: ingredient, strength, and physical form, for example, *"Naproxen 250 MG Oral Tablet"*. The branded normalized name includes the brand enclosed in square brackets at the end, e.g. *"Naproxen 250 MG Oral Tablet [Prosaid]"* [63].

A key element of RxNorm is the Concept Unique Identifier (RxCUI), which is assigned to each drug enity. The identifier points to a concept with a common meaning,

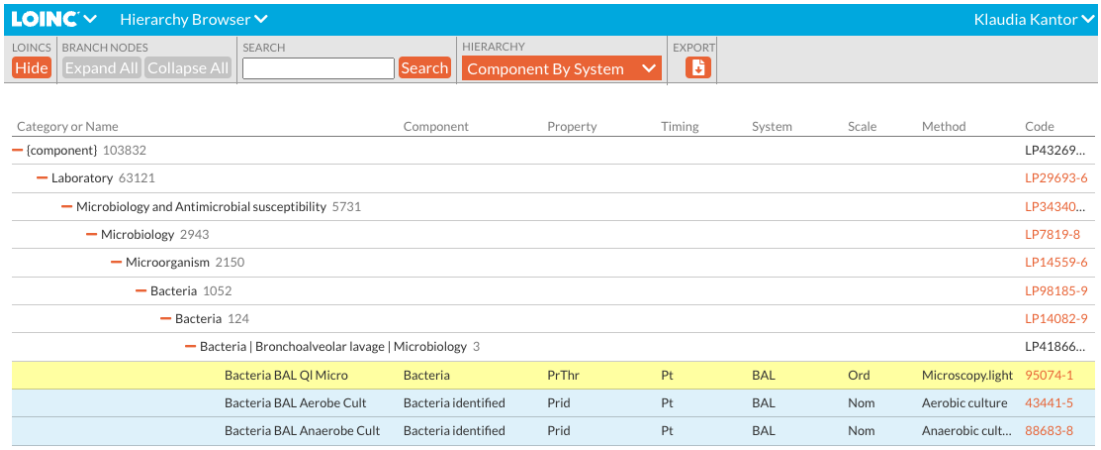


Figure 4.5: The LOINC hierarchy for a BAL test in lungs (screenshot from: https://loinc.org/search/).

allowing synonymous drug names to share the same RxCUI. RxNorm also maintains relationships between concepts, which indicate how substances relate to one another, such as *"Naproxen Pills" HAS_INGREDIENT "Naproxen"*, or *"Naproxen 250 MG Oral Tablet" IS_A "Naproxen Oral Tablet"*. Moreover, RxNorm serves as a source vocabulary for the Unified Medical Language System.

4.2.6 MEDCIN

MEDCIN is a standardized clinical terminology for encoding medical information in Electronic Health Records (EHRs), developed by the Medicomp Systems and released in 1978 [64]. It comprises over 400 000 clinical concepts that are linked through various relationships [65]. The MEDCIN vocabulary includes diagnoses, symptoms, patient histories, therapies, tests, and physical exams. Each concept within MEDCIN is a clinical statement meaningful to clinicians at medical sites.

MEDCIN structures its concepts into two distinct hierarchies: one that organizes terms from more general to more specific, and another that links concepts to specific diagnoses. Each concept is assigned a unique six-digit numeric identifier. Additionally, MEDCIN enhances its concepts with parameters such as prefixes, modifiers, and statuses. Prefixes help distinguish between different instances of a clinical concept, such as personal vs. family history of a diagnosis. Modifiers provide additional information about a condition's severity, for example, indicating *"stage IV"* in breast cancer. Statuses describe the condition's progression, such as stable, in remission, or worsening.

Fig. 4.6 illustrates how the same concept, *"Breast Cancer"*, can vary when different parameters are applied within MEDCIN.

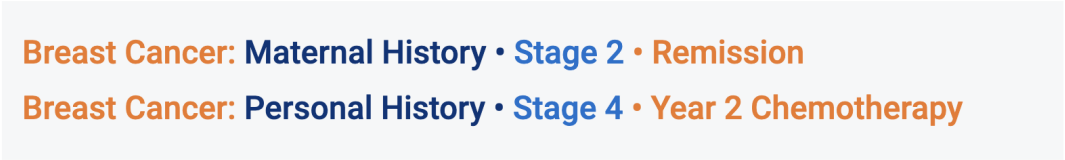


Figure 4.6: Variability in the representation of the *"Breast Cancer"* concept with different parameters in MEDCIN (screenshot from: https://medicomp.com/medcin-details/).

4.2.7 MedDRA

The Medical Dictionary for Regulatory Activities (MedDRA) is a standardized medical terminology designed for regulatory reporting, widely used by regulatory authorities, healthcare professionals, and pharmaceutical companies since 1999. It was developed by the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH), with significant contributions from the UK Medicines Control Agency [66]. MedDRA is applied in both pre-market and post-market activities, including registration, documentation, and safety monitoring of medical products.

MedDRA provides terminology for reporting a variety of adverse events, such as signs, symptoms, diagnoses, medical tests, procedures, patient histories, and device issues. It does not cover patient demographics, clinical trial design, numerical values, or qualifiers. The vocabulary is organized into a five-level hierarchy, where the most basic unit is the Preferred Term, a distinct descriptor for a unique condition represented in MedDRA. This hierarchy facilitates the linkage between synonyms and the relationship between more detailed and more general terms [67].

4.3 Pretrained language models

Another useful group of BioNLP resources are biomedical language models. In recent years, Transformer-based pretrained language models have significantly influenced the NLP field. A Transformer is a neural network architecture that utilizes an attention mechanism, which allows each token in a sequence to dynamically adjust its representation based on the relevant tokens in the input sequence, output sequence, or the sequence itself [68]. When the attention is computed based on the tokens in the same sequence, it is known as a self-attention. Transformer models include variations built on the Encoder, Decoder, or combined Encoder-Decoder architectures. The original Encoder architecture comprises six blocks, each containing a multi-head self-attention layer and a fully-connected feed-forward layer, while the Decoder includes an additional multi-head masked attention layer in its six blocks. Different models may incorporate varying numbers of Transformer blocks. This section focuses on Encoder models.

A key example of a Transformer Encoder model is BERT (Bidirectional Encoder

Representations from Transformers) [69]. BERT is a Masked Language Model (MLM) which randomly masks tokens in its input and is trained to correctly fill those gaps based on their context. A second training objective is Next Sentence Prediction (NSP), where BERT predicts whether a second sentence logically follows the first. BERT is pretrained on the general-domain corpuses, the English Wikipedia and the Book Corpus [70].

A major strength of Transformer-based models is their architecture, particularly the attention mechanism, that addresses the loss of contextual information in long sequences - an issue prevalent in the previously dominant Recurrent Neural Networks (RNNs). This attention mechanism has revolutionized NLP. It also introduced bidirectional representation learning, which enables models to incorporate context from both preceding and succeeding tokens - a capability not feasible with traditional sequential models such as RNNs. Another critical feature of Transformers is transfer learning, which mitigates the challenge of insufficient training data by leveraging knowledge acquired during pretraining. This underscores the importance of domain-specific pretrained models, which apply pre-acquired knowledge to specialized tasks without extensive training on large labeled datasets, a frequent limitation. Consequently, numerous domain-specific models, including biomedical models, have been developed. These models demonstrate higher performance on domain-specific tasks than general-domain models [71], but retaining general knowledge in these models can also provide benefits.

Biomedical pretraining leverages several key data sources:

- **PubMed:** This database contains abstracts of biomedical publications and links to full-text articles from PubMed Central (PMC), as discussed in Section 4.1.1.
- **MIMIC-III (Medical Information Mart for Intensive Care III):** This database includes Electronic Health Records (EHRs) from 58 976 hospital admissions involving over 38 000 patients, collected between 2001 and 2012 at the Beth Israel Deaconess Medical Center [72].
- **UMLS (Unified Medical Language System):** This repository contains biomedical terminologies, providing information on the synonyms and relationships of medical concepts, as detailed in Section 4.1.2.

There are two types of biomedical Transformer models, which differ in their pretraining setup: mixed-domain pretraining and domain-specific pretraining from scratch [71]. This section explores BERT-based models from both categories.

4.3.1 Mixed-domain pretrained language models

Mixed-domain pretrained models utilize a continual pretraining approach, beginning with weights from an existing pretrained model like BERT and continuing self-supervised learning on domain-specific corpora. This method allows models to retain general language knowledge from the initialization of the training process. Examples of biomedical language models trained with that approach include:

- **BioBERT:** Initialized with BERT's weights and further pretrained on PubMed abstracts and PMC full-text articles [73].
- **ClinicalBERT:** Initialized with BioBERT's weights and further pretrained on clinical notes from MIMIC-III [74].
- **BlueBERT:** Initialized with BERT's weights and further pretrained on both PubMed abstracts and MIMIC-III clinical notes [75].
- **CODER:** Initialized with PubMedBERT's weights (model referenced in Section 4.3.2) and further pretrained on UMLS concepts and relations [76].

4.3.2 Domain-specific language models pretrained from scratch

In contrast, domain-specific language models pretrained from scratch do not begin with any pretrained language model. They start their learning process with randomly initialized weights. This approach requires the models to acquire language knowledge directly from their domain-specific training data, since they do not have any general language understanding at the beginning. Examples of these models include:

- **SciBERT:** Utilizes a non-initialized BERT architecture and is pretrained on a random sample of full-text scientific publications from Semantic Scholar, with a mix of biomedical (82%) and computer science (18%) articles [77].
- **PubMedBERT:** Utilizes a non-initialized BERT architecture and is pretrained on PubMed abstracts and PMC full-text articles [71].

4.4 Generative AI

Currently, a particular group of language models has attracted significant attention from society. These are generative models based on the Transformer Decoder architecture, trained with the objective of next word prediction, also referred to as Causal Language Models. The latest GPT (Generative Pretrained Transformer) models, such as GPT-3 [78] and especially GPT-4 [79], have achieved state-of-the-art performance in numerous NLP tasks. Their results are comparable to human-level performance in natural language understanding and reasoning. The strength of these models lies in their vast training data and the number of their parameters - 175 billion parameters in GPT-3 and trillions of parameters in GPT-4 (exact number undisclosed). This is why these models are called large language models (LLMs). In contrast, BERT base model has 110 million parameters, and its large version has 340 million parameters. Additionally, the recent GPT models (GPT-3.5 and GPT-4) utilize Reinforcement Learning with Human Feedback (RLHF), where human reviewers score outputs from a pretrained model, and those scores are used in model fine-tuning [80]. This method helps guide the model towards more accurate and human-like outputs.

The GPT-3.5 model, a modified version of GPT-3, forms the core of the first ChatGPT version, released in 2022. It quickly gained interest of people around the world, allowing users to interact with it via its chatbot interface at chatgpt.com. ChatGPT, an LLM conversational tool, was trained on a huge corpus of text data, including books, articles and websites [81]. It has learnt various phrases, relations between them and contexts, hence it is able to generate human-like responses in conversations [82].

Recent generative models are capable of performing numerous NLP tasks, such as text classification, question answering, machine translation, and summarization, without additional training. They address the issue of limited task-specific training data. These models have these capabilities because they learn a vast amount of linguistic features and patterns from the data they are pretrained on, and they are able to generalize this knowledge to new tasks and domains.

Additionally, large language models offer a new mode of interaction known as prompt engineering, which is a technique of deliberately crafting an input to a language model in order to achieve a desired output. A prompt in generative models is the textual input provided by the user to the model, which leads to content generation [83]. It can be a simple question, or a detailed instruction with input data and examples. By giving the model a specific prompt, one can control and guide the output of the model to generate relevant and meaningful results. More precise prompts result in better outcomes, making prompt engineering an important field of research in AI. Many prompt patterns have been discovered, including *Persona*, *Alternative Approaches*, or *Flipped Interaction* patterns, which increase the accuracy of the generated output [84].

The development of GPT-3 and especially the ChatGPT tool started a new chapter in AI development, a Generative AI era. They have inspired new research directions in Artificial Intelligence and prompted the industry transformations, with more and more LLM-based systems being utilized in many applications. The release of GPT-4 in 2023, which surpasses human-level performance in many tasks, demonstrated that these models could serve as human assistants across various domains. Currently, an increasing number of competitors are releasing more LLMs. Fig. 4.7 illustrates the evolution of generative models and the rising number of new releases in recent years.

Despite the significant performance, large language models are not without limitations. Some of them include [83]:

- **Hallucinations:** They can generate incorrect content that sounds truthful but has no basis in facts.
- **Knowledge cut-off:** Their knowledge is limited to their training dataset; they do not have access to data that appeared after their pretraining.
- **Inconsistency in generated output:** As probabilistic models, their outputs may vary for the same prompt.
- **Limited context window:** Their architecture limits the number of tokens they can process in a single input, e.g. 16 385 tokens for GPT-3.5 and 128 000 tokens for GPT-4 [85].

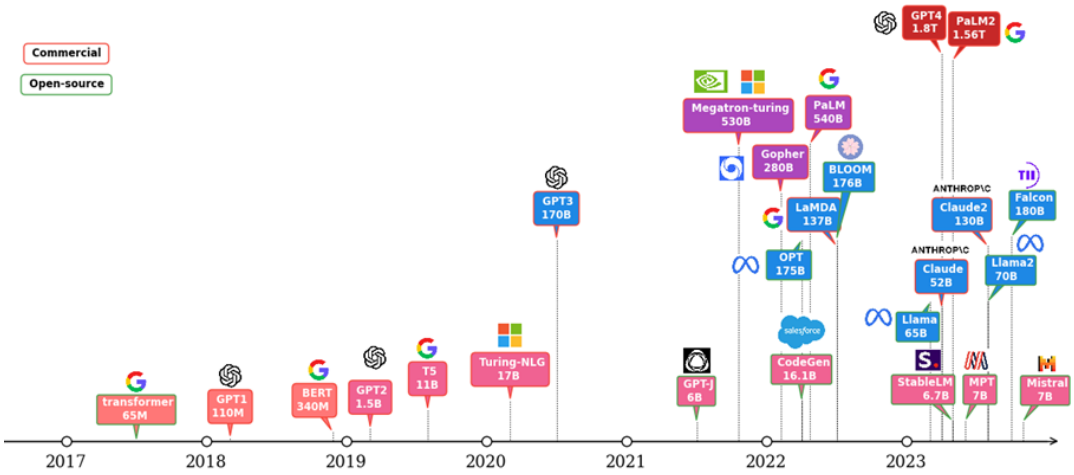


Figure 4.7: The evolution of large language models. Source: https://bit.ly/3WHHej8

- **No continuous retraining:** They are static models that do not learn from experience.
- **Costly execution:** Their large size (billions of parameters) results in suboptimal performance and may require substantial infrastructure.

A deep understanding on how LLMs work, combined with well-crafted prompts and careful checking of the generated output, can mitigate some of these limitations.

Chapter 5

Natural Language Processing in Clinical Trial Eligibility Criteria Parsing

As discussed in Chapter 2, one of the primary obstacles to the success of clinical trials is the recruitment process, which is highly inefficient, labor-intensive, time-consuming, and prone to human errors. Automating this process could significantly benefit patients and the healthcare industry. However, a major challenge in this automation is parsing the eligibility criteria. Using Machine Learning (ML) and Natural Language Processing (NLP) offers a potential solution, although the complexity of the syntax and semantics of the eligibility criteria presents significant challenges for existing ML and NLP methods.

In recent years, we have witnessed an explosion of ML and NLP models capable of streamlining the patient accrual process. This chapter presents the most comprehensive scoping review of the scientific literature on the use of machine learning and natural language processing for parsing eligibility criteria from clinical trial protocols. The survey, conducted strictly according to the PRISMA methodology, covers papers published between January 1st 2020 and July 2nd 2024. The search queries executed against four databases resulted in the initial pool of 9160 potentially relevant publications. The final scoping review contains 88 publications analyzed using 17 criteria. Full results of the scoping review are published as an interactive dashboard and available at https://drudis-d-23-00233r1.streamlit.app.

Section 5.1 discusses previous scoping reviews on this topic. Section 5.2 outlines the methods used to conduct this review, including the research questions, search strategy, study selection, and data charting. Section 5.3 presents the results of the publication selection process and the main findings from the review. Section 5.4 discusses the research questions in light of the study findings, lists the limitations of the review, and identifies potential future research directions.

5.1 Previous reviews

In the last few years, several reviews on the use of machine learning and artificial intelligence in clinical trial parsing have been published. Table 5.1 summarizes their scope and design protocols. Below we describe these studies, and, in particular, discuss their shortcomings that prompted us to prepare the current scoping review.

Idnay *et al.* [86] conducted a systematic review focused on the use of NLP systems in eligibility prescreening. They searched five databases and selected 11 publications that evaluate ready-to-use NLP systems. These studies were required to pertain to clinical eligibility prescreening tasks, involve real patient medical records, and include a non-NLP baseline for comparison. The authors assessed the publications' quality and their impact on the clinical process, noting that the limited number of studies might suggest this application of NLP is understudied.

Askin *et al.* [87] reviewed the application of AI and ML thorough all stages of clinical development - from pre-clinical research to trial design, recruitment, conduct, analysis, and regulatory documentation. They identified 48 publications and 9 regulatory documents from Europe and the USA, published between 2017 and 2021. Their review broadly examines the use of AI in clinical trials, discussing the opportunities, challenges, and implications for practice but does not delve into specific AI methods and technicalities.

Bhatnagar *et al.* [88] analyzed NLP applications in drug development and discovery, focusing on different NLP libraries and models. Beyond traditional publication databases, they also searched code repositories. They identified 20 possible solutions and stratified them according to drug development stages, which resulted in 12 applications used in the patient-trial matching area. The authors highlighted the need for more research into model explainability in drug development applications.

Kim and Quintana [89] investigated the evaluation metrics of NLP systems for clinical trial matching based on five publications. They noted difficulties in comparing tool performance due to the lack of a gold standard for trial eligibility matching applications. Of the systems reviewed, four were assessed manually by multiple raters. The review emphasized the need for creating a standardized benchmark for evaluating patient-trial matching algorithms.

Su *et al.* [90] reviewed over 60 papers on the use of computer technologies in processing eligibility criteria, categorizing them into four groups: natural language processing, patient pre-screening, eligibility criteria evaluation, and clinical trial query. Dominated by AI applications, the review identifies significant challenges, which include the need for interdisciplinary expertise, data availability, and large-scale testing.

A common limitation among these studies is their narrow focus and the small number of publications reviewed. The review by Idnay *et al.* [86] is the most relevant to the current work but it is much more narrow in scope and does not address the main question informing this study: does the current revolution in machine learning and natural language processing impact the field of clinical development?

Table 5.1: Previous reviews.

Title	Objective	Databases	# art.	Search date	Filters
A systematic review on natural language processing systems for eligibility prescreening in clinical research.	To review the use of NLP systems in eligibility prescreening.	PubMed, Embase, CINAHL, IEEE, ACM	11	Feb 2021	use of real patient data; non-NLP baseline; evaluation of NLP system ready to use in real-world setting; business task: eligibility prescreening
Artificial Intelligence Applied to clinical trials: opportunities and challenges.	To review the use of ML and AI in clinical trials.	PubMed, Scopus, International Pharmaceutical Abstracts, Google Scholar	48	Oct 2021	publication year: since 2017; research area: EU and US; referring to AI or ML applications in clinical trials
How can natural language processing help model informed drug development?: a review.	To review NLP applications in drug discovery and development.	PubMed, Google Scholar, GitHub	20	Feb 2022	publication year: since 2010; programming languages: Python, Java, Scala, C++, R; transformer models pre-trained on biomedical corpora; NLP systems not involving speech analysis or generation
Review of the Performance Metrics for Natural Language Systems for Clinical Trial Matching.	To review the performance metrics of NLP systems for clinical trial matching.	PubMed, ACM, IEEE, ScienceDirect	5	May 2021	publication year: since 2016; direct evaluation of CT matching system performance; definition of gold standard
A review of research on eligibility criteria for clinical trials.	To review the use of computer science in CT eligibility criteria processing.	Google Scholar, arXiv, Nature, NIH	> 60	not specified	not specified

5.2 Methods

Scoping reviews are considered a valid method of summarizing knowledge in science [91]. By presenting a broad overview of the existing literature, scoping reviews allow to identify key concepts, methods, datasets, and dominating research themes in a specific domain. More importantly, scoping reviews help clarify definitions and conceptual frameworks, thus allowing to identify gaps in the research. This is particularly useful in complex or emerging fields where it might be difficult to carry out a systematic review. Finally, scoping reviews can be done more rapidly than systematic reviews, making them practical for informing decision-making and research design. They serve as valuable tools in the scientific research process, providing a comprehensive and accessible summary of existing literature. This scoping review follows the PRISMA methodology outlined by Khan *et al.* [92].

5.2.1 Questions informing the scoping review

This study presents a broad scoping review on the use of machine learning and natural language processing tools for parsing clinical trial eligibility criteria. As it was already mentioned, the failure to match sufficient patient cohorts to eligibility criteria is the primary cause of unsuccessful clinical trials. We see machine learning and natural language processing as powerful tools that can address this problem. Recent advances in the area of machine learning, particularly the emergence of generative large language models, might indicate the beginning of a new chapter in automated patient-trial matching, significantly boosting the success rate of clinical trials.

In order to design the scoping review, a systematic approach must be developed for the selection of features extracted from the literature search results. One solution is to design a set of questions that will inform the scoping review, in particular, the selection of sources, the design of search queries, and the development of the data charting protocol. Questions selected for the presented scoping review are as follows:

- **Question 1**: How are recent generative language models being adopted for clinical trial eligibility criteria parsing?
- **Question 2**: What NLP methods, models, and tools are commonly used in patient-trial matching?
- **Question 3**: What resources are currently lacking or insufficient for the widespread adoption of ML and NLP techniques in patient-trial matching?

5.2.2 Search strategy

Four scientific literature databases have been selected as sources for potential publications: Scopus, EMBASE, Web of Science, and PubMed. The search query has been designed in such a way as to cover publications discussing ML or NLP in the context of clinical trials. A publication must have met all the following criteria to be considered for inclusion in the review:

- The paper must address patient-trial matching: title or abstract must include any of the following terms: *enrollment, enrolment, recruitment, screening, criteria, eligibility, matching, cohort selection, cohort ascertainment, accrual, prescreening, phenotyping, population enrichment*.
- The title or abstract must contain the word *trial*.
- The paper must mention ML or NLP: title or abstract includes any of the following terms: *transformer, bert, named entity recognition, NER, textual entailment, natural language inference, language model, semantic similarity, semantic textual similarity, entity linking, information extraction, natural language processing, NLP, deep learning, machine learning, attention mechanism, generative AI, GPT*.
- The paper must have been published after the year 2000.
- The paper must be written in English

5.2.3 Study selection

In order to select publications for inclusion in the scoping review, two annotators independently screened titles and abstracts using the Prodigy annotation tool and a custom-built annotation recipe, as illustrated in Figure 5.1. The interface displays the title, the DOI, and the abstract of a paper. Annotators can accept or reject the paper using buttons or keyboard shortcuts. They also have the capability to navigate backward in the annotation history to correct any previous annotations. For more clarity, keywords from the literature search string are highlighted. To avoid selection bias, the annotators were not given access to the publication's metadata, such as the list of authors or the venue of publication. After study selection the publications were randomly assigned to one of the annotators for data charting.

5.2.4 Data charting

The purpose of data charting is to establish and follow a protocol for extracting relevant information from the selected studies. This process involves the selection of features and the categorization of their values, which must be agreed between annotators. The selected features and their categorization are designed to assist in answering the research questions that define the scoping review. Table 5.2 summarizes the features extracted from the studies, provides the rationale behind each feature, and indicates the specific research question each feature addresses.

5.3 Results

5.3.1 Search results

Figure 5.2 presents the flowchart of the literature selection process. The initial search yielded 9160 papers, with the largest number of papers coming from the Scopus

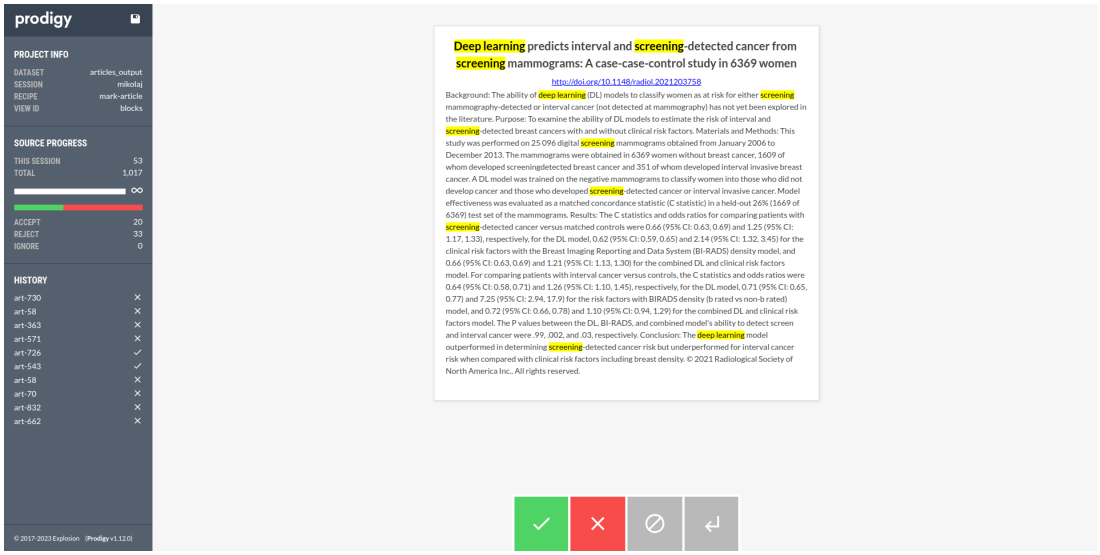


Figure 5.1: Annotation of publications in Prodigy.

database (3220), followed by Embase (2436), Web of Science (1934), and PubMed (1543). In the first step, papers lacking a DOI or an abstract were removed (480 positions), which resulted in 8680 remaining papers. In the subsequent deduplication step, which considered DOI, abstract, and a combination of title and publication year, the number was reduced to a final pool of 4249 publications viable for the scoping review screening.

Out of the pool of papers accepted for screening, 4,049 were rejected after title and abstract screening, leaving 200 papers to the full reading phase. The inter-annotator agreement, measured by Cohen's Kappa, was substantial ($\kappa = 0.767$) [93]. The selected papers were then divided between two researchers for data charting. During the full reading phase, 112 publications were rejected from the scoping review. The main reasons for paper rejection at this late stage were as follows:

- The publication was only available as a short abstract or conference communication.
- The publication did not contain any NLP methods.
- The publication did not contain any ML methods.
- Only the abstract was in English, the publication was in a language different than English.
- The publication did not mention clinical trial eligibility criteria.
- The text of the publication was behind a paywall (3 publications)

Finally, 88 publications (83 papers and 5 reviews) were accepted into the current scoping review.

Table 5.2: Features extracted from studies.

Feature	Description	Rationale	Question
year	year of publication	to observe the dynamics of publications on NLP in CT parsing	Q1
author country	the country of affiliation of the authors	to identify leading countries conducting research in the field of NLP in CT parsing	Q1
research country	the country where research was conducted	to identify leading countries conducting research in the field of NLP in CT parsing	Q1
therapeutic area	the area of medicine of clinical trials	to identify dominating medical areas of research	Q1
general type	the general characteristics of the paper (research, review, etc.)	to better understand research on NLP in CT parsing	Q1
business objective	the main medical goal of the research	to better understand the needs of medical community	Q1
NLP methods	main NLP techniques, tools, and tasks	to better understand research on NLP in CT parsing	Q2
metrics	main metrics used to evaluate and report research results	to better understand research on NLP in CT parsing	Q2
datasets	public datasets with EHRs, CTs, and ECs	to make inventory of NLP resources in the area of clinical trial parsing	Q3
protocol source	public databases of CTs	to make inventory of NLP resources in the area of CT parsing	Q3
auxiliary datasets	additional public databases of EHRs, CTs, and ECs	to make inventory of NLP resources in the area of CT parsing	Q3
knowledge bases	public databases of medical terms	to make inventory of NLP resources in the area of CT parsing	Q3
contribution	the main scientific contribution of the paper	to better understand research on NLP in CT parsing	Q2
ground truth	the method of creating ground truth in model evaluation	to make inventory of NLP resources in the area of CT parsing	Q3
preprocessing	main NLP methods of data preprocessing	to better understand research on NLP in CT parsing	Q2
shallow ML methods	main non-neural ML tools and techniques reported in the paper	to better understand research on NLP in CT parsing	Q2
neural models	pretrained and fine-tuned language models used for research	to better understand research on NLP in CT parsing	Q2

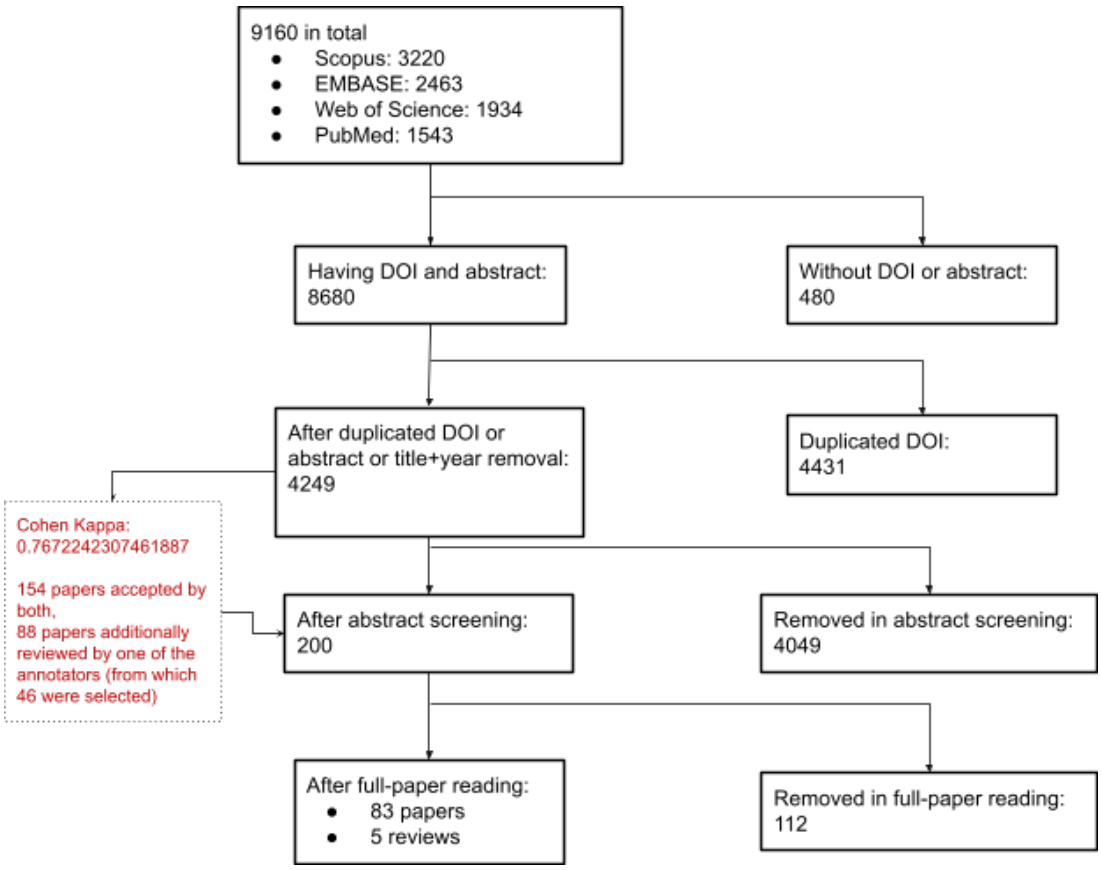


Figure 5.2: Flowchart of the literature selection process.

5.3.2 Study demographics

This subsection presents the general characteristics of the papers included in the review. It focuses on demographic data, such as the year of publication, country of origin of the studies, and the general character of the papers. This allows to assess the scope and the breath of the scoping review.

5.3.2.1 Year of the publication

Figure 5.3 illustrates the number of works included in the study published each year. The interest in applying modern ML & NLP techniques to clinical trial descriptions has steadily grown in recent years. A significant surge in the number of papers published in 2021 can be attributed to advancements in NLP, particularly the emergence of foundational models and the increased availability of domain-specific smaller models.

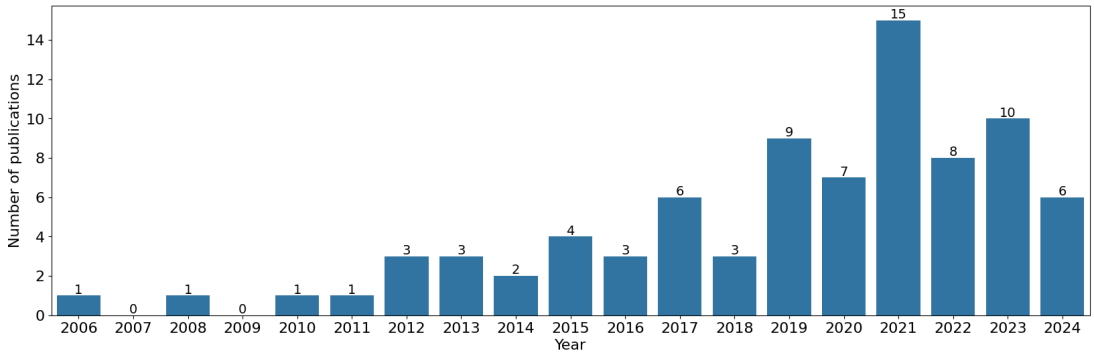


Figure 5.3: Number of publications per year.

5.3.2.2 Main contribution

Most studies focus on the presentation of new methods and algorithms (Figure 5.4), which aligns with general publication bias that encourages novelty elements in accepted studies. A large number of studies evaluate pretrained models or healthcare systems in narrow tasks (efficacy of pretrained BERT models in semantic similarity matching of eligibility criteria [94], the performance of IBM Watson in matching prospective subjects to a cancer study [95], etc.). Additionally, there are contributions in the form of dataset curation (Chia – annotated dataset of eligibility criteria [96], Leaf clinical trials corpus [97], the knowledge base of clinical eligibility criteria [98]).

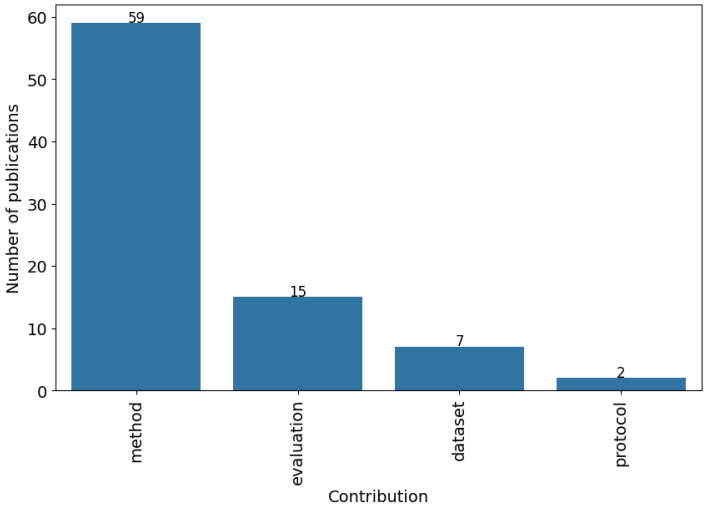


Figure 5.4: Main contribution of the study.

5.3.2.3 Author country

A majority of the studies are published by authors affiliated with American institutions (Figure 5.5), which confirms the USA's dominance in biomedical and AI re-

search. However, this distribution may be skewed by the review's selection criteria, which limit the scope to papers published in English. Notably, original works written in other languages, in particular in Chinese, were excluded from the review.

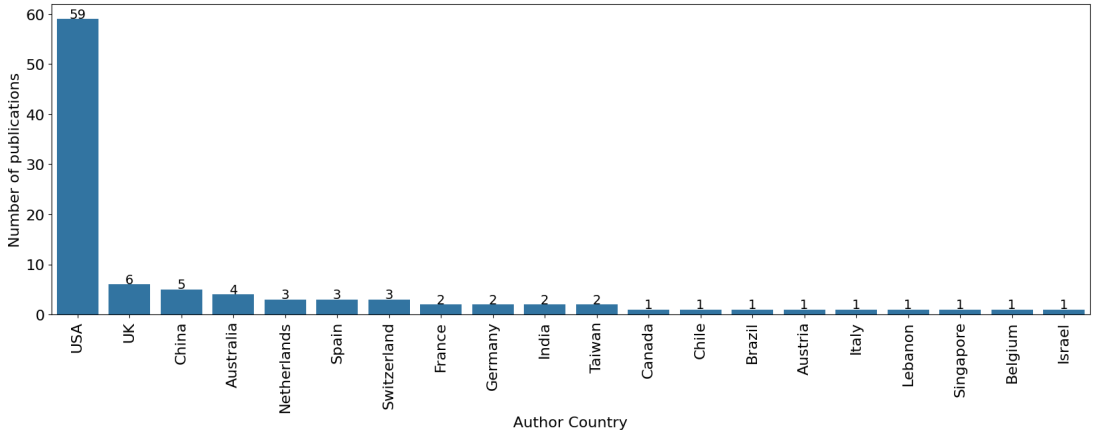


Figure 5.5: Country of affiliation of authors.

5.3.2.4 Therapeutic area

The papers span a broad range of therapeutic areas (Figure 5.6). Most discuss general-purpose methods that are not limited to a particular disease or area of medicine (39 papers). However, a significant number focus on developing methods, datasets, rules, and frameworks for narrowly defined areas. Oncology is the most frequently addressed field (25 papers in total), with papers particularly focusing on breast cancer (10 papers) and pediatric oncology (3 papers). Alzheimer's disease is also a common topic of study (9 papers). As the collection of papers for the review has been conducted shortly after the coronavirus pandemic, the review includes 5 papers related to COVID-19 clinical trials.

5.3.2.5 Research country

Figure 5.7 summarizes the countries in which the reviewed studies have been conducted. Most studies are universal and not limited to any specific country, or the papers do not explicitly limit their conclusions to a particular country. However, when geographic limitations are mentioned, the USA is frequently the country of research, most likely due to the particular format of Electronic Health Records processed in these studies.

5.3.2.6 General character of the paper.

Each paper included in the scoping review has been broadly classified based on its general character. As illustrated in Figure 5.8, the majority of publications (35 papers) are research papers that introduce new methods, algorithms, and solutions.

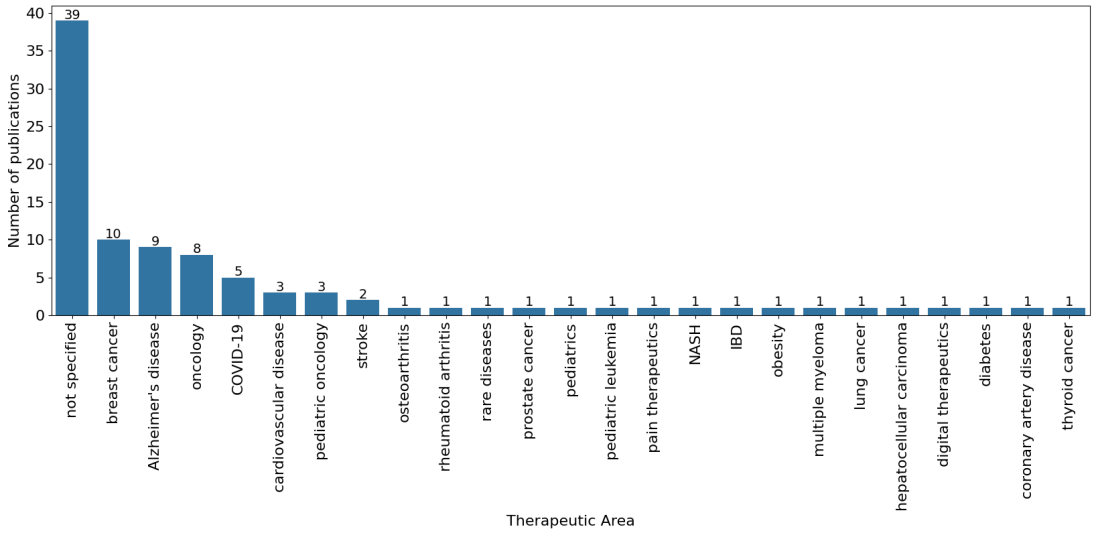


Figure 5.6: Therapeutic area discussed in the paper.

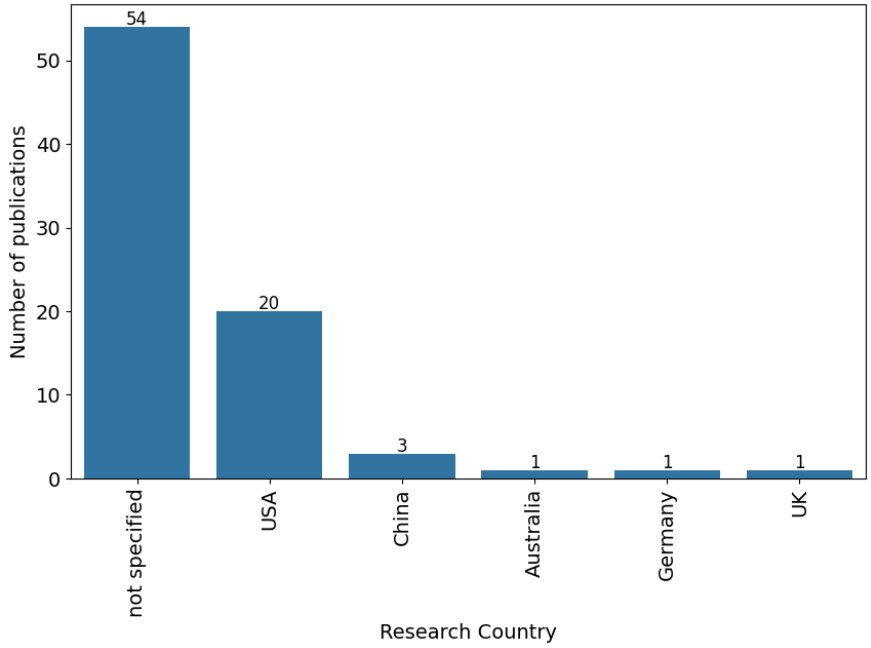


Figure 5.7: Country of research.

The second most frequent type of publication is model evaluation (24 papers), which typically involves applying a particular pretrained language model to clinical trial eligibility criteria. Additionally, a number of studies discuss the efficacy of specific software for EHR processing, patient-trial matching, medical text normalization, etc. These works were categorized as software presentations (18 papers). Finally, 6 papers introduce new datasets relevant to ML/NLP, such as annotated sets of eligibility criteria, new ontologies of medical concepts, and lexicons of relevant terms.

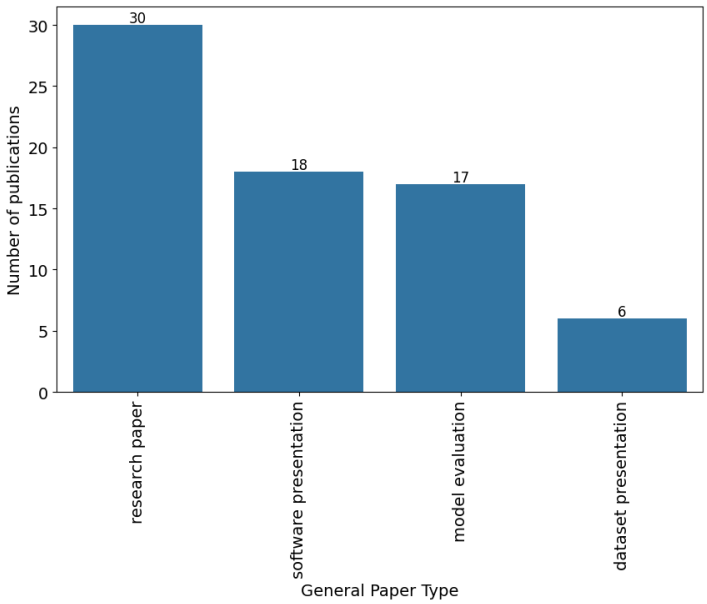


Figure 5.8: General character of the paper.

5.3.3 Study findings

This section presents the results from the data charting process, with features categorized as described in Table 5.2.

5.3.3.1 Business objective

A critical aspect of the practical application of machine learning is selecting the main business objective and the subsequent alignment of both training and evaluation metrics with the relevant business metrics. This scoping review assesses how modern NLP models and tools assist medical practitioners in successfully conducting clinical trials. Evaluation of the business objectives in the included works supports the hypothesis that the primary business goal in this area is to enhance the success rate of clinical studies by focusing on patient-trial matching. As illustrated in Figure 5.9, the majority of studies explicitly state that improving patient-trial matching is their main business objective. The second most common business objective is clinical trial description parsing, which is a broader objective serving multiple practical purposes. The third business objective supports medical practitioners in designing clinical trials, in particular, by assisting the designers to specify better eligibility criteria based on successes or failures of previous clinical trials.

5.3.3.2 NLP tasks

Modern NLP models address a wide range of tasks, including text classification, sentiment analysis, relational entailment, entity recognition, entity disambiguation, and linking, to name a few. Despite these capabilities, the adoption of modern NLP

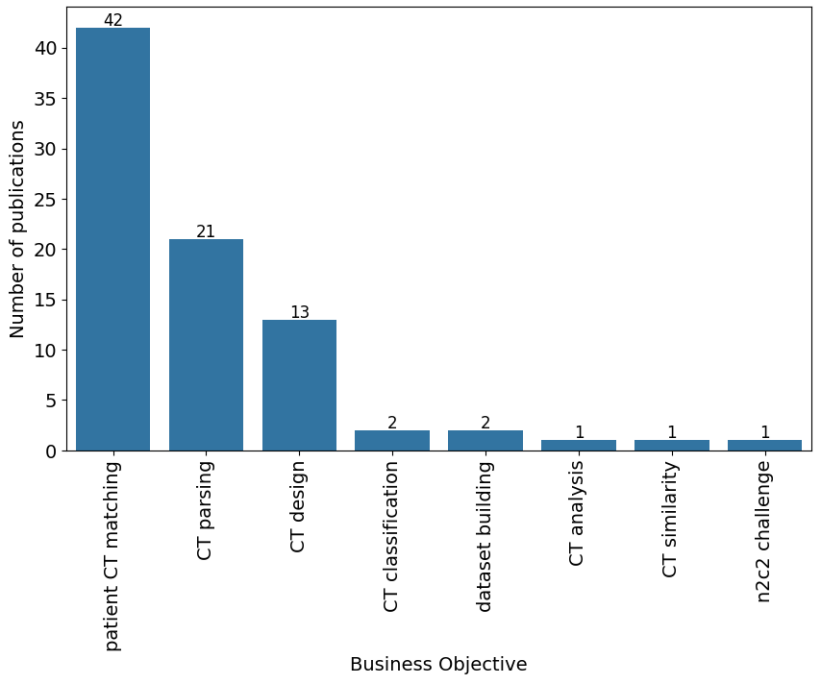


Figure 5.9: Main business objective.

in the reviewed publications is not widespread. As documented in Figure 5.10, the majority of studies focus on named entity recognition (NER) as the main NLP task. This is not surprising given the importance of identifying medical terms in free-form clinical trial descriptions. Many studies target the patient-trial matching objective through text normalization, aiming to identify relevant terms in trial descriptions, eligibility criteria, and Electronic Health Records, and perform matching based on this extracted knowledge. The second most frequent NLP task is general information retrieval (IR), which involves non-specific searches within clinical trial descriptions. More advanced NLP tasks include representation learning (RL), a task of using pre-trained language models to generate dense numerical text representations for better semantic matching, text classification, and entity linking (EL). However, these tasks are not the most common applications of NLP tools.

5.3.3.3 Datasets

The standardization of datasets with eligibility criteria is minimal, as illustrated in Figure 5.11. Most researchers either do not specify the datasets, or describe them as internal datasets. Usage of publicly available datasets, such as ELilE, Covance, or MUSC warehouse, is incidental. The datasets used across several studies are the Chia dataset and the N2C2 dataset.

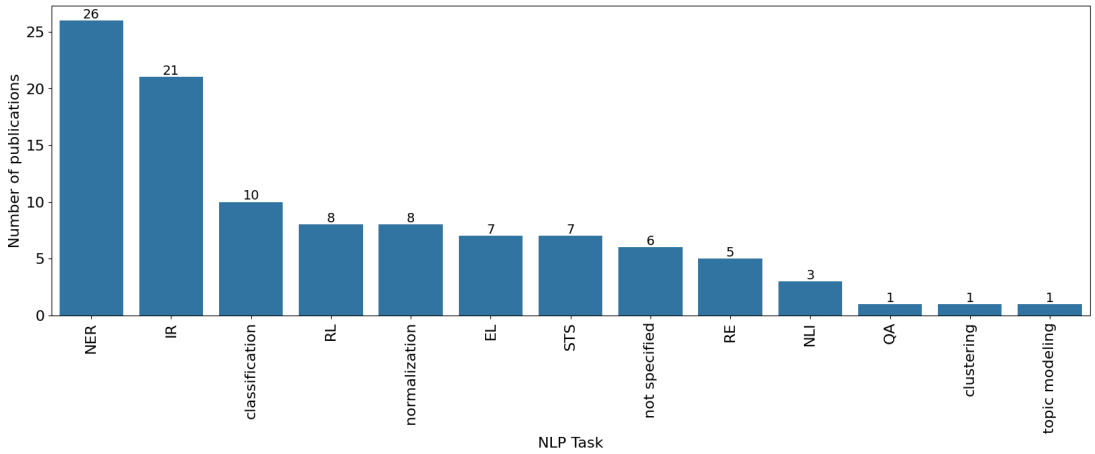


Figure 5.10: NLP tasks.

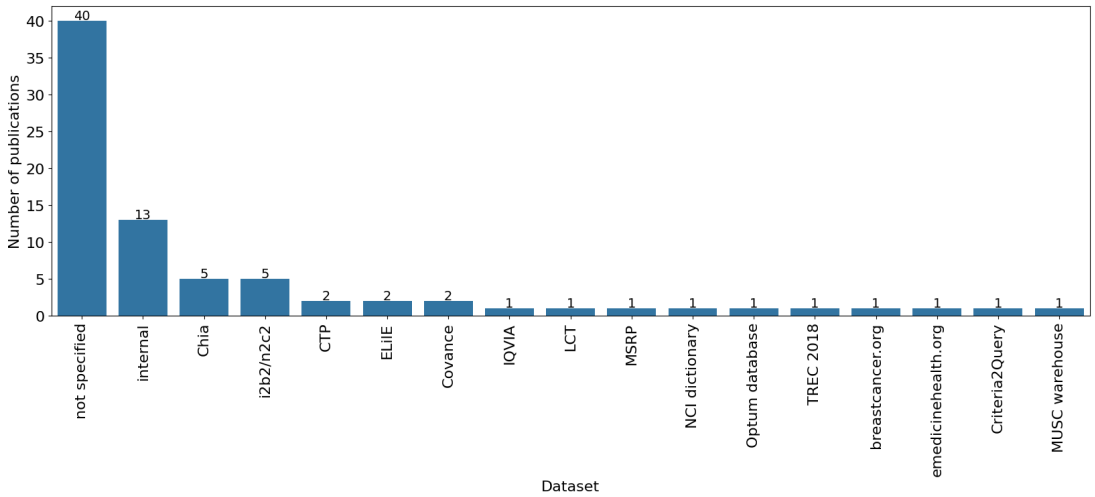


Figure 5.11: Datasets used in studies.

5.3.3.4 Protocol source

In contrast to the lack of annotated datasets, there is a high degree of standardization in the source of clinical trial protocols, as shown in Figure 5.12. The online database ClinicalTrials.org is the largest collection of clinical trial and observational study descriptions, and is the dominant source of clinical trial records. Clinical trials available in this database are semi-structured and significant work is required to extract more structured data, e.g., regarding particular types of inclusion or exclusion criteria. Therefore, some studies use internal collections of protocols that have been further preprocessed and annotated.

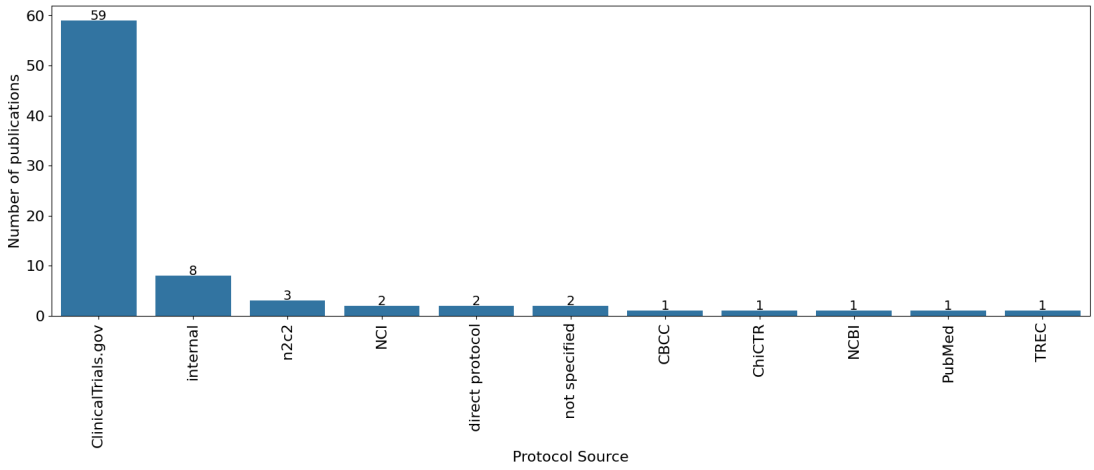


Figure 5.12: The source of clinical trial protocols.

5.3.3.5 Auxiliary datasets

Throughout the data charting process, various auxiliary datasets were identified, as depicted in Figure 5.13. Most of these datasets contain Electronic Health Records from different patient cohorts, including general EHR datasets and the MIMIC-III dataset. A few auxiliary datasets are utilized for normalizing medical terms, for instance, by providing additional information on drug-drug interactions or general drug information. Term normalization is also supported by domain-specific dictionaries the National Cancer Institute dictionary. Annotated examples of patient-trial matching, such as those provided by IQVIA, represent another category of auxiliary datasets. However, it is important to note that the use of these auxiliary datasets is quite limited in the existing literature.

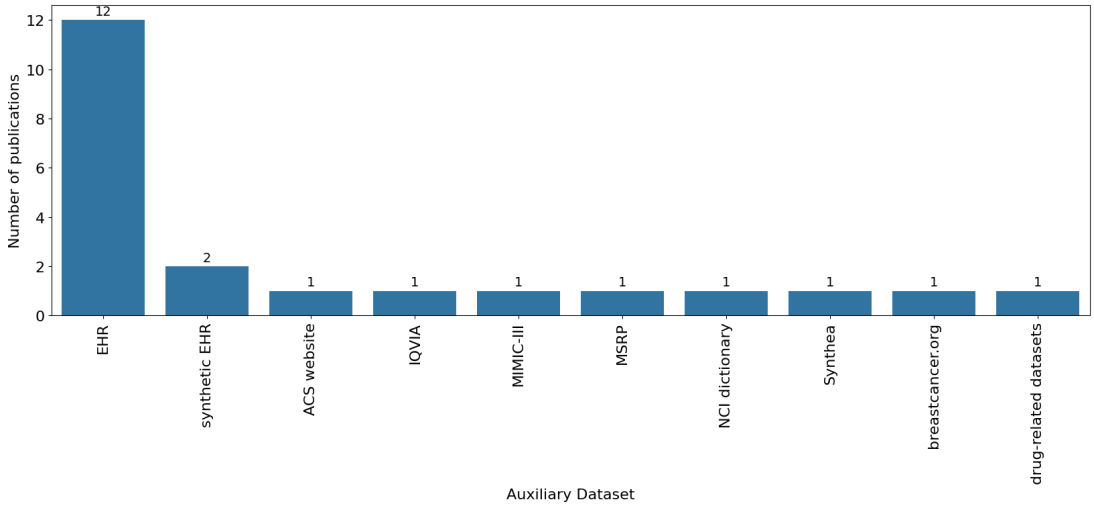


Figure 5.13: Auxiliary datasets used in studies.

5.3.3.6 Knowledge bases and ontologies

This section differentiates between auxiliary datasets, previously discussed, and the use of knowledge bases, which are highly structured repositories of normalized medical information, including ontologies and ontology-like vocabularies. These resources are utilized in many studies for tasks such as term disambiguation, entity linking, and semantic similarity evaluation. The most widely used knowledge base is the Unified Medical Language System (UMLS) mentioned in 18 studies (Figure 5.14). The SNOMED Clinical Terms collection, another comprehensive set of medical terms, codes, synonyms, and definitions, is the second most utilized resource. Other popular resources for processing clinical trial descriptions include DrugBank, the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), the International Classification of Diseases (ICD), and the RxNorm vocabulary of drug names.

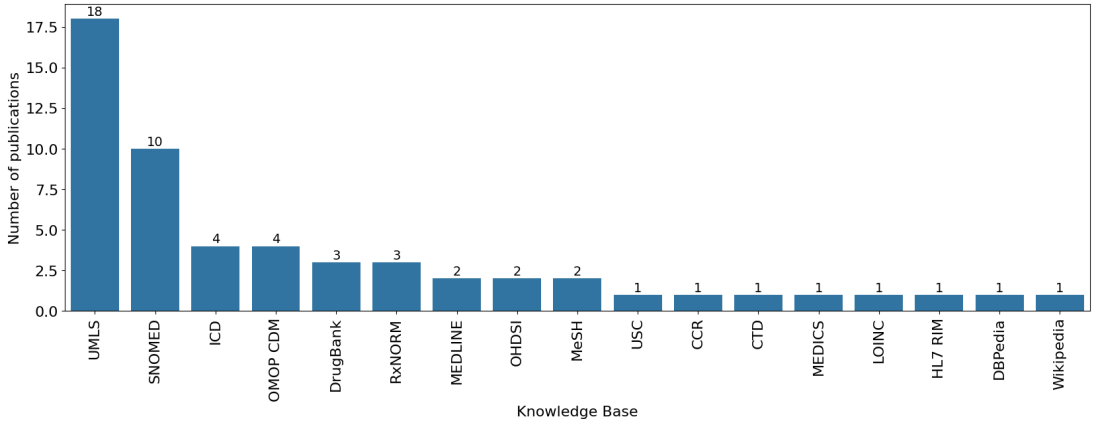


Figure 5.14: Knowledge bases and ontologies used in studies.

5.3.3.7 Ground truth

The dominant method for establishing ground truth in reported experiments is the manual curation of the test dataset (Figure 5.15). Evaluations are often performed through post hoc annotation of results, where the output of an algorithm is assessed by human annotators. Only a few studies use trial descriptions from ClinicalTrials.org as the source of ground truth. Some studies, typically originating from research groups affiliated with commercial entities, employ internal benchmarks as the ground truth.

5.3.3.8 Preprocessing

Almost every study included in the scoping review describes some form of text preprocessing. The most common form of preprocessing clinical trial descriptions is text and term normalization (Figure 5.16). *Text normalization* encompasses a variety of techniques, including stop-word removal, identification of measurement unit

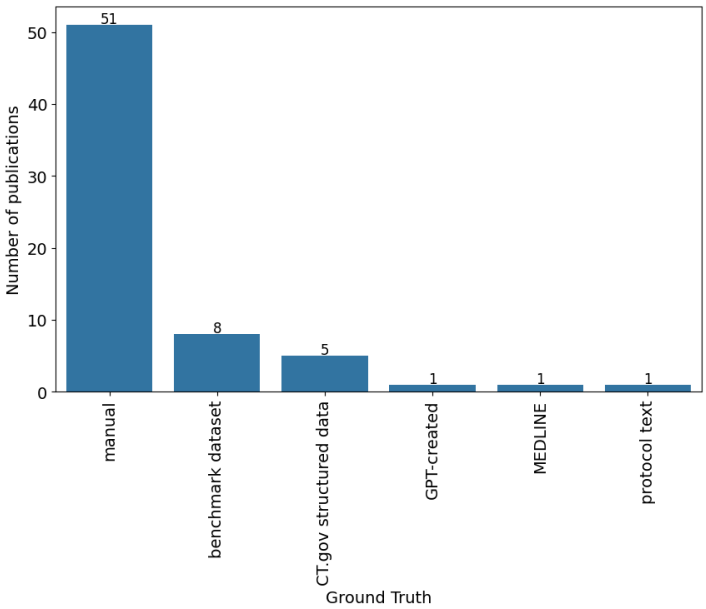


Figure 5.15: Methods for establishing ground truth.

markers, correcting the casing of names, converting numbers to a common format, expanding common abbreviations, etc. *Term normalization*, on the other hand, specifically refers to the disambiguation of medical terms - such as names of drugs, diseases, medical procedures, and conditions - often supported by external knowledge bases, ontologies, and dictionaries. Additionally, sentence segmentation (i.e. the discovery of sentence boundaries) is an often employed technique. Beyond the traditional tasks of part-of-speech tagging, text vectorization, and keyword filtering, the studies also commonly use regular expressions and identification of negation markers.

5.3.3.9 Shallow machine learning methods

A large variety of ML models have been utilized in the papers included in the study, but unquestionably the most common method can be described as custom rule-based models. This general term encompasses studies where the researchers present processing pipelines with significant manual feature and model engineering. This is quite surprising given recent advances in the design and development of end-to-end models, particularly in the field of NLP. Also, the common use of dated methods, such as Support Vector Machines (SVM) and logistic regression suggests that many studies in the field rely heavily on statistical software packages with ready implementations of these methods. These findings, evidenced in Figure 5.17, underscore a gap between cutting-edge research and practical application in the field.

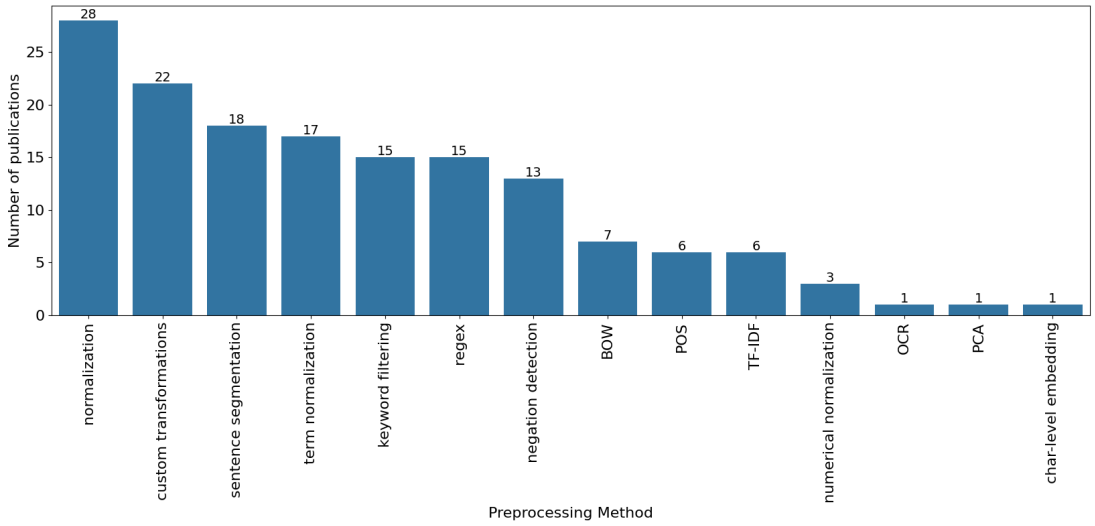


Figure 5.16: NLP preprocessing methods.

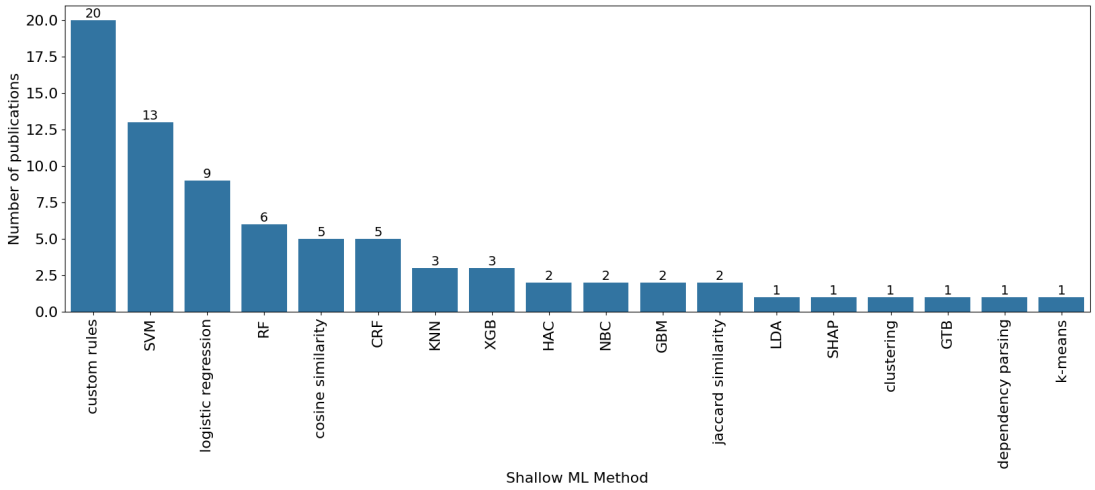


Figure 5.17: Shallow machine learning models.

5.3.3.10 Neural models

Before starting the scoping review, it was anticipated that there would be numerous studies utilizing modern neural language models due to a Cambrian-like explosion of pretrained language models (such as BERT, RoBERTa, XLNet), pretrained word embeddings (word2vec, GloVe, FastText), and large language models (GPT, Gopher, Codex), along with their versions fine-tuned for the medical domain (BioBERT, ClinicalBERT, PubMedBERT, BioRoBERTa, BioBERT-CRF, MedPaLM). However, it was found that neural models are relatively scarce in the literature (Figure 5.18), with their usage mostly limited to pretrained embeddings (most frequently word2vec) or to the original BERT, a model released back in 2018, whose capabilities have been long surpassed by much larger and more robust language models. Only 4 studies

used generative large language models such as Llama or GPT.

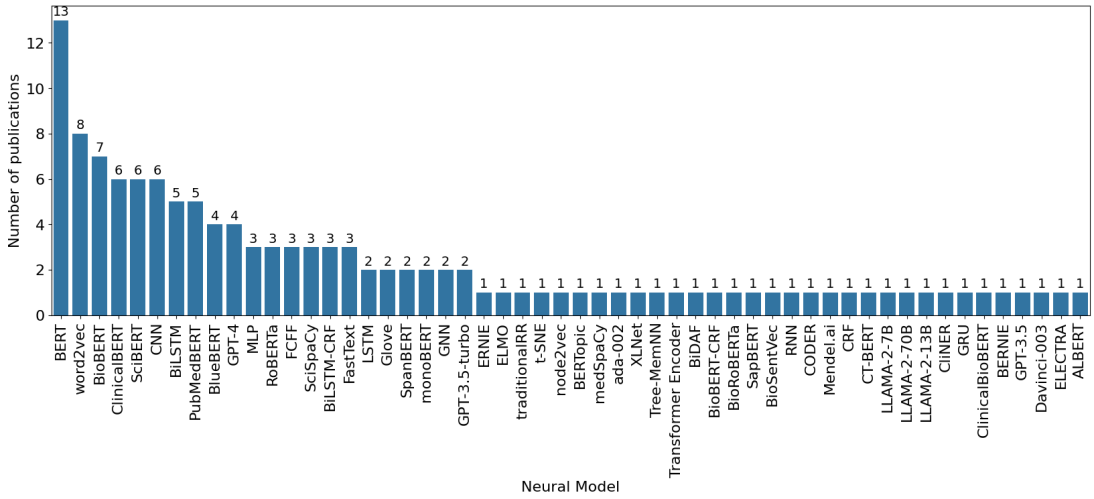


Figure 5.18: Deep neural models.

5.3.3.11 Number of protocols and criteria

Figures 5.19 and 5.20 display the distributions of the numbers of protocols and criteria processed in the analyzed studies, respectively. Three general types of studies emerge. Studies focusing on a narrowly defined domain (e.g. clinical trials on Alzheimer's disease) perform experiments on a small number of protocols (below 1000), usually selected via the API of *ClinicalTrials.gov*. Studies targeting a broader medical domain (e.g. oncological clinical studies) experiment with a larger sample of protocols (between 1000 and 10 000). Finally, studies introducing general methods, not bound to a particular medical domain, utilize large samples of clinical protocols (over 10 000). A similar pattern is observed with eligibility criteria: narrowly focused studies process only specific criteria (below 100), a group of studies addresses eligibility criteria in a specific domain (below 10 000), and the largest group of studies does not impose limits on the number of processed eligibility criteria.

5.4 Discussion

At the beginning of the survey, three questions were posed to define the scope of the survey and the data charting protocol. The following subsections return to these questions.

5.4.1 Adoption of recent generative language models

The scoping review reveals a surprising lack of common adoption of the large language models that have been published in recent years. Most identified works utilize models developed between 2013 (`word2vec` embeddings) and 2018 (`BERT` and

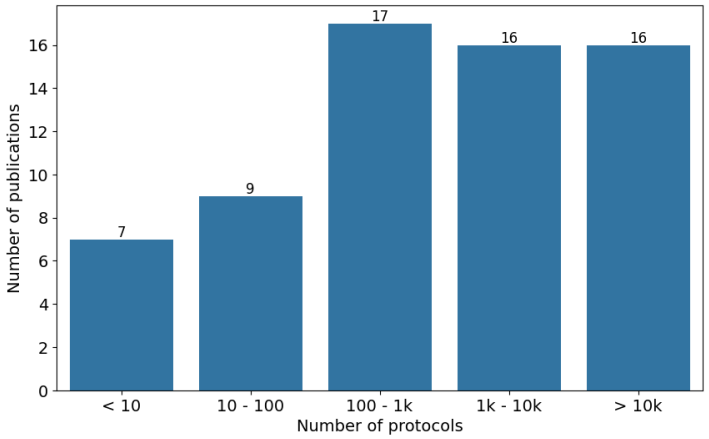


Figure 5.19: Number of clinical trial protocols.

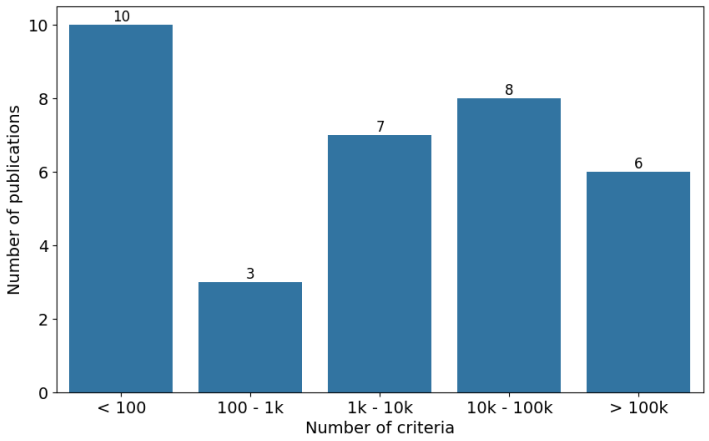


Figure 5.20: Number of eligibility criteria.

its many successors). Foundational large language models are rarely employed to extract information from clinical trial descriptions. There appears to be substantial opportunity for incorporating generative language models in the processing of clinical trials.

5.4.2 NLP methods, models, and tools in patient-trial matching

By far, the most common NLP method used in patient-trial matching is text and term normalization. This involves standardizing numbers, units, casing, and sentence segmentation to make source texts more coherent. Additionally, there is a need to align medical terms, such as names of drugs, medical procedures, medical conditions, etc. Learning of text representations is rare and limited to either pretrained word embeddings or even simpler vectorization methods like bag-of-words and TF-IDF. Regular expressions are frequently used by researchers to define custom text transformations, but these solutions often struggle to optimize both precision and recall

– precise matching rules result in a low recall, and universal matching rules result in low precision of transformations. An interesting finding of the scoping review is the popularity of the NegEx algorithm, a 20-year-old technique to detect markers of negations in text. Detecting negation in complex eligibility criteria is challenging. For instance, consider the following exclusion criterion for a clinical trial: "*Patients must not have a negative history of non-adherence to non-pharmacological interventions in cardiovascular-related treatments*". It is beyond non-trivial to judge if a patient, whose EHR states "*[. . .] has demonstrated consistent compliance with dietary changes and regular physical therapy sessions*", qualifies for the study.

5.4.3 Resources for ML and NLP in patient-trial matching

One of the most striking findings from the scoping review is the absence of common standardized benchmarks, which negatively affects the robustness and trustworthiness of reported results. Almost all evaluations are performed either manually, or via user surveys, making it extremely difficult to conduct meta-analyses or compare results across studies. The lack of benchmarks also influences the way results are presented – the researchers are allowed to cherry-pick metrics that produce the most optimistic results, leading to studies with custom-tailored, unusual metrics of evaluation. This issue is compounded by a prevalent positive publication bias afflicting all science, where nearly all studies included in the current review claim to report success or partial success, and none present unequivocally negative results.

The lack of benchmark datasets for model training and evaluation is the most consequential problem identified in this review. Researchers are forced to perform painstaking and tedious work on the same clinical trial descriptions downloaded from ClinicalTrials.gov, wasting time and resources. The absence of benchmark datasets impedes effective evaluation and comparison of studies. Significantly different results across similar studies due to non-standardized evaluation protocols make it difficult to draw actionable conclusions. Addressing the lack of benchmark datasets is crucial and should be prioritized by the research community. Benchmark datasets should cover the following NLP tasks:

- **Criterion identification**: Determining whether a given text contains an eligibility criterion.
- **Criterion similarity**: Deciding if a text, such as a criterion or fragment of an EHR, entails a given criterion.
- **Criterion NER**: Identifying named entities within the eligibility criterion.
- **Criterion entailment**: Given two criteria, assessing if one criterion subsumes, contradict, or has no logical relation to another.

5.4.4 Limitations

This review has notable limitations. Firstly, the search protocol explicitly excluded all works published in languages other than English. This causes the omission of

a significant research, particularly in Chinese. Anecdotally, a few papers from the review at the full-paper reading phase had to be rejected because, despite the titles and abstracts being written in English, the full texts of papers were in Chinese.

Secondly, the bibliographic search was limited to four major databases: Scopus, Embase, Web of Science, and PubMed. Sources like Google Scholar, arXiv, Science Direct, or bioRxiv were not included in the search to ensure the inclusion of only peer-reviewed publications. Distinguishing peer-reviewed papers from pre-prints, such as those on arXiv or bioRxiv, is challenging. Besides, it was assumed that a work published as a preprint, if accepted by a journal or conference, would be visible in the databases used in this search.

5.4.5 Future research directions

The most urgent need is the compilation of a normalized and standardized benchmark dataset for clinical trial descriptions and eligibility criteria. This represents the most significant obstacle impeding the development of research on automatic patient-trial matching. Given the vast number of published clinical trials and the non-standardized nature of trial descriptions, this task is very challenging, yet extremely important to address.

Another promising research direction, that only recently appears in the current research record, is exploring the capabilities of generative large language models for parsing clinical trial descriptions and eligibility criteria. These models demonstrate unprecedented versatility in extracting knowledge from free-form text. For instance, as shown by Gu *et al.* [99], large language models can be successfully used to extract complex terms from biomedical texts. This area of research is rapidly evolving, presenting many intriguing scientific and engineering challenges. As of the time of writing, we have identified only four published papers that address the use of the generative AI to parse clinical trials, but this is obviously subject to change in the incoming months.

Chapter 6

Biomedical Semantic Textual Similarity

Biomedical texts, including clinical trial eligibility criteria, encode semantics in domain vocabulary, extensive use of acronyms, proper nouns, named entities, and numerical values with implied meaning. This layer of complexity is often absent from the surface form of the text, making semantic textual similarity challenging for models trained on the general English corpora.

This chapter evaluates various techniques for sentence embedding and their effectiveness in semantic textual similarity searches within the biomedical domain. It compares static embeddings, transformer-based representations (with a focus on models fine-tuned to the biomedical domain), and sentence transformers. These techniques are also relevant for processing clinical trial eligibility criteria among many NLP tasks, as the text requires embedding before further algorithmic processing.

Section 6.1 discusses the biomedical semantic similarity task, its challenges, and the scope of this study. Section 6.2 lists studies and research related to semantic textual similarity in the biomedical field. Section 6.3 presents the data used in the experiments and Section 6.4 outlines the models being evaluated in this study. Section 6.5 describes the experimental setup, the evaluation metrics used, and directional expectation tests. Section 6.6 discusses the findings from the experiments and Section 6.7 summarizes the conclusions derived from the study.

This research was presented at the 20th International Conference on Artificial Intelligence in Medicine and published in the conference proceedings [100].

6.1 Study motivation and scope

The primary objective of semantic textual similarity (STS) is to determine the similarity score between two texts based on the likeness of meaning (semantics), rather than on lexical features of compared texts. To compute STS, one must select an appropriate text representation. Before the advent of neural networks, the most popular text representations included various bag-of-word models, such as TF-IDF, combined with lexical features like part-of-speech tags, dependency tags, named entities, and dictionary-based taggers. Recently, dense vector representations (embeddings) have

gained popularity. The results of STS are vital for many downstream tasks, such as information retrieval, recommendation systems, filtering, and document clustering, serving as an intrinsic method for evaluating the quality of text embeddings.

Due to the specialized language and terminology used in the biomedical field, STS presents significant challenges, including dealing with rare words, acronyms, and numerical values with specific semantic implications. While Transformer Encoder models trained on general English corpora are capable of solving general semantic similarity tasks, they often lack precision when dealing with specialized biomedical texts, such as eligibility criteria for clinical trials. Corpus annotation for supervised training of domain-adapted models is very expensive. The annotation task cannot be crowd-sourced, and medical experts do not have sufficient time to annotate corpora large enough to train language models. Consequently, the only viable solution is training models on general English corpora and performing unsupervised fine-tuning of these models.

This chapter examines the effectiveness of various unsupervised text representations in biomedical STS tasks, evaluates popular text preprocessing methods, and explores text representations including static embeddings, general language models, language models fine-tuned to the biomedical domain, and sentence transformers. It also demonstrates how auxiliary techniques like principal component reduction and word frequency weighting can improve the results of STS task for most models, when applied to embeddings. The experiments utilize two benchmark datasets for biomedical STS and explore how directional expectation tests can provide deeper insights into the characteristics of sentence representations for biomedical STS.

6.2 Related work

Blagec *et al.* [101] evaluated several sentence embedding models for biomedical semantic text similarity, including fastText [102], sent2vec [103], Skip-Thought [104], and Paragraph Vector [105]. They concluded that the Paragraph Vector model yields the best representations, achieving a Pearson correlation of 0.819. However, sent2vec, a more cost-effective model, also achieves a competitive Pearson correlation of 0.798. Their findings indicate a significant performance difference between the skip-gram and CBOW models (0.766 vs. 0.253 Pearson correlation), suggesting that rare domain-specific words pose a challenge to CBOW.

Ranasinghe *et al.* [106] explored contextualized word representations in the biomedical STS task. Their analysis compared ELMo [107], BioBERT, Flair [108], and a stacked ELMo+BioBERT model, using word2vec embeddings as a baseline. Evaluation metrics included the cosine similarity of averaged word vectors, Word Mover's Distance, and cosine similarity with Smooth Inverse Frequency (SIF) [109]. Interestingly, only the stacked ELMO+BioBERT model surpassed the baseline when using cosine similarity with SIF, and other models did not outperform the baseline.

Koroleva *et al.* [110] investigated the similarity of clinical trial outcome descriptions, framing it as a binary classification task. The models assessed whether two de-

scriptions related to the same medical concept. The study included BERT, BioBERT, and SciBERT, with BioBERT achieving the best F1-score without reliance on external knowledge sources like UMLS or WordNet. However, the error analysis revealed that some sentences required additional domain knowledge for accurate similarity judgement.

The current study differs from the works listed above as it evaluates and compares sentence representations derived from static embeddings, biomedical pretrained Transformers, and general-domain Sentence Transformers, enhanced with two modifications - word frequency weighting and principal component reduction. Additionally, it introduces a new method of evaluation - directional expectation tests.

6.3 Data

Currently, there are no benchmark datasets specifically for semantic textual similarity on clinical trial eligibility criteria. Therefore, this study utilizes two public benchmark datasets from the biomedical domain:

- **BIOSSES dataset:** This dataset consists of 100 sentence pairs extracted from biomedical literature [111]. The pairs have been manually annotated by five medical experts, with similarity scores ranging from 0.0 (unrelated) to 4.0 (semantically equivalent). Table 6.1 presents a sample of the BIOSSES dataset.

- **Clinical Outcomes (CO) dataset:** This dataset is composed of pairs of texts extracted from 3938 randomized controlled trials published in PMC [110]. Each pair includes one phrase from the primary outcomes and one phrase from the reported outcomes of the same trial. Medical experts manually annotated these pairs with a binary label to indicate whether both outcomes refer to the same medical concept. The training set includes 2108 pairs of unrelated texts and 616 pairs of similar texts, while the test set contains 226 pairs of unrelated texts and 78 pairs of similar texts. A sample of this dataset is presented in Table 6.2.

One of the main differences between these datasets is the length of the texts. The BIOSSES dataset consists of complete sentences, whereas CO includes only single words or phrases extracted from sentences. This variation suggests that contextualized models might perform better on the BIOSSES dataset, where they can leverage the full context of each sentence. In contrast, the CO dataset requires models to understand abbreviations or synonyms, since the context is often missing. A detailed text length comparison is provided in Table 6.3.

The experiment operates under the assumption that the results obtained on those datasets are transferable to clinical trial eligibility criteria. Nevertheless, directional expectation tests are conducted on texts from eligibility criteria to validate this assumption.

Table 6.1: BIOSSES dataset sample.

Sentence 1	Sentence 2	Score
The in vivo data is still preliminary and other potential roadblocks such as drug resistance have not been examined.	The GEM model used in this study retains wild-type Tp53, suggesting that the tumors successfully treated with bortezomib and fasudil might not be as aggressive as those in most NSCLC patients.	0.0
It has recently been shown that Craf is essential for Kras G12D-induced NSCLC.	It has recently become evident that Craf is essential for the onset of Kras-driven non-small cell lung cancer.	4.0
Three programs, PicTar, miRanda, and TargetScan, were used to predict the targets of miR-21.	The genes that decreased 2-fold or more were further screened for possible miR-372/3 target sites using a local version of the TargetScan algorithm.	2.4

Table 6.2: Clinical Outcomes dataset sample.

Sentence 1	Sentence 2	Label
FEV 1 (% predicted) in the two groups at 1, 5, 15, 30, 60 min after administration of the study drug	FEV1 (% predicted)	1
The proportion of children achieving a well clinical response and time to a well clinical response	Mortality	0
Blood pressure	Glycated hemoglobin A1c	0
ICP control	Uncontrollable intracranial pressure	1

Table 6.3: Text length comparison of the BIOSSES and the CO datasets.

Dataset	Median word count	Min. word count	Max. word count	Nb. of unique words
BIOSSES	22	7	49	1309
CO	4	1	47	1132

6.4 Models

The experiment evaluates static text embeddings and neural-based sentence representations. As a baseline for comparison, GloVe pretrained vectors [112] are employed. Three variants of GloVe vectors are examined: 50-dimension and 300-dimension GloVe models built on Wikipedia corpus, and 300-dimension model trained on the Common Crawl. Additionally, the BERT model [69] trained on BookCorpus and

English Wikipedia was chosen for neural-based sentence representations.

Recently, many new transformer models trained on biomedical texts have been introduced, as detailed in Section 4.3. This experiment evaluates six such models: BioBERT, SciBERT, PubMedBERT, BlueBERT, ClinicalBERT, and CODER. The training objective of BERT involves two related NLP tasks: predicting the masked word and predicting the next sentence. However, BERT is not designed to provide accurate sentence representations, as it focuses on token-level embeddings. Sentence vectors are typically generated by averaging token vectors or extracting the `[CLS]` token embedding, but these methods have been shown to be less effective than averaged GloVe embeddings.

Sentence-BERT (SBERT) [113] represents a new model specifically trained to generate semantically rich sentence representations. SBERT uses Siamese and triplet network architectures to produce meaningful sentence representations for downstream regression and classification tasks. This study examines the performance of four sentence transformers, SMiniLM (all-MiniLM-L6-v2), SRoBERTa (all-distilroberta-v1), SMPNet (all-mpnet-base-v2) and SBERT (multi-qa-distilbert-cos-v1). These models are based on different architectures including distilled RoBERTa [114], MiniLM [115], MPNet [116], and distilled BERT [114], respectively. They all were trained on general English corpora. Currently, there are no sentence transformers specifically fine-tuned to the biomedical domain.

6.5 Experiments

The experiments assess the performance of various text preprocessing techniques combined with different sentence representations on the biomedical semantic textual similarity task. Text preprocessing techniques includes lower-casing, lemmatization, word splitting at punctuation marks, stop-word removal and filtering of punctuation and numbers, implemented using Python libraries: spaCy and NLTK. Text representations are generated from language models through `[CLS]` token embedding, extraction from the last layer, the second to last layer, and the first layer. Sentence representations are then computed either by averaging or max-pooling of token embeddings.

6.5.1 Weighting embeddings by word frequency

This study introduces two techniques that can be easily applied to any embedding extracted from a language model. The first technique, inspired by Smooth Inverse Frequency [109], weights individual token embeddings by the token's relative frequency in general English.[1] Three modes of weighting embeddings are considered:

- **Simple weighting**: Token frequency is added as a weight when calculating the average sentence embedding.

[1]The wordfreq Python library is used which calculates token frequencies based on Google Books, Leeds Internet Corpus, Wikipedia and Para Crawl

- **Concatenation**: Tokens are split into rare and frequent tokens based on a frequency threshold, and average or max-pooled embeddings are calculated for these two sets of tokens before concatenating the resulting vectors.
- **Rare words**: Frequent tokens are filtered out before vectorization using a predefined frequency threshold.

The rationale behind using word-frequency weighting of embeddings is to lower the impact of common English words on the final sentence embedding, thus increasing the weight of in-domain rare words. Common English words may introduce misleading similarity in the biomedical domain when describing similar processes in semantically unrelated topics. For example, the sentences *"AST and ALT $\leq$ 2.5 x ULN with the following exception"* and *"Serum bilirubin $\leq$ 1.5 x ULN with the following exception"* involve unrelated medical concepts, despite similar structural formats.

6.5.2 Principal component reduction

The second technique for embedding fine-tuning is removing the first principal component of the embedding matrix. It has been observed [117, 118] that static embeddings have a large mean vector, and after subtracting this mean, the remaining mass of the embedding is concentrated in just a few dimensions. All vectors share the mean and these dominating dimensions, it is hypothesized that the information content of these dimensions encodes the general grammatical structure of the language (e.g., part-of-speech sequence, idiomatic expressions, syntactic rules of phrase composition). By subtracting the first principal component, the resulting vectors are enriched with more semantic information and reduced syntactic noise.

6.5.3 Evaluation metrics

For the BIOSSES dataset, STS techniques are evaluated by training a regression model using three metrics: Pearson's r (Pearson correlation coefficient), Spearman's ρ (Spearman's rank correlation coefficient), and mean squared error (MSE). These metrics are commonly used to evaluate the quality of unsupervised STS tasks in the literature.

The task in the CO dataset is the binary classification, so the continuous semantic similarity score returned by the distance function (either cosine or Jaccard) are binarized. The threshold is set to mean similarity scores between the first quartile of distances for similar pairs and the third quartile of cosine distances for unrelated pairs of sentences from the test set. This threshold is used to classify a given pair of sentences as either similar or unrelated. The quality of STS for the CO dataset is evaluated using precision, recall, and the F1-score.

All the models are run with different combinations of sentence embedding and text preprocessing methods, which results in 74 048 experiments (64 for the Jaccard distance, 768 for GloVe, 8064 for base transformers, 64 512 for base transformers with wordfreq, and 640 for sentence transformers). These experiments are repeated across both BIOSSES and CO datasets, giving 148 096 experiments in total.

6.5.4 Directional expectation tests

To evaluate the robustness of the models, three directional expectation tests (DETs) [119] are conducted, related to the medical concepts of hemoglobin level, neutrophil count, and age, derived from the clinical trial eligibility criteria. The idea of DETs is straightforward: given a test instance with a known expected outcome, one perturbs the test instance in such a way that there is an expectation of the direction and scale of change of the test result. For biomedical STS, one similar (or almost identical) sentence pair for each DET is chosen, and additional pairs are created by introducing minor changes in the second sentence. For example, the word *hemoglobin* is replaced with its abbreviation *Hb* or a different word *bilirubin*. It is expected that in the first scenario, the similarity score will not change significantly or may remain the same, while in the second scenario, the semantic similarity score is expected to decrease.

6.6 Results

6.6.1 Results on the BIOSSES dataset

Table 6.4 shows the evaluation results on the BIOSSES dataset. The baseline is established using $\text{GloVe}_{wiki300}$ vectors. Interestingly, even a simple Jaccard distance computed on the sets of unique tokens in compared sentences outperforms this baseline. All transformer models surpass the baseline, with those fine-tuned on the biomedical domain performing better than the vanilla BERT model. Among the transformers, the CODER model achieves the highest Pearson's r of 0.849. Sentence transformers, despite being trained on general-domain corpora without biomedical domain fine-tuning, show superior performance, with SRoBERTa achieving the highest Pearson's r of 0.878.

The impact of the sentence embedding method on the biomedical STS performance is also evaluated. The best performing strategy for GloVe embeddings is max-pooling of token vectors combined with the removal of the first principal component. In general, principal component reduction significantly boosts the performance of GloVe representations. For instance, a 0.220 increase in Pearson's r is observed for the $\text{GloVe}_{wiki300}$ model. Principal component reduction and embedding weighting by word frequency (mostly selecting rare words only) improve the results of nearly all BERT-based models, except for the BioBERT, where only principal component reduction is successful. A minimum increase of 0.1 in Pearson's r is noted for other models after applying principal component reduction. However, this technique does not benefit embeddings derived from sentence transformers, which already produce semantically rich representations with minimal syntactic information.

Different text preprocessing methods and their impact on the model performance are assessed. It is observed that splitting compound words joined by punctuation slightly improves the performance of all models. This improvement is likely because compound words are ubiquitous in the biomedical domain, yet often treated as out-

of-vocabulary tokens by language models. Other text preprocessing techniques do not show any deterministic changes in the model performance.

Table 6.4: Evaluation on the BIOSSES dataset.

Model	Embedding	Freq. weight	PCA	r	ρ	MSE
BERT base	first layer mean	concat	✓	0.777	0.783	0.153
BioBERT	first layer max	-	✓	0.795	0.799	0.166
SciBERT	first layer max	rare words	✓	0.813	0.808	0.143
PubMedBERT	first layer max	rare words	✓	0.803	0.803	0.139
BlueBERT	last layer max	rare words	✓	0.809	0.789	0.129
ClinicalBERT	last layer mean	rare words	✓	0.800	0.785	0.098
CODER	last layer max	rare words	✓	**0.849**	**0.834**	**0.096**
SMiniLM	-	-	-	0.842	0.813	0.029
SRoBERTa	-	-	-	**0.878**	**0.843**	**0.023**
SBERT	-	-	-	0.820	0.821	0.033
SMPNet	-	-	-	0.845	0.804	0.032
GloVe$_{wiki50}$	maximum	-	✓	0.624	0.639	0.161
GloVe$_{wiki300}$	maximum	-	✓	**0.775**	**0.775**	0.154
GloVe$_{crawl300}$	maximum	-	✓	0.757	0.742	**0.138**
Jaccard distance	-	-	N/A	**0.776**	**0.807**	**0.222**

6.6.2 Results on the CO dataset

As demonstrated in Table 6.5, CODER, enhanced with principal component reduction, outperforms other models in the CO dataset binary classification with an F1-score of 0.798. Again, GloVe embeddings serve as the baseline for the evaluation. Surprisingly, the second-best result is achieved by a simple Jaccard distance with a threshold. Fine-tuned transformers do not perform better than the vanilla BERT trained on the general English corpus. Sentence transformers perform comparably to traditional transformers. The best representations are created by average-pooling of embeddings from the first layer of the models. Embedding weighting by word frequency does not consistently improve results. However, principal component reduction enhances the performance of all examined models, with the most significant improvement noted for the GloVe$_{crawl300}$ model, showing an increase of 0.129 in the F1-score.

The results on the CO dataset significantly differ from those observed for the BIOSSES dataset. However, an additional step in the pipeline after the STS calculation needs to be stressed here. The similarity score must be converted into a binary label based on a dynamically selected threshold for the CO dataset. This conversion may be responsible for the loss of information and different behavior of models. Additionally, variations in vocabulary, syntax, and sentence length between the datasets impact the evaluation results.

Table 6.5: Evaluation on the Clinical Outcomes dataset.

Model	Embedding	Freq. weight	PCA	F1	Precision	Recall
BERT base	first layer max	-	✓	0.756	0.667	0.872
BioBERT	first layer max	-	✓	0.746	0.667	0.846
SciBERT	first layer max	-	✓	0.737	0.653	0.846
PubMedBERT	first layer max	-	✓	0.725	0.635	0.846
BlueBERT	first layer max	-	✓	0.739	0.663	0.833
ClinicalBERT	first layer max	-	✓	0.754	0.680	0.846
CODER	last layer max	-	✓	**0.798**	**0.710**	**0.910**
SMiniLM	-	-	-	0.719	0.640	0.821
SRoBERTa	-	-	-	0.705	0.598	**0.859**
SBERT	-	-	-	**0.729**	**0.641**	0.846
SMPNet	-	-	-	0.663	0.583	0.769
GloVe$_{wiki50}$	mean	-	✓	0.640	0.525	**0.821**
GloVe$_{wiki300}$	mean	-	✓	0.674	0.592	0.782
GloVe$_{crawl300}$	maximum	-	✓	**0.696**	**0.612**	0.808
Jaccard distance	-	-	N/A	**0.757**	**0.677**	**0.859**

6.6.3 Directional expectation tests

Directional expectation test are conducted on two best-performing models: SRoBERTa and CODER. As shown in Table 6.6, the overall performance of both models is satisfying because, in most cases, STS scores change accordingly to the expectations. However, there are examples where embeddings fail to preserve semantic differences, leading to incorrect STS scores. For instance, the cosine similarity score between the embeddings from SRoBERTa is higher for the pair (*Age* $\geq$ *18 years at the time of signing Informed Consent Form*, *Signed Informed Consent Form*) than for the pair: (*Age* $\geq$*18 years at the time of signing Informed Consent Form*, *Age* $\geq$ *18 years*). This suggests that the model may overly focus on the latter part of the sentence *signing Informed Consent Form* than the age limit. Similar behavior is observed in the sentence *No history of other diseases at the age* $\geq$ *18 years* where CODER focuses on the age limit, resulting in a very similar embedding to the base sentence. Another interesting finding is that CODER creates more similar embedding for *Adult* than for *Age* $<$ *60 years*, indicating a potential understanding of the semantic implication of *Age* $\geq$ *18*. There are also differences in STS scores between *Age* $\geq$ *18 years* and *Age* $<$ *60 years*. On the one hand, it might suggest that CODER has the basic ability to perform numerical inference and numerical comparisons. On the other hand, terms *age* and *18* may be common in training corpora as the age of majority, and this may result in the false perception of CODER's inference abilities. Furthermore, CODER effectively handles medical abbreviations, with only slight changes in similarity scores when substituting *Hemoglobin* with its abbreviation *Hb*, or *Absolute Neutrophil Count* with *ANC*. However, the differences between all the STS scores of CODER vectors are minimal. In contrast, SRoBERTa is better at distinguishing between sentences. DET scores vary significantly from 0.26 to 0.98 for SRoBERTa, whereas for CODER,

they range from 0.72 to 1.0. For instance, the scores for SRoBERTA representation of *Bilirubin* $\geq$ *9 g/dL* and *ECOG Performance Status* $\geq$ *1* drop significantly, which is the correct behavior as these terms are not related to hemoglobin. The neutrophils test further confirms that both models can correctly encode the meanings of medical terms like *neutrophil*, *WBC*, and *lymphocyte*.

6.7 Conclusions

This study experiments with various sentence embeddings to identify the best representation for biomedical semantic textual similarity. It assesses biomedical BERT models and several sentence transformers trained on general-domain corpus, introducing two preprocessing techniques that improve the expressiveness of embeddings. The first technique applies word frequency weights to word embeddings, hypothesizing that separating rare and frequent words reduces the impact of more generic language on the final sentence embedding. The second technique involves removing the first principal component from embeddings, effectively shifting the informational content of embeddings from encoding grammatical syntax to enhancing semantics understanding. This principal component reduction improves performance across all BERT models and static GloVe embeddings.

The algorithms are evaluated in two use cases: similarity score estimation and binary label prediction. Sentence transformers outperform other models. CODER comes in second, surpassing biomedical BERT models and achieving the highest F1-score in the binary label prediction task. This study points to sentence transformers as the most versatile and best performing models for biomedical STS. The representations extracted from SRoBERTa are more polarised and have larger expressiveness even for potentially similar biomedical sentences. This result is somewhat surprising because sentence transformers used in the experiments are trained on the general English corpora, in contrast to domain fine-tuned BERT models.

Additionally, directional expectation tests are utilized to assess the quality of sentence representations. These tests reveal non-obvious focus aspects of models - features that attract the models' attention but do not correspond to meaningful textual similarity. This behavior is attributed to the unsupervised nature of the training and the insufficient level of input text annotation.

Table 6.6: Results of directional expectation tests.

Sentence 1	**Sentence 2**	**CODER**	**SRoBERTa**
Hemoglobin $\geq$ 9 g/dL	Hemoglobin $\geq$ 10 g/dL	1.00	0.91
	Hemoglobin greater than or equal to 9 g/dL	0.95	0.96
	Hb $\geq$ 9 g/dL	0.98	0.75
	Hemgolobin $\geq$ 9 g/dL	0.89	0.69
	Bilirubin $\geq$ 9 g/dL	0.83	0.43
	ECOG Performance Status $\geq$ 1	0.83	0.10
Absolute neutrophil count $\geq$ 1500/μL without granulocyte colony-stimulating factor	Absolute neutrophil count $\geq$ 1500/μL	0.94	0.83
	ANC $\geq$ 1.5 x 109/L (1500/μL) without granulocyte colony-stimulating factor support	0.92	0.65
	Lymphocyte count $\geq$ 0.5 $\times$ 109/L (500/μL)	0.83	0.54
	WBC count >= 2.5 x 109/L (2500/μL)	0.79	0.29
	Neutrophil count normal	0.82	0.72
Age $\geq$ 18 years at the time of signing Informed Consent Form	Age $\geq$ 18 years	0.89	0.63
	Adult	0.76	0.41
	Age $\geq$ 18	0.88	0.62
	Signed Informed Consent Form	0.79	0.74
	Age >= 18	0.87	0.60
	Age $<$ 60 years	0.72	0.48
	No history of other diseases at the age >=18 years	0.77	0.28

Chapter 7

Named Entity Recognition in Eligibility Criteria

Complex formulations of inclusion and exclusion criteria are a primary reason for failing to meet recruitment quotas and delaying progress through the subsequent phases of clinical trials [120]. Named entity recognition (NER) is an essential approach in eligibility criteria parsing. By identifying and extracting information from clinical trial protocols, such as patient demographics, medical conditions, and treatments, NER enables researchers to quickly determine whether a patient meets the criteria for a particular clinical trial. This allows for faster and more efficient patient recruitment and potentially more accurate data collection.

NER is typically performed as a sequence tagging task (a.k.a. token classification). A deep learning model processes each token in the input sequence and assigns it to one of the predetermined categories. These categories come from a labeling scheme that may vary between models. Traditionally, NER models recognize categories (classes) such as persons, organizations, numbers, geographical locations, time expressions, dates, and organizations. More sophisticated NER models can tag works of art, events, geopolitical entities, currencies, and more. Before the advent of deep neural networks, NER models were implemented using conditional random fields [121] or hidden Markov models [122]. Later, sequence tagging became the domain of bi-directional LSTM [123] (long short-term memory) networks. Recently, BERT-based architectures have been the primary tool for implementing NER models [124].

Over the past decade, significant efforts have led to the development of dedicated models to extract information from biomedical texts. These models can identify and classify entities such as diseases, drugs, genes, proteins, and other related concepts. They are also capable of recognizing complex relationships between different entities in text and detecting patterns in the data. NER models, such as BioBERT, ClinicalBERT, or BioMedicalRoBERTa [125], have been successfully applied to various tasks, including literature-based discovery, clinical information extraction, and drug development.

The last four years have witnessed unprecedented development in the domain of language models. With the release of BERT, transformer-based language models

started to achieve state-of-the-art results in many natural language processing tasks. Since then, generative language models such as GPT-2 [126], T5 [127], GPT-3 and GPT-4 have been released, achieving even better results in a variety of tasks. These language models offer a new mode of interaction - prompt engineering. Prompt engineering is a technique of deliberately crafting an input to a language model in order to achieve a desired output. By giving the model a specific prompt or sentence, one can control and guide the output of the model to generate relevant and meaningful results. In particular, a prompt can be used to extract instances of named entity classes from a text.

This chapter examines application of generative language models in the named entity recognition task. Section 7.1 formulates the study objectives and outlines previous works. Section 7.2 describes the selected prompt engineering technique - in context learning. Section 7.3 discusses the dataset used for the experiments and evaluation metrics. Section 7.4 presents the experimentation results and Section 7.5 provides the conclusions.

7.1 Study objective

The main aim of this paper is to evaluate the effectiveness of prompt engineering in extracting named entities from the eligibility criteria of clinical trials. While generative large language models (LLMs) have demonstrated promise in biomedical NER, our focus is specifically on their application to eligibility criteria.

Previous research hase explored NER in the biomedical domain using various generative models and techniques. Zhou *et al*. [128] utilized GPT-3.5-turbo for zero-shot NER of biomedical terms, comparing it to fine-tuned biomedical BERT models. While BERT outperformed GPT-3.5, the study only tested a limited number of prompts, and the BERT models were fine-tuned on a sizable dataset. Sivarajkumar *et al.* [129] compared generative models with SciSpaCy [130] for biomedical NER and found GPT-3.5 to be the most accurate. Li *et al*. [131] proposed a few-shot NER method combining generative models with BERT embeddings and a retrieval module for example selection, achieving superior results with chain-of-thought prompts. Hu *et al*. [132] developed advanced prompts for GPT models in biomedical NER, finding that ClinicalBERT, fine-tuned on a substantial dataset, outperformed the GPT models.

Our study differs from these works as we aim to evaluate the performance of generative models against current state-of-the-art BERT models fine-tuned on a very limited dataset, using few examples for in-context learning. Additionally, we provide a simple prompt without additional guidelines for the GPT model. This setup mimics a scenario with limited annotated data and minimal support from domain experts.

The research hypothesis is that large language models have sufficient capacity not only to recognize domain-specific biomedical vocabulary, but also to differentiate between classes that describe various aspects of eligibility criteria. To test this hypothesis, we compare the effectiveness of the prompt engineering-based model

with state-of-the-art NER models.

7.2 Few-shot prompt engineering for entity recognition

Prompt engineering involves constructing sequences of tokens to improve the accuracy and generalization of large language models. By providing the model with a specific prompt, the model can focus its attention on particular language patterns or contexts. This approach aligns well with transfer learning, where a model trained on one task can be adapted to another. In our case, we aim to adapt a general-purpose language model trained for next-token prediction to the sequence tagging task (identifying tokens in the input sequence as either belonging to a given entity class or not).

Since individual eligibility criteria can be relatively long, and spans of tokens representing an entity can be quite short, simply generating an instruction such as "*[CRITERION]. List examples of drugs in this text*" (where the text of the original criterion is inserted verbatim in lieu of the `CRITERION` token) yields no usable results. We leverage the fact that LLMs are known to hook on atypical patterns in the text. By providing a template with several examples of eligibility criteria and extracted entities marked with marker tokens, we can guide the model in the desired direction.

After experimentation, we selected a template that allows few-shot learning within the scope of the prompt. For each entity class, we randomly select five eligibility criteria and explicitly list entities annotated in these criteria. Both positive examples (criteria where a given entity appears) and negative examples (criteria where there is no instance of a given entity) are equally important. These criteria serve as the blueprint for the generated output. An example of the prompt template for the entity class *cancer* is presented in Figure 7.1. This prompt is provided as input to the LLM. While the overall template structure remains consistent, the specific examples included in the prompt change according to the entity class being analyzed. For each entity class in the dataset, we build a separate prompt with different few-shot examples tailored to that class. Using the "cancer" prompt presented in Figure 7.1, the assumption is that the LLM will "understand" that it should generate the list of cancers mentioned in *"entities"*. The expected output generated by the model is *[medullary thyroid cancer (MTC), RET-altered solid tumor]*.

7.3 Dataset and metrics

In our experiments, we used the annotated dataset from the *Clinical Trial Parser* [133], which contains eligibility criteria for 3314 randomly selected interventional trials in the United States. The sample was downloaded from the Aggregate Analysis of ClinicalTrials.gov (AACT) Database using the daily static DB copy of 2020-04-16. The criteria were split by a new line character into 49 903 samples, and the annotation was done by professional annotators. This resulted in 120 906 labeled entities, with the distribution of labels presented in Table 7.1.

```
Find examples of cancer in the following criterion.
Your response should be a list of comma separated values, eg: 'foo, bar, baz'
If no examples are found, type 'None'. Return only the entities found in the criterion.

criterion: Participant has received no prior radiotherapy or chemotherapy for rhabdomyosarcoma
 (excluding steroids) unless an emergency situation requires local tumor treatment
entities: tumor

criterion: Aspartate aminotransferase (AST) (serum glutamic oxaloacetic transaminase [SGOT])
and alanine aminotransferase (ALT) (serum glutamate pyruvate transaminase [SGPT]) =< 2.5 x ULN
 (or =< 5 x ULN if liver metastases [mets])
entities: liver metastases [mets]

criterion: Patients must have histologically confirmed, BRAF-mutant (V600E/K) melanoma (
molecularly confirmed using validated, commercially available assay performed in a Clinical
Laboratory Improvement Act [CLIA]-approved laboratory) that is metastatic or unresectable and
for which standard curative measures do not exist or are no longer effective
entities: melanoma

criterion: loop recorder explanted within the past 12 months
entities: None

criterion: Other medical or psychiatric disorder placing the subject at undue risk for
treatment complications
entities: None

criterion: All patients treated at doses > 120 mg per day must have medullary thyroid cancer (
MTC), or a RET-altered solid tumor per local assessment of tumor tissue and/or blood
entities:
```

Figure 7.1: Example of a prompt template.

Table 7.2 provides examples of entity spans annotated in the *Clinical Trial Parser* dataset. One immediately notices that the annotated text is highly specialized and contains many abbreviations, domain-specific terms, proper names, etc. In our opinion, it is a very challenging dataset for any NER model.

In our experiments, we focused on five medical entities that are most relevant for parsing eligibility criteria: *treatment*, *chronic disease*, *clinical variable*, *cancer*, and *allergy name*. To create prompts for the `GPT-4-turbo` model, we selected 22 samples from 17 trials. To simulate a scenario with limited availability of annotated data, we randomly selected two subsets of data, a larger set of 100 trials and a smaller set of 27 trials, which were used for Transformer fine-tuning. We fine-tuned BERT-based models in two scenarios:

- 80 trials in the training dataset (1243 samples), 20 trials in the validation dataset (376 samples)
- 17 trials in the training dataset (448 samples), 10 trials in the validation dataset (213 samples)

The 17 trials used for the few-shot prompt were included only in the training datasets. All models (BERT-based and GPT) were evaluated on a hold-out test set containing 663 randomly selected trials, with 1106 samples in total. The distribution of the labels in the evaluation dataset is presented in Table 7.3.

Table 7.1: Entity distribution in the CTP dataset

Label	**Count**
treatment	30972
chronic_disease	26212
upper_bound	13967
lower_bound	13633
clinical_variable	13255
cancer	9344
gender	3661
pregnancy	2773
age	2616
allergy_name	1887
contraception_consent	1603
language_fluency	482
bmi	287
technology_access	132
ethnicity	82

Table 7.2: Examples of entity span annotations in the CTP dataset

Entity class	**Example**
allergy_name	Dipeptidyl peptidase-4 (DDP-4) inhibitors, Linagliptin, propofol, glycerol
clinical_variable	total bilirubin, CIWA-Ar score, ALT, AST, Cockcroft-Gault formula
chronic_disease	liver dysfunction, glaucoma, kidney dysfunction, psychiatric disorder
treatment	flutamide, nilutamide, bicalutamide, Prior androgen deprivation therapy
cancer	melanoma, bone marrow plasmacytosis, Philadelphia (Ph)+ ALL

Table 7.3: Entity distribution in the evaluation dataset

Label	**Count**
treatment	13541
chronic_disease	11362
clinical_variable	7205
cancer	5166
allergy_name	588

To evaluate NER models, we used traditional metrics of precision, recall, and F_1. The eligibility criteria are transformed into the BIO format, where each token in the input stream is labeled as `B-entity` (beginning token of the entity span), `I-entity` (inside token of the entity span), or `O` (outside any entity span). Evaluation followed the popular scheme for sequence-to-sequence learning tasks, where metrics represent the performance of the model for each of the BIO tokens. However, it should be noted that this evaluation scheme is detrimental to generative LLMs. Consider the following criterion: "*Histologically or cytologically confirmed diagnosis of gastric, lung, colorectal, or breast cancer on file*". For this criterion the GPT-4-turbo model generated the following list of entities: *[gastric cancer, lung cancer, breast cancer]*, which is a very good extraction. Unfortunately, neither "*lung cancer*" nor "*gastric cancer*" appears in the input criterion, making it impossible to align the answer of the LLM with the input sequence. Similarly, for the criterion containing "*cancer of the prostate*", the LLM generated the following entity: *[prostate cancer]*. This is also a very good answer, but according to the BIO evaluation scheme, the LLM failed to mark the input entity span correctly.

Typically, the problem of annotating complex, overlapping, and disjoint entity spans is addressed by Discontinuous Named Entity Recognition (DNER). DNER is an advanced technique designed to identify and categorize entities in text that are not contiguous but are semantically linked. Unlike traditional NER systems, which annotate continuous spans of text, DNER can precisely label entities consisting of multiple non-adjacent segments. This capability is particularly vital in complex domains such as biomedicine, where entities such as symptoms or drug effects can be described in a fragmented manner within a sentence.

The research on discontinuous and overlapping entities started 15 years ago with the investigation of nested entities [134], and since then has gathered significant scientific interest [135]. Various methods address the problem of disjoint entity spans, such as applying relation extraction techniques to identify spans belonging to a single entity, and then post-processing these spans to combine them into disjoint, nested, or overlapping sequences of token spans [136]. Another approach is to extend the basic set of entity tags (BIO, IOBES) with additional H and D tags to mark words shared by multiple mentions and parts of discontinuous mentions not shared by other mentions, respectively (BIOHD [137]). One can also introduce a degree of uncertainty in the assignment of entity tags, as seen in the FuzzyBIO entity labeling scheme [138]. Discontinuous entities may be discovered by applying complex post-processing to initial token tagging, e.g., modeling discontinuous entities as maximal cliques [139]. Recently, end-to-end neural models for discovering discontinuous entities have been proposed [140].

Unfortunately, these methods only partially address the problem. While they allow for the annotation of a source span like "*lung or breast cancer*" as containing two entities ([*lung cancer, breast cancer*]), they do not solve the problem of LLM responses that do not align with the source text at all. In other words, the sequence tagging paradigm is not well-suited for evaluating LLM responses. One might argue that it is possible to force the LLM to generate output identical to the input via prompting

(i.e., preventing the LLM from producing output token sequences not present in the input sequence), but this is hardly a solution, and there is no guarantee that the LLM will follow the prompt for every generation. Switching to the BIOHD evaluation would require re-annotating the entire dataset, which is prohibitively expensive.

In this study, we have decided to follow the traditional BIO labeling scheme and evaluation procedure, noting that this approach underevaluates the LLM. Given the size of the dataset used in our experiments, it is difficult to quantify what percentage of GPT-4-turbo predictions are correct but incompatible with the input stream. However, it is obvious that we urgently need a new evaluation metric for the generative NER task, similar to the BERTScore [141] for text similarity.

7.4 Experiments

For prompting, we selected the GPT-4-turbo model [79]. We compared this LLM with several BERT models: BERT uncased, Biomedical BERT NER [142], BioBERT, SciBERT, PubMedBERT, BlueBERT, ClinicalBERT, and CODER. All layers of the BERT models were unfrozen, and an additional linear layer was added on top for token classification. Early stopping was used in training, with patience set to 5. The number of epochs was set to 30, but all training processes were completed before the 15th epoch. The training parameters were as follows: a learning rate was set at $\eta = 1e-5$, batch size $bs = 8$ for both training and evaluation, and a weight decay was set at $\gamma = 0.01$. The learning rate scheduler used cosine with restarts, with 50 warm-up steps. Below, we present the results for CODER, the best performing BERT-based model. We note that SciBERT was a close second in terms of overall performance.

Table 7.4 presents the comparison of GPT-4-turbo and CODER models. The results marked as `CODER-27` represent the low-resource scenario, assuming only a small dataset is available for fine-tuning. In this scenario, eligibility criteria from only 27 randomly selected clinical trials were used for fine-tuning. `CODER-100` represents results obtained after fine-tuning the model on a random selection of 100 clinical trials. As shown, GPT-4-turbo is outperformed on every metric and BIO tag. Surprisingly, even a relatively small dataset for fine-tuning is highly beneficial. Annotating medical texts is extremely expensive as it cannot be outsourced to crowd-working platforms and must be performed by highly trained domain experts. Our results suggest that even a small investment in additional annotation for fine-tuning can significantly improve the quality of NER models.

Table 7.5 presents the comparison of the same two models on a simpler task of predicting entity span, where the models do not need to distinguish between the beginning and inside of an entity span. The results are very similar to the BIO evaluation scheme, with the CODER model significantly outperforming GPT-4-turbo, and more fine-tuning data leading to better performance.

Figure 7.2 presents the comparison of the confusion matrices for the models. An important observation is that the nature of errors differs between the two models. While GPT-4-turbo tends to confuse entities (particularly when using the BIO evalu-

Table 7.4: Comparison of GPT-4-turbo and CODER models on BIO NER (p-precision, r-recall, f-F_1 score

	GPT-4-turbo			CODER-27			CODER-100			
	p	**r**	**f**	**p**	**r**	**f**	**p**	**r**	**f**	**support**
B-CANCER	0,30	0,35	0,32	0,71	0,46	0,56	0,76	0,66	0,71	2093
I-CANCER	0,33	0,39	0,36	0,74	0,51	0,60	0,78	0,73	0,75	3073
B-TREATMENT	0,30	0,26	0,28	0,64	0,76	0,69	0,70	0,77	0,73	6209
I-TREATMENT	0,28	0,35	0,31	0,66	0,73	0,70	0,75	0,69	0,72	7332
B-CLINICAL_VARIABLE	0,32	0,47	0,38	0,73	0,72	0,72	0,84	0,68	0,75	2435
I-CLINICAL_VARIABLE	0,32	0,45	0,37	0,72	0,82	0,77	0,85	0,75	0,80	4770
B-ALLERGY_NAME	0,05	0,74	0,10	0,00	0,00	0,00	1,00	0,08	0,14	323
I-ALLERGY_NAME	0,02	0,35	0,03	0,00	0,00	0,00	0,89	0,06	0,12	265
B-CHRONIC_DISEASE	0,37	0,32	0,34	0,66	0,76	0,71	0,77	0,76	0,76	5115
I-CHRONIC_DISEASE	0,42	0,34	0,37	0,69	0,82	0,75	0,80	0,79	0,80	6247
Micro avg	0,27	0,34	0,30	0,68	0,72	0,70	0,77	0,73	0,75	37862
Macro avg	0,27	0,40	0,29	0,55	0,56	0,55	0,81	0,60	0,63	37862
Weighted avg	0,33	0,34	0,33	0,67	0,72	0,69	0,78	0,73	0,75	37862

Table 7.5: Comparison of GPT-4-turbo and CODER models on IO NER (p-precision, r-recall, f-F_1 score

	GPT-4-turbo			CODER-27			CODER-100			
	p	**r**	**f**	**p**	**r**	**f**	**p**	**r**	**f**	**support**
CLINICAL_VARIABLE	0,33	0,48	0,39	0,76	0,82	0,79	0,88	0,75	0,81	7205
CHRONIC_DISEASE	0,42	0,35	0,38	0,74	0,86	0,80	0,84	0,83	0,83	11362
TREATMENT	0,32	0,33	0,33	0,73	0,84	0,78	0,79	0,79	0,79	13541
CANCER	0,34	0,40	0,37	0,87	0,58	0,69	0,84	0,76	0,80	5166
ALLERGY_NAME	0,04	0,64	0,07	0,00	0,00	0,00	0,95	0,07	0,13	588
Micro avg	0,30	0,37	0,33	0,75	0,80	0,77	0,83	0,78	0,80	37862
Macro avg	0,29	0,44	0,31	0,62	0,62	0,61	0,86	0,64	0,67	37862
Weighted avg	0,36	0,37	0,35	0,75	0,80	0,76	0,83	0,78	0,80	37862

Table 7.6: Comparison of models in low-resource settings (precision, recall, and F_1 score computed for BIO and IO labeling schemes).

	BIO p	**BIO r**	**BIO f**	**IO p**	**IO r**	**IO f**
BERT	0,5948	0,0003	0,0006	0,6441	0,0003	0,0007
BioBERT	0,6501	0,0004	0,0007	**0,7818**	0,0004	0,0007
Biomedical NER	0,5418	0,2005	0,2907	0,6092	0,2226	0,3232
BlueBERT	0,6547	0,0006	0,0013	0,7079	0,0007	0,0014
ClinicalBERT	0,6429	0,0003	0,0007	0,6820	0,0004	0,0008
CODER	0,6722	**0,7197**	**0,6896**	0,7458	**0,7953**	**0,7635**
GPT4	0,3294	0,3408	0,3279	0,3556	0,3676	0,3550
PubMedBERT	0,4565	0,1401	0,2104	0,5310	0,1659	0,2492
SciBERT	**0,6764**	0,0210	0,0407	0,7427	0,0230	0,0447

Table 7.7: Comparison of models in high-resource settings (precision, recall, and F_1 score computed for BIO and IO labeling schemes).

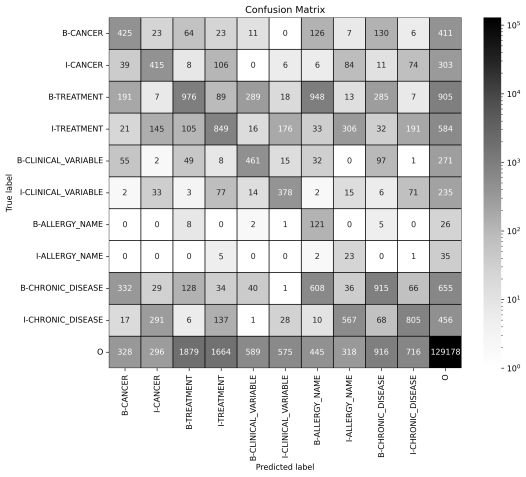
	BIO p	**BIO r**	**BIO f**	**IO p**	**IO r**	**IO f**
BERT	**0,7852**	0,0004	0,0008	0,7828	0,0004	0,0008
BioBERT	0,4426	0,0003	0,0007	0,7732	0,0003	0,0007
Biomedical NER	0,6645	0,0162	0,0317	0,7244	0,0177	0,0345
BlueBERT	0,3882	0,0004	0,0007	0,3933	0,0004	0,0007
ClinicalBERT	0,5777	0,0004	0,0007	0,7732	0,0004	0,0007
CODER	0,7780	0,7278	0,7454	0,8313	0,7798	0,7985
GPT4	0,3294	0,3408	0,3279	0,3556	0,3676	0,3550
PubMedBERT	0,7413	0,7468	0,7414	0,7958	**0,8036**	0,7983
SciBERT	0,7820	**0,7515**	**0,7648**	**0,8324**	0,8013	**0,8153**

ation scheme), the CODER model makes most of its mistakes with the `O` tag, meaning CODER does not confuse entity spans but primarily omits entities.

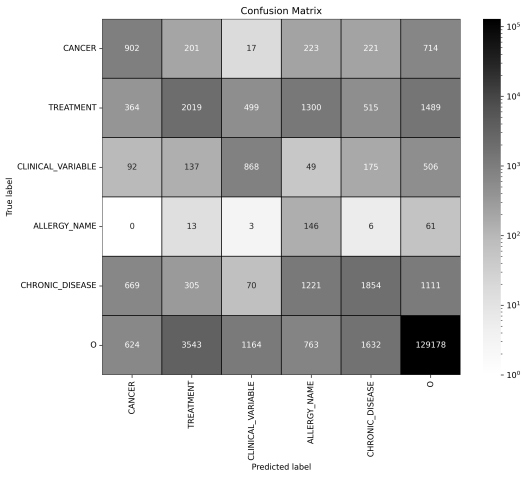
Table 7.6 shows the comparison of all models included in the study in a low-resource setting. Interestingly, CODER is the only BERT-based model that outperforms GPT-4-turbo in this setting. Other BERT-based models cannot operate effectively in the low-resource setting, and their pretrained capabilities are insufficient for meaningful work. Table 7.7 presents the comparison of the same models trained in a high-resource setting. As can be clearly seen, only when sufficient annotated data for fine-tuning is available can Transformer-based models correctly mark medical entities.

7.5 Conclusions

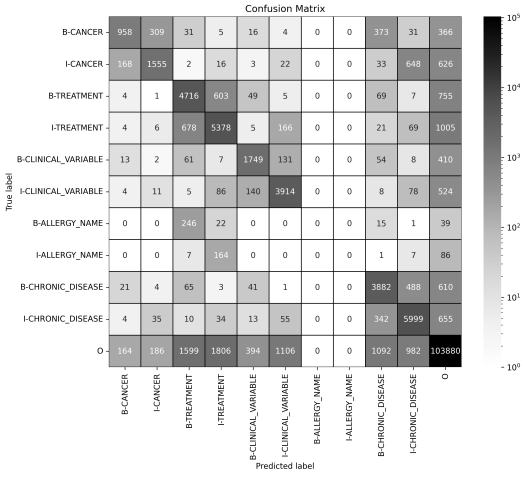
To our knowledge, this is the first attempt to examine the usefulness of a few-shot prompt engineering in the processing of eligibility criteria for clinical trials. We demonstrate that a simple few-shot prompting of the LLM can be used to perform



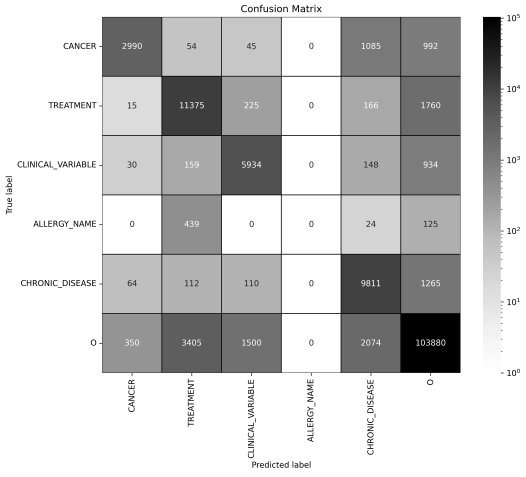
(a) GPT-4-turbo BIO



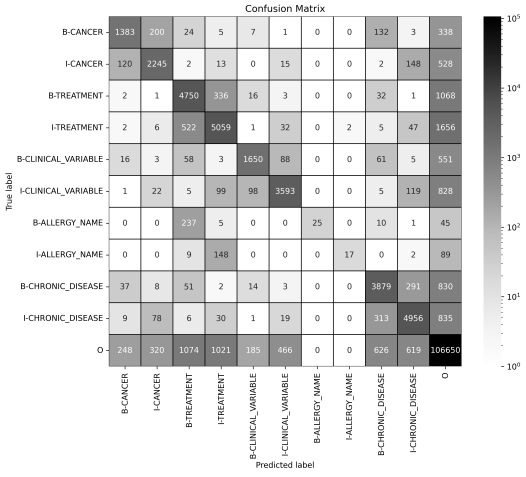
(b) GPT-4-turbo IO



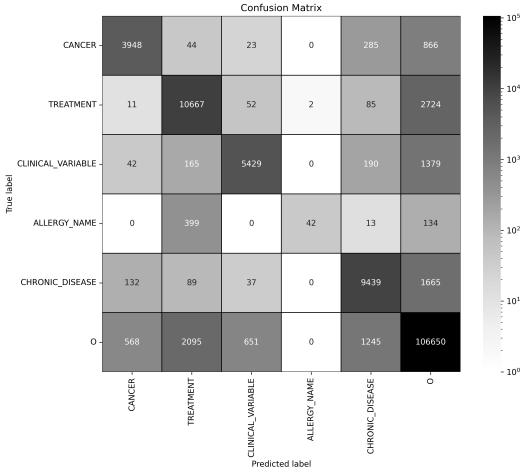
(c) `CODER-27` BIO



(d) `CODER-27` IO



(e) `CODER-100` BIO



(f) `CODER-100` IO

Figure 7.2: Confusion matrices for GPT-4-turbo and CODER models

named entity recognition, but this approach is most effective when no fine-tuning data is available. In general, BERT-based pretrained models, which are much smaller than GPT-4-turbo, perform better, especially when fine-tuning can be applied.

BERT-based models show superior performance over GPT-4-turbo on the NER task for eligibility criteria within the medical domain, largely due to the unique nature of these criteria. Eligibility criteria in medical literature represent a form of medical argot, a semi-language characterized by partial grammar and a vast number of specialized terms. BERT-based models excel at capturing relationships between closely placed tokens. This capability aligns well with the Begin-Inside-Outside (BIO) evaluation scheme commonly utilized in NER tasks, enabling these models to accurately identify and classify entities within such a dense and specialized lexicon. The distinction in CODER's efficacy stems from its pretraining on the Unified Medical Language System (UMLS) ontology, a comprehensive repository of medical terminology. This pretraining allows CODER to possess an exceptional ability to recognize and understand specialized terms present in medical eligibility criteria. In contrast, general language understanding and the broad comprehension of grammatical structure offered by larger language models like GPT-4-turbo do not directly contribute to the success of the NER task for medical eligibility criteria.

In conclusion, we propose possible future work directions. Our current prompting scheme is rudimentary and relies on the large language model's ability to notice hard-coded patterns in the prompt. The next step could involve creating an ensemble of prompts and aggregating generated outputs from the ensemble. Even more promising is transitioning from the domain of hard prompts (textual inputs) to soft prompts [143, 144] (dense numerical embeddings). The advantage of dense numerical embeddings is that they are trainable and can be fine-tuned to the specific task. Such trained prompts can, in theory, maximize the precision of information extraction from clinical trial protocols.

Another interesting research question is the feasibility of retrieval-augmented models for named entity recognition in biomedical texts. Biomedical ontologies are structured vocabularies that describe entities, relationships, and properties in a given knowledge domain. Examples of biomedical ontologies are described in Chapter 4. Retrieval-augmented models [145, 146] consist of two modules: a parametric language model responsible for generating output, and a neural retriever responsible for matching the input with data stored externally in ontologies or knowledge bases. Information stored in the non-parametric external storage is used at run-time to fine-tune the language model and provides additional information and context needed to process the input. We hypothesize that named entity recognition using a retrieval-augmented NER model might be a viable alternative to prompting a large language model directly.

Transitional Note

Having explored the scientific foundations in the preceding chapters, this thesis now shifts its focus toward the practical application of these concepts. This work has been conducted as part of an industrial PhD program designed to foster collaboration between academia and industry. The program facilitates partnerships between universities, PhD students, and their employers, with the ultimate goal of implementing the research outcomes within the companies involved.

The subsequent chapters are dedicated to the industrial implementation at Roche, the pharmaceutical company where I work. These chapters illustrate how the discussed theoretical aspects are put into practice, highlighting their real-world relevance and significance for Roche. They delve into the tools, methodologies, and practical considerations involved in implementing IT products in the pharmaceutical industry.

The remaining parts of this thesis focus on the work conducted within Roche. A key component of this work is a Proof of Concept (PoC) developed as part of the Clinical Trial Distribution Network (CTDN) initiative in which Roche participates. This PoC is currently under review, with the potential for future implementation in production.

Chapter 8

Prompt Engineering Hackathon at Roche

To further explore the potential of using large language models for eligibility criteria parsing, we organized a hackathon at Roche. The event brought together diverse expertise and generated innovative solutions for structuring conditions from criteria text. Organizing the hackathon was a collaborative effort of Roche employees. My role involved leading the initiative, coordinating and executing the event, defining objectives and tasks, selecting test examples and supporting its annotation, communicating with participants, and assessing submissions.

Section 8.1 outlines the objectives of the hackathon. Section 8.2 details the task to be solved and participation rules. Section 8.3 focuses on the organizational aspect of the event, detailing the steps to successfully organize the hackathon at Roche. Section 8.4 provides information on the participation rate and a general overview of the submissions. Section 8.5 describes the assessment process, including test dataset preparation and evaluation of submissions. Section 8.6 presents the main findings of the hackathon, demonstrating its impact on the parsing tool Proof of Concept (PoC) development. The parsing tool PoC is the primary deliverable of the implementation phase and will be described in Chapter 10.

8.1 Objectives

The hackathon was designed to achieve multiple goals focusing on two main aspects: technical and community-related. On the technical side, the primary objective was to advance the criteria parsing tool, while on the community side, the goal was to support professional growth among Roche employees and strengthen their sense of belonging.

Given the complexity of eligibility criteria parsing, diverse perspectives are invaluable. The hackathon aimed to gather innovative ideas and approaches to address specific challenges related to this task, thereby enhancing the parsing tool's accuracy. By leveraging the broad expertise of Roche colleagues from informatics and business areas, we sought to discover new methods for structuring eligibility criteria. An additional objective was to engage domain experts in the annotation process, ulti-

mately creating a small annotated dataset for testing submissions and aiding in the development of the eligibility criteria parsing tool. The outcomes and findings of the hackathon, along with lessons learned from participants, domain experts and out evaluation of submissions, were intended to directly inform the development of the PoC.

Equally important were the community-related goals. Roche is committed to fostering generative AI usage and enhancing employees' prompt engineering capabilities. The hackathon aligned with this purpose by supporting professional development through hands-on experience in crafting prompts and using an internal ChatGPT tool. The goal was to familiarize participants with the internal tool, provide an opportunity to learn about prompt engineering techniques, test them in real applications, and share experiences with others.

Moreover, the hackathon aimed to strengthen collaboration between people from different departments. It was designed to facilitate knowledge sharing, networking, and reinforcing the community spirit, which is a key focus of the company. It was intended to encourage people to take a break from their daily routines, engage in a stimulating activity, and delve into cutting-edge AI technologies. The event aimed to be both educational and enjoyable, motivating participants to explore technical topics while having fun.

Another objective of the hackathon was to spread awareness of the criteria parsing initiative within the organization. This was meant to ensure that the employees understood the project's focus, learned about the work of other departments, and gained insights into clinical trial protocol design, fostering a sense of unity and shared purpose within Roche. The combination of technical and community-related goals ensured that the hackathon was designed not only to influence the development of the parsing tool but also to promote the spirit of teamwork and innovation at Roche.

8.2 Task description and rules

The primary task of the hackathon was to develop prompt for the internal ChatGPT tool to transform clinical trial eligibility criteria text into a structured format. Based on the analysis of eligibility criteria and discussions with domain experts, two subtasks were identified for the hackathon. The participants had the option to tackle either one or both subtasks.

8.2.1 Subtask 1: Logical dependencies

The first subtask was related to logical dependencies between conditions. The named entity recognition approach, explored in Chapter 7, extracts a list of entities from criteria without preserving any relationships between them. However, for an AI-supported patient-trial matching system it is crucial to recognize which conditions are always required and which groups of conditions allow partial fulfillment. This requirement is also higlighted in the parsing tool requirements in Chapter 9.

Hackathon participants received a list of seven examples of single criteria statements with the desired structured output. They were instructed to focus on five condition types: biomarkers, stages, subtypes, metastases, and treatments. The goal of this subtask was to prompt the LLM chatbot to extract these conditions while maintaining the logical relations between them. The output needed to be a boolean statement using *AND*, *OR*, and *NO* operators, and parentheses where necessary, e.g.,

```
(condition_1 OR condition_2) AND condition_3 OR NO condition_4
```

To mimic a scenario with limited support from domain experts, no additional guidelines or examples were provided. An example of an input-output pair is included in Table 8.1.

8.2.2 Subtask 2: Temporal attributes

The second subtask addressed temporal attributes extraction, which is another requirement for the tool listed in Chapter 9. Analyzing the criteria content revealed that merely extracting conditions is insufficient. For determining patient eligibility, time-specific information provided in the criteria is crucial as it can significantly impact recruitment results.

Participants were asked to extract conditions along with their temporal specifications while also maintaining logical dependencies. They needed to prompt the LLM to use *"greater than"* (>) or *"less than"* (<) operators to state the temporal limitations of conditions, e.g.,

```
condition_1 > n weeks OR condition_2 > x days OR condition_3 < k days
```

Participants were provided with seven examples of criteria-output pairs for this subtask. An example of the input criterion and the desired output is provided in Table 8.1.

8.2.3 Participation rules

The hackathon was open to all Roche employees, regardless of their department. It was conducted fully virtually, allowing participants to work offline on their solutions for two weeks. Participants could work individually or in teams, based on their preferences. To participate, they needed access to the internal ChatGPT tool.

8.3 Event coordination

The successful organization of the hackathon required detailed planning and organization.

Table 8.1: Examples of input-output pairs for hackathon subtasks.

Subtask	Criterion	Desired output
#1	Patients must have metastatic and/or recurrent (distant or locoregionally recurrent) breast cancer and be HER2 non-over expressing per 2013 American Society of Clinical Oncology (ASCO)-College of American Pathologists (CAP) HER testing guidelines (0 or 1+ by immunohistochemistry [IHC]; and/or HER2 ratio < 2.0 and HER2 copy number < 4 signals/cell by in-situ hybridization [ISH])	(metastatic cancer OR recurrent cancer) AND (HER2 non-over expressing AND (0 IHC OR 1+ IHC OR (HER2 ratio < 2.0 AND HER2 copy number < 4 signals/cell))
#2	No prior chemotherapy, radiation therapy, or breast resection within 6 months of study entry	NO chemotherapy if < 6 months OR NO radiation if < 6 months OR NO breast resection if < 6 months

8.3.1 Promotion and kick-off

Initially, we needed to spread information about the event within Roche and encourage people to participate. Various communication channels were used to promote the hackathon and invite people to join the showcase and kick-off event.

We organized a virtual meeting to showcase the eligibility criteria parsing initiative and present the hackathon challenge. This included detailing the task, participation rules, and instructions on how to participate and submit solution. A FAQ (Frequently Asked Questions) document was created and shared with all registered participants to address common queries.

8.3.2 Communication and support

To facilitate ongoing communication, a Google Chat space was created for all hackathon participants and moderated by the organizers. This served as the main communication channel where participants could ask questions, share experiences, and receive updates from the organizing team. Reminders about deadlines, motivational posts, clarifications, and other useful information were regularly shared in this space.

Additionally, a final reminder email was sent to ensure participants did not miss the submission deadline. Upon receiving submissions, confirmation emails were sent to all participants.

8.3.3 Closing and recognition

Following the evaluation of submissions, a virtual closing meeting was held to announce the winners, gather feedback from the participants and highlight notable solutions recognized by the jury. This event provided an opportunity to celebrate the participants' efforts and share insights gained from the hackathon. At the end, the winning teams and individuals received awards.

8.4 Participation and submissions

The hackathon attracted active participation from various areas of Roche, bringing a wide variety of expertise, perspectives, and ideas. Eight teams took part, each consisting of one or two members from departments including Informatics (Data Science, Software Development, Business Analytics), Pharma, and Diagnostics. There were eight submissions for subtask 1 (logical dependencies) and six submissions for subtask 2 (temporal attributes).

During the hackathon, participants actively used the Google Chat space to communicate with the organizing team, ask questions, and respond to important announcements and reminders. This communication channel worked well. Moreover, the prepared guidelines and the FAQ document effectively clarified the task, as evidenced by the submission which demonstrated a correct understanding of the requirements. All participants followed the rules.

The solutions were developed using the internal ChatGPT tool with the GPT-3.5 model. Submissions included a complete conversation with the tool, starting with a prompt describing the task and providing instructions, followed by tests on the provided examples. The approaches and complexity of the submissions varied significantly and showed different methods to tackle the challenges.

Many solutions utilized in-context learning techniques, where a few examples with expected outputs were explicitly provided in the prompt. Additionally, a clear definition of the task was often presented at the beginning of the prompt. To enhance the solutions, participants employed a range of other strategies. Some used bullet points with rules to highlight aspects to focus on and the expected output format, while others included high-level guidelines for understanding components and dependencies. Additionally, several prompts incorporated the Persona Pattern, defining the model's role and the audience. Interestingly, explicitly stating the importance of the task and placing responsibility on the model also seemed to add value. While some solutions were capable of processing multiple criteria at once, others handled new examples one by one.

The solutions were highly innovative, demonstrating a wide range of creativity and ideas. Each team approached the problem from a unique perspective, using various methods. This diversity allowed us to evaluate which techniques were successful in specific applications and which were not. Each submission provided unique insights and potential ideas for further development of the parsing tool.

8.5 Assessment

The assessment of hackathon submissions followed a multi-step process designed to ensure fairness and reliability. A crucial aspect was the preparation of the test dataset, which was not revealed to the participants to allow a fair comparison of results and unbiased solutions, maintaining objectivity.

To create representative sets of examples for both subtasks, the criteria were carefully selected to maximize diversity and to present different challenges. Initially, examples were reviewed and proposed output boolean expressions were defined for each criterion. This preliminary versions of the annotated datasets were then shared with domain experts for verification and refinement. This approach was efficient and saved valuable time for the experts.

Once the test sets were finalized, the evaluation process started. To ensure impartiality, the submissions were de-identified so that the evaluators did not know the authors of the specific submissions during the assessment process. Each submission was thoroughly reviewed and assigned points according to predefined criteria.

The jury consisted of three members. Points ranged from 1 to 10 per example. Points were awarded based on accuracy of the boolean expression, including the correct usage of *AND/OR/NOT* operators, correct polarity of biomarkers, and the extraction of all relevant conditions using key terms only instead of full phrases or sentences. Additionally, innovative approaches were awarded extra points. Each evaluator worked individually without knowledge of others' assessment to avoid bias. After all evaluations were completed, the average score for each submission was calculated to determine the final score. The submission with the highest score was declared the winner. The subtasks were evaluated and awarded separately.

8.6 Key findings

The hackathon provided several significant insights that should be considered in the development of the eligibility criteria parsing tool. The valuable findings emerged not only from the winning solution but from all submissions, each offering unique perspectives. Additionally, the annotation process added further observations.

Firstly, the hackathon demonstrated that pretrained LLMs can interpret the text of eligibility criteria, extract meaningful information, and structure it into boolean expressions while maintaining logical dependencies and temporal attributes. The prompt plays a crucial role in this process, as even subtle differences can significantly impact the output. Utilizing in-context learning and incorporating input-output examples into the prompt effectively guides the model and increases accuracy. Moreover, clearly defining the model's role (Persona Pattern), the audience, and the task objectives in the prompt positively influences results. Another useful strategy is stating explicit guidelines for reasoning and inference in the prompt. Notably, a clear statement of model responsibility improves accuracy as well.

In addition to these findings from the submissions, the preparation of test sets and assessment process revealed areas for improvement in hackathon setup and task

definition. A predefined list of terms for boolean expressions could significantly enhance performance by ensuring the model uses consistent terminology, such as *"ER"* instead of synonyms like *"estrogen receptor"* or *"oestrogen receptor"*. Each relevant condition should be mapped to one specific term from a pre-existing list. This standardized list would also aid in the annotation process.

Furthermore, domain experts pointed out that some conditions are inferred from context rather than explicitly stated, such as *"metastatic breast cancer"* implying *"stage IV breast cancer"*. Defining these dependencies in guidelines would benefit annotators, participants, and the model. Discussion with experts also highlighted that eligibility criteria alone do not always provide a complete picture of population characteristic. References to the study title and objectives are necessary for full context, as criteria can sometimes be unclear without this additional information. During the test set validation, experts also noted that not all temporal attributes are equally important. Some attributes related to short periods before the study drug administration may not be critical, as patients can adjust behaviors or medications if they wish to participate in the trial. For example, extracting *"7 days before the first dose"* may be unnecessary. Therefore, it is important to define a threshold for the time period from which the tool should start extracting attributes.

In summary, the hackathon not only demonstrated the capabilities and potential of large language models in eligibility criteria parsing but also identified key areas to focus on when designing and developing the parsing tool.

Chapter 9

Requirements for an Eligibility Criteria Parsing Tool Using LLMs

As presented in previous chapters, the pharmaceutical industry faces significant challenges in managing and conducting clinical trials, particularly in terms of patient recruitment and achieving planned enrolment within a short timeframe. As clinical trials increase in complexity, precise structuring and analysis of trial eligibility criteria become essential for enhancing overall trial effectiveness.

However, the scoping review (Chapter 5) reveals a lack of generalizable and scalable tools for parsing eligibility criteria and a lack of comprehensive benchmarks. A few solutions utilize state-of-the-art NLP techniques, such us generative language models, but these are limited. This indicates a significant gap in technological advancements in this area. This gap has also been identified by Roche, the pharmaceutical company I work for. In the past, numerous projects aimed to develop a tool for eligibility criteria parsing, but the NLP techniques available at that time could not overcome obstacles such as the lack of annotated datasets, the absence of standardized nomenclature for criteria parsing, and the complexity of criteria semantics and syntax.

With the current advancements in technology, including the rapid development of large language models, building such a tool has become feasible. This chapter presents the functional and operational requirements for an eligibility criteria parsing tool using LLMs from Roche's perspective. The tool is designed to transform complex free-text eligibility criteria into a structured, machine-readable format, which can then be used in algorithms supporting trial design, optimizing trial operational efficiency, and matching patients to trials. The development of this tool directly responds to needs identified within Roche. The specifications presented in this chapter aim to meet technical demands and adhere to the specific requirements of a large healthcare company operating in a highly regulated environment.

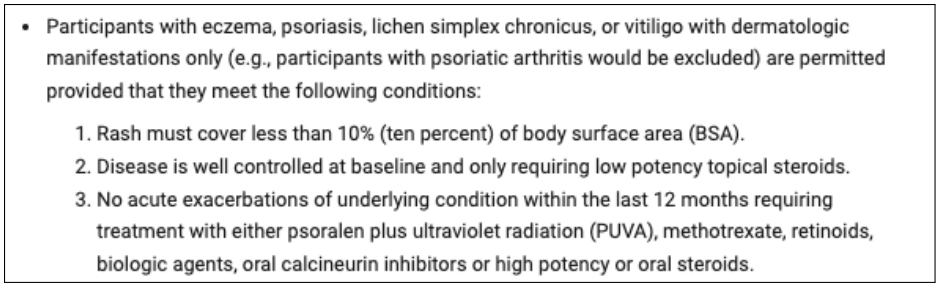
Section 9.1 outlines the technical specifications for the tool's input and output. Section 9.2 explains the data extraction and standardization processes, highlighting the need to align extracted data with medical terminology and preserve condition attributes and dependencies. Section 9.3 tackles the issue of limited annotated data and emphasizes the requirement for the tool to operate effectively without a large an-

notated dataset. Section 9.4 focuses on the technical best practices for managing and deploying ML systems efficiently and reliably. Section 9.5 emphasizes the importance of security and compliance, stressing the need to follow regulatory and internal standards to protect patient safety and maintain data confidentiality. Section 9.6 covers regular maintenance and user support topics, which ensure tool efficiency and user satisfaction. Section 9.7 discusses the need for improved workflow efficiency and reliable validation, highlighting the need for the tool to enhance business processes to be launched. Section 9.8 explains the significance of diverse expertise in tool development, underscoring the requirement for a multidisciplinary team to build reliable AI-driven solutions.

9.1 Input and output specification

A crucial aspect of the tool's design is specifying its input and output, which is highly dependent on its objective. This tool is designed to process raw eligibility criteria from a trial protocol and transform them into a machine-readable format. Therefore, the tool must accept unstructured text of inclusion and exclusion criteria for a single trial as input.

While reviewing previous works, it was observed that many solutions processed individual criteria one by one. However, this approach is not ideal, as it fails to maintain the context between individual criteria, which is essential in real conditions. Additionally, separating the criteria section into distinct units can be challenging due to the complex structure of this section. There are two frequently used techniques for separating the criteria into single units: splitting by the new line character or splitting by sentences. Both methods have drawbacks, since isolating sentences or lines can lead to losing the full context and meaning of a criterion. For example, Figure 9.1 presents a fragment of exclusion criteria, where the first sentence (and first line as well) states that the subsequent lines describe conditions allowed for participation. Without this sentence, the algorithm may misinterpret the subsequent criteria as exclusions rather than exceptions.



- Participants with eczema, psoriasis, lichen simplex chronicus, or vitiligo with dermatologic manifestations only (e.g., participants with psoriatic arthritis would be excluded) are permitted provided that they meet the following conditions:
    1. Rash must cover less than 10% (ten percent) of body surface area (BSA).
    2. Disease is well controlled at baseline and only requiring low potency topical steroids.
    3. No acute exacerbations of underlying condition within the last 12 months requiring treatment with either psoralen plus ultraviolet radiation (PUVA), methotrexate, retinoids, biologic agents, oral calcineurin inhibitors or high potency or oral steroids.

Figure 9.1: An example of a multi-line exception in the exclusion criteria (source: ClinicalTrials.gov).

There are also even more complex criteria that describe different requirements for various trial cohorts. For instance, an inclusion criteria section may begin with

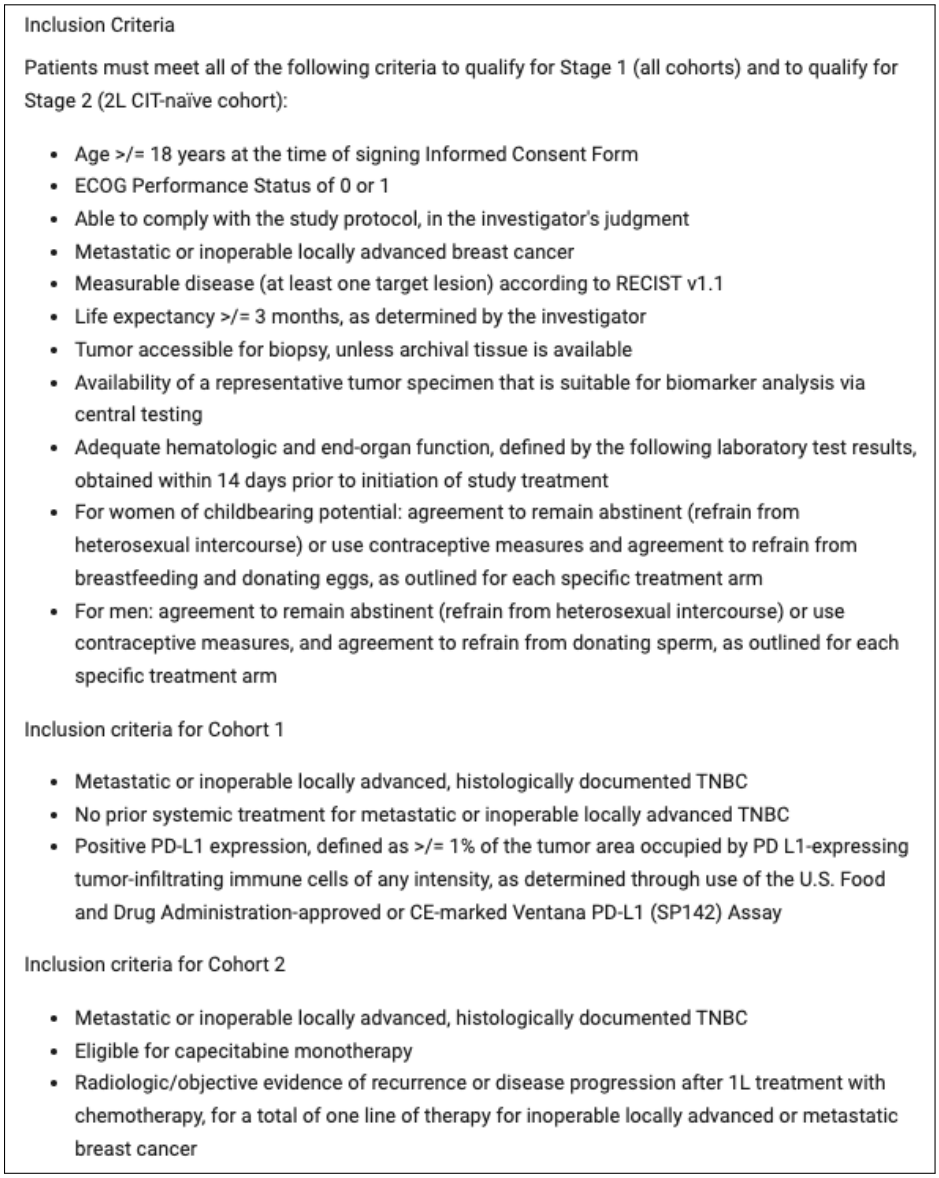
general criteria applicable to all sub-populations, followed by subsections with headers specifying the cohort and a few lines defining the requirements. An example of such criteria is shown in Figure 9.2. In this case, processing each line, sentence, or paragraph individually is not sufficient. Therefore, a key specification for this tool is to process the free-text from both sections (inclusion and exclusion criteria) at once to maintain the context between them as well as between individual lines and sentences.

Moreover, the tool must output structured data in a format that is easily usable in further analysis. The output data should retain information about individual required or excluded conditions. Conditions that are allowed but not required are not as important in this application unless they specify exceptions for exclusions. The output should list the condition name and some attributes detailing the inclusions or exclusions. For example, the output for a condition excluding metastases should detail the sites of metastases, if defined, or the required laboratory test for creatinine should be connected with its numerical threshold and unit in the output. This indicates that a simple list of named entities is inadequate for this task. A more complex structure, such as JSON or XML, is required to appropriately capture and represent the data.

9.2 Data extraction and standardization

In order to ensure the tool's output is usable in other systems, relevant data needs to be extracted and standardized. Standardization involves aligning the structured format with existing medical terminologies or at least providing a dictionary to map terms to medical terminologies. This feature enhances the tool's interoperability and applicability. Moreover, to maintain accuracy in further processing, the output should preserve condition attributes and dependencies between conditions.

Clinical concepts in eligibility criteria are expressed in various ways using different terms. Extracting them directly from text without standardization would result in an unnecessary complex structure. This can be simplified by merging synonymous terms into a single concept used in the final output. Given the input is free-text written by different people for different trial objectives, the variety of conditions and related aspects that can be extracted is vast. Therefore, it is important to select specific conditions and attributes for the tool to focus on. This helps define a limited scope for extraction and improves the readability of the output. A very detailed structure would not be much easier to read than the unstructured criteria text. Thus, a clearly defined list of output fields (e.g., condition, upper limit threshold, unit) and a list of allowed terms are required to build the tool. Additionally, not all conditions are equally important for further processing. For example, if the eligibility criteria state that the patient needs to sign a consent form, extracting this information would not benefit other systems and can be ignored during parsing. A predefined list of condition types (e.g., laboratory tests, cancer biomarkers, pregnancy) that need to be included in the output structure should be created. All these lists may vary for



Inclusion Criteria

Patients must meet all of the following criteria to qualify for Stage 1 (all cohorts) and to qualify for Stage 2 (2L CIT-naïve cohort):

- Age >/= 18 years at the time of signing Informed Consent Form
- ECOG Performance Status of 0 or 1
- Able to comply with the study protocol, in the investigator's judgment
- Metastatic or inoperable locally advanced breast cancer
- Measurable disease (at least one target lesion) according to RECIST v1.1
- Life expectancy >/= 3 months, as determined by the investigator
- Tumor accessible for biopsy, unless archival tissue is available
- Availability of a representative tumor specimen that is suitable for biomarker analysis via central testing
- Adequate hematologic and end-organ function, defined by the following laboratory test results, obtained within 14 days prior to initiation of study treatment
- For women of childbearing potential: agreement to remain abstinent (refrain from heterosexual intercourse) or use contraceptive measures and agreement to refrain from breastfeeding and donating eggs, as outlined for each specific treatment arm
- For men: agreement to remain abstinent (refrain from heterosexual intercourse) or use contraceptive measures, and agreement to refrain from donating sperm, as outlined for each specific treatment arm

Inclusion criteria for Cohort 1

- Metastatic or inoperable locally advanced, histologically documented TNBC
- No prior systemic treatment for metastatic or inoperable locally advanced TNBC
- Positive PD-L1 expression, defined as >/= 1% of the tumor area occupied by PD L1-expressing tumor-infiltrating immune cells of any intensity, as determined through use of the U.S. Food and Drug Administration-approved or CE-marked Ventana PD-L1 (SP142) Assay

Inclusion criteria for Cohort 2

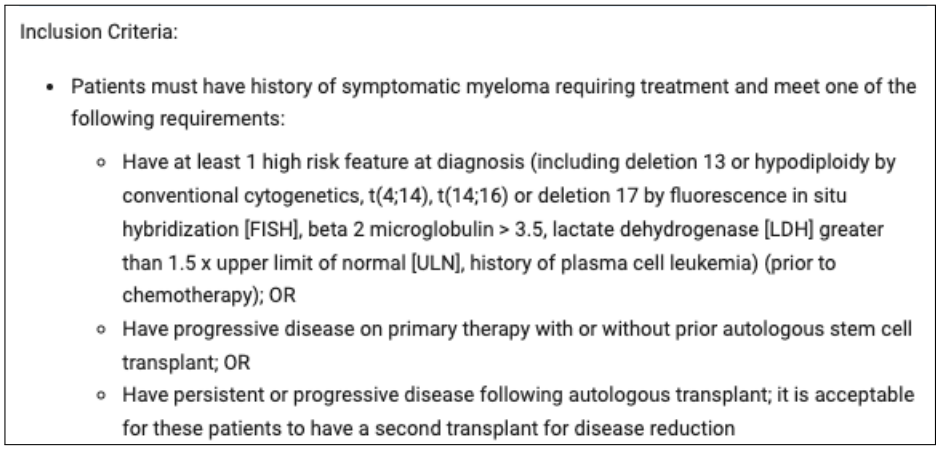
- Metastatic or inoperable locally advanced, histologically documented TNBC
- Eligible for capecitabine monotherapy
- Radiologic/objective evidence of recurrence or disease progression after 1L treatment with chemotherapy, for a total of one line of therapy for inoperable locally advanced or metastatic breast cancer

Figure 9.2: An example of inclusion criteria with sub-populations (source: ClinicalTrials.gov).

different therapeutic areas. For instance, breast cancer trials have different inclusions and exclusions compared to Alzheimer's disease trials.

Maintaining logical relationships between conditions is another important aspect. Not all conditions mentioned in the criteria need to be fulfilled to include the patient in the study. Sometimes, criteria sections list alternative conditions, where only one or a subset must be fulfilled for eligibility. An example of such criteria is illustrated in Figure 9.3. A similar situation occurs with exceptions for exclusions (Figure 9.1), where not all conditions are mandatory for inclusion in the study. As a consequence, a flat structure of extracted conditions is insufficient for this task. It needs to re-

flect whether a condition is mandatory or if one of the conditions from a group is mandatory.


Inclusion Criteria:

- Patients must have history of symptomatic myeloma requiring treatment and meet one of the following requirements:
  - Have at least 1 high risk feature at diagnosis (including deletion 13 or hypodiploidy by conventional cytogenetics, t(4;14), t(14;16) or deletion 17 by fluorescence in situ hybridization [FISH], beta 2 microglobulin > 3.5, lactate dehydrogenase [LDH] greater than 1.5 x upper limit of normal [ULN], history of plasma cell leukemia) (prior to chemotherapy); OR
  - Have progressive disease on primary therapy with or without prior autologous stem cell transplant; OR
  - Have persistent or progressive disease following autologous transplant; it is acceptable for these patients to have a second transplant for disease reduction

Figure 9.3: An example of alternative conditions in inclusion criteria (source: ClinicalTrials.gov).

Furthermore, eligibility criteria often specify the temporal attributes of conditions, such as the time period of absence or presence of a condition, usually expressed in relation to the screening date or first drug dose. Adding this temporal data to the output is crucial, as it can significantly modify the inclusive or exclusive meaning of a criterion. For example, a trial excluding patients with a history of cardiovascular disorders in past five years does not disqualify a patient who had an aorta disease 20 years ago.

Some requirements have a very short time period for excluding or including certain conditions, like a week or two weeks before the first drug dose. These requirements are relatively easy to meet because patients can often stop taking specific medications, avoid alcohol, or cease concurrent therapies for a short time before the trial participation if they wish to enroll in the study. Criteria with such short temporal requirements are less important from a trial design optimization or patient-trial matching perspective. Therefore, a time threshold should be defined to skip these conditions in the final structured output of the tool.

Another important aspect is whether a condition is related to the current state or the past. Eligibility criteria often describe the current health status of the patient, but some conditions refer the patient's health history. For example, the exclusion criterion *"History of high blood pressure"* excludes patients who have had blood pressure in the past, even if they no longer suffer from it. Hence, the structured output of the parser should include a field defining whether a condition relates to the present or past.

The output should also specify the medical characteristics (e.g. disease stage, metastases, biomarkers) along with their context. Most criteria describe the main indication of a trial, e.g. breast cancer, specifying the required stage, subtype, biomarkers, etc. However, similar condition types may be used to characterize other con-

comitant disorders. For instance, in the exclusion criteria of a breast cancer trial, there might be a statement about *"other metastatic or locally advanced tumors"*, while the trial accepts metastatic breast cancer. It is important that the structured output distinguishes that these subtypes relate to other tumors, not breast cancer, to avoid incorrect patient disqualification.

Moreover, the previous works, reviewed in Chapter 5, focus on recognizing negations in the criteria. This is crucial because not all conditions in the exclusion criteria are true exclusions, and similarly, some excluded conditions are mentioned in the inclusion criteria. As highlighted several times in this thesis, there are many exceptions mentioned in the eligibility criteria, as well as some conditions are explicitly expressed with negations, such as *"No CNS metastases"* in the inclusion criteria. Extracting only the *"CNS metastases"* entity is inaccurate, as it does not retain the full meaning. Therefore, incorporating negation information into the output structure is necessary. This will also help distinguish between conditions extracted from inclusion and exclusion criteria.

As presented above (Figure 9.2), some eligibility criteria are highly complex, defining different sets of inclusions and exclusions for different cohorts. It adds another level of complexity to the final output of the tool, since it changes the meaning of the criteria significantly. This aspect can be maintained using logical relationships, transforming subpopulations into logical sequences of conditions, such as *"Patients must have either non-small cell lung cancer and brain metastases, or breast cancer and leptomeningeal disease"*. In this example, the first subpopulation is lung cancer patients, and the second is breast cancer patients.

In summary, the tool should extract only the conditions that are important for further processing and express them with standardized predefined terms to make them maximally useful for other applications. Each condition type should be associated with its essential attributes, defining patient eligibility. Several aspects must be included in the output structure, such as numerical attributes, units, logical dependencies, temporal attributes, current/past flag, negations, and subpopulations. This all indicates that despite reducing the amount of extracted information, the final structure returned by the tool is still complex.

9.3 Limited annotated data

Traditional NLP methods require training or fine-tuning a language model for a downstream task to achieve accurate results. This training process, as well as further model evaluation, demands a substantial annotated dataset. However, the pharmaceutical domain lacks adequate biomedical datasets for such purposes. In fact, there is no reliable dataset available for developing and testing a tool for eligibility criteria parsing that meets the requirements outlined in previous sections. Creating such a dataset from scratch is also very challenging and labour intensive. It requires the involvement of subject matter experts and extensive manual effort to generate enough examples for model training and evaluation.

In a large pharmaceutical company like Roche, it is practically impossible to pull employees away from their daily responsibilities to engage in such a burdensome task without any guarantee of future gains from the tool. Using publicly available datasets, such as the *Chia* or *Clinical Trial Parser* dataset, is not feasible due to the custom requirements set for this tool. However, recent developments in generative AI and in-context learning, evaluated in Chapter 7, along with the use of Retrieval-Augumented Generation, might offer a breakthrough.

Therefore, a key specification for the eligibility criteria parsing tool is to be developed without the need for a substantial annotated dataset. While it is clear that a limited dataset is required for tool evaluation, it should be no more than a dozen examples. The tool can also utilize predefined guidelines for information retrieval from eligibility criteria, thereby minimizing the annotation effort. Additionally, incorporating new trial types or therapeutic areas should not require a large annotated sample. It is essential that the tool can generalize well with only a few new examples and instructions.

9.4 MLOps best practices

In industrial IT applications, effectively managing the entire lifecycle of products is crucial, and machine learning applications are no exception. Robust operational practices ensure that these solutions are developed, deployed, and maintained efficiently. MLOps (Machine Learning Operations) addresses this need by integrating practices from machine learning, data engineering, and DevOps. By adopting MLOps, processes related to ML system deployment and maintenance can be standardized, optimized, and simplified. Importantly, MLOps not only focuses on deployment but also covers other phases of the ML application lifecycle: research, development, and post-deployment.

Recently, a new branch of MLOps has emerged - LLMOps, which focuses on the operationalization of large language models. LLM-based applications differ from standard ML applications due to several factors. First, many LLM solutions utilize pretrained hosted models and API connections (like OpenAI APIs), meaning the model itself is not deployed. In such cases, there is no model training or fine-tuning; instead, the experimentation phase focuses on prompt engineering, sometimes involving in-context learning. In this setup, experiment tracking includes prompt templates used, custom pipeline parameters (such as preprocessing input data and post-processing model output), the generated content, and metadata of the pretrained model.

LLM-based applications also present unique challenges compared to standard ML solutions, even when they involve training, fine-tuning, or deploying self-hosted models, due to the significant model size. LLMs have billions of parameters and require vast amounts of data for training, which needs adequate computational resources. Additionally, evaluating LLMs is more complex than other models because they generate natural text, which is harder to validate accurately. Furthermore, the

use of LLMs and Retrieval-Augmented Generation introduces another component to the MLOps architecture: a vector database. Nevertheless, it is important to note that the LLMOps field is continually evolving, with new tools and frameworks being introduced to support LLM operationalization. This area remains in the research phase.

To develop the eligibility criteria parsing tool that uses LLMs, it is essential to ensure effective development throughout its lifecycle, so that the solution is efficient, reproducible and scalable. This sections describes the most important strategies and best practices to achieve these goals.

9.4.1 Reproducibility

A first important aspect of MLOps is ensuring ML reproducibility, which means the ability to recreate the ML workflow executed previously and produce exactly the same results. Reproducibility includes version control, proper experiment tracking, randomness reduction, and reliable environment management. In software development, using a robust version control ecosystem (e.g. GitLab, GitHub) is essential. It tracks code changes, enables rollback to previous versions, allows parallel work on the code, and supports code reviews. Developers must ensure that all changes are committed and pushed. They should work on specific branches following a defined branch workflow and naming convention, avoid committing directly to the main branch, and precede merges with a code review. The project repository should contain proper documentation, including a README.md file that summarizes the project and provides instructions on how to install, develop and execute the code. Moreover, all credentials (passwords, API keys) must not be hard-coded in the project files and pushed to Git repository but managed as environment variables. According to Roche guidelines, data files cannot be stored in Git as well. Separate data storage solutions are required to ensure data security and compliance.

To enforce rigid reproducibility of the environment, the project should specify all its dependencies in a file listing all required packages along with pinned versions. Using containerization is a good practice to ensure a consistent environment. Hence, the project should also include a Dockerfile with specified commands to create the Docker image. Container registries should be used to manage and distribute the Docker images.

Reproducible ML involves systematic experiment tracking, which means logging all executions of ML pipelines. All metadata of the experiment, including configuration parameters, dataset version, model hyperparameters, preprocessing and post-processing parameters, artifacts, and evaluation results, should be stored in order to reproduce the selected experiments when needed. In LLM applications, it is also important to track the prompt templates and samples used for in-context learning. Templates should be stored along with the generated output and, ideally, with some evaluation indicators to assess and filter the experiments more effectively. This filtering functionality is supported by many dedicated experiment tracking tools, which

can be easily integrated with ML pipelines, such us Neptune.ai,[1] Comet,[2] Weights & Biases,[3] for instance. Using experiment tracking tools instead of tracking in files or spreadsheets significantly increases development efficiency. They provide interfaces for experiment browsing, comparing, visualizing, and even creating dashboards that can be shared with other stakeholders.

9.4.2 Code quality

MLOps also focuses on code quality and proper documentation. Developers should ensure that the code is well-crafted with modules, classes, and functions, and minimizes repetitions and complexity. Code should be documented with docstrings, and type hints should be used to make development efficient and error-robust. Basic unit tests should be implemented and successfully executed before code commits. Achieving a defined minimum code coverage, which means covering a significant portion of the code with unit tests, is a good practice.

Using code linter tools is also recommended. These tools perform static analysis of the code syntax and check for discrepancies with existing code standards. They help prevent errors, typos, bad formatting, and unused code. Some tools list the code lines to be changed, while others apply the changes automatically. A very useful feature is using linters as pre-commit hooks, which inspect the code about to be commited before executing the commit according to predefined rules. They run automatically each time the developer attempts to commit the code and do not allow the commit until all discrepancies are fixed. This ensures the consistent quality of the code coming from different developers in the team.

9.4.3 CI/CD

Best practices in software development include implementation of CI/CD, which stand for Continuous Integration and Continuous Delivery (or Deployment). A Continuous Integration pipeline is triggered by code changes and executed on the source code in the repository. Depending on the setup, it involves running various tests (unit tests, integration tests, etc.) and building the application. This process prevents breaking existing functionalities with new changes in the repository and ensures that the code is tested in a consistent environment. Continuous Deployment automates the deployment of solutions to different stages - development, test, and production - and reduces manual effort. It checks ML model compatibility with the infrastructure, verifies if the Docker image is created without errors, and employs inference service testing in an isolated environment before deployment. A successful build can be deployed automatically or triggered by the developer. Implementing CI/CD in a project supports the automatic building,

[1] https://neptune.ai/
[2] https://www.comet.com/
[3] https://wandb.ai/site

testing, and deployment of IT products. It enables quicker releases of new features, ensures high quality, and reduces the risk of failures.

9.4.4 Data management

Effective data management is a critical component of MLOps, as data is the fuel of all ML systems. This includes robust data storage solutions and rigorous validation processes. An ML project needs storage systems that provide seamless access to training and inference data and can handle large datasets efficiently. The storage solutions must be selected and configured according to data security and compliance polices. Separate storage locations are required for raw data, development/test data, and production data. Additionally, datasets should be versioned to ensure reproducibility and traceability.

Furthermore, it is crucial that all data processing steps are automated and well-documented. Data pipelines must be designed to be scalable and reliable, efficiently handling data ingestion, processing, and transformation without manual intervention. Relying on manual processes is inefficient, prone to errors, and can lead to reproducibility issues.

Moreover, defining an input data schema and implementing validation mechanisms to check if the input data aligns with the expected format are crucial. This is needed because sometimes data inconsistencies, such as incorrect formats or missing fields, may not trigger any errors in the ML pipeline, resulting in completely incorrect predictions. Without a validation step, such discrepancies are difficult to detect. The model execution should occur only after validation functions confirm data quality and availability.

9.4.5 Model registry

A reliable ML project includes maintaining a model registry where all models are stored and versioned along with their metadata. This practice ensures effective tracking of models, easy access, and efficient deployment. It also plays a crucial role in addressing any issues in the inference service or drops in prediction quality, allowing easy rollback to a specific version of the model. The registry needs to contain metadata for each model, including training datasets, preprocessing methods, model hyperparameters, training configurations, and performance metrics. This information can be managed by linking to records in an experiment tracking tool, which stores all these details as well.

In the case of an LLM application, not only the model but also the utilized prompt template significantly impacts output quality. Therefore, it is important to maintain a registry that includes models and prompt templates, along with their metadata and configurations. This registry should capture parameters such as the temperature setting, which determines the randomness of the generated output, and other hyperparameters like maximum token length, top-k sampling, and top-p sampling.

From an operational perspective, the model registry should be integrated with

CI/CD pipelines to streamline the deployment of models. Moreover, access control cannot be overlooked, as it ensures that unauthorized individuals cannot modify or remove models.

9.4.6 Scalability

Maintaining tool efficiency as data volume increases and more users interact with it is a crucial aspect of scalability. Scalability highly influences the choice of infrastructure, architecture design, and memory utilization. Leveraging cloud platforms can be beneficial due to their scalable infrastructure services. Implementing caching mechanisms is another effective practice, as it reduces the load on the backend and enhances response time. Additionally, using a modular code structure and conducting regular performance profiling helps identify bottlenecks, allowing for optimizations that keep the system efficient at scale.

An effective approach in machine learning to improve service latency is to avoid processing all operations in real-time upon user request. Instead, precalculated results, for instance, embeddings or historical predictions, can be stored in a database. This way, handling a user request does not trigger complex model execution unless absolutely necessary. Often, if the user does not input additional data for calculation, on-demand execution is not needed - and this is precisely the case with the eligibility criteria parser.

The tool will utilize data from published clinical trial protocols, which are not expected to change frequently, allowing the model to process this data in advance and store the structured output in the database. This reduces latency because no LLM query needs to be executed in real time; rather, the data is simply retrieved from the database. Consequently, the concern of high execution times for the structuring algorithm is less critical in terms of scalability. Data from clinical trials will be parsed once, and the output will be saved and subsequently utilized in the browsing tool and other applications.

Nevertheless, stress tests should be conducted on the browsing application to measure the response time under heavy user load and ensure that the solution remains responsive and meets latency requirements.

9.4.7 Monitoring and logging

Implementing logging mechanisms integrated with a monitoring solution is fundamental for any IT product. The system needs to log all information that might be useful for monitoring purposes, such as request and response data, process metadata (start and end time, user ID, system version), intermediate statuses, execution times, hardware usage, and error messages. In the ML field, logging the model version is also helpful. Any information that can support system inspection should be saved in logs. These logs should then be ingested into a system health monitoring tool, which displays service availability, latency, and other critical data for developers. In machine learning, this could include evaluation metrics to assess the accuracy of the

predictions and the quality of the model. Useful metrics also include the number of predictions per second (in batch inference) or time per prediction.

Adding alerting mechanisms is also a good practice, ensuring that developers are notified about any decreases in model accuracy, service unavailability, or high latency. These alerts enable developers to take appropriate actions to address the issues. For example, decreases in model performance can be caused by data drift and may require modifications in training examples or, in generative AI, adjustments in prompts and guidelines for large language models. Additionally, implementing automatic actions in case of issues, such as rolling back to a previous working version of the system, triggering model retraining, retrying API connections (relevant when using hosted LLMs), or at least displaying a message to users, can further enhance system reliability.

9.4.8 Flexibility

Flexibility is a very important feature of an AI system in a pharmaceutical company where the environment and demands are highly dynamic due to the research nature of the organization. It ensures that the tool not only responds to current needs but is also adaptable to future requirements, handling upcoming changes and expansions.

An AI tool must be easily modifiable to meet growing business requirements. It needs to efficiently manage increasing amounts of data, expand beyond its initial scope, and handle a growing number of tasks. For the eligibility criteria parsing tool, this means it must be relatively simple to incorporate new therapeutic ares, condition types, or guidelines. Future advancements should not require a complete redesign. Instead, minor modifications should be sufficient to accommodate new needs.

Furthermore, the tool must generalize well beyond the scope of its training data examples. This capability is particularly lacking in the use of regular expressions or rule-based algorithms - approaches commonly employed for eligibility criteria parsing, as identified in the scoping review (Chapter 5). Highly specific rules, especially those handling exceptions, may not generalize effectively with an increased volume of data, new output specifications, or new vocabulary used in the eligibility criteria. While some challenges, such as data drift, cannot always be overcome even by more complex models, the models should be designed to generalize as much as possible during the development phase, ensuring that tuning them for new use cases does not require a fundamental rebuild.

9.5 Security and compliance

Considering the tool will be used in a highly regulated environment within a pharmaceutical company, ensuring security and compliance is crucial. The tool must adhere to stringent regulatory standards and internal guidelines to prevent errors that could impact patient safety and to protect confidential information. Implementing robust mechanisms for data accuracy, access controls, and secure data handling is essential to prevent mistakes that could have serious consequences. Additionally, the

tool must undergo an internal validation process before being approved for production. This ensures that the tool meets regulatory requirements and maintains high standards of accuracy and confidentiality.

With advancements in AI applications in the healthcare industry, researchers have started focusing on the responsible and ethical implementation of AI in the biomedical domain and creating relevant guidelines [147, 148, 149, 150]. Moreover, legislative bodies worldwide have begun publishing AI regulatory acts, such as the European Union's Artificial Intelligence Act, to establish formal frameworks ensuring the ethical and safe use of AI technologies. This has become even more important in the era of generative AI, where LLM applications are spreading rapidly and are prone to security and noncompliance issues. Recognizing this problem, researchers and organizations, with the World Health Organization (WHO) at the forefront [151], have started developing ethical guidelines specifically for the use of LLMs. Consequently, the eligibility criteria parsing tool using LLMs must adhere to these ethical guidelines to be implemented at Roche. There are several aspects that need consideration even before starting the development of such a tool.

A fundamental principle in the deployment of AI tools at Roche is accountability. The company has established a comprehensive governance framework for the design and application of AI. This framework includes assigning technical and validation leads to each product, conducting system risk assessments, adhering to specific procedures, maintaining detailed system documentation, ensuring tool transparency, and implementing system monitoring. Each application undergoes a thorough validation process to evaluate its business and regulatory impact versus its risk before production deployment. Additionally, Roche has designated specialized teams responsible for ensuring that the AI tools align with these established principles. This rigorous approach not only mitigates potential risks but also enhances the reliability and effectiveness of AI applications within the company.

Moreover, according to WHO recommendations, humans should control healthcare systems and medical decisions. Therefore, the target solution should be a decision support tool, not an autonomous one, as will be elaborated in one of the following sections.

An essential ethical consideration in AI use is data privacy, which is highly at risk with cloud-based models (accessed via API) that require all data (including sensitive ones) to be transmitted and processed by third-party providers, such as OpenAI GPT, Google Gemini [152], or Anthropic Claude [153] models. When using these APIs, all prompts and data entered into the model leave company systems and are exposed to unauthorized parties. Consequently, such insecure use of models on confidential data is forbidden at Roche. All employees are obliged to protect sensitive personal data, company secrets, and Intellectual Property. Patentable information or trade secrets cannot be revealed outside. One option to ensure this, when building an LLM-based solution, is to use publicly available data only, but this is not always feasible. Another solution is to use on-premises models that can be downloaded and hosted on the company infrastructure, such as Meta LLama [154], Mistral [155], or the Huggingface Transformers [156]. These models are ideal for scenarios demand-

ing data privacy, offline usage, and complete control over the model environment. However, according to the latest version of the Holistic Evaluation of Language Models leaderboard [157, 158], hosted models like Claude 3 Opus, GPT-4o, GPT-4, Gemini 1.5 Pro, and GPT-4-turbo rank highest in language understanding tasks, which indicates their superior performance. Therefore, Roche has established cooperation with OpenAI and Microsoft Azure to provide secured access to OpenAI models. The company recommends using the GPT models or self-hosted models in LLM-based solutions to ensure data privacy.

Another important aspect to consider when implementing an LLM-based solution is the phenomenon of hallucinations. While generative language models are very powerful, they may produce incorrect responses that are contextually inappropriate, factually incorrect, or unintended. Because large language models generate text based on probability estimation, it is possible that the generated information appears reliable, but it is not grounded in reality. Additionally, these models have a knowledge cut-off, meaning they were trained on data available up to a specific point in time. For example, the GPT-4-turbo model's knowledge cut-off is December 2023, so it is not aware of information that became publicly available after that time. Hence, it is important to acknowledge and address the risk of receiving inaccurate output. When considering the application of an LLM-driven parser, it is crucial to select an appropriate use case, where the consequences of incorrect information are less severe or where a margin of error can be tolerated. If it is not possible to easily verify the output's correctness or the application requires high accuracy, then it is a very risky use case, which is not accepted by Roche. There should always be a human in the loop to review the tool's output and assess its quality.

Fairness and minimization of bias are also critical AI ethics principles, especially for a pharmaceutical company making crucial decisions on drug development and clinical trials, which impact human subjects. One of best practices to ensure AI fairness is the careful selection of datasets used for model training. The data should be inclusive and representative to avoid discrimination based on gender, race, ethnicity, sexuality, political views, or religion. However, with pretrained LLM, the selection of datasets is already done by third-party developers. If the training data for a model is biased, there is a high probability that the tool will reproduce or even amplify these biases. Moreover, when developing a solution using a pretrained LLM, it is essential to be cautious about the content of the prompt. Discriminatory language in prompts may result in biased responses. Additionally, the developer of the eligibility criteria parser can influence the selection of examples for in-context learning or the documents for a RAG application. These should be created cautiously to ensure the tool's fairness, so it will generate text that respects diversity and inclusion. The tool should be also continuously monitored and adapted to correct any biases.

A very useful feature of AI tools is the explainability of their decisions. This capability helps avoid treating the machine learning model as an opaque "black box" by providing insights into why the model produced a specific output. With simpler models, such as linear regression or decision trees, it was possible to understand their functioning and even perform their calculations manually. However, as algo-

rithms have become increasingly complex, they have also become less transparent and more difficult to interpret. Nevertheless, the mechanism of action of the implemented AI technologies should be understandable to developers to minimize the risk of improper functioning when deployed in production. Where feasible, the tool should include a module that explains the reasoning behind the algorithm's decisions. While not all models can be explained, having this capability is a significant advantage. It not only aids in system risk assessment but also increases user confidence in the final decisions made with the tool's assistance.

Furthermore, environmental sustainability is an important consideration in the design of all IT products, though it is one of several critical factors to be balanced. Resource and energy consumption should be carefully evaluated to minimize ecological impact, yet this must be weighed alongside other technical and business priorities. Developers are encouraged to optimize the use of algorithms and thoughtfully select those that offer a reduced environmental footprint without compromising performance. In this context, environmental sustainability should also be an important consideration in the selection of computational infrastructure.

In conclusion, the use of AI, particularly large language models, can significantly improve productivity and decision-making processes in the pharmaceutical industry. However, it is essential to use these tools responsibly to maintain ethical standards, ensure data privacy, and avoid other risks associated with generative AI. The tool for eligibility criteria parsing is no exception and should adhere to these principles.

9.6 Maintenance and support

Regular maintenance of IT products and user support are fundamental for ensuring the continued availability of tools for users. Companies must ensure that responsible personnel promptly address any issues to minimize tool downtime. It is also important to continuously monitor and optimize the solutions even after deployment to maintain the highest possible efficiency.

The field of generative models is highly dynamic, with new phenomena and challenges being identified regularly. Consequently, regulations must adapt to these observations. To keep the system compliant with all regulatory laws and industry standards, adjustments may be required. Rapid and effective responses to these updates are essential to prevent any serious consequences arising from inappropriate use of the tool.

In software development, ensuring user satisfaction and effectiveness in using the product is crucial. To fully leverage the capabilities of the developed tool, users should attend a comprehensive training session once the product is released. This session should present the tool's functionalities, demonstrate the most effective ways to use its features, share best practices, and openly communicate any limitations. This is especially important in the pharmaceutical industry, where the processing of sensitive data and making critical decisions require that users are well-informed about system deficiencies, so they can use the tool responsibly. This importance

is amplified when using AI tools that employ generative language models, which come with several risk considerations, as outlined in Section 9.5. Additionally, users should have access to complete documentation of the tool, providing detailed usage instructions for ongoing reference.

Beyond initial training, continuous support must be available to users, with a dedicated team ready to resolve any issues that arise during system use. Users should have clear contact points for assistance, and they should be promptly supported by a team capable of efficiently addressing and resolving their problems.

Another vital component of user support is the collection of feedback after the tool has been in use for some time. Users are the primary stakeholders of the tool, and their satisfaction is fundamental. To enhance the user experience, feedback should be systematically collected, analysed, and addressed, leading to system improvements tailored to their needs.

To conclude, ensuring regular maintenance and robust user support significantly impacts user satisfaction and tool reliability in the long term. This can be achieved by promptly addressing issues, providing comprehensive training sessions, and consistently gathering and acting on user feedback.

9.7 Efficiency and validation

An essential requirement for implementing an AI-based tool is increased efficiency in performing target task compared to current methods. The criteria parsing tool is no exception. As there are no existing solutions to compare with, the baseline for this tool is the manual effort to transform the criteria into the required structure. Since the tool is designed to work in the healthcare sector, it will not operate autonomously but will be used as a support tool to avoid adverse consequences on peoples' health and lives. This important consideration determines that the tool does not need to have extremely high accuracy, as the user will have a chance to correct its output. However, an algorithm that provides highly incorrect extractions would not be beneficial either, because domain experts would spend significant time correcting the parsed conditions. Thus, while high accuracy is not the most critical factor, it is important to provide a sufficient level of accuracy to add value to the overall workflow. If the tool significantly reduces the time spent on criteria parsing, including correction time, compared to the non-AI-assisted manual effort, it can be considered for production.

This requirement introduces another: a correction interface. Subject matter experts need access to an easy-to-use application that allows them to browse the parser's outputs and correct them if needed. Before using the parsing results in other systems, they should be reviewed and corrected to ensure the high accuracy of the source data for those systems, such as tools for patient-trial matching.

Given the very limited annotated dataset and the lack of reliable methods to evaluate the output of such a parser (and the output from generative models), it is infeasible to precisely measure the accuracy of the tool in an automated manner.

Therefore, the correctness of the structured output should be evaluated manually. Due to the complex structure of the parsed conditions described in Section 9.2, the final solution must be equipped with a component that enables manual validation of the extractions. The user should be presented with the inclusion and exclusion criteria alongside the parser output and be able to interact with the interface to assess the extractions.

From the perspective of a patient-trial matching application, a valuable feature of the tool is to avoid overly limiting the trial population. It is better to overlook some mandatory conditions than to extract conditions incorrectly, which may disqualify eligible patients from a trial. Too stringent a formulation of a condition in the output is a more serious error than too wide a criterion because trials already struggle to find enough participants. A tool that incorrectly excludes eligible patients would exacerbate this issue, thereby hindering trial optimization.

Ultimately, the tool must enhance efficiency by reducing the time and effort required for criteria parsing while maintaining satisfactory accuracy to be valuable. By providing a robust validation interface and prioritizing precision over recall, the tool can effectively support patient-trial matching and optimize trial operations without compromising the inclusion of eligible participants.

9.8 Diverse expertise in tool development

In developing advanced technological tools, particularly those used in specialized fields like the pharmaceutical industry, integrating diverse expertise is not only highly beneficial but also a critical requirement for creating reliable solutions. This is especially important for AI-driven systems that learn from domain-specific datasets and must provide meaningful insights based on them. Including various stakeholders such as project managers, validation leads, technical developers, domain experts, UX designers, and end users is essential to ensure the development of an optimal tool.

In the initial phase of each project, the role of project manager is crucial. Setting clear project goals and defining key performance indicators (KPIs) at the beginning ensures that developers understand the requirements and end users are aware of the expected outcomes. Project managers often lead the coordination of different stakeholders, collecting their requirements and translating them into specific, measurable goals. Setting numerical thresholds that need to be met to add value to business processes and release the tool is a good practice. This allows for continuous assessment of whether the tool has the potential to achieve these thresholds or if the project should be discontinued early to avoid wasting time and resources on solutions that will not be deployed in production.

When creating an AI-based system, involving AI experts in an obvious step. They are responsible for the technical development of the tool. However, AI experts, with their deep knowledge of algorithm functionality, machine learning techniques, and data processing methods, often lack insights into the complex biomedical aspects

necessary to design the system accurately and assess its functionality. They rely on support from domain experts and their guidance throughout the development process to shape as effective a tool as possible. Domain experts should be involved from the start, helping data scientists understand the business processes and guiding them in selecting the most appropriate datasets for training and testing the solution, as well as supervising data preprocessing.

However, it is also important to emphasize that while AI experts are involved in the technical development, they also need to acquire some domain knowledge to understand business challenges and user perspectives. Not all responsibilities can be delegated to domain experts. It is crucial that technical team members do not focus solely on technologies without contact with the business world, which may result in a lack of practical usability or relevance to users' needs. Employing advanced AI technologies, which are very interesting for AI experts, without any practical impact on end users is not advisable. There should be a balance between applying cutting-edge technologies and ensuring they provide valuable impacts on business processes.

In the development of the eligibility criteria parser, domain experts should participate in defining the scope of the tool, for instance, selecting relevant therapeutic areas or types of clinical trials. They are essential in choosing the types of conditions and attributes that need to be extracted and creating the dictionary of terms used in the tool's final output. Furthermore, they should assist in preparing the guidelines and instructions for the tool, as well as selecting the most informative and representative examples for in-context learning. They can also review the entire prompt for the LLM.

As the project progresses, both domain experts and end users (if they are not the same) should engage in the evaluation of the tool. In scenarios where generative language models are used and no benchmarks are available, the manual review of the tool's output is crucial to asses its effectiveness and verify if it meets business needs. During the validation stage, feedback from domain experts is invaluable and can be incorporated to adjust the tool - for example, modifying the prompt. Domain experts could use the validation interface described in Section 9.7 to review and correct structured conditions, and the time spent on this task can be measured to verify the tool's value.

Regular presentations of progress to stakeholders are vital to maintain transparency and ensure the product aligns with user expectations. AI experts should provide updates on the development process, intermediate results, and observations, so that the end users are aware of any obstacles or changes in the design. It is important to openly communicate any limitations of the tool as early as possible, allowing domain experts to react and possibly bring valuable suggestions for technical adjustments.

UX Designers also play a critical role in the development of a decision-support tool. They recognize user needs and engage in defining the system functionalities required. This task is challenging - UX Designers strive to bridge the gap between technical and business worlds, effectively ensuring that the tool is user-friendly and

meets the practical demands of end users. Engaging users in the design process through mockups, asking them to select the most suitable version, and requesting their feedback and suggestions for changes makes them feel involved in the process and significantly influences the final design of the product. This approach ensures that the product is effectively modified according to their needs, saving time and resources, and avoiding the risk of rebuilding the solution from scratch at a later stage. Feedback sessions and prototype testing are excellent examples of user-centric development techniques that can be employed to improve the tool's usability. An iterative design process results in effective tool refinement and adaptation to changing user requirements and technical advancements.

In conclusion, developing an AI-driven tool that supports human work requires much more than just AI experts. It demands a multidisciplinary team and an iterative, user-centric approach to enhance the quality and applicability of the tool. Additionally, this results in a more effective implementation within the pharmaceutical company.

Chapter 10

Implementation of the Eligibility Criteria Parsing Tool: A Proof of Concept

In previous chapters, we discussed the challenges the pharmaceutical industry encounters in structuring and analysing clinical trial eligibility criteria, which impact trial recruitment and overall trial success. We reviewed the limitations of existing solutions, explored the potential of large language models for information extraction from eligibility criteria, and outlined the requirements for a parsing tool from Roche's perspective.

This chapter focuses on the technical implementation of the proof of concept (PoC) for the eligibility criteria parsing tool. The development of this tool employed a multidisciplinary approach, combining domain expert support, advanced AI technologies, robust software engineering practices, and user-centric design.

Section 10.1 defines the goal of the PoC and its scope. Section 10.2 describes the data collection and preparation processes, along with the model selection and prompt engineering. Section 10.3 details the PoC architecture, including data handling, inference pipeline, model monitoring, user interface, and MLOps practices. Section 10.4 outlines the evaluation experiment design. Section 10.5 discusses the findings and their implications. Section 10.6 identifies limitations of the solution and areas for future work. Section 10.7 provides a summary of the proof of concept findings.

10.1 Objective and scope

The main objective of this tool is to transform complex, free-text eligibility criteria into a structured, machine-readable format. This structured format includes a set of conditions along with any information that is relevant for a patient-trial matching algorithm. It should clearly indicate which conditions are required or excluded to qualify a patient for a trial. The tool aims to provide standardized data points, in line with the requirements detailed in Section 9.2, while balancing the inclusion of

detailed condition specification without over-complicating the structure.

The primary function of the tool is to parse eligibility criteria and create a structured output that can be utilized by other applications, such as a patient-trial matching algorithm - a use case that we focus on in this PoC. The designated workflow for matching patients to trials involves comparing survey responses, which assess patients' demographics and health status, with a database of structured trial eligibility criteria. The process filters out trials that do not match the specified characteristics and provides a list of relevant trials to the user. This reduces the number of trials to review and significantly improves the current method, which involves browsing a long list of trials in specific disease areas and manually interpreting complex criteria. This method is highly ineffective and often unfeasible without medical knowledge. Moreover, it is impractical to expect doctors to extensively focus on a single patient and browse numerous trials to find a suitable match. One major benefit of such a patient-trial matching algorithm is reducing the number of trials to review—either by a patient or a medical representative—thereby increasing the chance of finding a relevant trial faster.

When developing the parsing tool for eligibility criteria, it is important to recognize that the matching algorithm does not need to achieve perfect precision. A partial narrowing of the trial set is beneficial, as the goal is to reduce the pool of trials, without being overly restrictive. It is preferable to have broader criteria that might include some irrelevant trials than to risk excluding relevant ones, which could prevent a patient from accessing a potentially suitable treatment. However, given the medical context, full automation is not advisable. The parsing tool should act as an assistant rather than an autonomous solution, with human supervision remaining essential to ensure that the output is accurate and does not inadvertently exclude life-saving treatment options.

As a result, the objective of this PoC is to create a parsing algorithm alongside a validation interface, allowing humans to review and edit the structured output before it is used in other applications. To be accepted and integrated into the workflow, the tool must add significant value by reducing manual effort and time spent on parsing eligibility criteria. This improvement in efficiency is critical for the tool's adoption within the company.

To quickly verify the tool's capabilities and build a prototype, it is important to focus on a manageable subset of data. Given the vast number of existing trials published on ClinicalTrials.gov and the numerous condition types in eligibility criteria, narrowing the scope for the proof of concept and testing the approach on a smaller specified sample is essential before extending it to other areas. For this PoC, the focus is on breast cancer trials, as they represent the largest disease area for trials conducted by Roche. However the trials used in the experiments include both Roche and non-Roche trials to enable some generalization from the beginning. With the help of domain experts, we selected three types of conditions that are most important for a patient-trial matching algorithm: biomarkers, breast cancer stages, and subtypes. These conditions significantly narrow the pool of relevant trials and are well-known to patients after diagnosis, enabling them to provide relevant responses

in the survey. It is also important to note that the survey and the patient-trial matching algorithm are out of scope for this PoC. The focus is strictly on the development of the parsing tool and the validation interface.

10.2 ML methodology

The successful development of the eligibility criteria parsing tool requires a comprehensive and structured approach. The machine learning methodology applied in this proof of concept includes several key components: input/output definition, data preparation, model selection and prompt engineering.

10.2.1 Input and output definition

The development of the tool began with clearly defining its expected input and output. As outlined in Chapter 9, the input consists of the full text of eligibility criteria section along with the trial ID for reference. The tool is designed to operate in batch inference mode, processing multiple trials at once. Consequently, the input is structured as a JSON array containing JSON objects for each trial, which include the criteria and trial ID pairs. The input template is presented in Figure 10.1. The expected output is also a JSON array, where each element is a JSON object representing the parsing result for a corresponding input trial. These objects include five fields:

- ***nct_id:*** trial ID,
- ***inclusion_criteria:*** text of inclusion criteria,
- ***exclusion_criteria:*** text of exclusion criteria,
- ***output:*** full response generated by the model,
- ***condition_list:*** a JSON array containing conditions required by the trial, extracted from *"output"* field. This inner array contains JSON objects with two fields: *"condition"*, which is a boolean expression defining a mandatory condition, and *"phrases"*, which is a list of phrases that were used to create the expression.

The output template is illustrated in Figure 10.2. The *"condition"* field is a boolean expression indicating a mandatory criterion that must be met by a patient to qualify for the trial. If there are alternative criteria where only a part is needed for eligibility, they are combined in a single object using boolean *"OR"* statements. We recognize that trial requirements are often complex, and the boolean expression may include many *"AND"* and *"OR"* statements within a single element. However, structuring the conditions this way helps clearly define separate requirements and facilitates quick assessment of patient eligibility. The *"phrases"* field contains a list of phrases that the model used to derive the boolean expression in the *"condition"* field. This enhances the explainability of the model's decisions and assists domain experts in

```
[
   {
      "trial_id": "NCT01234",
      "eligibility_criteria": "Inclusion Criteria: ... Exclusion Criteria: ..."
   },
   {
      "trial_id": "NCT56789",
      "eligibility_criteria": "Inclusion Criteria: ... Exclusion Criteria: ..."
   }
]
```

Figure 10.1: Input JSON format.

quickly assessing the correctness of the defined conditions. Discrepancies can be immediately identified based on the provided phrases. For instance, if the model extracts *"NOT metastatic"* as a breast cancer subtype but the provided phrases state *"Patients with any other metastatic tumours are excluded"*, the domain expert can swiftly recognize that the extraction is incorrect because the subtype relates to other tumors, not breast cancer. Moreover, these phrases can justify experts' decisions and be useful in additional validations. Since the tool is a decision support system, its final output includes user adjustments.

```
[
   {
      "nct_id": "NCT01234",
      "inclusion_criteria": "...",
      "exclusion_criteria": "...",
      "output": "...",
      "condition_list": [
         {
            "condition": "Condition1",
            "phrases": ["Phrase1"]
         },
         {
            "condition": "Condition2 OR (Condition3 AND Condition4)",
            "phrases": ["Phrase2", "Phrase3, Phrase4"]
         }
      ]
   },
   {
      "nct_id": "NCT56789",
      "inclusion_criteria": "...",
      "exclusion_criteria": "...",
      "output": "...",
      "condition_list": [
         {
            "condition": "Condition5 OR Condition6 OR Condition7 OR Condition8" ,
            "phrases": ["Phrase5", "Phrase6"]
         }
      ]
   }
]
```

Figure 10.2: Output JSON format.

10.2.2 Data preparation

The data preparation process started with defining a list of standardized condition names that are allowed to be used in the output. For each selected condition type (biomarkers, subtypes, stages), we created a separate list of terms based on various breast cancer resources and observations from eligibility criteria content. These lists were then verified and adjusted by experts, resulting in the final dictionary presented in Figure 10.3.

```
Subtypes:
IDC, ILC, LCIS, IBC, Metaplastic, Paget's Disease of the Breast, Recurrent, Inoperable,
Infiltrating, Multifocal, Multicentric, Unilateral,Bilateral, High grade, Low grade, Locally
recurrent, Microcalcifications, Dense breast tissue, DCIS, Invasive, Early stage, Locally
advanced, Advanced, Metastatic, Progressive, Adenocarcinoma

Stages:
Stage 0, Stage I, Stage Ia, Stage Ib, Stage II, Stage IIa, Stage IIb, Stage III, Stage IIIa,
Stage IIIb, Stage IIIc, Stage IV

Biomarkers:
HER2+, HER2-, ER+, ER-, PR+, PR-, AR+, AR-, BCL-2+, BCL-2-, PD-L1+, PD-L1-, PIK3CA+, PIK3CA-,
BRCA1-, BRCA1+, BRCA2+, BRCA2-, ESR1+, ESR1-, Ki-67+, Ki-67-
```

Figure 10.3: A dictionary of terms allowed to be used in boolean expressions.

Furthermore, domain experts helped develop a set of instructions to be included when prompting the large language model. These guidelines, written in concise natural language, reflect the rules that experts use when interpreting eligibility criteria. For instance, they use a dependency that *"locally advanced breast cancer"* translates to *"stage IIIb or stage IIIc"*. We believe such knowledge can significantly improve the model's reasoning. The complete set of guidelines is presented in Figure 10.4.

```
Instructions for Grouped Biomarker Terms:
"Triple-negative": Should be explicitly shown as ER- AND PR- AND HER2-
"HR+": Should be explicitly shown as ER+ OR PR+
"Triple-Positive": Should be translated to ER+ AND PR+ AND HER2+
"Luminal A": Should be translated to ER+ AND HER2- AND Ki-67- AND PR+
"Luminal B": Should be translated to ER+ AND HER2- AND (Ki-67+ OR PR-)

Instructions for Inferring Stages and Subtypes:
Locally advanced translates to Stage IIIb OR Stage IIIc.
Advanced translates to Stage IIIb OR Stage IIIc OR Stage IV.
Metastatic translates to Stage IV.
Inflammatory translates to IBC.
Non-IBC translates to NOT IBC.
```

Figure 10.4: Guidelines for inferring conditions from eligibility criteria.

The final step in data preparation involved creating examples for in-context learning of the LLM. These examples include the input criteria and the desired output to demonstrate what is expected from the model. We selected four representative studies from ClinicalTrials.gov, which encompass various dependencies and relationships between criteria. Some criteria were modified, removed, or inserted artificially based

on experimentation results. To minimize the manual effort of domain experts, we annotated the criteria, and these annotations were subsequently validated by the experts. For the purpose of model prompting, the examples are written in natural language without maintaining a JSON structure. The input and output transformations to the desired format are applied using pre- and post-processing pipelines.

The first trial contains use cases for applying the defined guidelines (Figure 10.4), such as translating grouped biomarker terms (*"HR+"*), subtype-stage mapping (*"metastatic"*, *"locally advanced"*), and inferring biomarker polarity (*"mutation(s) of PIK3CA"*). It also contains an excluded subtype that should be negated in the output (*"Metaplastic breast cancer"* in Exclusion Criteria). This trial also includes conditions with allowed terms that should not be extracted by the model:

- condition related to past medical history (*History of stage I-IIIb cancer*),
- condition specifying a metastasis site, especially in exclusion criteria, which does not mean the cancer should not be metastatic (*Known and untreated, or active CNS metastases*),
- condition related to other diseases, not breast cancer (*inflammatory bowel disease*).

The input and output for this trial are shown in Figure 10.5.

The second trial also covers examples for using the instructions for grouped biomarker terms (*"triple-negative"*, *"luminal B"*) and inferring stages (*"metastatic"*) and subtypes (*inflammatory*). Importantly, it contains a complex mandatory condition with four alternatives, which is not even explicitly stated in the text but inferred from the knowledge that the cancer cannot be both triple-negative and luminal B. It also includes two conditions that are allowed but not required (*Subjects with progressive adenocarcinoma are eligible*, and *Patients in a metastatic situation can be included regardless of the therapeutic line.*), so they should be skipped by the model. The output for this trial includes only one condition, represented by a highly complex boolean expression, as visible in Figure 10.6.

The third trial presents an example of multiple subtypes in one phrase (*advanced metastatic adenocarcinoma*), which should be transformed into separate conditions or conditions connected by *AND*. It also contains criteria types that should not be parsed by the model due to a different meaning, irrelevant to the task:

- conditions present in medical history (*History of DCIS*),
- treatments and other conditions that contain allowed terms but do not specify required or disallowed subtypes, stages, or biomarkers (*PIK3CA inhibitors*),
- conditions with unspecified biomarker status (*Known hormon receptor status*),
- conditions specifying metastasis site (*Known brain metastases unless treated and stable*).

Figure 10.7 presents the input and expected output for this trial.

```
Inclusion criteria:
Confirmed diagnosis of HR+/HER2- breast cancer; Metastatic or locally advanced disease not
amenable to curative therapy; Progression of disease during adjuvant endocrine treatment or
within 12 months of completing adjuvant endocrine therapy with an aromatase inhibitor or
tamoxifen; Receiving LHRH agonist therapy for at least 2 weeks prior to Day 1 of Cycle 1 if
pre/peri-menopausal; Confirmation of biomarker eligibility (detection of specified mutation(s)
 of PIK3CA via specified test); Consent to provide fresh or archival tumor tissue specimen;
Measurable disease per Response Evaluation Criteria in Solid Tumors, Version 1.1 (RECIST v1.1)
; evaluable bone-only disease is not eligible; bone-only disease with at least one measurable,
 soft-tissue component, even if considered disease that is limited to bone but has lytic or
mixed lytic/blastic lesions and at least one measurable soft-tissue component per RECIST v1.1
may be eligible; Eastern Cooperative Oncology Group Performance Status of 0 or 1; Life
expectancy of > 6 months; Adequate hematologic and organ function within 14 days prior to
initiation of study treatment;

Exclusion criteria:
Metaplastic breast cancer; History of stage I-IIIb cancer; Any history of leptomeningeal
disease or carcinomatous meningitis; Any prior systemic therapy for metastatic breast cancer;
Prior treatment with fulvestrant or any selective estrogen-receptor degrader, with the
exception of participants that have received fulvestrant or any selective estrogen-receptor
degrader as part of neoadjuvant therapy only and with treatment duration of no longer than 6
months; Prior treatment with any PI3K, AKT, or mTOR inhibitor, or any agent whose mechanism of
 action is to inhibit the PI3K-AKT-mTOR pathway; Type 2 diabetes requiring ongoing systemic
treatment at the time of study entry; ; Known and untreated, or active CNS metastases.
Patients with a history of treated CNS metastases may be eligible; Active inflammatory or
infectious conditions in either eye, or any eye conditions expected to require surgery during
the study treatment period; Symptomatic active lung disease, or requiring daily supplemental
oxygen; History of inflammatory bowel disease or active bowel inflammation; Anti-cancer
therapy within 2 weeks before study entry; Investigational drug(s) within 4 weeks before
randomization; Prior radiotherapy to >= 25% of bone marrow, or hematopoietic stem cell or bone
 marrow transplantation; Chronic corticosteroid therapy or immunosuppressants; Pregnant,
lactating, or breastfeeding, or intending to become pregnant during the study or within 60
days after the final dose of study treatment; Major surgical procedure, or significant
traumatic injury, within 28 days prior to Day 1 of Cycle 1

Output:
- Condition: "ER+ OR PR+" , Phrases: "Confirmed diagnosis of HR+/HER2- breast cancer"
- Condition: "HER2-" , Phrases: "Confirmed diagnosis of HR+/HER2- breast cancer"
- Condition: "Stage IV OR Stage IIIb OR Stage IIIc" , Phrases: "Metastatic or locally advanced
 disease not amenable to curative therapy"
- Condition: "PIK3CA+" , Phrases: "Confirmation of biomarker eligibility (detection of
specified mutation(s) of PIK3CA)"
- Condition: "NOT Metaplastic", Phrases: "Metaplastic breast cancer"
```

Figure 10.5: Input criteria and target output for the first trial example.

The fourth trial contains an example of TNM notation (*T2, T3, T4a, T4b, T4c, T4d which is node negative or node positive (histologically or cytologically confirmed) or; Any T with lymph node positive disease (histologically or cytologically confirmed)*) to hint to the model that it should resolve and translate these to specific stages. It also includes a condition with unspecified biomarker status (*Oestrogen and progesterone receptor status known prior to study entry*) and a condition related to past medical history (*History of any other malignancy within the past 5 years, with the exception of non-melanoma skin cancer, in situ carcinoma of the breast (ductal or lobular) or carcinoma-in-situ of the cervix*). Additionally, it presents an interesting example of a not-required but allowed condition, impacting other requirements: *Patients with triple-negative breast cancer (TNBC) are allowed if they meet all other mandatory criteria*. In this case, we

```
Inclusion criteria:
ECOG (Eastern Cooperative Oncology Group) performance status <= 2; Signature of the
participation consent to the study,; Affiliation to a social security scheme; Subjects with
progressive adenocarcinoma are eligible; Major woman with:; inflammatory metastatic triple-
negative (TN) breast cancer, histologically proven before treatment and high grade, receiving
neoadjuvant chemotherapy and having, after treatment, a breast residue of at least 15 mm on
the specimen. The mammary residue will measure at least 15 mm on the mammography performed at
the end of neoadjuvant treatment; inflammatory metaplastic triple-negative (TN) breast cancer,
 histologically proven before treatment and high grade, treated by primary surgery with a
tumor size of at least 15 mm on the specimen.; IBC TN breast cancer (T4d), histologically
proven prior to treatment, receiving neoadjuvant chemotherapy and having, after treatment, a
breast residue of at least 15 mm on the specimen. The mammary residue will measure at least 15
 mm on the mammography performed at the end of the neoadjuvant treatment.; inflammatory
Luminal B breast cancer (LB), histologically proven prior to treatment, receiving neoadjuvant
chemotherapy and having, after treatment, a mammary residue of at least 30 mm on the specimen.
 The mammary residue will measure at least 15 mm on the mammography performed at the end of
the neoadjuvant treatment.; Patients in a metastatic situation can be included regardless of
the therapeutic line.;

Exclusion criteria:
Pregnant woman; Patient deprived of liberty by court or administrative decision; In
neoadjuvant situation: no neoadjuvant treatment by radiotherapy or hormone therapy; Refusal to
 participate in the study

Output:
- Condition: "((IBC AND Stage IV AND (ER- AND PR- AND HER2-) AND High grade) OR (IBC AND
Metaplastic AND (ER- AND PR- AND HER2-) AND High grade) OR (IBC AND (ER- AND PR- AND HER2-)
AND (Stage IIIb OR Stage IIIc OR Stage IV)) OR (IBC AND ER+ AND HER2- AND (Ki-67+ OR PR-)))" ,
 Phrases: "inflammatory metastatic triple-negative (TN) breast cancer, histologically proven
before treatment and high grade", "inflammatory metaplastic triple-negative (TN) breast cancer
, histologically proven before treatment and high grade", "IBC TN breast cancer (T4d),
histologically proven prior to treatment", "inflammatory Luminal B breast cancer (LB),
histologically proven prior to treatment"
```

Figure 10.6: Input criteria and target output for the second trial example.

cannot state that TNBC is required, but given the study requires HER2+, the final mandatory condition should be defined as *HER2+ OR (HER2- AND ER- AND PR-)*. We show this example to the model in order to make it aware of such situations and handle them correctly. The fourth trial example is included in Figure 10.8.

The presented examples were carefully selected and adjusted to cover a broad scope of different criteria aspects. They were inspired by the set of criteria of real trials but were modified explicitly for the purpose of eligibility criteria parsing. It was a conscious decision to artificially add some more complex patterns and remove some redundant sentences to make the prompt full of important information only. It was observed in the experimentation phase that including more examples results in more accurate model outputs, which is not the case with expanding the task description or parsing instructions. The model learns more for few-shot examples than from guidelines. Therefore, it is very important to craft the examples consciously to make them maximally representative and meaningful.

```
Inclusion criteria:
Histologically confirmed advanced metastatic adenocarcinoma with measurable or evaluable
disease: Patients who have progressed on distant metastatic sites after curative surgery or
have stage IV breast cancer at diagnosis; Patients with inflammatory disease are eligible; Age
 > 19 years; ECOG performance status 0 - 2; Patient has HER2-negative breast cancer with IHC
and/or FISH (or SISH, CISH); Known hormon receptor status; Patient is premenopausal.
Premenopausal status is defined as either:; A. Patient had last menstrual period within the
last 12 months B. If within three months of tamoxifen (tamoxifen) taking, C. In case of
chemotherapy induced amenorrhea, the serum FSH <=40IU/l; A. Patient who have stage IV breast
cancer at diagnosis, allow disease that progressed after 1st line chemotherapy. B. Patient who
 have stage IV breast cancer at diagnosis, allow disease that progressed after tamoxifen or
goserelin. C. In case of recur/metastatic breast cancer, allow disease that progressed after
12 month of completion of neo/adjuvant chemotherapy .; Urine or serum HCG test must be
negative.; Adequate bone marrow function (>= ANC 1,500/ul, >= platelet 100,000/ul, >=
Hemoglobin 9.0 g/dl); Adequate renal function (<= serum creatinine 1.5 mg/dl or CCr >= 50 ml/
min); Adequate liver function (<= serum bilirubin 1.5 mg/dl, <= AST/ALT x 3 upper normal limit
); Patients who were already established on bisphosphonate therapy may continue on
bisphosphonates.; Patients agreed to use effective contraception or not of childbearing
potential; Written informed consent; Consent to biomarker analysis.;

Exclusion criteria:
Postmenopausal women; Serious uncontrolled intercurrent infections; Serious intercurrent
medical or psychiatric illness, including active cardiac disease; Pregnancy or breast feeding;
 Second primary malignancy(except in situ carcinoma of the cervix or adequately treated
nonmelanomatous carcinoma of the skin or resected thyroid papillary carcinoma or other
malignancy treated at least 5 years previously with no evidence of recurrence); History of
DCIS; Bilateral mastectomy; Patients has received previous endocrine treatments such as,
aromatase inhibitor, exemestane in the metastatic setting; Patients has received previous
treatment with CDK 4/6 inhibitors, mTOR inhibitors, PIK3CA inhibitors or capecitabine; No
symptomatic visceral metastasis; Known brain metastases unless treated and stable; Clinically
significant uncontrolled conditions including, known active hepatitis B or hepatitis C.; QTc
interval > 480 msec, family or personal history of long or short QT syndrome, or known history
 of QTc prolongation or Torsade de Pointes.; Known positive testing for human immunodeficiency
 virus or acquired immune deficiency syndrome.; Unable to swallow and retain oral medication.;
 Treatment radiotherapy within 4 weeks of the study; Use of any investigational drug within 4
weeks of the study; Treatment with chemotherapy within 3 weeks or hormone therapy within 2
weeks of the study.

Output:
- Condition: "Stage IV", Phrases: "Histologically confirmed advanced metastatic adenocarcinoma
 with measurable or evaluable disease", "Patients who have progressed on distant metastatic
sites after curative surgery or have stage IV breast cancer at diagnosis"
- Condition: "Adenocarcinoma", Phrases: "Histologically confirmed advanced metastatic
adenocarcinoma with measurable or evaluable disease"
- Condition: "HER2-", Phrases: "Patient has HER2-negative breast cancer with IHC and/or FISH (
or SISH, CISH)"
```

Figure 10.7: Input criteria and target output for third trial example.

10.2.3 Model selection

Structuring eligibility criteria is a challenging task that requires a model which can address this complexity. Recent advancements in generative AI and our successful experiments with using large language models for named entity recognition have led us to choose an LLM for this task. This decision aligns with the requirements for the parsing tool, which include functioning effectively with limited annotated data, accepting full text of criteria as input, following parsing guidelines, and extracting meaningful information from complex domain-specific text. The task demands ad-

```
Inclusion criteria:
Written informed consent obtained prior to any study-related procedures; Age > 18 years;
Histologically proven breast cancer, for which neo-adjuvant chemotherapy and trastuzumab is
considered a valid therapeutic strategy.; Patients with the following TNM stages (refer to
AJCC 7th Edition - Appendix M) of breast cancer are eligible:; T2, T3, T4a, T4b, T4c, T4d
which is node negative or node positive (histologically or cytologically confirmed) or; Any T
with lymph node positive disease (histologically or cytologically confirmed); Patients with
multifocal tumours are not excluded; T stage assignment must be based on the largest tumour.;
Patients with bilateral breast cancer are not eligible; Tumour HER2/neu positive (3+ by IHC or
 fluorescence in situ hybridization (FISH) positive); Oestrogen and progesterone receptor
status known prior to study entry; Patients with triple-negative breast cancer (TNBC) are
allowed if they meet all other mandatory criteria; ECOG performance status score < or equal to
 1; Cardiac ejection fraction >= 50% as measured by echocardiogram or MUGA scan within 3
months prior to randomisation. Women of child-bearing potential and men must agree to use
adequate contraception (non-hormonal or barrier method of birth control, abstinence or a
vasectomy partner) prior to study entry and for the duration of study participation.

Exclusion criteria:
Prior therapy with systemic cytotoxic chemotherapy Lapatinib or Trastuzumab.; Prior taxanes;
Radiotherapy (Except for radiotherapy localised to radiotherapy to a primary squamous or basal
 cell skin cancer).; Patients with metastatic disease (M1).; Concurrent therapy with any other
 non-protocol anti-cancer therapy; History of any other malignancy within the past 5 years,
with the exception of non-melanoma skin cancer, in situ carcinoma of the breast (ductal or
lobular) or carcinoma-in-situ of the cervix.; Current therapy with any hormonal agent such as
raloxifene, tamoxifen, or other selective oestrogen receptor modulators (SERMs), either for
osteoporosis or prevention of breast cancer. Patients must have discontinued these agents 14
days prior to enrolment.; Concurrent treatment with ovarian hormonal replacement therapy.
Prior treatment must be stopped prior to enrolment.; Pre-existing motor or sensory
neurotoxicity of a severity >= Grade 2 by NCI-CTCAE version 4.0.; Poorly controlled
hypertension (e.g. systolic >180mm Hg or diastolic >100mm Hg.); Any history of myocardial
infarction, angina pectoris or congestive heart failure. Patients on current therapy for
arrythmias are excluded.

Output:
- Condition: "Stage II OR Stage IIIa OR Stage IIIb OR Stage IIIc", Phrases: "T2, T3, T4a, T4b,
 T4c, T4d which is node negative or node positive (histologically or cytologically confirmed)
or; Any T with lymph node positive disease (histologically or cytologically confirmed)"
- Condition: "NOT Bilateral", Phrases: "Patients with bilateral breast cancer are not eligible
"
- Condition: "(HER2+ OR (HER2- AND ER- AND PR-)", Phrases: "Tumour HER2/neu positive (3+ by
IHC or fluorescence in situ hybridization (FISH) positive)", "Patients with triple-negative
breast cancer (TNBC) are allowed if they meet all other mandatory criteria"
- Condition: "NOT Stage IV", Phrases: "Patients with metastatic disease (M1)"
```

Figure 10.8: Input criteria and target output for fourth trial example.

vanced reasoning, a large context window, broad knowledge, and the capability to generate text not present in the input, making traditional NER approaches unsuitable.

Given the rapidly evolving area of generative AI, many new models are frequently released. However, selecting the right LLM involves considering factors such as model size, context window length, maximal output length, and performance on benchmarks. For Roche, security is also a critical consideration, requiring the use of either a self-hosted model or a model hosted in the Roche cloud that is allowed to handle confidential data. Currently, only OpenAI models meet these criteria among Roche cloud-hosted options. According to the Stanford's leaderboard on the accuracy on the MMLU (Massive Multitask Language Understanding) benchmark [158],

the GPT-4o model from OpenAI ranks second with other OpenAI models and self-hosted models ranked lower. As stated in the documentation [85], the GPT-4o model has a context window of 128,000 tokens, is two times faster and 50% cheaper than its predecessor, GPT-4-turbo. The maximum token length of response is not revealed. When using an OpenAI model, there is no need to focus on model size, because it is hosted by third parties, not on the project side. The important aspects are the low cost of usage and a balance between model accuracy and processing speed. This balance is secured by GPT-4o, which is also ranked eighth in the MMLU inference runtime leaderboard, where the models listed above it are not even in top 10 for accuracy in the task. Additionally, the cost is reasonable given the model's performance, making it acceptable for Roche. Therefore, the GPT-4o model is selected for the PoC.

We use a pretrained model without any additional fine-tuning. To minimize creativity and ensure accuracy in information retrieval, the model temperature parameter is set to 0.0 for inference, which forces the model to generate the most likely next token.

10.2.4 Prompt engineering

When using pretrained large language models, prompt engineering is even more important than the model selection and setup. The model's output can vary significantly depending on the prompt's precision. Less precise prompts lead to less accurate outputs, and specific words and phrases can greatly impact the model's performance, as observed in our experiments.

Since the model uses an in-context learning approach, the prompt must include not only the parsing guidelines but also the four selected examples (detailed in Section 10.2.2). The final prompt version resulted from an iterative experimentation process and prompt tuning using five trial examples specifically chosen for this purpose. This process was supported by a Streamlit application we developed to facilitate prompt engineering. The application connects with GPT-4o via the OpenAI API and allows users to test different prompts and see immediate outputs for a selected trial. It provides a user-friendly interface and effective visualisation of the input and output data, enabling quick execution of tuned prompts on specific eligibility criteria. The interface of this application is presented in Figure 10.9.

The final prompt for the eligibility criteria parser includes the following components:

- **Task description:** Defines the model's role (using the Persona Pattern), outlines the task's goal, and highlights crucial aspects identified through experimental results. For instance, discrepancies were observed in the model's output when handling optional criteria, that are not required for participation. To address this, we added a sentence to the task description to emphasize the rule for extracting only the required or disallowed conditions. The full task description is presented in Figure 10.10.

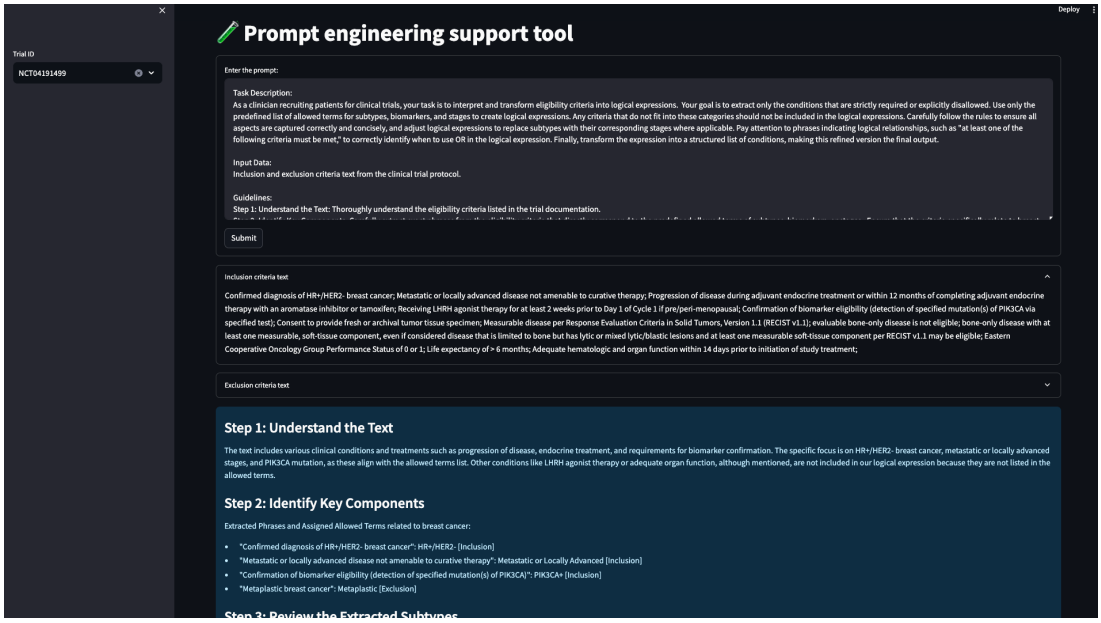


Figure 10.9: The application supporting prompt engineering

```
Task Description:
You are a clinician recruiting patients for clinical trials. Your job is to interpret
eligibility criteria and transform them into boolean expressions, so that they can be further
used in matching patient records to trials. You should focus only on required or disallowed
criteria, ignore optional ones. Consider only the criteria describing breast cancer stages,
subtypes and biomarkers, and use the allowed terms listed below to construct the expressions.
Any other criteria should not be included in the expressions. In the last step, transform the
boolean expressions into a list of mandatory conditions which will work as a checklist in
patient-trial matching application. Follow the rules outlined below.
```

Figure 10.10: Task description included in the model prompt.

- **Input data description:** Explains what the model will receive as input - specifically, the inclusion and exclusion criteria from the clinical trial protocol.
- **Parsing guidelines:** Provides step-by-step instructions for transforming the input data into the desired output, which details the reasoning process the model should follow. These instructions evolved during experimentation, with some rules added or modified for clarity. The guidelines are shown in Figure 10.11.
- **Allowed terms:** Lists terms describing biomarkers, subtypes, and stages allowed for use in creating boolean expression (as defined in Figure 10.3).
- **Instructions for inferring conditions:** Includes guidelines created with domain experts to support the model's reasoning with medical knowledge for interpreting specific phrases (as defined in Figure 10.4).
- **Responsibility statement:** Stresses the importance of adhering to the guidelines due to their potential impact on patient lives. During the hackathon, we

```
Rules:
Step 1: Identify phrases and conditions: Extract exact phrases from the eligibility criteria
that directly correspond to the allowed terms of breast cancer subtypes, biomarkers, or stages
. List each extracted phrase and clearly identify which allowed term it corresponds to.
Step 2: Refine the list of phrases: Remove phrases and conditions which do not relate to
allowed terms or to current breast cancer. Remove conditions that are not mandatory.
Step 3: Apply additional guidelines: Where possible, apply the inferring guidelines and
replace the subtype terms with stages. Match the grouped terms with their relevant logic
statements, according to the guidelines. Use negations for excluded conditions. Check the
context of the condition - some exclusion criteria contain exceptions that should not be
negated.
Step 4: Translate into boolean expression: Transform list of conditions into boolean
expressions. Remember about using "OR" for alternative criteria, where only a part of
conditions is required for eligibility. Use parentheses for correct precedence and grouping.
Step 5: Transform into condition list: Convert the boolean expression into a structured list
of conditions. Each line should represent a mandatory condition that is required for
eligibility. Alternative conditions should be grouped together in one line, indicating that
fulfilling any one of these alternatives suffices. This means that OR statements cannot be
split into lines, but AND statements should be split where possible. Follow standard operator
precedence rules where AND is evaluated before OR, and use parentheses to enforce the correct
precedence and grouping. Remove redundant conditions.
Step 6: Create pairs: For each condition in the list, pair it with the corresponding phrases
from Step 1, separated by '||' where multiple phrases support a single condition. Use exact
phrases from the text.
```

Figure 10.11: Parsing guidelines included in the prompt.

noticed that emphasizing responsibility improved the model's accuracy and reliability. This statement is: *"Please stick to these rules as your response will be used in a patient trial matching system, so it can even save patients' lives."*

- **Examples for in-context learning:** Includes input and output text for four trials (presented in Figures 10.5- 10.8) to provide examples of expected outputs. During experimentation, we discovered that the model's output improves when examples include the full thought process, not just input-output pairs. Therefore, we use a chain-of-thought approach, detailing each example step-by-step according to the parsing guidelines (Figure 10.11), with additional comments to clarify specific decisions. It also contains negative examples - incorrect extractions and refining steps, to explain the model why some of the conditions should not be included in the final output (e.g., *"History of stage I-IIIb cancer: stage I OR stage II OR stage IIIa OR stage IIIb [Exclusion] - stage related to history, not current state, should be removed"*). We observed that this feature enhances model reasoning, in the inference model corrects some of the mistaken extraction in the refining step. Figure 10.12 and 10.13 show example chain of thoughts for two selected trials.

```
Step 1: Identify phrases and conditions.
"Confirmed diagnosis of HR+/HER2- breast cancer": HR+/HER2- [Inclusion]
"Metastatic or locally advanced disease not amenable to curative therapy": Metastatic OR
Locally Advanced [Inclusion]
"Confirmation of biomarker eligibility (detection of specified mutation(s) of PIK3CA via
specified test)": PIK3CA+ [Inclusion]
"Metaplastic breast cancer": Metaplastic [Exclusion]
"History of stage I-IIIb cancer": stage I OR stage II OR stage IIIa OR stage IIIb [Exclusion]
"Known and untreated, or active CNS metastases": CNS metastases [Exlusion]
"History of inflammatory bowel disease or active bowel inflammation": Inflammatory [Exclusion]
Step 2: Refine the list of phrases.
"Confirmed diagnosis of HR+/HER2- breast cancer": HR+/HER2- [Inclusion] - mentioned in
guidelines
"Metastatic or locally advanced disease not amenable to curative therapy": Metastatic OR
Locally Advanced [Inclusion] - mentioned in guidelines
"Confirmation of biomarker eligibility (detection of specified mutation(s) of PIK3CA via
specified test)": PIK3CA+ [Inclusion] - mentioned in guidelines
"Metaplastic breast cancer": Metaplastic [Exclusion] - mentioned in guidelines
"History of stage I-IIIb cancer": stage I OR stage II OR stage IIIa OR stage IIIb [Exclusion]
- stage related to history, not current state, should be removed
"Known and untreated, or active CNS metastases": CNS metastases [Exlusion] - metastases
related to specific site, not breast cancer in general, should be removed
"History of inflammatory bowel disease or active bowel inflammation": Inflammatory [Exclusion]
 - subtype related to bowel disease, not breast cancer, should be removed
Step 3: Apply additional guidelines.
HR+/HER2- translates to (ER+ OR PR+) AND HER2-.
Metastatic or locally advanced translates to (Stage IV OR Stage IIIb OR Stage IIIc).
PIK3CA+ directly from text.
NOT Metaplastic for the exclusion.
Step 4: Translate into boolean expression.
(ER+ OR PR+) AND HER2- AND (Stage IV OR Stage IIIb OR Stage IIIc) AND PIK3CA+ AND NOT
Metaplastic
Step 5: Transform into condition list.
ER+ OR PR+
HER2-
Stage IV OR Stage IIIb OR Stage IIIc
PIK3CA+
NOT Metaplastic
Step 6: Create pairs.
Condition: "ER+ OR PR+" , Phrases: "Confirmed diagnosis of HR+/HER2- breast cancer" /
Condition: "HER2-", Phrases: "Confirmed diagnosis of HR+/HER2- breast cancer" / Condition: "
Stage IV OR Stage IIIb OR Stage IIIc" , Phrases: "Metastatic or locally advanced disease not
amenable to curative therapy" / Condition: "PIK3CA+" , Phrases: "Confirmation of biomarker
eligibility (detection of specified mutation(s) of PIK3CA via specified test)" / Condition: "
NOT Metaplastic", Phrases: "Metaplastic breast cancer"
```

Figure 10.12: Chain of thought for the first trial example included in the prompt.

- **Inclusion criteria:** Contains the text of the inclusion criteria for a trial to be parsed in the inference. It starts with the heading *"Inclusion criteria:"* followed by a new line.
- **Exclusion criteria:** Contains the text of the exclusion criteria for a trial to be parsed in the inference. It starts with heading *"Exclusion criteria:"* followed by a new line.
- **Triggering phrase:** Ends the prompt with a phrase that triggers the model to follow the chain of though demonstrated in the examples. The phrase used is: *"Step 1: Understand the Text:"*

```
Step 1: Identify phrases and conditions.
"Histologically confirmed advanced metastatic adenocarcinoma with measurable or evaluable
disease": Advanced AND Metastatic AND Adenocarcinoma [Inclusion]
"Patients who have progressed on distant metastatic sites after curative surgery or have stage
 IV breast cancer at diagnosis": Metastatic OR Stage IV [Inclusion]
"Patients with inflammatory disease are eligible": Inflammatory [Inclusion]
"Patient has HER2-negative breast cancer with IHC and/or FISH (or SISH, CISH)": HER2- [
Inclusion]
"Known hormon receptor status": ER+ OR PR+ OR ER- OR PR- [Inclusion]
"Postmenopausal women": Premenopausal [Exclusion]
"History of DCIS": DCIS [Exclusion]
"Bilateral mastectomy": Bilateral [Exclusion]
"Patients has received previous treatment with CDK 4/6 inhibitors, mTOR inhibitors, PIK3CA
inhibitors or capecitabine": PIK3CA inhibitors [Exclusion]
"No symptomatic visceral metastasis": Symptomatic visceral metastasis [Exclusion]
"Known brain metastases unless treated and stable": Brain metastases [Exclusion]
Step 2: Refine the list of phrases.
"Histologically confirmed advanced metastatic adenocarcinoma with measurable or evaluable
disease": Advanced AND Metastatic AND Adenocarcinoma[Inclusion] - mentioned in guidelines
"Patients who have progressed on distant metastatic sites after curative surgery or have stage
 IV breast cancer at diagnosis": Metastatic OR Stage IV [Inclusion] - mentioned in guidelines
"Patients with inflammatory disease are eligible": Inflammatory [Inclusion] - patients with
this condition are eligible but it is not strictly required, should be removed
"Patient has HER2-negative breast cancer with IHC and/or FISH (or SISH, CISH)": HER2- [
Inclusion] - mentioned in guidelines
"Known hormon receptor status": ER+ OR PR+ OR ER- OR PR- [Inclusion] - hormon receptor status
not specified, should be removed
"Postmenopausal women": Premenopausal [Exclusion] - not mentioned in guidelines, should be
removed
"History of DCIS": DCIS [Exclusion] - related to history, not current state, should be removed
"Bilateral mastectomy": Bilateral [Exclusion] - related to mastectomy, does not specify
required biomarker, stage or subtype, should be removed
"Patients has received previous treatment with CDK 4/6 inhibitors, mTOR inhibitors, PIK3CA
inhibitors or capecitabine": PIK3CA inhibitors [Exclusion] - related to inhibitors, does not
specify required biomarker, stage or subtype, should be removed
"No symptomatic visceral metastasis": Symptomatic visceral metastasis [Exclusion] - not
mentioned in guidelines, should be removed
"Known brain metastases unless treated and stable": Brain metastases [Exclusion] - not
mentioned in guidelines, should be removed
Step 3: Apply additional guidelines.
Advanced translates to (Stage IIIb OR Stage IIIc OR Stage IV) and Metastatic translates to
Stage IV, so Advanced AND Metastatic AND Adenocarcinoma translates to Stage IV AND
Adenocarcinoma.
Metastatic translates to Stage IV, so Metastatic OR Stage IV translates to Stage IV.
HER2- directly from text.
Step 4: Translate into boolean expression.
Stage IV AND Adenocarcinoma AND HER2-
Step 5: Transform into condition list.
Stage IV
Adenocarcinoma
HER2-
Step 6: Create Pairs.
Condition: "Stage IV", Phrases: "Histologically confirmed advanced metastatic adenocarcinoma
with measurable or evaluable disease" || "Patients who have progressed on distant metastatic
sites after curative surgery or have stage IV breast cancer at diagnosis" / Condition: "
Adenocarcinoma", Phrases: "Histologically confirmed advanced metastatic adenocarcinoma with
measurable or evaluable disease" / Condition: "HER2-", Phrases: "Patient has HER2-negative
breast cancer with IHC and/or FISH (or SISH, CISH)"
```

Figure 10.13: Chain of thought for the third trial example included in the prompt.

The full text of the prompt is included in Appendix C.

10.3 Tool implementation

To create a functional system, the implementation of the eligibility criteria parsing tool integrates various components: data processing and inference pipelines, robust logging mechanisms, model monitoring, and user interface development. The overall architecture design ensures seamless work of these elements. Moreover, the DevOps and MLOps practices are applied to make the development process and the tool itself efficient and reliable.

10.3.1 Tool architecture

The architecture of the eligibility criteria parsing tool PoC follows MLOps best practices to ensure effective processing, robust logging, and monitoring. Additionally, the tool is designed to enable quick delivery, which is crucial for the PoC. The overall architecture diagram is illustrated in Figure 10.14.

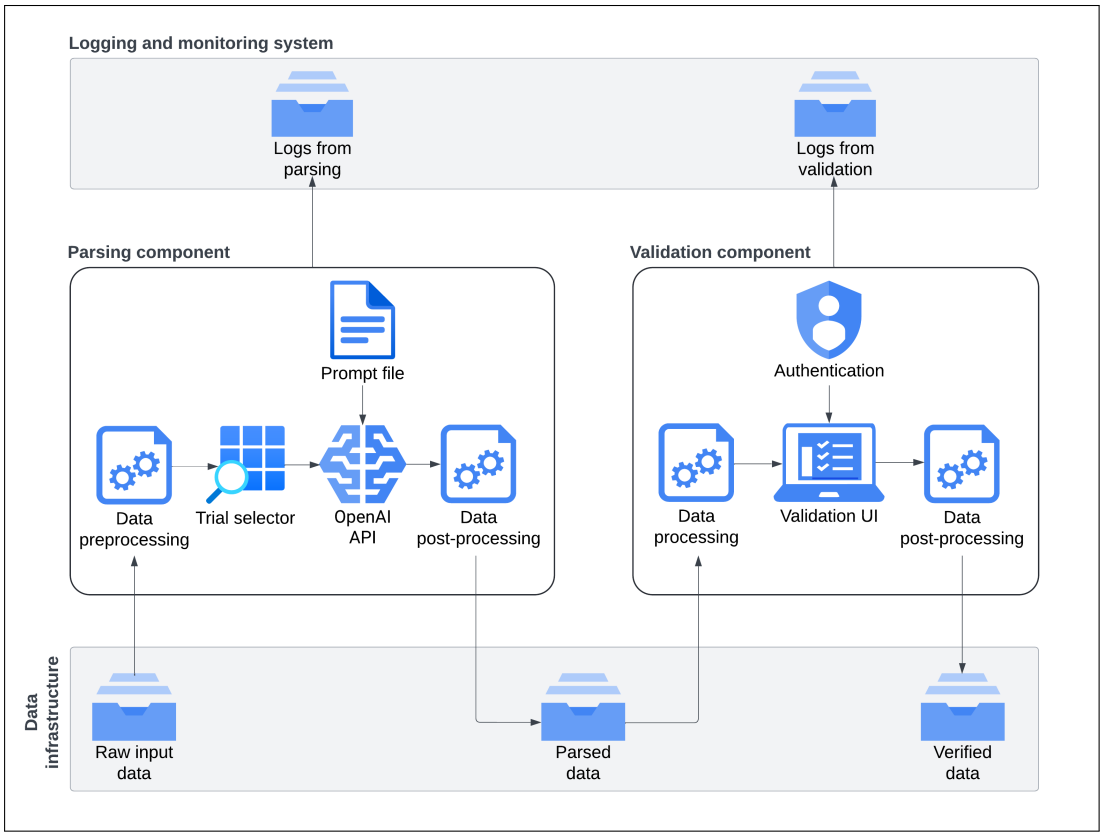


Figure 10.14: Architecture of the eligibility criteria parsing tool.

The architecture includes four main elements: the parsing component, validation component, data storage, and logging and monitoring system. Each element is

crucial for the tool's functionality.

Data storage contains three main objects:

- **Raw input data:** A JSON file containing an array of JSON objects, each with two fields: *"nct_id"* (trial ID) and *"eligibility_criteria"* (the eligibility criteria section from the trial protocol or ClinicalTrials.gov entry). Each JSON object relates to a single trial.
- **Parsed data:** A JSON file created after criteria parsing, which contains an array structured as shown in Figure 10.2. This is the output file from the parsing component.
- **Verified data:** A JSON file that is the final validated version of the parsed criteria. It has the same structure as the parsed data but includes modifications made by experts during validation. This is the output of the validation component.

For the PoC, on-premises storage is used, and all input data consists of public records from ClinicalTrials.gov.

The parsing component includes the following elements:

- **Data preprocessing:** Processes the inclusion criteria texts from the input JSON array, cleans the text from section headers, and splits it into *"inclusion_criteria"* and *"exclusion_criteria"* fields using regular expressions. The new fields are added to each JSON object, and the *"eligibility_criteria"* field is removed.
- **Trial selector:** Selects a single JSON object from the array for inference.
- **Prompt file:** Contains the prompt template described in Section 10.2.4, and included in the Appendix C.
- **Inference service:** Replaces placeholders in the prompt with inclusion and exclusion criteria for a single trial, connects with the pretrained GPT-4o model via the OpenAI API to generate output, and adds the generated text to the JSON array as an *"output"* field. The model output includes the full chain of thought for criteria parsing, similar to the examples provided in the prompt. The inference service uses LangChain library to facilitate prompt creation and API interactions.
- **Data post-processing:** Extracts the condition list from the generated output, processes the text to create an array *"condition_list"* with JSON objects including *"condition"* and *"phrases"* fields (Figure 10.2, and saves the main JSON array as a file (parsed data) to data storage. This processing is done using regular expressions.

The parsing component operates independently of the validation tool, saving resulting files for later validation. Domain experts use the validation component to check parsing results when needed.

The validation component consists of:

- **Data preprocessing:** Prepares parsed data for display in the UI by joining lists of phrases into comma-separated strings for better readability. It also creates annotated inclusion and exclusion criteria objects for the Streamlit HTML component.

- **Authentication:** Requires users to log in with their first and last names and a password defined in app secrets. The name fields require at least three characters in each. Successful authentication grants access to the first page of the UI.

- **Validation UI:** A Streamlit application deployed on the public Streamlit Cloud, accessible via a link for authorized users. It displays parsing outputs, including eligibility criteria and a table of extracted conditions and relevant phrases. The phrases are also highlighted in the criteria text to enhance the validation process. Users can edit the table to make necessary adjustments. To ensure consistency in validation, the UI contains guidelines for parsing. The UI details are described in Section 10.3.2. Approved structured conditions are passed to the post-processing pipeline.

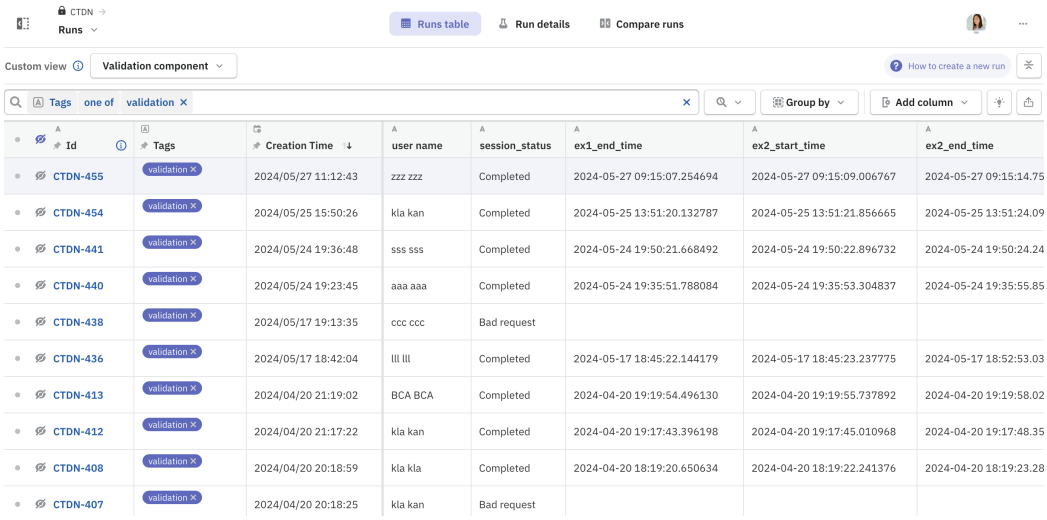
- P**ost-processing:** Converts the comma-separated strings in the phrases column back into lists, transforms table rows into JSON obects with *"condition"* and *"phrases"* pairs, and saves the updated array as a verified data JSON file. This file can then be used for other applications, such as patient-matching use case.

The parsing and validation components are integrated with Neptune.ai for logging and monitoring. Both components log common metrics such as start and end times, standard output and error, pipeline step statuses, CPU and memory usage, and final processing statuses. Additionally, the parsing component logs raw input data ID, inference input data, prompt file ID, model name, temperature parameter value, and parsed data. The validation component logs user details upon successful authentication, input parsed data ID, intermediate files with user modifications and the final verified data.

Integration with Neptune.ai allows each execution of the parsing or validation component to add a new record to the registry. Neptune's UI enables easy browsing of runs, filtering by logged parameters, and adding custom tags, which is crucial for error inspection and verifying the correct operation of the application. Figures 10.15 and 10.16 illustrate the runs table and detailed logs view in Neptune.ai, respectively.

10.3.2 User interface design

The user interface (UI) of the eligibility criteria parser is a fundamental component that enables domain experts to validate parsing results and make necessary corrections. It is built using Streamlit, which allows for seamless integration with data and ML pipelines, and quick delivery, which is one of the PoC goals. Using Streamlit eliminates the need for a dedicated frontend developer to build a simple UI.



Id	Tags	Creation Time	user name	session_status	ex1_end_time	ex2_start_time	ex2_end_time
CTDN-455	validation	2024/05/27 11:12:43	zzz zzz	Completed	2024-05-27 09:15:07.254694	2024-05-27 09:15:09.006767	2024-05-27 09:15:14.75
CTDN-454	validation	2024/05/25 15:50:26	kla kan	Completed	2024-05-25 13:51:20.132787	2024-05-25 13:51:21.856665	2024-05-25 13:51:24.09
CTDN-441	validation	2024/05/24 19:36:48	sss sss	Completed	2024-05-24 19:50:21.668492	2024-05-24 19:50:22.896732	2024-05-24 19:50:24.24
CTDN-440	validation	2024/05/24 19:23:45	aaa aaa	Completed	2024-05-24 19:35:51.788084	2024-05-24 19:35:53.304837	2024-05-24 19:35:55.85
CTDN-438	validation	2024/05/17 19:13:35	ccc ccc	Bad request			
CTDN-436	validation	2024/05/17 18:42:04	lll lll	Completed	2024-05-17 18:45:22.144179	2024-05-17 18:45:23.237775	2024-05-17 18:52:53.03
CTDN-413	validation	2024/04/20 21:19:02	BCA BCA	Completed	2024-04-20 19:19:54.496130	2024-04-20 19:19:55.737892	2024-04-20 19:19:58.02
CTDN-412	validation	2024/04/20 21:17:22	kla kan	Completed	2024-04-20 19:17:43.396198	2024-04-20 19:17:45.010968	2024-04-20 19:17:48.35
CTDN-408	validation	2024/04/20 20:18:59	kla kla	Completed	2024-04-20 18:19:20.650634	2024-04-20 18:19:22.241376	2024-04-20 18:19:23.28
CTDN-407	validation	2024/04/20 20:18:25	kla kan	Bad request			

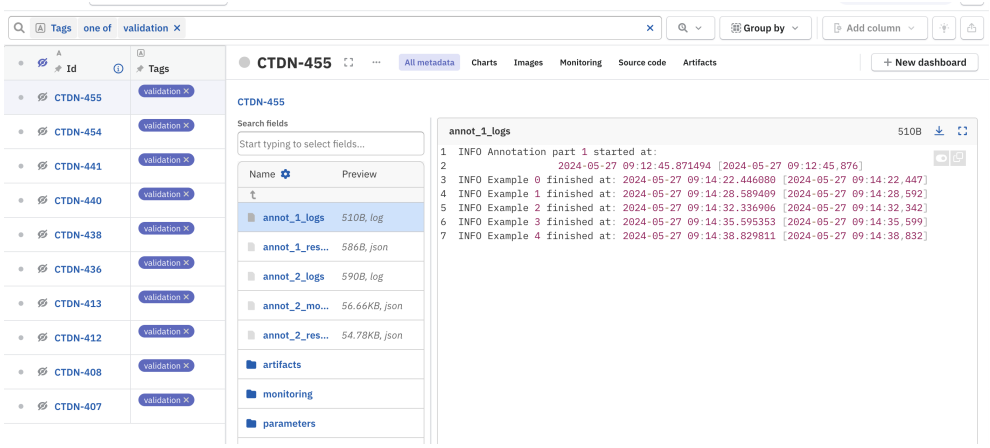
Figure 10.15: Main view of the validation component logs in Neptune.ai

The interface is designed to be intuitive and user-friendly, ensuring it can be easily operated by non-technical users and providing all necessary information for domain experts to perform validation tasks accurately. This section outlines the various elements and functionalities of the UI.

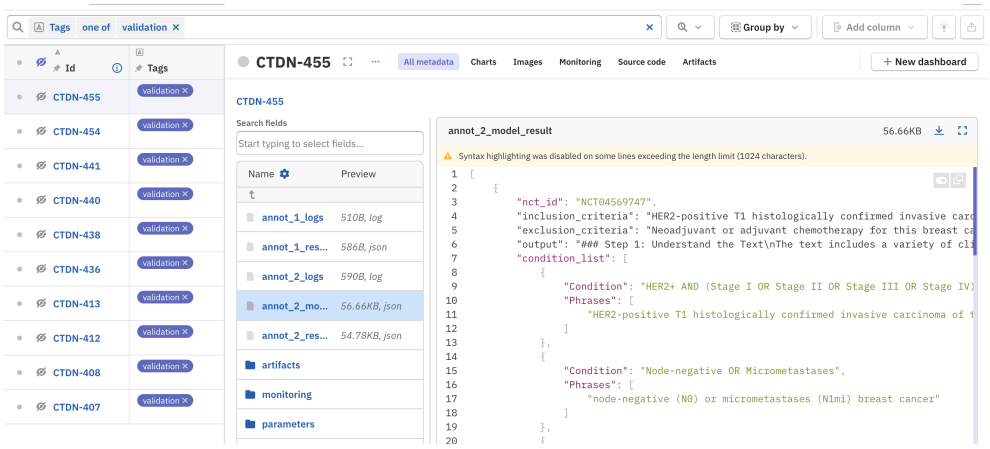
The first page of the interface is the authorization page. It contains a header: *"Please enter your name and the secret key to access the validation app"* and a form with three fields: *"Your first name"*, *"Your surname"*, and *"Secret key"* (Figure 10.17). Users must provide their names, which must be at least three characters long. Any shorter input results in a yellow warning *"First name and surname must be at least 3 characters long."* (Figure 10.17(a)) and prevents proceeding. The *"Secret key"* is a password-type field where input characters are obscured. The provided value is compared against a string defined in a secret TOML file. If there is a mismatch, the user sees a red error: *"Secret key incorrect"*. (Figure 10.17(b)). The user can log in after providing names that meet the three-characters condition and a correct password.

The next page is the welcome page with instructions for the user on how to use the application and validate the parsed data (Figure 10.18). After familiarizing themselves with the guidelines, the user clicks the green *"Start"* button to proceed.

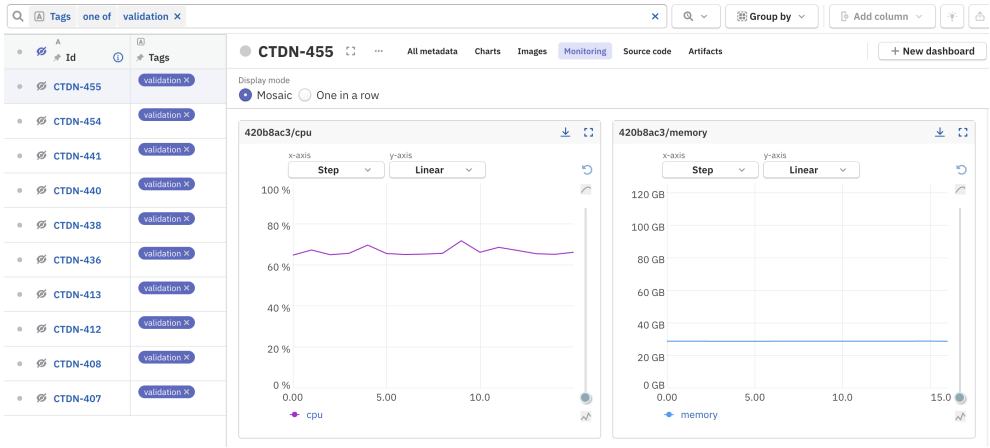
The validation page provides the interface for correcting the parsed data for a single trial (Figure 10.19). It is divided into four parts. The top two are scrollable text containers for inclusion and exclusion criteria text. The bottom left contains a table with conditions and phrases, which can be modified. It enables text corrections, adding, or deleting rows. To increase validation efficiency, each phrase from the table is highlighted in the inclusion and exclusion criteria text, and the highlighted spans have tags indicating the relevant conditions (Figure 10.20). Adding a new phrase to the table causes it to be highlighted it in the text, and removing a phrase removes the highlighted span. The highlighting feature is case-insensitive but requires matching syntax, punctuation and spelling.



(a) Pipeline statuses and execution times.

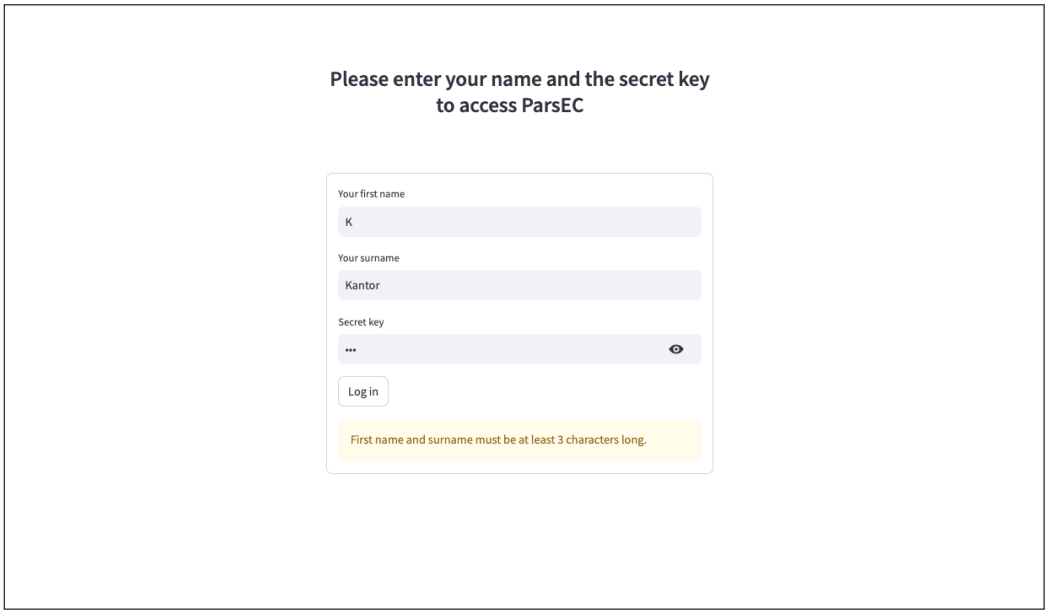


(b) JSON file with parsed data.

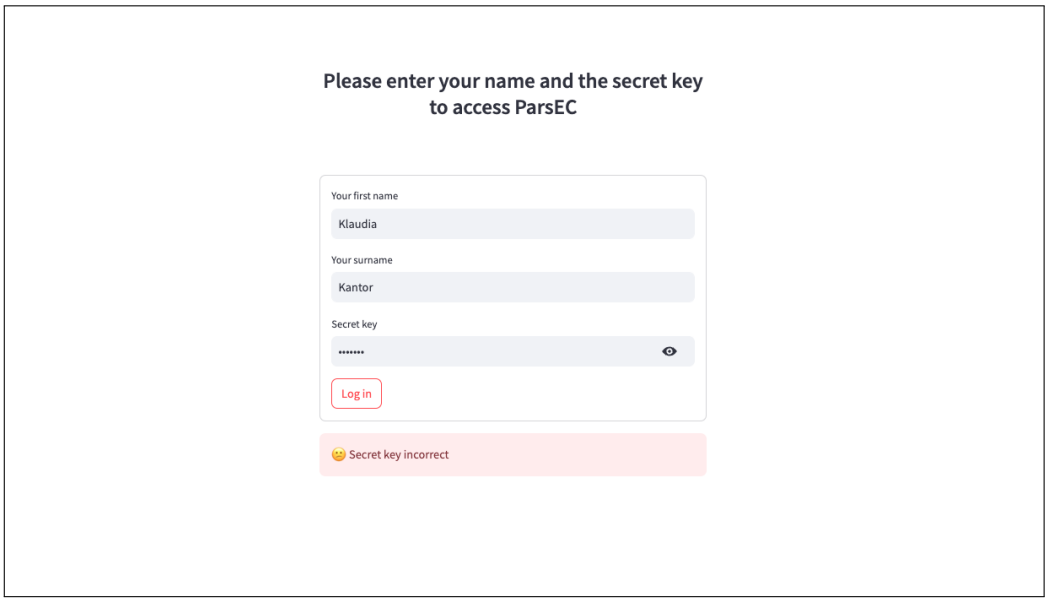


(c) Hardware monitoring.

Figure 10.16: Detailed view of an individual run's logs in Neptune.ai.



(a) Authorization with incorrect name value.



(b) Authorization with incorrect password.

Figure 10.17: Authorization page of the validation UI.

The bottom right sections contains parsing guidelines to ensure consistent validation, providing users with reminders on the rules for boolean expression creation. This section includes expanders with allowed terms, rules for defining required, disallowed, or alternative conditions, and instructions for separating different phrases in a single row (Figure 10.21). When the user finishes validation, they click the green *"Next"* button to move on to a new example.

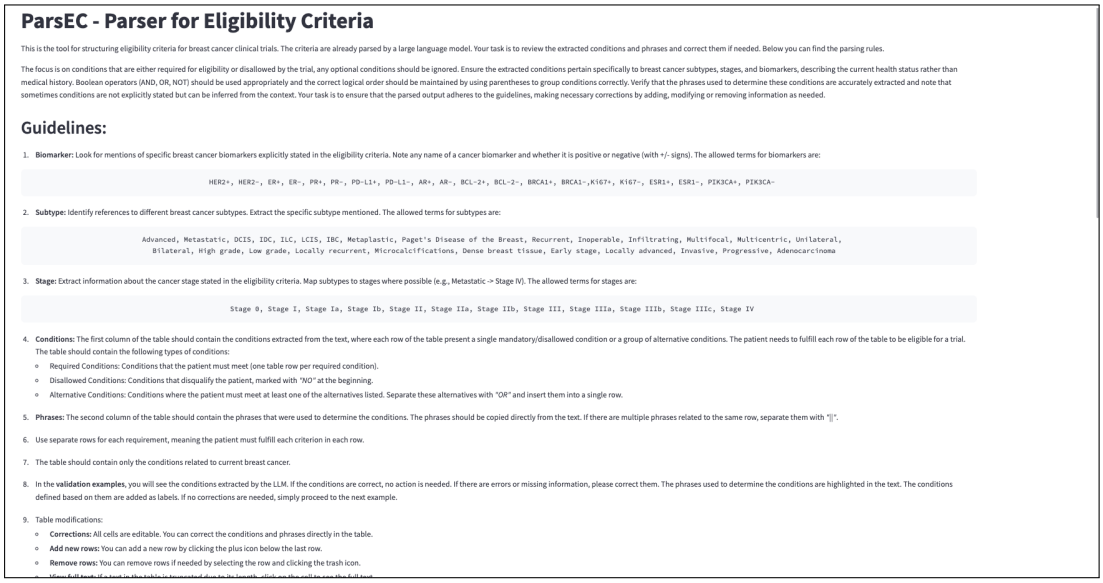


Figure 10.18: Welcome page of the validation UI.

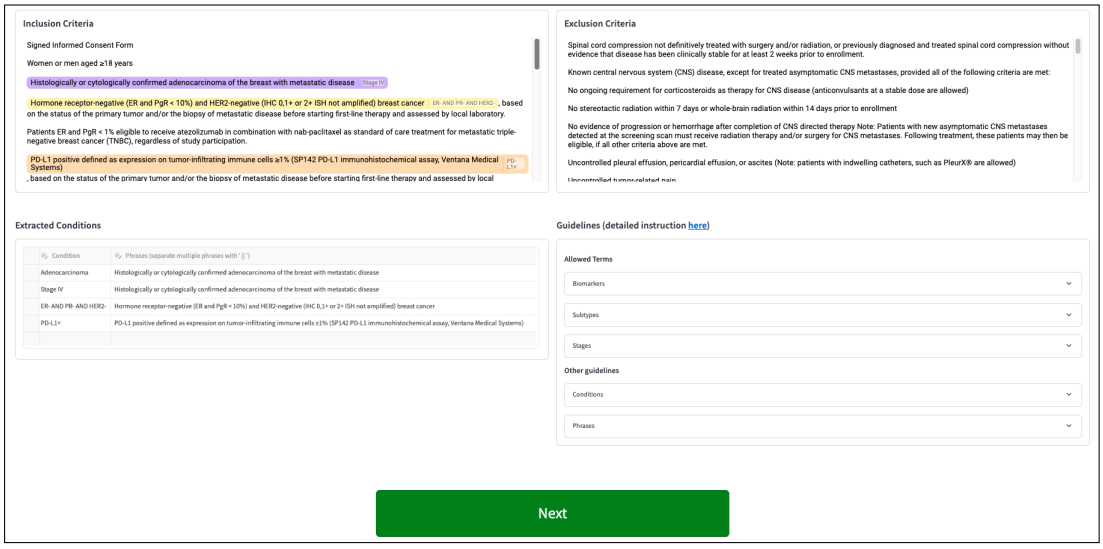


Figure 10.19: Validation page of the validation UI.

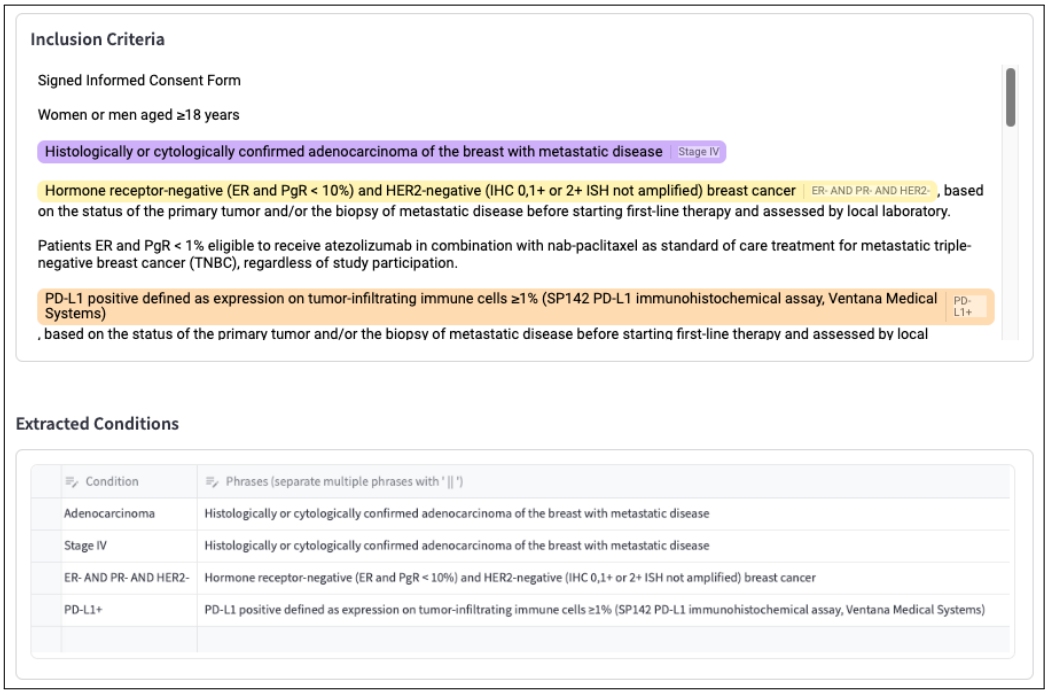


Figure 10.20: Phrase highlighting feature of the validation UI.

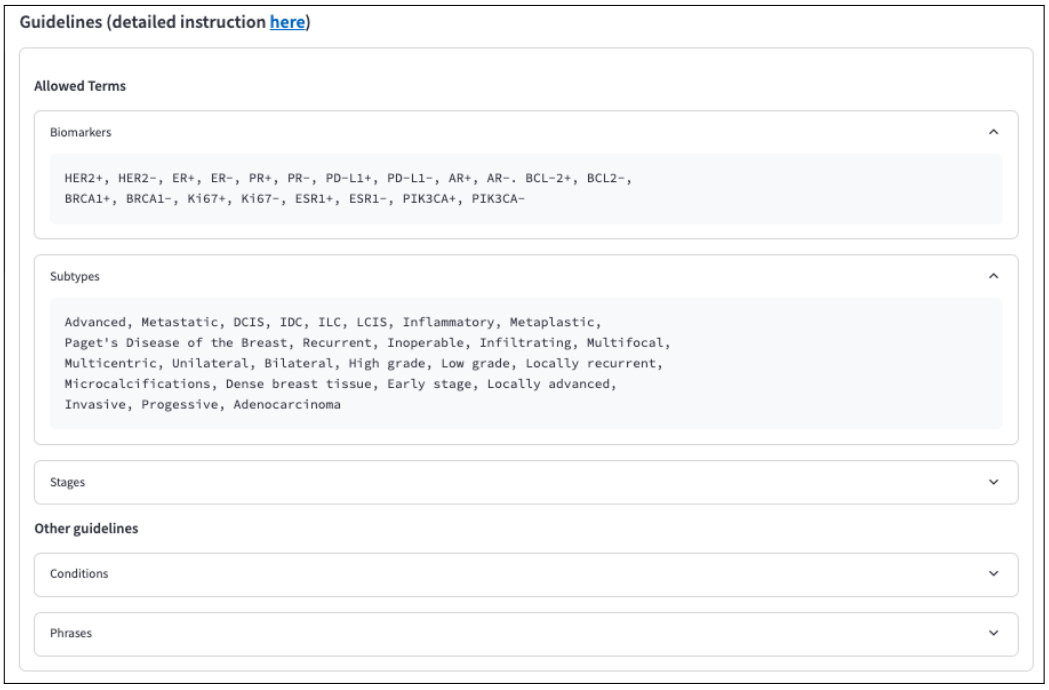


Figure 10.21: Guidelines section of the validation UI.

When all examples are validated, the user is presented with the final page of the application - the *"Thank you"* page, which includes a message of gratitude for completing the task and instruction on whom to contact in case of questions (Figure 10.22). To exit the app, the user simply closes the tab in the browser.

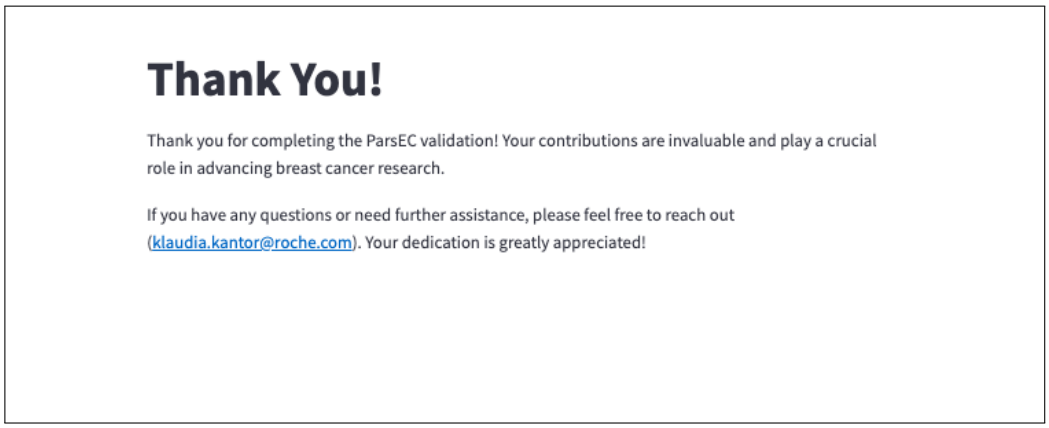


Figure 10.22: Final page of the validation UI.

10.3.3 Explainability

One of the valuable features of any ML product, especially in the medical domain, is its capability to explain the model decisions. The eligibility criteria parser provides such functionality through the use of a chain-of-thought approach. The examples provided in the prompt contain the complete reasoning behind the decisions made in parsing. They present a detailed process, include comments with explanations, and show some intermediate transformations. By utilizing in-context learning, the model learns to follow the same approach when generating responses for new examples in inference. This process is triggered by a phrase added at the end of the prompt, which mimics the start of the chain of thought presented in the examples.

This technique was tested in experimentation, and we observed that the final parsed outcomes are more accurate when the model is fed with chain-of-thought examples and follows this reasoning in its output. Additionally, providing more examples with chains of thought has better impact on the accuracy, than expanding the guidelines and adding some extra rules, exceptions etc. Moreover, including examples with incorrect extractions on purpose, and refining them after with some explanations, improves the accuracy even more. In the inference, we see that the model folows this process in exactly the same way, some of the mistakes are fixed in the intermediate steps, resulting in a correct final output. Moreover, this generated output with full reasoning can function as an explanation feature. To illustrate how it works, we will focus on a specific example of trial NCT00486668 and its inference results. The inclusion and exclusion criteria for this trial are shown in Figure 10.23.

The trial criteria were processed by the parsing component, where they were inserted into the placeholders of the tool prompt (presented in Appendix C) and

the prompt was passed to the GPT-4o model. The output generated by the model is shown in Figure 10.24. From the output, it is possible to inspect step-by-step how the model arrived at the final output. For instance, we see that four conditions were extracted from exclusion criteria, so they are negated in the boolean expression, which is visible in Step 4. Moreover, in Step 3, the term *"metastatic"* is translated to *"(Stage IV)"*, according to the guidelines. Additionally, we see that in first step it extracts also a sentence *"Prior history of breast cancer, including DCIS (Patients with a history of LCIS are eligible)"*, which is related to the past, not current state, but in the Step 2 it correctly refines it and explains *"related to history, not current state, should be removed"* and this condition is not further used by the model.

Additionally, the output from the model includes phrases that were responsible for creating the condition list, which is also very supportive in interpretation and validation. For example, we know that *"cN2b"* and *"cN3"* were mapped to *"Stage IIIb"* and *"Stage IIIc"*, which makes these extractions easy to be verified by the domain experts with the provided phrases. We also see that the model correctly merged two phrases related to *"NOT Stage IV"* into one list, separated by *"||"*, as specified in the instruction.

Currently, the full explanations are only stored in logs and available from the Neptune.ai tool, but they can be easily integrated into the validation app if required by users. In the current version, the users are only presented with the conditions and relevant phrases.

```
Inclusion criteria:
Female; 18 years or older; ECOG performance status of 0 or 1; Primary breast tumor palpable
and measures greater than or equal to 2.0 cm by physical exam; Diagnosis of invasive
adenocarcinoma made by core needle biopsy; Breast cancer determined to be HER2-positive; LVEF
assessment by MUGA scan or ECG within 3 months prior to randomization; Blood counts must meet
the following criteria:; ANC greater than or equal to 1200/mm3; Platelet count greater than or
 equal to 100,000/mm3; Hemoglobin greater than or equal to 10 g/dL; Serum creatinine less than
 or equal to ULN for the lab; Adequate hepatic function by these criteria:; Total bilirubin
less than or equal to the ULN for the lab unless the patient has a bilirubin elevation greater
 than ULN to 1.5 x ULN resulting from Gilbert's disease or similar syndrome due to slow
conjugation of bilirubin; and; Alkaline phosphatase less than or equal to 2.5 x ULN; and; AST
less than or equal to 1.5 x ULN for the lab.; If skeletal pain present or alkaline phosphatase
 greater than ULN (but less than or equal to 2.5 x ULN), bone scan or PET scan must not
demonstrate metastatic disease; If AST or alkaline phosphatase greater than ULN , liver
imaging (CT, MRI or PET scan) must not demonstrate definitive metastatic disease and the
requirements in criterion for hepatic function must be met; Able to swallow oral medications;

Exclusion criteria:
FNA alone to diagnose the primary tumor; Excisional biopsy or lumpectomy was performed prior
to randomization; Surgical axillary staging procedure prior to randomization. Exceptions: 1)
FNA or core biopsy of an axillary node for any patient, and 2) although not recommended, a pre
-neoadjuvant therapy SN biopsy for patients with clinically negative axillary nodes.; Tumors
clinically staged as T4; Ipsilateral cN2b or cN3 disease (Patients with cN1 or cN2a disease
are eligible); Definitive clinical or radiologic evidence of metastatic disease; Synchronous
bilateral invasive breast cancer; Requirement for chronic use of any of the medications or
substances specified in the protocol; Treatment including RT, chemotherapy, and/or targeted
therapy for the currently diagnosed breast cancer prior to randomization; Any sex hormonal
therapy, e.g., birth control pills, ovarian hormone replacement therapy, etc. (These patients
are eligible if therapy is discontinued prior to randomization); Continued therapy with any
hormonal agent such as raloxifene, tamoxifen, or other SERM. (Patients are eligible only if
these medications are discontinued prior to randomization); Prior history of breast cancer,
including DCIS (Patients with a history of LCIS are eligible); Prior therapy with
anthracyclines, taxanes, trastuzumab, or lapatinib for any malignancy; Other malignancies
unless the patient is considered to be disease-free for 5 or more years prior to randomization
 and is deemed by her physician to be at low risk for recurrence. Patients with the following
cancers are eligible if diagnosed and treated within the past 5 years: carcinoma in situ of
the cervix, carcinoma in situ of the colon, melanoma in situ, and basal cell and squamous cell
 carcinoma of the skin.; Cardiac disease that would preclude the use of the drugs included in
the B-41 treatment regimens. This includes but is not confined to:; Active cardiac disease:;
angina pectoris requiring the use of anti-anginal medication;; ventricular arrhythmias except
for benign premature ventricular contractions controlled by medication;; conduction
abnormality requiring a pacemaker;; supraventricular and nodal arrhythmias requiring a
pacemaker or not controlled with medication; and; clinically significant valvular disease.;
History of cardiac disease:; myocardial infarction;; congestive heart failure; or;
cardiomyopathy.; Uncontrolled hypertension, defined as blood pressure greater than 150/90 mm/
Hg on antihypertensive therapy; History of or current symptomatic interstitial pneumonitis or
pulmonary fibrosis or definitive evidence of interstitial pneumonitis or pulmonary fibrosis
described on CT or chest x-ray in asymptomatic patients; Sensory/motor neuropathy greater than
 or equal to grade 2, as defined by the NCI's CTCAE v3.0; Malabsorption syndrome, ulcerative
colitis, resection of the stomach or small bowel, or other disease significantly affecting
gastrointestinal function; Other non-malignant systemic disease that would preclude treatment
with any of the treatment regimens or would prevent required follow-up; Conditions that would
prohibit administration of corticosteroids; Administration of any investigational agents
within 30 days before randomization; Pregnancy or lactation
```

Figure 10.23: Inclusion and exclusion criteria for trial NCT00486668.

```
### Step 1: Identify phrases and conditions.
"Diagnosis of invasive adenocarcinoma made by core needle biopsy": Invasive AND Adenocarcinoma
 [Inclusion]
"Breast cancer determined to be HER2-positive": HER2+ [Inclusion]
"Tumors clinically staged as T4": Stage IV [Exclusion]
"Ipsilateral cN2b or cN3 disease (Patients with cN1 or cN2a disease are eligible)": Stage IIIb
 OR Stage IIIc [Exclusion]
"Definitive clinical or radiologic evidence of metastatic disease": Metastatic [Exclusion]
"Synchronous bilateral invasive breast cancer": Bilateral [Exclusion]
"Prior history of breast cancer, including DCIS (Patients with a history of LCIS are eligible)
": DCIS [Exclusion]
### Step 2: Refine the list of phrases.
"Diagnosis of invasive adenocarcinoma made by core needle biopsy": Invasive AND Adenocarcinoma
 [Inclusion] - mentioned in guidelines
"Breast cancer determined to be HER2-positive": HER2+ [Inclusion] - mentioned in guidelines
"Tumors clinically staged as T4": Stage IV [Exclusion] - mentioned in guidelines
"Ipsilateral cN2b or cN3 disease (Patients with cN1 or cN2a disease are eligible)": Stage IIIb
 OR Stage IIIc [Exclusion] - mentioned in guidelines
"Definitive clinical or radiologic evidence of metastatic disease": Metastatic [Exclusion] -
mentioned in guidelines
"Synchronous bilateral invasive breast cancer": Bilateral [Exclusion] - mentioned in
guidelines
"Prior history of breast cancer, including DCIS (Patients with a history of LCIS are eligible)
": DCIS [Exclusion] - related to history, not current state, should be removed
### Step 3: Apply additional guidelines.
Invasive AND Adenocarcinoma directly from text.
HER2+ directly from text.
Stage IV directly from text.
Stage IIIb OR Stage IIIc directly from text.
Metastatic translates to Stage IV.
Bilateral directly from text.
### Step 4: Translate into boolean expression.
(Invasive AND Adenocarcinoma AND HER2+) AND NOT (Stage IV OR Stage IIIb OR Stage IIIc OR
Bilateral)
### Step 5: Transform into condition list.
Invasive
Adenocarcinoma
HER2+
NOT Stage IV
NOT Stage IIIb
NOT Stage IIIc
NOT Bilateral
### Step 6: Create pairs.
Condition: "Invasive", Phrases: "Diagnosis of invasive adenocarcinoma made by core needle
biopsy"
Condition: "Adenocarcinoma", Phrases: "Diagnosis of invasive adenocarcinoma made by core
needle biopsy"
Condition: "HER2+", Phrases: "Breast cancer determined to be HER2-positive"
Condition: "NOT Stage IV", Phrases: "Tumors clinically staged as T4" || "Definitive clinical
or radiologic evidence of metastatic disease"
Condition: "NOT Stage IIIb", Phrases: "Ipsilateral cN2b or cN3 disease (Patients with cN1 or
cN2a disease are eligible)"
Condition: "NOT Stage IIIc", Phrases: "Ipsilateral cN2b or cN3 disease (Patients with cN1 or
cN2a disease are eligible)"
Condition: "NOT Bilateral", Phrases: "Synchronous bilateral invasive breast cancer"
```

Figure 10.24: Output from the GPT-4o model for trial NCT00486668.

10.3.4 Implementation of MLOps practices

The development of the eligibility criteria parsing tool incorporated numerous MLOps practices to ensure robust performance, reproducibility, reliable results, and effective monitoring. As this is a Proof of Concept, not all requirements stated in Chapter 9 were implemented. However, the implemented practices made the experimentation process and tool evaluation more efficient. Furthermore, they ensure that this PoC design will not need to be rebuilt from scratch if approved for production but can serve as a solid foundation.

Firstly, Git is used for version control of the code. There are two repositories, one for the parsing component and another for the validation component. The history of updates is maintained, and all modifications are tracked. Each change is developed within a separate branch, and changes are integrated into the main branch through merge requests. The code is documented, with all functions containing docstrings and type hints. The repositories include *README.md* files with project descriptions and instructions on how to execute the code.

Care is taken to ensure code quality. The code is well-crafted and organized into modules and functions with a narrow scope. Code duplication is avoided, and confidential variables are stored as environment variables rather than hard-coded. The Python *logging* module is used for logging. Pre-commit hooks such as *isort* and *Flake8* are implemented to ensure consistent code formatting and adherence to software development standards. Code that does not meet these standards cannot be committed. Automated testing is implemented with basic unit tests. The test coverage for both projects is above 80%.

The project environment is reproducible and maintained with *Poetry*. Project dependencies are defined in the *pyproject.toml* file along with their specific versions. In the validation component repository, they are also listed in the *requirements.txt* file, which is needed for Streamlit application deployment.

All pipelines are automated, covering data loading, preprocessing, inference, post-processing, and data saving. There are no manual intermediate steps. For the PoC scope, only the parsing component pipeline needs to be started manually. The tool design is modular and flexible, allowing new functionalities to be easily added to the pipelines. For instance, a new therapeutic area can be included by slightly modifying the prompt, such as updating medical instructions or providing additional examples for in-context learning.

The data files are stored on-premise. However, in the case of the validation component, the input dataset also needs to be added to the Git repository to be loaded in the deployed application in Streamlit Cloud. This is a temporary solution that can be easily replaced if another infrastructure is selected over Streamlit Cloud. Notably, the datasets used contain public data, so this solution does not violate any security rules.

Moreover, all experiments are tracked in Neptune.ai, which is integrated with the pipelines of both components. Neptune.ai stores the prompt templates used and serves as a reliable repository of experimental configurations. This allows for

easy comparison of experiments and selection of the best configuration, ensuring reproducible results. Neptune.ai also functions as a monitoring tool, storing logs from the final solution and tracking all necessary data for assessing tool performance and error inspection.

Despite working only with public data, the validation application includes authorization, requiring users to enter a password to log in. This ensures that only authorized users can access and modify the data, which is aligned with security practices. Additionally, the Streamlit Cloud app is connected to the Git repository, enabling continuous deployment of changes. Any modification to the main branch results in an update to the validation application.

If accepted for production at Roche, additional MLOps practices will need to be implemented, for instance, containerization and CI/CD. Streamlit Cloud should be replaced with an internal cloud solution, and a recommended data storage option must be chosen. Furthermore, the logging storage and monitoring tool should be changed and separated from the experiment tracking.

10.4 Evaluation

Evaluating the output from a generative model in the complex task of information extraction is challenging due to a lack of robust, standardized methods. Therefore, the criteria parsing tool is evaluated manually in a specially designed experiment to ensure a valid assessment of its effectiveness.

The experiment involves four participants: a domain expert with extensive knowledge of breast cancer eligibility criteria, a doctor with general medical knowledge, and two less experienced individuals familiar with basic medical terminology but not experts. The objective is to compare the time efficiency of manual parsing against LLM-supported parsing. A dedicated Streamlit application has been developed for the experiment, accessible only to authorized users with the correct password.

Upon authorization, users are presented with detailed parsing instructions and a link to open these instructions in a separate tab for reference during the experiment. The instructions include a list of allowed terms for the extraction tables and a link to a document with three examples of pre-parsed criteria, to show the participants what is expected from them. These instructions are available in Appendix D.

The experiment begins with a training phase where users work through four examples of eligibility criteria: two for manual parsing and two for LLM-supported parsing. This phase allows participants to familiarize themselves with the tool, but the results are not included in the final analysis. Examples are presented on separate screens, one after another. In the manual parsing screen, users manually fill in the table with conditions and phrases, then click *"Finish example"* to proceed. The manual parsing page is presented in Figure 10.25. In the AI-supported parsing screen, tables are pre-filled with the parsed output from the LLM, and relevant phrases are highlighted in the text, as in the validation app (Figure 10.19). The user's task is to verify and correct these conditions.

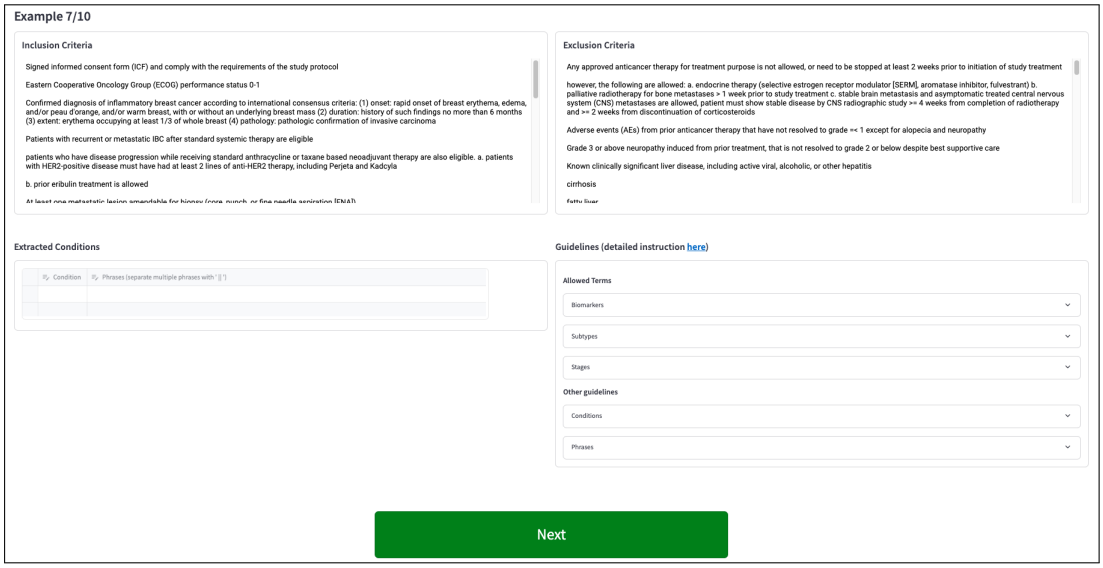


Figure 10.25: UI for manual parsing in the evaluation experiment.

The second phase is the core of the experiment, where results are logged and used for performance analysis. This phase mirrors the structure of the training phase but includes ten examples—five for manual parsing and five for AI-supported parsing. The examples from both groups are alternated to reduce the impact of the time reduction per example correlated with the increasing experience in the app and total time spent in the app. The examples include real eligibility criteria text from ten breast cancer trials sourced from ClinicalTrials.gov. These examples were carefully selected to ensure similar lengths of criteria texts and a comparable number of extracted rows in both manual and AI-assisted parsing examples, which enables a fair comparison between the two approaches.

The examples are separated by intermittent screens to offer users the chance to take breaks. Before starting this experiment phase, users are informed that the time spent on each example is measured and are asked to avoid disturbances during annotation. The time spent on each trial, along with the final versions of the extraction tables, is logged to Neptune.ai in the background.

At the end of the experiment, the average time taken for manual parsing versus LLM-supported parsing is analyzed for each participant. It is anticipated that AI-supported parsing will significantly reduce the time required compared to the manual approach.

10.5 Results

The experiment results reveal varied outcomes, likely influenced by differences in participants' expertise and their interaction with the parsing tool. Table 10.1 shows the time taken by each participant—identified as Expert, Doctor, Novice 1, and Novice 2—to complete both manual and LLM-supported parsing across all ten exam-

Table 10.1: Time taken by participants for manual and LLM-supported parsing (in minutes).

	Expert	**Doctor**	**Novice 1**	**Novice 2**
#1 - manual	**2.996**	3.138	5.975	5.450
#1 - LLM	4.520	**0.968**	**0.920**	**1.783**
#2 - manual	6.815	7.407	8.845	**6.678**
#2 - LLM	**4.417**	**6.372**	**5.766**	8.884
#3 - manual	6.900	4.452	**5.372**	**3.463**
#3 - LLM	**2.199**	**2.574**	5.461	4.883
#4 - manual	**4.798**	7.382	5.658	6.137
#4 - LLM	9.836	**4.829**	**2.697**	**2.427**
#5 - manual	**3.881**	8.201	6.310	**5.806**
#5 - LLM	5.463	**2.904**	**4.184**	6.692
Manual Avg.	**5.078**	6.116	6.432	5.507
LLM Avg.	5.287	**3.529**	**3.806**	**4.934**

ples. These examples were paired based on the criteria text length and the expected number of rows in the extraction table.

The Expert's results show notable variability between tasks. The average time for manual parsing was 5.078 minutes, while the LLM-supported took slightly longer, averaging 5.287 minutes. This indicates that the parsing tool did not consistently reduce the overall time for the Expert. However, we there is considerable inconsistency in the results for this participant. For instance, in pair #3, the LLM-supported parsing time was significantly shorter than manual parsing, and in pair #2, manual parsing took almost 2.5 minutes longer than the LLM validation. In contrast, the Expert achieved better times in manual parsing than in the LLM-supported parsing for the remaining pairs. This could suggest that the Expert's experience makes them very efficient manually, or that the LLM output required more time to verify and correct. Additionally, while the Expert was expected to achieve the shortest time in all trials, this was not consistently observed. For example, in #1 - LLM, the Expert was much slower than the others, which is also noticeable in #3 - manual and #4 - LLM. In the fourth pair, the Expert took almost 10 minutes, compared to around 5 minutes for the Doctor and about 2.5 minutes for the Novices. This suggests that other factors, such as distractions, technical issues, or loss of focus, may have influenced the Expert's performance.

For the other participants, the average time results show a significant improvement with LLM-supported parsing. For the Doctor and Novice 1, the time was approximately 1.7 times shorter with LLM support. The Doctor's average time dropped from 6.116 minutes to 3.529 minutes, with the most substantial reduction observed in pair #1, where LLM parsing took only 0.968 minutes compared to 3.138 minutes for manual parsing. Similarly, for the Novice 1, LLM support reduced the average time from 6.432 to 3.806 minutes. The results for these two participants were consistent across examples, with almost every LLM-supported example being faster than

its manual counterpart. The only exception was in pair #3, where the Novice 1 was slightly faster manually, but the difference was just 0.089 minutes (approximately 5 seconds).

For the Novice 2, there was greater variability in the results, which is expected given this participant's lower experience level. In pairs #1 and #4, LLM-assisted parsing was significantly faster, with a time reduction of about 4 minutes. However, in other examples, the Novice 2 was more efficient with manual parsing, although the differences were smaller (1-2 minutes). Nonetheless, the average time for AI-assisted parsing was still lower by 0.573 minutes (approximately 34 seconds).

To better understand these time results, it is also necessary to analyze the individual parsing outputs. While time efficiency provides valuable insights into the potential benefits of LLM-supported parsing, accuracy of the tool is also important. In this analysis, we treat the Expert's responses as the ground truth and use them to evaluate the LLM extractions as well as the outputs of the other participants. We focus mostly on the extracted conditions, but the detailed outputs from the LLM-assisted parsing, including the phrases, are available in Appendix E.

In example #1 - LLM, presented in Table 10.2, no differences were observed between the outputs of all four participants and the LLM-generated output, which means that no changes were made to the provided extractions. This is reflected in the Doctor's and Novice 1's times, which were below 1 minute. However, it is surprising that this example took the Expert more time than the manual parsing of a similar example, even though there were no corrections from the Expert.

Table 10.3 presents the outputs for example #2 - LLM. In this case, the LLM's output was not fully correct, as two rows were removed by the Expert. Additionally, the Expert modified the logical expression in the first row by adding the value from the third row, *"IBC"*, although the Expert did not follow the instruction correctly, using the synonym *"Inflammatory"*, which in not listed in the allowed terms. Interestingly, the Doctor's response was identical to the Expert's but correctly used the *"IBC"* term. The Novice 1 also removed the second row from the model output but did not fully grasp the logical meaning of the *"IBC"* condition. The Novice 2's output contained even more mistakes, including the addition of two incorrect rows *NOT DCIS OR NOT LCIS* and *"NOT HER2+"*, the latter of which contradicts the inclusion condition *"HER2+"* in the previous row. Surprisingly, this example took the Expert less time, despite involving more corrections to the LLM output. This additional burden is reflected in the longer times for the other participants, particularly the Novice 2, who took almost 9 minutes—over 2 minutes longer than the corresponding manual example.

In the third example (Table 10.4), the responses from the Expert, Doctor, and Novice 1 were identical, with no changes made to the model output. However, the Novice 2 accepted all the rows from LLM but incorrectly added four additional lines, two of which were redundant (*"HER2-"* and *"ER- AND PR-"*) as they duplicated conditions already mentioned in the second row. Additionally, the Novice 2 misinterpreted the phrase *"More than one prior line of chemotherapy in the locally advanced unresectable or metastatic setting."* as indicating an excluded breast cancer type (*"NOT*

Stage IIIb OR NOT Stage IIIc"), when it actually refers to an excluded treatment type. A similar situation was observed in the phrase assigned by Novice 2 to the condition *"Stage IV"*: *"Treatment with < 1 prior line of systemic therapy in the metastatic setting or adjuvant/neoadjuvant setting if metastatic recurrence within 12 months of treatment"*, which was incorrectly interpreted as a required metastatic cancer subtype, when in fact it refers to a required treatment type. In this example we also notice a significant time reduction in LLM-supported parsing for both the Expert and the Doctor, reflecting the absence of changes changes made to the model output by these participants. However, the Novice 2, who added several unnecessary rows, took about 1.5 minutes longer than in the corresponding manual example.

Table 10.5 shows the conditions extracted in example #4 - LLM. In this case, the model's output was mostly correct, but it missed three additional conditions that were identified only by the Expert. These conditions required highly specialized knowledge to infer, such as determining *Stage I OR Stage II OR Stage III* from the phrase *Primary breast tumor palpable and measures greater than or equal to 2.0 cm by physical exam*, or *"NOT Stage IIIc"* from *"Ipsilateral cN2b or cN3 disease (Patients with cN1 or cN2a disease are eligible)"*. Additionally, the Expert extracted the *"NOT Recurrent"* condition based on the exclusion criterion *"Prior history of breast cancer, including DCIS (Patients with a history of LCIS are eligible)"*, which has a very indirect meaning. Furthermore, the model, Expert, and Novice 2 included only *"NOT Bilateral"* in the table, inferred from the phrase *"Synchronous bilateral invasive breast cancer"*, but the term *"Invasive"* should also have been added, which was correctly done by the Doctor and Novice 1. Interestingly, despite the Expert spending significantly more time on this example (almost 10 minutes) than the other participants, their extraction was not entirely correct. This example highlights the significant difference in time between the Expert and the rest, as well as between the Expert's manual and LLM-supported parsing (4.798 vs. 9.836 minutes).

In the fifth example, the model's output was entirely correct, with no changes made by the Expert or the Doctor. The Novice 1 added one incorrect row, *"Metastatic"*, based on a phrase that referred to treatment characteristics rather than breast cancer type. The Novice 2, however, introduced four additional rows, including a duplicated *"Stage IV"* row and an incorrect *"NOT Stage IV"* row that contradicted the other conditions. For example, the *"Stage IV"* was inferred from the phrase *"the biopsy of metastatic disease before starting first-line therapy and assessed by local laboratory"* which is related to the required treatment, not to the cancer subtype. The Novice 2 also misinterpreted the condition *"ER+ AND PGR+"* based on a phrase *"Patients ER and PgR < 1% eligible to receive atezolizumab in combination with nab-paclitaxel as standard of care treatment for metastatic triple-negative breast cancer (TNBC), regardless of study participation"*, that is related to treatment eligibility rather than required cancer biomarkers. Moreover, the level *"<1%"* means negative polarity, not positive, as stated in the Novice 2' extraction. Interestingly, the Novice 2 took almost 7 minutes for this example, which is longer than the corresponding manual task. The Expert also took more time on this example than on the manual parsing, which is surprising given that no changes were made to the model output in the Expert's response. The

Doctor and Novice 2, however, had relatively low times, 3.529 and 3.806 minutes, respectively, which is consistent with the lack of any corrections needed.

In conclusion, the accuracy assessment of the model shows that the LLM-supported parsing tool performs well, with high accuracy overall. The most significant error occurred when the model added two incorrect rows in second example, which could have resulted in limiting the pool of eligible participants—a more serious issue compared to missing conditions like in the fourth example. However, the analysis also highlights inconsistencies in the Expert's time performance, particularly in cases where the LLM-supported parsing took longer than expected despite no corrections being made to the model's output. This anomaly suggests that additional factors, such as distractions or technical issues, may have influenced the Expert's performance. This is a limitation of conducting experiments in non-laboratory conditions, where full control over participants' environments is not possible.

Despite these inconsistencies, when we consider the average and overall time performance across all participants, there is a clear benefit to using the parsing tool. The Doctor's results, for instance, demonstrate a significant time reduction with LLM support, coupled with a high level of agreement with the model's output. The Novice 2, on the other hand, introduced many incorrect modifications, which not only increased their time but also reduced the accuracy of the results. This contrasts with the Doctor's performance, which underscores the advantage of having a medical background when using the tool.

Finally, one complex task required highly specialized knowledge in breast cancer eligibility criteria, which only the Expert could accurately address. This suggests that integrating more domain-specific rules and knowledge into the prompt, potentially with the Expert's involvement in prompt engineering, could further enhance the model's performance in such specialized tasks.

Overall, the experiment shows that the LLM-supported parsing tool offers substantial benefits in both time efficiency and accuracy, particularly when used by individuals with relevant expertise.

Table 10.2: Parsing results for example #1 - LLM.

LLM	Expert	Doctor	Novice 1	Novice 2
Stage IIIb OR Stage IIIc OR Stage IV	Stage IIIb OR Stage IIIc OR Stage IV	Stage IIIb OR Stage IIIc OR Stage IV	Stage IIIb OR Stage IIIc OR Stage IV	Stage IIIb OR Stage IIIc OR Stage IV
HER2+	HER2+	HER2+	HER2+	HER2+
Progressive	Progressive	Progressive	Progressive	Progressive

Table 10.3: Parsing results for example #2 - LLM.

LLM	Expert	Doctor	Novice 1	Novice 2
Stage II OR Stage IIIa OR Stage IIIb OR Stage IIIc	Stage II OR Stage IIIa OR Stage IIIb OR Stage IIIc OR Inflammatory	Stage II OR Stage IIIa OR Stage IIIb OR Stage IIIc OR IBC	Stage II OR Stage IIIa OR Stage IIIb OR Stage IIIc	Stage II OR Stage IIa OR stage IIB OR Stage IIIa OR Stage IIIb OR Stage IIIc
Stage IIIb OR Stage IIIc	-	-	-	-
IBC	-	-	IBC	IBC
HER2+	HER2+	HER2+	HER2+	HER2+
NOT Stage IV	NOT Stage IV	NOT Stage IV	NOT Stage IV	NOT Stage IV
NOT Multicentric	NOT Multicentric	NOT Multicentric	NOT Multicentric	NOT Multicentric
NOT Bilateral	NOT Bilateral	NOT Bilateral	NOT Bilateral	NOT Bilateral
-	-	-	-	NOT DCIS OR NOT LCIS
-	-	-	-	NOT HER2+

Table 10.4: Parsing results for example #3 - LLM.

LLM	Expert	Doctor	Novice 1	Novice 2
Stage IV OR Stage IIIb OR Stage IIIc	Stage IV OR Stage IIIb OR Stage IIIc	Stage IV OR Stage IIIb OR Stage IIIc	Stage IV OR Stage IIIb OR Stage IIIc	Stage IV OR Stage IIIb OR Stage IIIc
ER- AND PR- AND HER2-	ER- AND PR- AND HER2-	ER- AND PR- AND HER2-	ER- AND PR- AND HER2-	ER- AND PR- AND HER2-
PD-L1+	PD-L1+	PD-L1+	PD-L1+	PD-L1+
-	-	-	-	HER2-
-	-	-	-	ER - AND PR -
-	-	-	-	Stage IV
-	-	-	-	NOT Stage IIIb OR NOT Stage IIIc

Table 10.5: Parsing results for example #4 - LLM.

LLM	Expert	Doctor	Novice 1	Novice 2
Invasive	Invasive	Invasive	Invasive	Invasive
Adenocarcinoma	Adenocarcinoma	Adenocarcinoma	Adenocarcinoma	Adenocarcinoma
HER2+	HER2+	HER2+	HER2+	HER2+
NOT Stage IV	NOT Stage IV	NOT Stage IV	NOT Stage IV	NOT Stage IV
NOT Bilateral	NOT Bilateral	NOT (Bilateral AND Invasive)	NOT (Bilateral AND Invasive)	NOT Bilateral
-	Stage I OR Stage II OR Stage III	-	-	-
-	NOT Stage IIIc	-	-	-
-	NOT Recurrent	-	-	-

Table 10.6: Parsing results for example #5 - LLM.

LLM	**Expert**	**Doctor**	**Novice 1**	**Novice 2**
Adenocarcinoma	Adenocarcinoma	Adenocarcinoma	Adenocarcinoma	Adenocarcinoma
Stage IV	Stage IV	Stage IV	Stage IV	Stage IV
ER- AND PR- AND HER2-	ER- AND PR- AND HER2-	ER- AND PR- AND HER2-	ER- AND PR- AND HER2-	ER- AND PR- AND HER2-
PD-L1+	PD-L1+	PD-L1+	PD-L1+	PD-L1+
-	-	-	Metastatic	Stage IV
-	-	-	-	ER+ AND PGR+
-	-	-	-	Stage IV
-	-	-	-	NOT Stage IV

10.6 Limitations and future directions

While the development and application of the parsing tool represent significant advancement in processing the eligibility criteria, we identified several limitations and areas for improvement during this study.

Firstly, the tool currently extracts only three types of conditions: biomarkers, subtypes, and stages, while ignoring other crucial criteria related to treatments, metastases, or laboratory tests, among others. It also focuses on the patient's current health status, neglecting the medical history, which is frequently mentioned in eligibility criteria and could disqualify many patients. Additionally, the tool is limited to breast cancer clinical trials, even though there are other therapeutic areas covering numerous trials that have different sets of eligibility criteria and vocabularies. The three condition types selected for this tool are specific to oncology trials and do not apply to other therapeutic areas. However, this was a intentional choice to focus on a small subset of criteria to explore the tool's potential. Breast cancer trials constitute a significant portion of Roche's trials and these three condition types were selected by domain experts as the most critical for this Proof of Concept. If the tool is accepted, it is designed to be easily extendable with more condition types or adaptable to other therapeutic areas.

Moreover, the current version of the tool does not consider *"if-then"* relationships in logical expressions, which are common in eligibility criteria. These relationships often apply criteria to specific groups of patients, significantly changing the meaning. For example, a criterion like *"For patients with TNBC, at least one line of prior systemic therapy is required"* could be expressed as *"IF (PR- AND ER- AND HER2-) THEN (chemotherapy OR immunotherapy)"*, and it is important to include this logic in future versions of the tool.

Domain experts also highlighted that there are often indirect duplications in inclusion and exclusion criteria, that should be correctly handled by the tool. For instance, requiring early-stage cancer in the inclusion criteria while excluding metastatic or advanced cancer in the exclusion criteria essentially express the same information. Hence, it may be sufficient to include only the inclusion criterion in the extraction table, as the exclusions are redundant and not needed for matching with patient records.

Additionally, experts identified other redundancies in the model's output. Some extracted conditions overlap with others, as they define subgroups of more general conditions. For instance, *"IDC"* (Invasive Ductal Carcinoma) and *"ILC"* (Invasive Lobular Carcinoma) are both forms of invasive cancer, so adding *"Invasive"* as a separate condition is redundant. Similarly, *"Adenocarcinoma"* is a broader category that includes IDC and ILC, so it does not need to be listed separately. A post-extraction step to remove such redundant conditions could be added to the final parsing pipeline.

There are also some technical concepts that need improvement. For instance, the current UI accepts any text in the *"Condition"* column, allowing users to enter terms that are not listed in the allowed terms or to make typos. This was a conscious

decision in the PoC phase to avoid restricting users from adding important conditions that may have been skipped in the model's guidelines. However, implementing a verification feature to prevent such issues would be beneficial in future versions of the tool. Additionally, the current UI highlights phrases in the text and associates them with only one tag from the table, even though a single phrase might relate to multiple conditions. This situation is not currently handled by the tool currently but should be improved so that relevant phrases can be associated with multiple tags. As this is a PoC, there are also MLOps considerations, such as CI/CD, dockerization, and integration with Roche;s cloud infrastructure and recommended data storage, that need to be addressed before deployment to production.

The evaluation experiment also requires adjustments to provide more reliable results. The current results serve as initial check of the tool's potential, but future steps should include refining the tool and conducting a more rigours experiment under controlled conditions with more resources. Performing the experiment in laboratory conditions would help eliminate distractions that could affect time measurements. Additionally, involving more experts and assessing the alignment between them is crucial, as inconsistencies in expert responses were observed even in this experiment. To obtain more convincing results, the experiment should include a larger and more diverse set of trials and examples. The current accuracy evaluation focuses only on LLM-assisted parsing without considering manual parsing examples. Establishing a ground truth for manual examples would also be beneficial, as participants in AI-assisted task might be influenced by the model's output. Finally, to fully evaluate the tool, it is essential to assess how the final output affects the population of patients eligible for the trial, determining whether the model is too restrictive or too lenient compared to manually parsed criteria. A control group of experts performing manual parsing versus those using AI-assisted parsing on the same set of trials would provide valuable insights into the tool's effectiveness.

10.7 Conclusions

In conclusion, the development of the parsing tool is a significant step forward in automating the extraction of eligibility criteria for clinical trials, particularly within the domain of breast cancer. It provides a structured form of inclusion and exclusion criteria, with maintaining logical relations between them. The initial version of the tool focusing on key conditions—biomarkers, subtypes, and stages—has demonstrated its potential to automate the trial eligibility determination process. Despite some limitations, the tool has shown promising results in both accuracy and time efficiency. The evaluation experiment demonstrated the tool's effectiveness and proved that it is beneficial in the criteria parsing process. The tool not only reduces the time required for parsing but also maintains high accuracy, making it a valuable asset for future clinical trial screening processes. With continued development and broader application, this tool has the potential to significantly improve patient-trial matching and trial efficiency across various therapeutic areas.

Chapter 11

Conclusion

This thesis explores the potential of neural language models, particularly generative large language models (LLMs), in parsing clinical trial eligibility criteria—a complex task within the biomedical natural language processing (NLP) that is crucial for the pharmaceutical industry. Given the challenges that clinical trial face in patient recruitment, the findings from this research offer promising solutions that could streamline the process and contribute to more effective delivery of therapies to the market.

The primary objective of this thesis was to investigate the application of language models in automating the parsing of clinical trial eligibility criteria. The research successfully demonstrated that this goal is achievable with the use of LLMs, such as the GPT-4 model evaluated in the named entity recognition task (Chapter 7) and the GPT-4o model utilized in the parsing tool (Chapter 10). The experiments revealed that these models are extremely powerful, especially in the context of eligibility criteria, which are challenging to process with simpler models. We also discovered that plain named entity recognition is insufficient for patient-trial matching applications and a more complex parsing output is required, which further justifies the use of generative models over BERT-based models.

Key achievements of this thesis include the development of a prototype tool that accurately transforms eligibility criteria into a machine-readable format required by a trial screening algorithm, the evaluation of various NLP models in biomedical applications, and the practical insights gained from implementing the tool within a real-world pharmaceutical context. Chapter by chapter, this thesis explores both the scientific and implementation aspects of the criteria parsing problem. The scientific section lays the foundation for practical applications, starting with an evaluation of the impact of eligibility criteria and other trial features on trial performance. Subsequent chapters explore the NLP resources for biomedical domain and provide a deep review of previous studies on eligibility criteria parsing using ML and NLP. The thesis also describes two experiments—one evaluating various text embedding techniques in the biomedical semantic textual similarity task, and the second evaluating different language models, including GPT-4, in named entity recognition from eligibility criteria. Building on the promising results from these studies, the thesis further focuses on the implementation at Roche. It describes a prompt engineering hackathon organized within the company, states the conceptual and technical re-

quirements for the parsing tool from Roche's perspective, and culminates in the successful design and implementation of a Proof of Concept tool. Chapter 10 provides the technical details of the tool—both the backend and frontend—and evaluates the parser's performance. It demonstrates the feasibility and potential value of utilizing LLMs in a clinical trial management system.

The research hypothesis stated that neural language models could enhance the parsing of clinical trial eligibility criteria compared to the traditional manual methods. The findings support this hypothesis, as the experiments showed that LLMs can correctly interpret and structure the criteria, even in low-resource settings. We observed a significant gain in efficiency when using LLM support for criteria parsing compared to manual work. The tool's ability to handle the complex semantics of eligibility criteria and provide correct logical structures emphasizes its potential to revolutionize the patient recruitment process.

This thesis makes several original contributions to the field of biomedical NLP. To our knowledge this thesis presents the most extensive and up-to-date scoping review of NLP and ML solutions for parsing eligibility criteria, which could serve as a valuable resource for future research. The evaluation of sentence embeddings and the comparison of GPT-4 model with BERT-based models providee additional insights into the performance of these models in biomedical named entity recognition. Additionally, this thesis presents a combination of advanced prompt engineering techniques, such as chain-of-though and few-shot prompting, along with negative examples, which have shown very promising results and significant improvement in LLM performance with limited annotated data. The proposed parsing tool presents a practical application of these findings and offers a prototype of a solution that might improve patient screening for clinical trials.

Despite the promising results, we acknowledge the limitations of this thesis, particularly concerning the parsing tool developed. First, this thesis focuses only on one part of the patient-trial matching process—eligibility criteria parsing—and evaluates it as an isolated task. In the proposal of a patient screening application, patient data will be provided by the survey filled out by the patient. We did not verify whether all the conditions we focus on are conditions that patients are knowledgeable about and can answer accurately in the survey. Another limitation is that the evaluation of the tool does not consider patient data and the tool's impact on patient selection. Furthermore, the tool focuses on breast cancer trials, and we did not assess whether the findings from the thesis apply to other therapeutic areas. Additionally, while the hypothesis was supported by the evaluation experiment, the experiment itself requires further refinement to provide more reliable results. The evaluation of the GPT model in named entity recognition also presented challenges, as it was approached in a traditional way better suited for encoder models than for generative models.

Future research should focus on developing reliable methods for evaluating generative models in information retrieval tasks. The examples from the PoC experiment were evaluated manually, which is not scalable for larger samples. We also encounter the lack of a proper evaluation method when assessing NER results from the LLM. Another research direction involves creating a benchmark dataset for parsing eligi-

bility criteria, since the absence of such datasets was identified as a significant issue during the scoping review of previous studies. Currently, there are few datasets available for comparing experiments in this area. We also recognize that there are limited studies that use the cutting-edge NLP techniques, such as LLMs, in eligibility criteria parsing. Many recent studies still utilize outdated techniques for text processing. There is a pressing need to explore recent NLP advancements in the clinical development area. Additionally, as LLMs become increasingly popular and represent the state-of-the-art in many applications, there is a need to focus on the industrial and technical aspects of implementing LLM-based solutions. A branch of MLOps—LLMOps—has begun to develop, but it is still under-studied. Given the rapid development of LLMs, this should also be a priority research subject. Moreover, the PoC tool developed in this research serves as an initial version of a parsing system that requires further improvement—specifically extending to other criteria types and therapeutic areas. More domain experts should be involved in the development of this tool to provide more insights, support prompt engineering with additional guidelines, and conduct deeper evaluations. The tool also requires further MLOps development.

The findings of this thesis highlight the transformative potential of LLMs in the clinical development area. By achieving significant performance in eligibility criteria parsing, the research offers a promising solution to one of the most challenging aspects of clinical trials: patient recruitment. While there are limitations to address, the successful implementation of the Proof of Concept demonstrates that combining AI, NLP, and domain expertise can lead to substantial advancements in the biomedical field.

Bibliography

1. H.-H. Wong, A. Jessup, A. Sertkaya, A. Birkenbach, A. Berlind, and J. Eyraud, "Examination of clinical trial costs and barriers for drug development," *Office of the Assistant Secretary for Planning and Evaluation, US Department of Health & Human Services*, pp. 1–92, 2014. (cited on page 6)

2. O. J. Wouters, M. McKee, and J. Luyten, "Estimated research and development investment needed to bring a new medicine to market, 2009-2018," *Jama*, vol. 323, no. 9, pp. 844–853, 2020. (cited on page 6)

3. Y. Ni, S. Kennebeck, J. W. Dexheimer, C. M. McAneney, H. Tang, T. Lingren, Q. Li, H. Zhai, and I. Solti, "Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department," *Journal of the American Medical Informatics Association*, vol. 22, no. 1, pp. 166–178, 2015. (cited on page 6)

4. L. Penberthy, R. Brown, F. Puma, and B. Dahman, "Automated matching software for clinical trials eligibility: measuring efficiency and flexibility," *Contemporary clinical trials*, vol. 31, no. 3, pp. 207–217, 2010. (cited on page 7)

5. T. E. Brunekreef, H. G. Otten, S. C. van den Bosch, I. E. Hoefer, J. M. van Laar, M. Limper, and S. Haitjema, "Text mining of electronic health records can accurately identify and characterize patients with systemic lupus erythematosus," *ACR open rheumatology*, vol. 3, no. 2, pp. 65–71, 2021. (cited on page 7)

6. L. T. Penberthy, B. A. Dahman, V. I. Petkov, and J. P. DeShazo, "Effort required in eligibility screening for clinical trials," *Journal of Oncology Practice*, vol. 8, no. 6, pp. 365–370, 2012. (cited on page 8)

7. K. D. Stensland, R. B. McBride, A. Latif, J. Wisnivesky, R. Hendricks, N. Roper, P. Boffetta, S. J. Hall, W. K. Oh, and M. D. Galsky, "Adult cancer clinical trials that fail to complete: an epidemic?," *Journal of the National Cancer Institute*, vol. 106, no. 9, p. dju229, 2014. (cited on page 8)

8. B. Carlisle, J. Kimmelman, T. Ramsay, and N. MacKinnon, "Unsuccessful trial accrual and human subjects protections: an empirical analysis of recently closed trials," *Clinical trials*, vol. 12, no. 1, pp. 77–83, 2015. (cited on page 8)

9. G. M. Ibrahim, C. Chung, and M. Bernstein, "Competing for patients: an ethical framework for recruiting patients with brain tumors into clinical trials," 2011. (cited on page 8)

10. C. E. Lawrence, V. N. M. Bruce, L. D. Salberg, T. Edwards, C. Morales, M. Palm, and G. R. Bernard, "Quantitative assessment of the impact of standard agreement templates on multisite clinical trial start up time," *Journal of Clinical and Translational Science*, vol. 7, no. 1, p. e204, 2023. (cited on page 8)

11. E. Fink, P. K. Kokku, S. Nikiforou, L. O. Hall, D. B. Goldgof, and J. P. Krischer, "Selection of patients for clinical trials: an interactive web-based system," *Artificial intelligence in medicine*, vol. 31, no. 3, pp. 241–254, 2004. (cited on page 8)

12. C. P. Somkin, L. Ackerson, G. Husson, V. Gomez, T. Kolevska, D. Goldstein, and L. Fehrenbacher, "Effect of medical oncologists' attitudes on accrual to clinical trials in a community setting," *Journal of oncology practice*, vol. 9, no. 6, pp. e275–e283, 2013. (cited on page 8)

13. T. C. Haddad, J. Helgeson, K. Pomerleau, M. Makey, P. Lombardo, S. Coverdill, A. Urman, M. Rammage, M. P. Goetz, and N. LaRusso, "Impact of a cognitive computing clinical trial matching system in an ambulatory oncology practice.," 2018. (cited on page 8)

14. C. Bieganek, C. Aliferis, and S. Ma, "Prediction of clinical trial enrollment rates," *Plos one*, vol. 17, no. 2, p. e0263193, 2022. (cited on page 8)

15. D. Kelly, A. Spreafico, and L. L. Siu, "Increasing operational and scientific efficiency in clinical trials," *British Journal of Cancer*, vol. 123, no. 8, pp. 1207–1208, 2020. (cited on pages 13 and 15)

16. M. Rosenblatt, "The large pharmaceutical company perspective," *New England Journal of Medicine*, vol. 376, no. 1, pp. 52–60, 2017. (cited on page 13)

17. M. Allison, "Reinventing clinical trials," *Nature biotechnology*, vol. 30, no. 1, pp. 41–49, 2012. (cited on page 13)

18. K. Wu, E. Wu, M. DAndrea, N. Chitale, M. Lim, M. Dabrowski, K. Kantor, H. Rangi, R. Liu, M. Garmhausen, *et al.*, "Machine learning prediction of clinical trial operational efficiency," *The AAPS Journal*, vol. 24, no. 3, p. 57, 2022. (cited on page 13)

19. B. Craven, C. Balioussis, S. Hitzig, C. Moore, M. Verrier, L. Giangregorio, and M. Popovic, "Use of screening to recruitment ratios as a tool for planning and implementing spinal cord injury rehabilitation research," *Spinal Cord*, vol. 52, no. 10, pp. 764–768, 2014. (cited on pages 14 and 15)

20. S. Blanton, D. M. Morris, M. G. Prettyman, K. McCulloch, S. Redmond, K. E. Light, and S. L. Wolf, "Lessons learned in participant recruitment and retention: the excite trial," *Physical therapy*, vol. 86, no. 11, pp. 1520–1533, 2006. (cited on pages 14 and 15)

21. T. M. Harris-Brown and D. L. Paterson, "Reporting of pre-enrolment screening with randomized clinical trials: A small item that could impact a big difference," *Perspectives in Clinical Research*, vol. 6, no. 3, pp. 139–143, 2015. (cited on pages 14 and 15)

22. S. J. Keith *et al.*, "Evaluating characteristics of patient selection and dropout rates," *Journal of Clinical Psychiatry*, vol. 62, pp. 11–16, 2001. (cited on pages 14 and 15)

23. D. B. Fogel, "Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review," *Contemporary clinical trials communications*, vol. 11, pp. 156–164, 2018. (cited on pages 14 and 15)

24. H.-G. Eichler and F. Sweeney, "The evolution of clinical trials: Can we address the challenges of the future?," *Clinical trials*, vol. 15, no. 1_suppl, pp. 27–32, 2018. (cited on page 15)

25. B. Smuck, P. Bettello, K. Berghout, T. Hanna, B. Kowaleski, L. Phippard, D. Au, and K. Friel, "Ontario protocol assessment level: clinical trial complexity rating tool for workload planning in oncology clinical trials," *Journal of Oncology Practice*, vol. 7, no. 2, pp. 80–84, 2011. (cited on page 15)

26. K. M. Cunanan, M. Gonen, R. Shen, D. M. Hyman, G. J. Riely, C. B. Begg, and A. Iasonos, "Basket trials in oncology: a trade-off between complexity and efficiency," *Journal of Clinical Oncology*, vol. 35, no. 3, p. 271, 2017. (cited on page 15)

27. L. Malik and D. Lu, "Increasing complexity in oncology phase i clinical trials," *Investigational new drugs*, vol. 37, pp. 519–523, 2019. (cited on page 15)

28. G. Yuan, L. Wang, J. Li, H. Feng, J. Ji, W. Gu, and B. Shan, "Complexity in clinical trials: blind spots, misleading criteria, winners and losers," *Clinical Cancer Drugs*, vol. 7, no. 1, pp. 3–15, 2020. (cited on page 15)

29. K. Getz, "Improving protocol design feasibility to drive drug development economics and performance," *International journal of environmental research and public health*, vol. 11, no. 5, pp. 5069–5080, 2014. (cited on page 15)

30. R. B. Giffin, Y. Lebovitz, R. A. English, *et al.*, "Transforming clinical research in the united states: challenges and opportunities: workshop summary," 2010. (cited on page 15)

31. A. M. McDonald, R. C. Knight, M. K. Campbell, V. A. Entwistle, A. M. Grant, J. A. Cook, D. R. Elbourne, D. Francis, J. Garcia, I. Roberts, *et al.*, "What influences recruitment to randomised controlled trials? a review of trials funded by two uk funding agencies," *Trials*, vol. 7, no. 1, pp. 1–8, 2006. (cited on page 15)

32. K. A. Getz, "Characterizing the real cost of site regulatory compliance," *Applied Clinical Trials*, vol. 24, no. 6/7, p. 18, 2015. (cited on page 15)

33. S. R. Thadani, C. Weng, J. T. Bigger, J. F. Ennever, and D. Wajngurt, "Electronic screening improves efficiency in clinical trial recruitment," *Journal of the American Medical Informatics Association*, vol. 16, no. 6, pp. 869–873, 2009. (cited on page 15)

34. G. Frank, "Current challenges in clinical trial patient recruitment and enrollment," *SoCRA Source*, vol. 2, no. February, pp. 30–38, 2004. (cited on page 15)

35. R. A. Kadam, S. U. Borde, S. A. Madas, S. S. Salvi, and S. S. Limaye, "Challenges in recruitment and retention of clinical trial subjects," *Perspectives in clinical research*, vol. 7, no. 3, pp. 137–143, 2016. (cited on page 15)

36. D. M. Dilts, A. B. Sandler, M. Baker, S. K. Cheng, S. L. George, K. S. Karas, S. McGuire, G. S. Menon, J. Reusch, D. Sawyer, *et al.*, "Processes to activate phase iii clinical trials in a cooperative oncology group: the case of cancer and leukemia group b," *Journal of Clinical Oncology*, vol. 24, no. 28, pp. 4553–4557, 2006. (cited on page 15)

37. K. A. Getz, J. Wenger, R. A. Campo, E. S. Seguine, and K. I. Kaitin, "Assessing the impact of protocol design changes on clinical trial performance," *American journal of therapeutics*, vol. 15, no. 5, pp. 450–457, 2008. (cited on page 15)

38. S. Ross, A. Grant, C. Counsell, W. Gillespie, I. Russell, and R. Prescott, "Barriers to participation in randomised controlled trials: a systematic review," *Journal of clinical epidemiology*, vol. 52, no. 12, pp. 1143–1156, 1999. (cited on page 15)

39. K. Boericke and B. Gwinn, "Planned to perfection," *Int Clin Trials*, vol. 17, no. 8, pp. 26–30, 2010. (cited on page 15)

40. J. W. Andersen, R. Fass, and C. van der Horst, "Factors associated with early study discontinuation in aactg studies, dacs 200," *Contemporary clinical trials*, vol. 28, no. 5, pp. 583–592, 2007. (cited on page 15)

41. T. Yitong, M. Salkola, M. Ahmed, K. Anuj, and A. Freddy, "Information extraction of clinical trial eligibility criteria," *arXiv preprint arXiv:2006.07296*, 2020. (cited on page 15)

42. H. Liu, Y. Chi, A. Butler, Y. Sun, and C. Weng, "A knowledge base of clinical trial eligibility criteria," *Journal of biomedical informatics*, vol. 117, p. 103771, 2021. (cited on page 15)

43. C. H. Wong, K. W. Siah, and A. W. Lo, "Estimation of clinical trial success rates and related parameters," *Biostatistics*, vol. 20, no. 2, pp. 273–286, 2019. (cited on page 16)

44. B. Munos, J. Niederreiter, and M. Riccaboni, "Improving the prediction of clinical success using machine learning," *medRxiv*, pp. 2021–02, 2021. (cited on page 16)

45. F. Feijoo, M. Palopoli, J. Bernstein, S. Siddiqui, and T. E. Albright, "Key indicators of phase transition for clinical trials through machine learning," *Drug discovery today*, vol. 25, no. 2, pp. 414–421, 2020. (cited on page 16)

46. E. H. Houssein, R. E. Mohamed, and A. A. Ali, "Machine learning techniques for biomedical natural language processing: a comprehensive review," *IEEE Access*, vol. 9, pp. 140628–140653, 2021. (cited on page 25)

47. E. W. Sayers, E. E. Bolton, J. R. Brister, K. Canese, J. Chan, D. C. Comeau, R. Connor, K. Funk, C. Kelly, S. Kim, *et al.*, "Database resources of the national center for biotechnology information," *Nucleic acids research*, vol. 50, no. D1, p. D20, 2022. (cited on page 25)

48. N. L. of Medicine, "About - pubmed," 2023. Accessed on April 30, 2024. (cited on page 25)

49. O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004. (cited on page 26)

50. N. L. of Medicine, "Unified medical language system (umls)," 2021. Accessed on April 30, 2024. (cited on page 26)

51. D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, "Drugbank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic acids research*, vol. 34, no. suppl_1, pp. D668–D672, 2006. (cited on page 26)

52. C. Knox, M. Wilson, C. M. Klinger, M. Franklin, E. Oler, A. Wilson, A. Pon, J. Cox, N. E. Chin, S. A. Strawbridge, *et al.*, "Drugbank 6.0: the drugbank knowledgebase for 2024," *Nucleic Acids Research*, vol. 52, no. D1, pp. D1265–D1275, 2024. (cited on page 27)

53. K. Liu, W. R. Hogan, and R. S. Crowley, "Natural language processing methods and systems for biomedical ontology learning," *Journal of biomedical informatics*, vol. 44, no. 1, pp. 163–179, 2011. (cited on page 27)

54. C. Gaudet-Blavignac, V. Foufi, M. Bjelogrlic, and C. Lovis, "Use of the systematized nomenclature of medicine clinical terms (snomed ct) for processing free text in health care: systematic scoping review," *Journal of medical Internet research*, vol. 23, no. 1, p. e24594, 2021. (cited on page 27)

55. S. International, "Snomed ct. the case for investment," 2021. Accessed on May 7, 2024. (cited on page 27)

56. N. L. of Medicine, "Introduction to mesh," 2023. Accessed on April 30, 2024. (cited on page 28)

57. J. E. Harrison, S. Weber, R. Jakob, and C. G. Chute, "Icd-11: an international classification of diseases for the twenty-first century," *BMC medical informatics and decision making*, vol. 21, pp. 1–10, 2021. (cited on page 29)

58. WHO, "Icd-11 2022 release," 2022. Accessed on May 7, 2024. (cited on page 30)

59. W. H. Organisation, "Icd-11 reference guide," 2023. Accessed on May 7, 2024. (cited on page 30)

60. M. Stram, T. Gigliotti, D. Hartman, A. Pitkus, S. M. Huff, M. Riben, W. H. Henricks, N. Farahani, and L. Pantanowitz, "Logical observation identifiers names and codes for laboratorians: potential solutions and challenges for interoperability," *Archives of pathology & laboratory medicine*, vol. 144, no. 2, pp. 229–239, 2020. (cited on page 31)

61. C. J. McDonald, S. M. Huff, J. G. Suico, G. Hill, D. Leavelle, R. Aller, A. Forrey, K. Mercer, G. DeMoor, J. Hook, *et al.*, "Loinc, a universal standard for identifying laboratory observations: a 5-year update," *Clinical chemistry*, vol. 49, no. 4, pp. 624–633, 2003. (cited on page 31)

62. W. R. Hogan and M. M. Wagner, "Information technology standards in biosurveillance," *Handbook of Biosurveillance*, pp. 439–452, 2006. (cited on page 31)

63. N. L. of Medicine, "Rxnorm overview," 2024. Accessed on May 7, 2024. (cited on page 31)

64. S. H. Brown, S. T. Rosenbloom, B. A. Bauer, D. Wahner-Roedler, D. A. Froehling, K. R. Bailey, M. J. Lincoln, D. Montella, E. M. Fielstein, and P. L. Elkin, "Direct comparison of medcin® and snomed ct® for representation of a general medical evaluation template," in *AMIA Annual Symposium Proceedings*, vol. 2007, p. 75, American Medical Informatics Association, 2007. (cited on page 32)

65. M. Systems, "Medcin graph database technology." Accessed on May 7, 2024. (cited on page 32)

66. E. G. Brown, L. Wood, and S. Wood, "The medical dictionary for regulatory activities (meddra)," *Drug safety*, vol. 20, no. 2, pp. 109–117, 1999. (cited on page 33)

67. J. Harrison and P. Mozzicato, "Meddra®: the tale of a terminology: side effects of drugs essay," in *Side effects of drugs annual*, vol. 31, pp. xxxiii–xli, Elsevier, 2009. (cited on page 33)

68. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. (cited on page 33)

69. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. (cited on pages 34 and 64)

70. Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE international conference on computer vision*, pp. 19–27, 2015. (cited on page 34)

71. Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021. (cited on pages 34 and 35)

72. A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016. (cited on page 34)

73. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020. (cited on page 35)

74. E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," *arXiv preprint arXiv:1904.03323*, 2019. (cited on page 35)

75. Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets," *arXiv preprint arXiv:1906.05474*, 2019. (cited on page 35)

76. Z. Yuan, Z. Zhao, H. Sun, J. Li, F. Wang, and S. Yu, "Coder: Knowledge-infused cross-lingual medical term embedding for term normalization," *Journal of biomedical informatics*, vol. 126, p. 103983, 2022. (cited on page 35)

77. I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," *arXiv preprint arXiv:1903.10676*, 2019. (cited on page 35)

78. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020. (cited on page 35)

79. J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023. (cited on pages 35 and 79)

80. P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *Advances in neural information processing systems*, vol. 30, 2017. (cited on page 35)

81. M. Liebrenz, R. Schleifer, A. Buadze, D. Bhugra, and A. Smith, "Generating scholarly content with chatgpt: ethical challenges for medical publishing," *The lancet digital health*, vol. 5, no. 3, pp. e105–e106, 2023. (cited on page 36)

82. P. P. Ray, "Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet of Things and Cyber-Physical Systems*, 2023. (cited on page 36)

83. X. Amatriain, "Prompt design and engineering: Introduction and advanced methods," *arXiv preprint arXiv:2401.14423*, 2024. (cited on page 36)

84. J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt, "A prompt pattern catalog to enhance prompt engineering with chatgpt," *arXiv preprint arXiv:2302.11382*, 2023. (cited on page 36)

85. OpenAI, "Models - openai api." Accessed on May 7, 2024. (cited on pages 36 and 125)

86. B. Idnay, C. Dreisbach, C. Weng, and R. Schnall, "A systematic review on natural language processing systems for eligibility prescreening in clinical research," *Journal of the American Medical Informatics Association*, vol. 29, no. 1, pp. 197–206, 2022. (cited on page 40)

87. S. Askin, D. Burkhalter, G. Calado, and S. El Dakrouni, "Artificial intelligence applied to clinical trials: opportunities and challenges," *Health and Technology*, vol. 13, no. 2, pp. 203–213, 2023. (cited on page 40)

88. R. Bhatnagar, S. Sardar, M. Beheshti, and J. T. Podichetty, "How can natural language processing help model informed drug development?: a review," *JAMIA open*, vol. 5, no. 2, p. ooac043, 2022. (cited on page 40)

89. J. Kim and Y. Quintana, "Review of the performance metrics for natural language systems for clinical trials matching," in *MEDINFO 2021: One World, One Health–Global Partnership for Digital Innovation*, pp. 641–644, IOS Press, 2022. (cited on page 40)

90. Q. Su, G. Cheng, and J. Huang, "A review of research on eligibility criteria for clinical trials," *Clinical and Experimental Medicine*, pp. 1–13, 2023. (cited on page 40)

91. H. Arksey and L. O'Malley, "Scoping studies: towards a methodological framework," *International journal of social research methodology*, vol. 8, no. 1, pp. 19–32, 2005. (cited on page 42)

92. K. S. Khan, R. Kunz, J. Kleijnen, and G. Antes, "Five steps to conducting a systematic review," *Journal of the royal society of medicine*, vol. 96, no. 3, pp. 118–121, 2003. (cited on page 42)

93. J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977. (cited on page 44)

94. J. Patricoski, K. Kreimeyer, A. Balan, K. Hardart, J. Tao, V. Anagnostou, and T. B. and, "An evaluation of pretrained BERT models for comparing semantic similarity across unstructured clinical trial texts," in *Studies in Health Technology and Informatics*, IOS Press, jan 2022. (cited on pages 47 and 186)

95. T. Haddad, J. M. Helgeson, K. E. Pomerleau, A. M. Preininger, M. C. Roebuck, I. Dankwa-Mullan, G. P. Jackson, and M. P. Goetz, "Accuracy of an artificial intelligence system for cancer clinical trial eligibility screening: Retrospective pilot study," *JMIR Medical Informatics*, vol. 9, p. e27767, mar 2021. (cited on pages 47 and 186)

96. F. Kury, A. Butler, C. Yuan, L. heng Fu, Y. Sun, H. Liu, I. Sim, S. Carini, and C. Weng, "Chia, a large annotated corpus of clinical trial eligibility criteria," *Scientific Data*, vol. 7, aug 2020. (cited on pages 47 and 186)

97. N. J. Dobbins, T. Mullen, Özlem Uzuner, and M. Yetisgen, "The leaf clinical trials corpus: a new resource for query generation from clinical trial eligibility criteria," *Scientific Data*, vol. 9, aug 2022. (cited on pages 47 and 187)

98. X. Liu, G. L. Hersch, I. Khalil, and M. Devarakonda, "Clinical trial information extraction with BERT," in *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, IEEE, aug 2021. (cited on pages 47 and 186)

99. Y. Gu, S. Zhang, N. Usuyama, Y. Woldesenbet, C. Wong, P. Sanapathi, M. Wei, N. Valluri, E. Strandberg, T. Naumann, and H. Poon, "Distilling large language models for biomedical knowledge extraction: A case study on adverse drug events," 2023. (cited on page 60)

100. K. Kantor and M. Morzy, "Biomedical semantic textual similarity: Evaluation of sentence representations enhanced with principal component reduction and word frequency weighting: Student paper submission," in *International Conference on Artificial Intelligence in Medicine*, pp. 393–403, Springer, 2022. (cited on page 61)

101. K. Blagec and *et al.*, "Neural sentence embedding models for semantic similarity estimation in the biomedical domain," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–10, 2019. (cited on page 62)

102. P. Bojanowski and *et al.*, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016. (cited on page 62)

103. M. Pagliardini and *et al.*, "Unsupervised learning of sentence embeddings using compositional n-gram features," *arXiv preprint arXiv:1703.02507*, 2017. (cited on page 62)

104. R. Kiros and *et al.*, "Skip-thought vectors," 2015. (cited on page 62)

105. Q. V. Le and *et al.*, "Distributed representations of sentences and documents," 2014. (cited on page 62)

106. T. Ranashinghe and *et al.*, "Enhancing unsupervised sentence similarity methods with deep contextualised word representations," RANLP, 2019. (cited on page 62)

107. Q. Jin and *et al.*, "Probing biomedical embeddings from language models," *arXiv preprint arXiv:1904.02181*, 2019. (cited on page 62)

108. A. Akbik and *et al.*, "Flair: An easy-to-use framework for state-of-the-art nlp," in *NAACL-HLT (Demonstrations)*, 2019. (cited on page 62)

109. S. Arora and *et al.*, "A simple but tough-to-beat baseline for sentence embeddings," in *International Conference on Learning Representations*, 2017. (cited on pages 62 and 65)

110. A. Koroleva and *et al.*, "Measuring semantic similarity of clinical trial outcomes using deep pre-trained language representations," *J. of Biomedical Informatics: X*, vol. 4, 2019. (cited on pages 62 and 63)

111. G. Soğancıoğlu and *et al.*, "Biosses: a semantic sentence similarity estimation system for the biomedical domain," *Bioinformatics*, vol. 33, no. 14, pp. i49–i58, 2017. (cited on page 63)

112. J. Pennington and *et al.*, "Glove: Global vectors for word representation," in *Proc of the 2014 EMNLP*, pp. 1532–1543, 2014. (cited on page 64)

113. N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019. (cited on page 65)

114. V. Sanh and *et al.*, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019. (cited on page 65)

115. W. Wang and *et al.*, "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," 2020. (cited on page 65)

116. K. Song and *et al.*, "Mpnet: Masked and permuted pre-training for language understanding," 2020. (cited on page 65)

117. J. Mu, S. Bhat, and P. Viswanath, "All-but-the-top: Simple and effective postprocessing for word representations," *arXiv preprint arXiv:1702.01417*, 2017. (cited on page 66)

118. V. Raunak and *et al.*, "Effective dimensionality reduction for word embeddings," in *Proc. of the RepL4NLP*, pp. 235–243, 2019. (cited on page 66)

119. M. T. Ribeiro and *et al.*, "Beyond accuracy: Behavioral testing of nlp models with checklist," *arXiv preprint arXiv:2005.04118*, 2020. (cited on page 67)

120. I. Kola and J. Landis, "Can the pharmaceutical industry reduce attrition rates?," *Nature reviews Drug discovery*, vol. 3, no. 8, pp. 711–716, 2004. (cited on page 73)

121. A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," 2003. (cited on page 73)

122. G. Zhou and J. Su, "Named entity recognition using an hmm-based chunk tagger," in *Proceedings of the 40th annual meeting of the association for computational linguistics*, pp. 473–480, 2002. (cited on page 73)

123. Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015. (cited on page 73)

124. J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 50–70, 2020. (cited on page 73)

125. S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," *arXiv preprint arXiv:2004.10964*, 2020. (cited on page 73)

126. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019. (cited on page 74)

127. C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020. (cited on page 74)

128. H. Zhou, R. Austin, S.-C. Lu, G. M. Silverman, Y. Zhou, H. Kilicoglu, H. Xu, and R. Zhang, "Complementary and integrative health information in the literature: its lexicon and named entity recognition," *Journal of the American Medical Informatics Association*, vol. 31, no. 2, pp. 426–434, 2024. (cited on page 74)

129. S. Sivarajkumar, M. Kelley, A. Samolyk-Mazzanti, S. Visweswaran, and Y. Wang, "An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study," *JMIR Medical Informatics*, vol. 12, p. e55318, 2024. (cited on page 74)

130. M. Neumann, D. King, I. Beltagy, and W. Ammar, "Scispacy: fast and robust models for biomedical natural language processing," *arXiv preprint arXiv:1902.07669*, 2019. (cited on page 74)

131. M. Li, H. Zhou, H. Yang, and R. Zhang, "Rt: a retrieving and chain-of-thought framework for few-shot medical named entity recognition," *Journal of the American Medical Informatics Association*, 2024. (cited on page 74)

132. Y. Hu, Q. Chen, J. Du, X. Peng, V. K. Keloth, X. Zuo, Y. Zhou, Z. Li, X. Jiang, Z. Lu, *et al.*, "Improving large language models for clinical named entity recognition via prompt engineering," *Journal of the American Medical Informatics Association*, 2024. (cited on page 74)

133. Y. Tseo, M. Salkola, A. Mohamed, A. Kumar, and F. Abnousi, "Information extraction of clinical trial eligibility criteria," *arXiv preprint arXiv:2006.07296*, 2020. (cited on page 75)

134. J. R. Finkel and C. D. Manning, "Nested named entity recognition," in *Proceedings of the 2009 conference on empirical methods in natural language processing*, pp. 141–150, 2009. (cited on page 78)

135. Y. Wang, H. Tong, Z. Zhu, and Y. Li, "Nested named entity recognition: a survey," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 16, no. 6, pp. 1–29, 2022. (cited on page 78)

136. F. Li, Z. Lin, M. Zhang, and D. Ji, "A span-based model for joint overlapped and discontinuous named entity recognition," *arXiv preprint arXiv:2106.14373*, 2021. (cited on page 78)

137. B. Tang, Q. Chen, X. Wang, Y. Wu, Y. Zhang, M. Jiang, J. Wang, and H. Xu, "Recognizing disjoint clinical concepts in clinical text using machine learning-based methods," in *AMIA annual symposium proceedings*, vol. 2015, p. 1184, American Medical Informatics Association, 2015. (cited on page 78)

138. A. Dirkson, S. Verberne, and W. Kraaij, "Fuzzybio: A proposal for fuzzy representation of discontinuous entities," in *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pp. 77–82, 2021. (cited on page 78)

139. Y. Wang, B. Yu, H. Zhu, T. Liu, N. Yu, and L. Sun, "Discontinuous named entity recognition as maximal clique discovery," *arXiv preprint arXiv:2106.00218*, 2021. (cited on page 78)

140. X. Dai, S. Karimi, B. Hachey, and C. Paris, "An effective transition-based model for discontinuous ner," *arXiv preprint arXiv:2004.13454*, 2020. (cited on page 78)

141. T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019. (cited on page 79)

142. S. Raza, D. J. Reji, F. Shajan, and S. R. Bashir, "Large-scale application of named entity recognition to biomedicine and epidemiology," *PLOS Digital Health*, vol. 1, no. 12, p. e0000152, 2022. (cited on page 79)

143. X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," *arXiv preprint arXiv:2101.00190*, 2021. (cited on page 83)

144. G. Qin and J. Eisner, "Learning how to ask: Querying lms with mixtures of soft prompts," *arXiv preprint arXiv:2104.06599*, 2021. (cited on page 83)

145. K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in *International conference on machine learning*, pp. 3929–3938, PMLR, 2020. (cited on page 83)

146. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020. (cited on page 83)

147. W. Guidance, "Ethics and governance of artificial intelligence for health," *World Health Organization*, 2021. (cited on page 107)

148. Novartis, "Novartis' commitment to the ethical and responsible use of artificial intelligence (ai) systems.," 2020. Accessed on June 1, 2024. (cited on page 107)

149. J. Zhang and Z.-m. Zhang, "Ethics and governance of trustworthy medical artificial intelligence," *BMC medical informatics and decision making*, vol. 23, no. 1, p. 7, 2023. (cited on page 107)

150. T. U. S. Food and D. Administration, "Using artificial intelligence machine learning in the development of drug and biological products," 2024. Accessed on June 1, 2024. (cited on page 107)

151. W. H. Organization *et al.*, "Ethics and governance of artificial intelligence for health: guidance on large multi-modal models," 2024. (cited on page 107)

152. G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023. (cited on page 107)

153. M. Enis and M. Hopkins, "From llm to nmt: Advancing low-resource machine translation with claude," *arXiv preprint arXiv:2404.13813*, 2024. (cited on page 107)

154. H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models (2023)," *arXiv preprint arXiv:2302.13971*, 2023. (cited on page 107)

155. A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023. (cited on page 107)

156. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019. (cited on page 107)

157. P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, *et al.*, "Holistic evaluation of language models," *arXiv preprint arXiv:2211.09110*, 2022. (cited on page 108)

158. T. C. for Research on Foundation Models (Stanford Institute for Human-Centered Artificial Intelligence), "Helm leaderboard," 2024. Accessed on June 2, 2024. (cited on pages 108 and 124)

159. D. Lonsdale, C. Tustison, C. Parker, and D. W. Embley, "Formulating queries for assessing clinical trial eligibility," in *Natural Language Processing and Information Systems*, pp. 82–93, Springer Berlin Heidelberg, 2006. (cited on page 185)

160. D. Lonsdale, C. Tustison, C. Parker, and D. Embley, "Assessing clinical trial eligibility with logic expression queries," *Data & Knowledge Engineering*, vol. 66, pp. 3–17, jul 2008. (cited on page 185)

161. S. Kiritchenko, B. de Bruijn, S. Carini, J. Martin, and I. Sim, "ExaCT: automatic extraction of clinical trial characteristics from journal publications," *BMC Medical Informatics and Decision Making*, vol. 10, sep 2010. (cited on page 185)

162. S. W. Tu, M. Peleg, S. Carini, M. Bobak, J. Ross, D. Rubin, and I. Sim, "A practical method for transforming free-text eligibility criteria into computable criteria," *Journal of Biomedical Informatics*, vol. 44, pp. 239–250, apr 2011. (cited on page 185)

163. I. Korkontzelos, T. Mu, and S. Ananiadou, "ASCOT: a text mining-based web-service for efficient search and assisted creation of clinical trials," *BMC Medical Informatics and Decision Making*, vol. 12, apr 2012. (cited on page 185)

164. Y. Wu, M. A. Levy, C. M. Micheel, P. Yeh, B. Tang, M. J. Cantrell, S. M. Cooreman, and H. Xu, "Identifying the status of genetic lesions in cancer clinical trial documents using machine learning," *BMC Genomics*, vol. 13, dec 2012. (cited on page 185)

165. A. Restificar and S. Ananiadou, "Inferring appropriate eligibility criteria in clinical trial protocols without labeled data," in *Proceedings of the ACM sixth international workshop on Data and text mining in biomedical informatics*, ACM, oct 2012. (cited on page 185)

166. S. Bhattacharya and M. N. Cantor, "Analysis of eligibility criteria representation in industry-standard clinical trial protocols," *Journal of Biomedical Informatics*, vol. 46, pp. 805–813, oct 2013. (cited on page 185)

167. R. Miotto, S. Jiang, and C. Weng, "eTACTS: A method for dynamically filtering clinical trial search results," *Journal of Biomedical Informatics*, vol. 46, pp. 1060–1067, dec 2013. (cited on page 185)

168. M. R. Boland, R. Miotto, J. Gao, and C. Weng, "Feasibility of feature-based indexing, clustering, and search of clinical trials," *Methods of Information in Medicine*, vol. 52, no. 05, pp. 382–394, 2013. (cited on page 185)

169. I. Soto-Rey, B. Trinczek, T. Karakoyun, M. Dugas, and F. Fritz, "Protocol feasibility workflow using an automated multi-country patient cohort system," in *e-Health–For Continuity of Care*, pp. 985–989, IOS Press, 2014. (cited on page 185)

170. A. Bucur, J. Van Leeuwen, N.-Z. Chen, B. Claerhout, K. De Schepper, D. Perez-Rey, R. Alonso-Calvo, L. Pugliano, and K. Saini, "Supporting patient screening to identify suitable clinical trials," in *e-Health–For Continuity of Care*, pp. 823–827, IOS Press, 2014. (cited on page 185)

171. A. Taweel, S. Miles, B. C. Delaney, and R. Bache, "An eligibility criteria query language for heterogeneous data warehouses," *Methods of Information in Medicine*, vol. 54, no. 01, pp. 41–44, 2015. (cited on page 185)

172. Y. Ni, S. Kennebeck, J. W. Dexheimer, C. M. McAneney, H. Tang, T. Lingren, Q. Li, H. Zhai, and I. Solti, "Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department," *Journal of the American Medical Informatics Association*, vol. 22, pp. 166–178, jul 2014. (cited on page 185)

173. Y. Ni, J. Wright, J. Perentesis, T. Lingren, L. Deleger, M. Kaiser, I. Kohane, and I. Solti, "Increasing the efficiency of trial-patient matching: automated clinical trial eligibility pre-screening for pediatric oncology patients," *BMC Medical Informatics and Decision Making*, vol. 15, apr 2015. (cited on page 185)

174. C. Shivade, C. Hebert, M. Lopetegui, M.-C. de Marneffe, E. Fosler-Lussier, and A. M. Lai, "Textual inference for eligibility criteria resolution in clinical trials," *Journal of Biomedical Informatics*, vol. 58, pp. S211–S218, dec 2015. (cited on page 185)

175. J. Hoxha, G. Jiang, and C. Weng, "Automated learning of domain taxonomies from text using background knowledge," *Journal of Biomedical Informatics*, vol. 63, pp. 295–306, oct 2016. (cited on page 185)

176. B. MacKellar and C. Schweikert, "Patterns for conflict identification in clinical trial eligibility criteria," in *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*, IEEE, sep 2016. (cited on page 185)

177. T. Hao, H. Liu, and C. Weng, "Valx: A system for extracting and structuring numeric lab test comparison statements from text," *Methods of Information in Medicine*, vol. 55, no. 03, pp. 266–275, 2016. (cited on page 185)

178. Y. Si and C. Weng, "An omop cdm-based relational database of clinical research eligibility criteria," *Studies in health technology and informatics*, vol. 245, p. 950, 2017. (cited on page 185)

179. K. Zhang and D. Demner-Fushman, "Automated classification of eligibility criteria in clinical trials to facilitate patient-trial matching for specific patient populations," *Journal of the American Medical Informatics Association*, vol. 24, pp. 781–787, feb 2017. (cited on page 185)

180. B. MacKellar and C. Schweikert, "Conflict discovery and analysis for clinical trials," in *Proceedings of the 2017 International Conference on Digital Health*, ACM, jul 2017. (cited on page 185)

181. T. Kang, S. Zhang, Y. Tang, G. W. Hruby, A. Rusanov, N. Elhadad, and C. Weng, "EliIE: An open-source information extraction system for clinical trial eligibility criteria," *Journal of the American Medical Informatics Association*, vol. 24, pp. 1062–1071, apr 2017. (cited on page 185)

182. G. Dietrich, M. Ertl, G. Fette, M. Kaspar, J. Krebs, D. Mackenrodt, S. Störk, and F. Puppe, "Extending the query language of a data warehouse for patient recruitment.," in *GMDS*, pp. 152–156, 2017. (cited on page 185)

183. V. Claveau, L. E. S. Oliveira, G. Bouzillé, M. Cuggia, C. M. C. Moro, and N. Grabar, "Numerical eligibility criteria in clinical protocols: Annotation, automatic detection and interpretation," in *Artificial Intelligence in Medicine*, pp. 203–208, Springer International Publishing, 2017. (cited on page 185)

184. C.-H. Chuan, "Classifying eligibility criteria in clinical trials using active deep learning," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, dec 2018. (cited on page 185)

185. D. Calaprice-Whitty, K. Galil, W. Salloum, A. Zariv, and B. Jimenez, "Improving clinical trial participant prescreening with artificial intelligence (AI): A comparison of the results of AI-assisted vs standard methods in 3 oncology trials," *Therapeutic Innovation & Regulatory Science*, p. 216847901881545, jan 2019. (cited on page 185)

186. A. Bustos and A. Pertusa, "Learning eligibility in cancer clinical trials using deep neural networks," *Applied Sciences*, vol. 8, p. 1206, jul 2018. (cited on page 185)

187. S. M. Meystre, P. M. Heider, Y. Kim, D. B. Aruch, and C. D. Britten, "Automatic trial eligibility surveillance based on unstructured clinical data," *International Journal of Medical Informatics*, vol. 129, pp. 13–19, sep 2019. (cited on page 185)

188. I. Spasic, D. Krzeminski, P. Corcoran, and A. Balinsky, "Cohort selection for clinical trials from longitudinal patient records: Text mining approach," *JMIR Medical Informatics*, vol. 7, p. e15980, oct 2019. (cited on page 185)

189. I. Segura-Bedmar and P. Raez, "Cohort selection for clinical trials using deep learning models," *Journal of the American Medical Informatics Association*, vol. 26, pp. 1181–1188, sep 2019. (cited on page 185)

190. C. Yuan, P. B. Ryan, C. Ta, Y. Guo, Z. Li, J. Hardin, R. Makadia, P. Jin, N. Shang, T. Kang, and C. Weng, "Criteria2query: a natural language interface to clinical databases for cohort definition," *Journal of the American Medical Informatics Association*, vol. 26, pp. 294–305, feb 2019. (cited on page 185)

191. C. Liu, C. Yuan, A. M. Butler, R. D. Carvajal, Z. R. Li, C. N. Ta, and C. Weng, "DQueST: dynamic questionnaire for search of clinical trials," *Journal of the American Medical Informatics Association*, vol. 26, pp. 1333–1343, aug 2019. (cited on page 185)

192. M. A. A. H. Khan, N. Dimitrova, M. Shamsuzzaman, S. A. Hasan, M. S. Sorower, J. Liu, V. Datla, M. Milosevic, G. Mankovich, and R. van Ommering, "Improving disease named entity recognition for clinical trial matching," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, nov 2019. (cited on page 185)

193. "Abstract supplement 2019 ACR/ARP annual meeting," *Arthritis & Rheumatology*, vol. 71, pp. 1–5420, oct 2019. (cited on page 186)

194. Y. Sun and K. Loparo, "Information extraction from free text in clinical trials with knowledge-based distant supervision," in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, IEEE, jul 2019. (cited on page 186)

195. C.-J. Chen, N. Warikoo, Y.-C. Chang, J.-H. Chen, and W.-L. Hsu, "Medical knowledge infused convolutional neural networks for cohort selection in clinical trials," *Journal of the American Medical Informatics Association*, vol. 26, pp. 1227–1236, aug 2019. (cited on page 186)

196. K. Zeng, Z. Pan, Y. Xu, and Y. Qu, "An ensemble learning strategy for eligibility criteria text classification for clinical trial recruitment: Algorithm development and validation," *JMIR Medical Informatics*, vol. 8, p. e17832, jul 2020. (cited on page 186)

197. J. T. Beck, M. Rammage, G. P. Jackson, A. M. Preininger, I. Dankwa-Mullan, M. C. Roebuck, A. Torres, H. Holtzen, S. E. Coverdill, M. P. Williamson, Q. Chau, K. Rhee, and M. Vinegra, "Artificial intelligence tool for optimizing eligibility screening for clinical trials in a large community cancer center," *JCO Clinical Cancer Informatics*, pp. 50–59, nov 2020. (cited on page 186)

198. X. Liu, L. A. Finelli, G. L. Hersch, and I. Khalil, "Attention-based LSTM network for COVID-19 clinical trial parsing," in *2020 IEEE International Conference on Big Data (Big Data)*, IEEE, dec 2020. (cited on page 186)

199. H.-J. Dai, F.-D. Wang, C.-W. Chen, C.-H. Su, C.-S. Wu, and J. Jonnagaddala, "Cohort selection for clinical trials using multiple instance learning," *Journal of Biomedical Informatics*, vol. 107, p. 103438, jul 2020. (cited on page 186)

200. X. Zhang, C. Xiao, L. M. Glass, and J. Sun, "DeepEnroll: Patient-trial matching with deep embedding and entailment prediction," in *Proceedings of The Web Conference 2020*, ACM, apr 2020. (cited on page 186)

201. M. Alexander, B. Solomon, D. L. Ball, M. Sheerin, I. Dankwa-Mullan, A. M. Preininger, G. P. Jackson, and D. M. Herath, "Evaluation of an artificial intelligence clinical trial matching system in australian lung cancer patients," *JAMIA Open*, vol. 3, pp. 209–215, may 2020. (cited on page 186)

202. E. Jung, H. Jain, A. P. Sinha, and C. Gaudioso, "Building a specialized lexicon for breast cancer clinical trial subject eligibility analysis," *Health Informatics Journal*, vol. 27, p. 146045822198939, jan 2021. (cited on page 186)

203. Y. Sun, A. Butler, L. A. Stewart, H. Liu, C. Yuan, C. T. Southard, J. H. Kim, and C. Weng, "Building an OMOP common data model-compliant annotated corpus for COVID-19 clinical trials," *Journal of Biomedical Informatics*, vol. 118, p. 103790, jun 2021. (cited on page 186)

204. J. Du, Q. Wang, J. Wang, P. Ramesh, Y. Xiang, X. Jiang, and C. Tao, "COVID-19 trial graph: a linked graph for COVID-19 clinical trials," *Journal of the American Medical Informatics Association*, vol. 28, pp. 1964–1969, apr 2021. (cited on page 186)

205. H. Dhayne, R. Kilany, R. Haque, and Y. Taher, "EMR2vec: Bridging the gap between patient data and clinical trial," *Computers & Industrial Engineering*, vol. 156, p. 107236, jun 2021. (cited on page 186)

206. Z. He, A. Erdengasileng, X. Luo, A. Xing, N. Charness, and J. Bian, "How the clinical research community responded to the COVID-19 pandemic: an analysis of the COVID-19 clinical studies in ClinicalTrials.gov," *JAMIA Open*, vol. 4, apr 2021. (cited on page 186)

207. S. Tian, P. Yin, H. Zhang, A. Erdengasileng, J. Bian, and Z. He, "Parsing clinical trial eligibility criteria for cohort query by a multi-input multi-output sequence labeling model," *medRxiv*, nov 2021. (cited on page 186)

208. M. E. Elkin and X. Zhu, "Predictive modeling of clinical trial terminations using feature engineering and embedding learning," *Scientific Reports*, vol. 11, feb 2021. (cited on page 186)

209. H. Zong, J. Yang, Z. Zhang, Z. Li, and X. Zhang, "Semantic categorization of chinese eligibility criteria in clinical trials using machine learning methods," *BMC Medical Informatics and Decision Making*, vol. 21, apr 2021. (cited on page 186)

210. J. Li, Q. Wei, O. Ghiasvand, M. Chen, V. Lobanov, C. Weng, and H. Xu, "Study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora," in *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, IEEE, aug 2021. (cited on page 186)

211. J. Li, Q. Wei, O. Ghiasvand, M. Chen, V. Lobanov, C. Weng, and H. Xu, "A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora," *BMC Medical Informatics and Decision Making*, vol. 22, sep 2022. (cited on page 186)

212. Y. Xiong, W. Peng, Q. Chen, Z. Huang, and B. Tang, "A unified machine reading comprehension framework for cohort selection," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, pp. 379–387, jan 2022. (cited on page 186)

213. Y. Fang, B. Idnay, Y. Sun, H. Liu, Z. Chen, K. Marder, H. Xu, R. Schnall, and C. Weng, "Combining human and machine intelligence for clinical trial eligibility querying," *Journal of the American Medical Informatics Association*, vol. 29, pp. 1161–1171, apr 2022. (cited on page 186)

214. T. Fu, K. Huang, C. Xiao, L. M. Glass, and J. Sun, "HINT: Hierarchical interaction network for clinical-trial-outcome predictions," *Patterns*, vol. 3, p. 100445, apr 2022. (cited on page 186)

215. S. Wang, S. Šuster, T. Baldwin, and K. Verspoor, "Predicting publication of clinical trials using structured and unstructured data: Model development and validation study," *Journal of Medical Internet Research*, vol. 24, p. e38859, dec 2022. (cited on page 186)

216. S. Jana, A. Naskar, T. Dasgupta, and L. Dey, "Syntactic and semantic knowledge-aware paraphrase detection for clinical data," in *Data Management, Analytics and Innovation*, pp. 169–183, Springer Singapore, sep 2021. (cited on page 187)

217. M. Rybinski, V. Nguyen, and S. Karimi, "A self-learning resource-efficient reranking method for clinical trials search," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, ACM, Oct. 2023. (cited on page 187)

218. S. Kaskovich, K. D. Wyatt, T. Oliwa, L. Graglia, B. Furner, J. Lee, A. Mayampurath, and S. L. Volchenboum, "Automated matching of patients to clinical trials: A patient-centric natural language processing approach for pediatric leukemia," *JCO Clinical Cancer Informatics*, July 2023. (cited on page 187)

219. P. P. Tun, J. Luo, J. Xie, S. Wibowo, and C. Hao, "Automatic assessment of patient eligibility by utilizing nlp and rule-based analysis," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine amp; Biology Society (EMBC)*, IEEE, July 2023. (cited on page 187)

220. Z. Sun and C. Tao, "Named entity recognition and normalization for alzheimer's disease eligibility criteria," in *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, IEEE, June 2023. (cited on page 187)

221. W. Kusa, E. Mendoza, P. Knoth, G. Pasi, and A. Hanbury, "Effective matching of patients to clinical trials using entity extraction and neural re-ranking," *Journal of Biomedical Informatics*, vol. 144, p. 104444, Aug. 2023. (cited on page 187)

222. J. S. Peterson, D. Plana, D. S. Bitterman, S. B. Johnson, H. J. W. L. Aerts, and B. H. Kann, "Growth in eligibility criteria content and failure to accrue among national cancer institute (nci)-affiliated clinical trials," *Cancer Medicine*, vol. 12, pp. 4715–4724, nov 2022. (cited on page 187)

223. E. Kavalci and A. Hartshorn, "Improving clinical trial design using interpretable machine learning based prediction of early trial termination," *Scientific Reports*, vol. 13, jan 2023. (cited on page 187)

224. S. M. Meystre, P. M. Heider, A. Cates, G. Bastian, T. Pittman, S. Gentilin, and T. J. Kelechi, "Piloting an automated clinical trial eligibility surveillance and provider alert system based on artificial intelligence and standard data models," *BMC Medical Research Methodology*, vol. 23, apr 2023. (cited on page 187)

225. B. P. Theodorou, C. Xiao, and J. Sun, "Treement: Interpretable patient-trial matching via personalized dynamic tree-based memory network," in *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '23, ACM, Sept. 2023. (cited on page 187)

226. S.-K. Chang, D. Liu, J. Mitchem, C. Papageorgiou, J. Kaifi, and C.-R. Shyu, "Understanding common key indicators of successful and unsuccessful cancer drug trials using a contrast mining framework on ClinicalTrials.gov," *Journal of Biomedical Informatics*, vol. 139, p. 104321, mar 2023. (cited on page 187)

227. S. Datta, K. Lee, H. Paek, F. J. Manion, N. Ofoegbu, J. Du, Y. Li, L.-C. Huang, J. Wang, B. Lin, H. Xu, and X. Wang, "Autocriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models," *Journal of the American Medical Informatics Association*, vol. 31, p. 375–385, Nov. 2023. (cited on page 187)

228. B. Y. Miao, M. Sushil, A. Xu, M. Wang, D. Arneson, E. Berkley, M. Subash, R. Vashisht, V. Rudrapatna, and A. J. Butte, "Characterisation of digital therapeutic clinical trials: a systematic review with natural language processing," *The Lancet Digital Health*, vol. 6, p. e222–e229, Mar. 2024. (cited on page 187)

229. J. Park, Y. Fang, C. Ta, G. Zhang, B. Idnay, F. Chen, D. Feng, R. Shyu, E. R. Gordon, M. Spotnitz, and C. Weng, "Criteria2query 3.0: Leveraging generative large language models for clinical trial eligibility query generation," *Journal of Biomedical Informatics*, vol. 154, p. 104649, June 2024. (cited on page 187)

230. M. Nievas, A. Basu, Y. Wang, and H. Singh, "Distilling large language models for matching patients to clinical trials," *Journal of the American Medical Informatics Association*, Apr. 2024. (cited on page 187)

231. B. Idnay, J. Liu, Y. Fang, A. Hernandez, S. Kaw, A. Etwaru, J. Juarez Padilla, S. O. Ramírez, K. Marder, C. Weng, and R. Schnall, "Sociotechnical feasibility of natural language processing-driven tools in clinical trial eligibility prescreening for alzheimer's disease and related dementias," *Journal of the American Medical Informatics Association*, vol. 31, p. 1062–1073, Mar. 2024. (cited on page 187)

232. J. Beattie, S. Neufeld, D. Yang, C. Chukwuma, A. Gul, N. Desai, S. Jiang, and M. Dohopolski, "Utilizing large language models for enhanced clinical trial matching: A study on automation in patient screening," *Cureus*, May 2024. (cited on page 187)

Appendices

A Scoping review search strings

Below are the detailed search strings employed to query the bibliographic databases included in the scoping review.

EMBASE

```
('enrollment':ti,ab,kw OR 'enrolment':ti,ab,kw OR 'recruitment':ti,ab,kw
    OR 'screening':ti,ab,kw OR 'criteria':ti,ab,kw OR 'eligibility':ti,ab,kw
    OR 'matching':ti,ab,kw OR 'cohort selection':ti,ab,kw
    OR 'cohort ascertainment':ti,ab,kw OR 'accrual':ti,ab,kw
    OR 'prescreening':ti,ab,kw OR 'phenotyping':ti,ab,kw
    OR 'population enrichment':ti,ab,kw)
AND 'trial*':ti,ab,kw
AND ('transformer*':ti,ab,kw OR 'bert':ti,ab,kw
    OR 'named entity recognition':ti,ab,kw OR 'ner':ti,ab,kw
    OR 'textual entailment':ti,ab,kw
    OR 'natural language inference':ti,ab,kw
    OR 'language model*':ti,ab,kw OR 'semantic similarity':ti,ab,kw
    OR 'semantic textual similarity':ti,ab,kw OR 'entity linking':ti,ab,kw
    OR 'information extraction':ti,ab,kw
    OR 'natural language processing':ti,ab,kw
    OR 'nlp':ti,ab,kw OR 'deep learning':ti,ab,kw
    OR 'machine learning':ti,ab,kw
    OR 'attention mechanism':ti,ab,kw)
    OR 'generative AI':ti,ab,kw)
    OR 'GPT':ti,ab,kw)
AND [2000-2024]/py
AND [english]/lim
```

Scopus

```
TITLE-ABS-KEY ( (
    "enrollment" OR "enrolment" OR "recruitment"
    OR "screening" OR "criteria" OR "eligibility"
    OR "matching" OR "cohort selection" OR "cohort ascertainment"
    OR "accrual" OR "prescreening" OR "phenotyping"
    OR "population enrichment" )
AND "trial*"
AND ( "transformer*" OR "bert"
    OR "named entity recognition" OR "NER"
    OR "textual entailment" OR "natural language inference"
    OR "language model*" OR "semantic similarity"
    OR "semantic textual similarity" OR "entity linking"
    OR "information extraction" OR "natural language processing"
    OR "NLP" OR "deep learning" OR "machine learning"
    OR "attention mechanism" OR "generative AI" OR "GPT") )
AND PUBYEAR > 1999
AND ( LIMIT-TO ( LANGUAGE , "English" ) )
```

PubMed

```
((("enrollment"[Title/Abstract] OR "enrolment"[Title/Abstract]
    OR "recruitment"[Title/Abstract] OR "screening"[Title/Abstract]
    OR "criteria"[Title/Abstract] OR "eligibility"[Title/Abstract]
    OR "matching"[Title/Abstract] OR "cohort selection"[Title/Abstract]
    OR "cohort ascertainment"[Title/Abstract] OR "accrual"[Title/Abstract]
    OR "prescreening"[Title/Abstract] OR "phenotyping"[Title/Abstract]
    OR "population enrichment"[Title/Abstract])
AND "trial*"[Title/Abstract]
AND ("transformer*"[Title/Abstract] OR "bert"[Title/Abstract]
    OR "named entity recognition"[Title/Abstract] OR "NER"[Title/Abstract]
    OR "textual entailment"[Title/Abstract]
    OR "natural language inference"[Title/Abstract]
    OR "language model*"[Title/Abstract]
    OR "semantic similarity"[Title/Abstract]
    OR "semantic textual similarity"[Title/Abstract]
    OR "entity linking"[Title/Abstract]
    OR "information extraction"[Title/Abstract]
    OR "natural language processing"[Title/Abstract]
    OR "NLP"[Title/Abstract] OR "deep learning"[Title/Abstract]
    OR "machine learning"[Title/Abstract]
    OR "attention mechanism"[Title/Abstract]
    OR "generative ai"[Title/Abstract]
    OR "gpt"[Title/Abstract]))
```

```
AND (("2000/01/01"[Date - Publication] : "3000"[Date - Publication]))
AND (English[Language])
```

Web of Science

```
("enrollment" OR “enrolment” OR "recruitment"
    OR "screening" OR “criteria” OR “eligibility”
    OR “matching” OR “cohort selection” OR “cohort ascertainment”
    OR “accrual” OR “prescreening” OR “phenotyping”
    OR “population enrichment”)
AND "trial*"
AND ("transformer*" OR "bert" OR "named entity recognition"
    OR "NER" OR "textual entailment"
    OR "natural language inference" OR "language model*"
    OR "semantic similarity" OR "semantic textual similarity"
    OR "entity linking" OR “information extraction”
    OR "natural language processing" OR “NLP”
    OR "deep learning" OR "machine learning"
    OR “attention mechanism” OR "generative AI"
    OR "GPT") (Topic)
AND English (Language)
AND Timespan: 2000-01-01 to 2024-02-07 (Publication Date)
```

B Scoping review data extraction table

The details of data extraction are presented in three separate tables to enhance readability of the study:

- Table 1 presents the title, bibliographic reference, DOI, year of publication, and a broad description of each paper included in the review.
- Table 2 provides detailed demographic data of each paper, including the title, DOI, countries from authors' affiliations, countries where each study was conducted, the therapeutic area covered by the study, and the general character of the paper.
- Table 3 details the extraction of features from each study, including DOI, main business objective, NLP methods used, metrics for evaluation, datasets used, number of protocols and eligibility criteria utilized in experiments, the source of clinical protocol data, auxiliary datasets, the method of establishing the ground truth, preprocessing techniques, shallow ML methods described, neural models used in experimentation, knowledge bases, and the main contribution of the paper.

Table 1: Papers included in the scoping review.

Article	DOI	Year	Description
Formulating queries for assessing clinical trial eligibility [159]	10.1007/11765448_8	2006	To extract semantic information from eligibility criteria for better patient-trial matching.
Assessing clinical trial eligibility with logic expression queries [160]	10.1016/j.datak.2007.07.005	2008	To extract semantic information reflecting eligibility criteria from clinical trial descriptions and formulate queries that can match criteria against medical data in patient records.
ExaCT: automatic extraction of clinical trial characteristics from journal publications [161]	10.1186/1472-6947-10-56	2010	To develop an information exctraction system for extracting key trial characteristics from full-text journal articles
A practical method for transforming free-text eligibility criteria into computable criteria [162]	10.1016/j.jbi.2010.09.007	2011	To develop a method of transforming free text eligibility criteria into a form suitable for SPARQL/SQL querying
ASCOT: A text mining-based web-service for efficient search and assisted creation of clinical trials [163]	10.1186/1472-6947-12-S1-S3	2012	To develop an efficient search tool for filtering clinical trial descriptions.
Identifying the status of genetic lesions in cancer clinical trial documents using machine learning [164]	10.1186/1471-2164-13-S8-S21	2012	To develop a machine learning system for identifying mentions of genes and genetic lesions in cancer clinical trials.
Inferring appropriate eligibility criteria in clinical trial protocols without labeled data [165]	10.1145/2390068.2390074	2012	To develop an unsupervised method of eligibility criteria identification in clinical trial texts.
Analysis of eligibility criteria representation in industry-standard clinical trial protocols [166]	10.1016/j.jbi.2013.06.001	2013	To develop a method of standarization of eligibility criteria across different clinical trial databases.
ETACTS: A method for dynamically filtering clinical trial search results [167]	10.1016/j.jbi.2013.07.014	2013	To evaluate the eTACTS system for advanced querying of clinical trial descriptions.
Feasibility of Feature-based Indexing, Clustering, and Search of Clinical Trials [168]	10.3414/ME12-01-0092	2013	To explore the feasibility of feature-based indexing, clustering, and search of clinical trials.
Protocol Feasibility Workflow Using an Automated Multi-country Patient Cohort System [169]	10.3233/978-1-61499-432-9-985	2014	To build a system for querying patients eligible for clinical trial participation across many countries' registries.
Supporting Patient Screening to Identify Suitable Clinical Trials [170]	10.3233/978-1-61499-432-9-823	2014	To develop a solution for flexible formalization of criteria and other trial metadata and for effective management of those representations
An eligibility criteria query language for heterogeneous data warehouses [171]	10.3414/ME13-02-0027	2015	To develop a clinical-readable query language for heterogenous warehouses of medical information.
Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department [172]	10.1136/amiajnl-2014-002887	2015	To evaluate an automated eligibility screening approach to clinical trials using different machine learning methods.
Increasing the efficiency of trial-patient matching: automated clinical trial eligibility Pre-screening for pediatric oncology patients [173]	10.1186/s12911-015-0149-3	2015	To identify patients who meet core eligibility characteristics of an oncology clinical trial.
Textual inference for eligibility criteria resolution in clinical trials [174]	10.1016/j.jbi.2015.09.008	2015	To build a dataset of clinical texts annotated with eligibility criteria entailment.
Automated learning of domain taxonomies from text using background knowledge [175]	10.1016/j.jbi.2016.09.002	2016	To create a framework for unsupervised ontology learning from clinical trial texts.
Patterns for conflict identification in clinical trial eligibility criteria [176]	10.1109/HealthCom.2016.7749519	2016	To develop a method for automated identification of potential treatment conflicts between trials.
Valx: A System for Extracting and Structuring Numeric Lab Test Comparison Statements from Text [177]	10.3414/ME15-01-0112	2016	To develop and evaluate an automated method for extracting and structuring numeric comparison statements in trial eligibility criteria text.
An OMOP CDM-Based Relational Database of Clinical Research Eligibility Criteria [178]	10.3233/978-1-61499-830-3-950	2017	To build a database of clinical trials for easy querying and filtering of trials using multiple criteria.
Automated classification of eligibility criteria in clinical trials to facilitate patient-trial matching for specific patient populations [179]	10.1093/jamia/ocw176	2017	To develop automated classification methods for eligibility criteria to facilitate patient-trial matching for specific populations such as persons living with HIV or pregnant women.
Conflict discovery and analysis for clinical trials [180]	10.1145/3079452.3079494	2017	To match medical treatments to exclusion criteria in clinical trial text.
EliIE: An open-source information extraction system for clinical trial eligibility criteria [181]	10.1093/jamia/ocx019	2017	To develop a system for parsing free text clinical trial descriptions.
Extending the Query Language of a Data Warehouse for Patient Recruitment [182]	10.3233/978-1-61499-808-2-152	2017	To develop a data warehouse querying language for matching patient electronic health records with clinical trial eligibility criteria
Numerical Eligibility Criteria in Clinical Protocols: Annotation, Automatic Detection and Interpretation [183]	10.1007/978-3-319-59758-4_22	2017	To create a model for the detection of complex numerical eligibility criteria in clinical trial texts.
Classifying Eligibility Criteria in Clinical Trials Using Active Deep Learning [184]	10.1109/ICMLA.2018.00052	2018	To automatically identify eligibility criteria that can be evaluated by patients without the help of medical professionals.
Improving Clinical Trial Participant Prescreening With Artificial Intelligence (AI): A Comparison of the Results of AI-Assisted vs Standard Methods in 3 Oncology Trials [185]	10.1177/2168479018815454	2018	To evaluate the efficacy of Mendel.ai in enrolling eligible patients in clinical trials.
Learning Eligibility in Cancer Clinical Trials Using Deep Neural Networks [186]	10.3390/app8071206	2018	To automatically predict whether short clinical statements were considered inclusion or exclusion criteria.
Automatic trial eligibility surveillance based on unstructured clinical data [187]	10.1016/j.ijmedinf.2019.05.018	2019	To develop an algorithm for automatic identification of patients eligible for clinical trial participation.
Cohort Selection for Clinical Trials From Longitudinal Patient Records: Text Mining Approach [188]	10.2196/15980	2019	To evaluate a system for patient-trial matching.
Cohort selection for clinical trials using deep learning models [189]	10.1093/jamia/ocz139	2019	To evaluate the efficiency of deep neural network architectures in cohort selection tasks.
Criteria2Query: A natural language interface to clinical databases for cohort definition [190]	10.1093/jamia/ocy178	2019	To develop a pipeline for eligibility criteria parsing and converting into CDM-based cohort queries.
DQueST: dynamic questionnaire for search of clinical trials [191]	10.1093/jamia/ocz121	2019	To develop a patient-trial matching system based on dynamic questionaire generation.
Improving Disease Named Entity Recognition for Clinical Trial Matching [192]	10.1109/BIBM47256.2019.8983421	2019	To develop a named entity recognition model for clinical trial texts.

Improving the efficiency of clinical trial recruitment using electronic health record data, natural language processing, and machine learning [193]	10.1002/art.41108	2019	To evaluate if ensemble machine learning algorithm can improve the efficiency of eligibility screening.
Information Extraction from Free Text in Clinical Trials with Knowledge-based Distant Supervision [194]	10.1109/COMPSAC.2019.00158	2019	To develop a method for extracting medical concepts from free text clinical trial descriptions.
Medical knowledge infused convolutional neural networks for cohort selection in clinical trials [195]	10.1093/jamia/ocz128	2019	To embed electronic health records using convolutional neural networks for better patient-trial matching.
An Ensemble Learning Strategy for Eligibility Criteria Text Classification for Clinical Trial Recruitment: Algorithm Development and Validation [196]	10.2196/17832	2020	To evaluate the efficacy of model ensembling in the task of eligibility criteria classification.
Artificial Intelligence Tool for Optimizing Eligibility Screening for Clinical Trials in a Large Community Cancer Center [197]	10.1200/CCI.19.00079	2020	To evaluate the performance of WCTM tool in patient data intake and matching porcesses.
Attention-Based LSTM Network for COVID-19 Clinical Trial Parsing [198]	10.1109/BigData50022.2020.9378451	2020	To investigate different versions of Att-BiLSTM models to extract entities from COVID-19 elgibility criteria
Chia, a large annotated corpus of clinical trial eligibility criteria [96]	10.1038/s41597-020-00620-0	2020	To create a large annotated corpus of eligibility criteria extracted from clinical trial descriptions.
Cohort selection for clinical trials using multiple instance learning [199]	10.1016/j.jbi.2020.103438	2020	To verify the usefulness of multiple instance learning paradigm in patient-trial matching task.
DeepEnroll: Patient-Trial Matching with Deep Embedding and Entailment Prediction [200]	10.1145/3366423.3380181	2020	To develop a deep neural network for patient-trial matching using cross-modal representation of eligibility criteria and patient electronic health records.
Evaluation of an artificial intelligence clinical trial matching system in Australian lung cancer patients [201]	10.1093/jamiaopen/ooaa002	2020	To evaluate the performance of IBM Watson for matching patients to clinical trials.
A Scalable Ai Approach For Clinical Trial Cohort Optimization [98]	10.1007/978-3-030-93733-1_36	2021	To inform clinical trial design by automatic extraction of eligibility criteria.
A knowledge base of clinical trial eligibility criteria [98]	10.1016/j.jbi.2021.103771	2021	To build a standardized knowledge base of elgibility criteria.
Accuracy of an Artificial Intelligence System for Cancer Clinical Trial Eligibility Screening: Retrospective Pilot Study [95]	10.2196/27767	2021	To evaluate Watson CDSS system for matching patients to clinical trials
Building a specialized lexicon for breast cancer clinical trial subject eligibility analysis [202]	10.1177/1460458221989392	2021	To develop the specialized lexicon of medical terms.
Building an OMOP common data model-compliant annotated corpus for COVID-19 clinical trials [203]	10.1016/j.jbi.2021.103790	2021	To build an annotated dataset of eligibility criteria for COVID-19 clinical trials.
COVID-19 trial graph: a linked graph for COVID-19 clinical trials [204]	10.1093/jamia/ocab078	2021	To create a graph database of structured and unstructured medical information fromm COVID-19 clinical trials
Classification And Extraction Of Medical Clinical Trial Screening Standard Texts Based On Bi-Lstm And Attention Mechanism [204]	10.1088/1755-1315/632/5/052088	2021	To develop a model for eligibility criteria classification.
Clinical Trial Information Extraction with BERT [98]	10.1109/ICHI52183.2021.00092	2021	To evaluate the effectiveness of BERT embeddings in medical information extraction.
EMR2vec: Bridging the gap between patient data and clinical trial [205]	10.1016/j.cie.2021.107236	2021	To develop a system for patient-trial matching.
How the clinical research community responded to the COVID-19 pandemic: an analysis of the COVID-19 clinical studies in ClinicalTrials.gov [206]	10.1093/jamiaopen/ooab032	2021	To conduct a systematic analysis of clinical trials on COVID-19
Parsing clinical trial eligibility criteria for cohort query by a multi-input multi-output sequence labeling model [207]	10.1101/2021.11.18.21266533	2021	To evaluate a sequence labeling model in parsing and extracting medical entities in eligibility criteria.
Predictive modeling of clinical trial terminations using feature engineering and embedding learning [208]	10.1038/s41598-021-82840-x	2021	To predict clinical trial terminations and identify main factors influencing the terminations.
Semantic categorization of Chinese eligibility criteria in clinical trials using machine learning methods [209]	10.1186/s12911-021-01487-w	2021	To extract semantic categories from eligibility criteria in Chinese clinical trials.
Study of Pre-trained Language Models for Named Entity Recognition in Clinical Trial Eligibility Criteria from Multiple Corpora [210]	10.1109/ICHI52183.2021.00095	2021	To explore transformer-based models for clinical term extraction from eligibility criteria
Transformer-Based Named Entity Recognition for Parsing Clinical Trial Eligibility Criteria [207]	10.1145/3459930.3469560	2021	To extract named entities in eligibility criteria using transformer-based model.
A Comparative Study Of Pre-Trained Language Models For Named Entity Recognition In Clinical Trial Eligibility Criteria From Multiple Corpora [211]	10.1186/s12911-022-01967-7	2022	To compare multiple pre-trained language models in the task of named entity recognition in clinical trial descriptions.
A Unified Machine Reading Comprehension Framework for Cohort Selection [212]	10.1109/JBHI.2021.3095478	2022	To develop a machine reading comprehension (MRC) framework for cohort selection and evaluate different MRC algorithms, in particular, the use of the cross-criteria attention.
An Evaluation of Pretrained BERT Models for Comparing Semantic Similarity Across Unstructured Clinical Trial Texts [94]	10.3233/SHTI210848	2022	To evaluate the efficacy of BERT models in assessing semantic similarity of clinical trial descriptions
Combining Human And Machine Intelligence For Clinical Trial Eligibility Querying [213]	10.1093/jamia/ocac051	2022	To develop a system (Criteria2Query 2.0) for automatic conversion of eligibility criteria into cohort queries.
Hint: Hierarchical Interaction Network For Clinical-Trial-Outcome Predictions [214]	10.1016/j.patter.2022.100445	2022	To predict the success of a clinical trial using deep neural networks.
Predicting Publication Of Clinical Trials Using Structured And Unstructured Data: Model Development And Validation Study [215]	10.2196/38859	2022	To develop a system for clinical trial publication

Syntactic and Semantic Knowledge-Aware Paraphrase Detection for Clinical Data [216]	10.1007/978-981-16-2937-2_13	2022	To develop a knowledge-aware neural network model for paraphrase detection in eligibility criteria texts.
The Leaf Clinical Trials Corpus: A New Resource For Query Generation From Clinical Trial Eligibility Criteria [97]	10.1038/s41597-022-01521-0	2022	To build an annotated corpus of clinical trial eligibility criteria.
A Self-Learning Resource-Efficient Re-Ranking Method For Clinical Trials Search [217]	10.1145/3583780.3615174	2023	To develop a CT ranking method to facilitate trial search.
Automated Matching Of Patients To Clinical Trials: A Patient-Centric Natural Language Processing Approach For Pediatric Leukemia [218]	10.1200/CCI.23.00009	2023	To create a patient-trial matching tool for pediatric leukemia.
Automatic Assessment Of Patient Eligibility By Utilizing Nlp And Rule-Based Analysis [219]	10.1109/EMBC40787.2023.10340494	2023	To develop a model for fully automatic selection of patients for clinical trials.
Distilling Large Language Models For Matching Patients To Clinical Trials [220]	10.1109/ICHI57859.2023.00100	2023	To develop a pipeline for information extraction from AD clinical trial eligibility criteria.
Effective Matching Of Patients To Clinical Trials Using Entity Extraction And Neural Re-Ranking [221]	10.1016/j.jbi.2023.104444	2023	To develop a method for clinical trial retrieval for the purpose of patient-trial matching
Growth In Eligibility Criteria Content And Failure To Accrue Among National Cancer Institute (Nci)-Affiliated Clinical Trials [222]	10.1002/cam4.5276	2023	To investigate the impact of eligibility criteria on trial accrual and identify criteria assosiated with accrual failure
Improving Clinical Trial Design Using Interpretable Machine Learning Based Prediction Of Early Trial Termination [223]	10.1038/s41598-023-27416-7	2023	To inform clinical trial design using interpretable machine learning models for predicting early trial terminations.
Piloting An Automated Clinical Trial Eligibility Surveillance And Provider Alert System Based On Artificial Intelligence And Standard Data Models [224]	10.1186/s12874-023-01916-6	2023	To improve patient-trial matching by electronic health record normalization and to build efficient user interface for clinical trial selection.
Treement: Interpretable Patient-Trial Matching Via Personalized Dynamic Tree-Based Memory Network [225]	10.1145/3584371.3612998	2023	To develop a model for interpretable patient trial matching.
Understanding Common Key Indicators Of Successful And Unsuccessful Cancer Drug Trials Using A Contrast Mining Framework On Clinicaltrials.Gov [226]	10.1016/j.jbi.2023.104321	2023	To evaluate the effectiveness of contrastive learning paradigm in the prediction of clinical trial success.
Autocriteria: A Generalizable Clinical Trial Eligibility Criteria Extraction System Powered By Large Language Models [227]	10.1093/jamia/ocad218	2024	To build an LLM-based tool for information extraction frm eligibility criteria without a need of labeled training data.
Characterisation Of Digital Therapeutic Clinical Trials: A Systematic Review With Natural Language Processing [228]	10.1016/S2589-7500(23)00244-3	2024	To explore digital therapeutics clinical trials.
Criteria2Query 3.0: Leveraging Generative Large Language Models For Clinical Trial Eligibility Query Generation [229]	10.1016/j.jbi.2024.104649	2024	To create a tool leveraging LLMs for information extraction from eligbility criteria.
Distilling Large Language Models For Matching Patients To Clinical Trials [230]	10.1093/jamia/ocae073	2024	To compare the efficacy of proprietary vs. open-source LLMs in patient-trial matching task.
Sociotechnical Feasibility Of Natural Language Processing-Driven Tools In Clinical Trial Eligibility Prescreening For Alzheimer'S Disease And Related Dementias [231]	10.1093/jamia/ocae032	2024	To assess feasibility of existing NLP tools for eligibility prescreening for AD trials.
Utilizing Large Language Models For Enhanced Clinical Trial Matching: A Study On Automation In Patient Screening [232]	10.7759/cureus.60044	2024	To apply LLMs for automatic trial aligibility screening.

Table 2: Study demographics.

Title	DOI	Country (author)	Country (research)	Therapeutic area	General character
Formulating queries for assessing clinical trial eligibility	10.1007/11765448_8	USA	not specified	not specified	software presentation
Assessing clinical trial eligibility with logic expression queries	10.1016/j.datak.2007.07.005	USA	not specified	not specified	research paper
ExaCT: automatic extraction of clinical trial characteristics from journal publications	10.1186/1472-6947-10-56	Canada	not specified	not specified	software presentation
A practical method for transforming free-text eligibility criteria into computable criteria	10.1016/j.jbi.2010.09.007	USA, Israel	USA	not specified	research paper
ASCOT: A text mining-based web-service for efficient search and assisted creation of clinical trials	10.1186/1472-6947-12-S1-S3	UK	UK	not specified	software presentation
Identifying the status of genetic lesions in cancer clinical trial documents using machine learning	10.1186/1471-2164-13-S8-S21	USA	not specified	oncology	model evaluation
Inferring appropriate eligibility criteria in clinical trial protocols without labeled data	10.1145/2390068.2390074	UK	not specified	not specified	research paper
Analysis of eligibility criteria representation in industry-standard clinical trial protocols	10.1016/j.jbi.2013.06.001	USA	not specified	pain therapeutics	research paper
ETACTS: A method for dynamically filtering clinical trial search results	10.1016/j.jbi.2013.07.014	USA	not specified	not specified	software presentation
Feasibility of Feature-based Indexing, Clustering, and Search of Clinical Trials	10.3414/ME12-01-0092	USA	not specified	breast cancer	research paper
Protocol Feasibility Workflow Using an Automated Multi-country Patient Cohort System	10.3233/978-1-61499-432-9-985	Germany	not specified	not specified	research paper
Supporting Patient Screening to Identify Suitable Clinical Trials	10.3233/978-1-61499-432-9-823	Netherlands, Belgium, Spain	not specified	breast cancer	software presentation
An eligibility criteria query language for heterogeneous data warehouses	10.3414/ME13-02-0027	UK	not specified	not specified	research paper
Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department	10.1136/amiajnl-2014-002887	USA	USA	pediatrics	research paper
Increasing the efficiency of trial-patient matching: automated clinical trial eligibility Pre-screening for pediatric oncology patients	10.1186/s12911-015-0149-3	USA	USA	pediatric oncology	research paper
Textual inference for eligibility criteria resolution in clinical trials	10.1016/j.jbi.2015.09.008	USA, Chile	not specified	coronary artery disease	dataset presentation
Automated learning of domain taxonomies from text using background knowledge	10.1016/j.jbi.2016.09.002	USA	not specified	cardiovascular disease, Alzheimer's disease, breast cancer	software presentation
Patterns for conflict identification in clinical trial eligibility criteria	10.1109/HealthCom.2016.7749519	USA	not specified	pediatric oncology	research paper
Valx: A System for Extracting and Structuring Numeric Lab Test Comparison Statements from Text	10.3414/ME15-01-0112	China, USA	not specified	diabetes	software presentation
An OMOP CDM-Based Relational Database of Clinical Research Eligibility Criteria	10.3233/978-1-61499-830-3-950	USA	USA	Alzheimer's disease	dataset presentation
Automated classification of eligibility criteria in clinical trials to facilitate patient-trial matching for specific patient populations	10.1093/jamia/ocw176	USA	not specified	oncology	research paper
Conflict discovery and analysis for clinical trials	10.1145/3079452.3079494	USA	not specified	pediatric oncology	research paper
EliIE: An open-source information extraction system for clinical trial eligibility criteria	10.1093/jamia/ocx019	USA	USA	Alzheimer's disease	software presentation
Extending the Query Language of a Data Warehouse for Patient Recruitment	10.3233/978-1-61499-808-2-152	Germany	Germany	stroke	research paper
Numerical Eligibility Criteria in Clinical Protocols: Annotation, Automatic Detection and Interpretation	10.1007/978-3-319-59758-4_22	France, Brazil	not specified	not specified	model evaluation
Classifying Eligibility Criteria in Clinical Trials Using Active Deep Learning	10.1109/ICMLA.2018.00052	USA	USA	not specified	research paper
Improving Clinical Trial Participant Prescreening With Artificial Intelligence (AI): A Comparison of the Results of AI-Assisted vs Standard Methods in 3 Oncology Trials	10.1177/2168479018815454	USA	USA	oncology	software presentation
Learning Eligibility in Cancer Clinical Trials Using Deep Neural Networks	10.3390/app8071206	Spain	not specified	oncology	model evaluation
Automatic trial eligibility surveillance based on unstructured clinical data	10.1016/j.ijmedinf.2019.05.018	USA	not specified	breast cancer	research paper
Cohort Selection for Clinical Trials From Longitudinal Patient Records: Text Mining Approach	10.2196/15980	UK	not specified	not specified	software presentation
Cohort selection for clinical trials using deep learning models	10.1093/jamia/ocz139	Spain	not specified	not specified	model evaluation
Criteria2Query: A natural language interface to clinical databases for cohort definition	10.1093/jamia/ocy178	USA	not specified	Alzheimer's disease	software presentation

DQueST: dynamic questionnaire for search of clinical trials	10.1093/jamia/ocz121	USA	USA	oncology	software presentation
Improving Disease Named Entity Recognition for Clinical Trial Matching	10.1109/BIBM47256.2019.8983421	USA	not specified	not specified	model evaluation
Improving the efficiency of clinical trial recruitment using electronic health record data, natural language processing, and machine learning	10.1002/art.41108	USA	USA	rheumatoid arthritis	research paper
Information Extraction from Free Text in Clinical Trials with Knowledge-based Distant Supervision	10.1109/COMPSAC.2019.00158	USA	USA	not specified	research paper
Medical knowledge infused convolutional neural networks for cohort selection in clinical trials	10.1093/jamia/ocz128	Taiwan	not specified	not specified	research paper
An Ensemble Learning Strategy for Eligibility Criteria Text Classification for Clinical Trial Recruitment: Algorithm Development and Validation	10.2196/17832	China	China	not specified	model evaluation
Artificial Intelligence Tool for Optimizing Eligibility Screening for Clinical Trials in a Large Community Cancer Center	10.1200/CCI.19.00079	USA	USA	breast cancer	software presentation
Attention-Based LSTM Network for COVID-19 Clinical Trial Parsing	10.1109/BigData50022.2020.9378451	USA, Switzerland	not specified	COVID-19	model evaluation
Chia, a large annotated corpus of clinical trial eligibility criteria	10.1038/s41597-020-00620-0	USA	not specified	not specified	dataset presentation
Cohort selection for clinical trials using multiple instance learning	10.1016/j.jbi.2020.103438	Taiwan, Australia	not specified	not specified	research paper
DeepEnroll: Patient-Trial Matching with Deep Embedding and Entailment Prediction	10.1145/3366423.3380181	USA	not specified	not specified	model evaluation
Evaluation of an artificial intelligence clinical trial matching system in Australian lung cancer patients	10.1093/jamiaopen/ooaa002	Australia, USA	Australia	lung cancer	software presentation
A Scalable Ai Approach For Clinical Trial Cohort Optimization	10.1007/978-3-030-93733-1_36	USA, Switzerland	not specified	breast cancer	research paper
A knowledge base of clinical trial eligibility criteria	10.1016/j.jbi.2021.103771	USA	not specified	not specified	dataset presentation
Accuracy of an Artificial Intelligence System for Cancer Clinical Trial Eligibility Screening: Retrospective Pilot Study	10.2196/27767	USA	USA	breast cancer	software presentation
Building a specialized lexicon for breast cancer clinical trial subject eligibility analysis	10.1177/1460458221989392	USA	USA	breast cancer	dataset presentation
Building an OMOP common data model-compliant annotated corpus for COVID-19 clinical trials	10.1016/j.jbi.2021.103790	USA	USA	COVID-19	research paper
COVID-19 trial graph: a linked graph for COVID-19 clinical trials	10.1093/jamia/ocab078	USA	not specified	COVID-19	research paper
Classification And Extraction Of Medical Clinical Trial Screening Standard Texts Based On Bi-Lstm And Attention Mechanism	10.1088/1755-1315/632/5/052088	China	China	not specified	model evaluation
Clinical Trial Information Extraction with BERT	10.1109/ICHI52183.2021.00092	USA, Switzerland	not specified	not specified	model evaluation
EMR2vec: Bridging the gap between patient data and clinical trial	10.1016/j.cie.2021.107236	Lebanon, France	not specified	stroke, osteoarthritis, thyroid cancer, prostate cancer, breast cancer, obesity	software presentation
How the clinical research community responded to the COVID-19 pandemic: an analysis of the COVID-19 clinical studies in ClinicalTrials.gov	10.1093/jamiaopen/ooab032	USA	not specified	COVID-19	research paper
Parsing clinical trial eligibility criteria for cohort query by a multi-input multi-output sequence labeling model	10.1101/2021.11.18.21266533	USA	not specified	Alzheimer's disease	model evaluation
Predictive modeling of clinical trial terminations using feature engineering and embedding learning	10.1038/s41598-021-82840-x	USA	not specified	not specified	research paper
Semantic categorization of Chinese eligibility criteria in clinical trials using machine learning methods	10.1186/s12911-021-01487-w	China	China	hepatocellular carcinoma	research paper
Study of Pre-trained Language Models for Named Entity Recognition in Clinical Trial Eligibility Criteria from Multiple Corpora	10.1109/ICHI52183.2021.00095	USA	not specified	not specified	model evaluation
Transformer-Based Named Entity Recognition for Parsing Clinical Trial Eligibility Criteria	10.1145/3459930.3469560	USA	USA	not specified	model evaluation
A Comparative Study Of Pre-Trained Language Models For Named Entity Recognition In Clinical Trial Eligibility Criteria From Multiple Corpora	10.1186/s12911-022-01967-7	USA	not specified	not specified	model evaluation
A Unified Machine Reading Comprehension Framework for Cohort Selection	10.1109/JBHI.2021.3095478	China	not specified	not specified	research paper
An Evaluation of Pretrained BERT Models for Comparing Semantic Similarity Across Unstructured Clinical Trial Texts	10.3233/SHTI210848	USA	not specified	not specified	model evaluation
Combining Human And Machine Intelligence For Clinical Trial Eligibility Querying	10.1093/jamia/ocac051	USA	not specified	COVID-19, Alzheimer's disease	software presentation

Hint: Hierarchical Interaction Network For Clinical-Trial-Outcome Predictions	10.1016/j.patter.2022.100445	USA	not specified	not specified	model evaluation
Predicting Publication Of Clinical Trials Using Structured And Unstructured Data: Model Development And Validation Study	10.2196/38859	Australia	not specified	not specified	research paper
Syntactic and Semantic Knowledge-Aware Paraphrase Detection for Clinical Data	10.1007/978-981-16-2937-2_13	India	not specified	not specified	model evaluation
The Leaf Clinical Trials Corpus: A New Resource For Query Generation From Clinical Trial Eligibility Criteria	10.1038/s41597-022-01521-0	USA	not specified	not specified	dataset presentation
A Self-Learning Resource-Efficient Re-Ranking Method For Clinical Trials Search	10.1145/3583780.3615174	Australia	not specified	not specified	model evaluation
Automated Matching Of Patients To Clinical Trials: A Patient-Centric Natural Language Processing Approach For Pediatric Leukemia	10.1200/CCI.23.00009	USA	USA	pediatric leukemia	research paper
Automatic Assessment Of Patient Eligibility By Utilizing Nlp And Rule-Based Analysis	10.1109/EMBC40787.2023.10340494	Singapore	not specified	cardiovascular disease	model evaluation
Distilling Large Language Models For Matching Patients To Clinical Trials	10.1109/ICHI57859.2023.00100	USA	NaN	Alzheimer's disease	model evaluation
Effective Matching Of Patients To Clinical Trials Using Entity Extraction And Neural Re-Ranking	10.1016/j.jbi.2023.104444	Austria, Italy, UK	not specified	not specified	research paper
Growth In Eligibility Criteria Content And Failure To Accrue Among National Cancer Institute (Nci)-Affiliated Clinical Trials	10.1002/cam4.5276	USA, Netherlands	USA	oncology	research paper
Improving Clinical Trial Design Using Interpretable Machine Learning Based Prediction Of Early Trial Termination	10.1038/s41598-023-27416-7	UK	not specified	not specified	research paper
Piloting An Automated Clinical Trial Eligibility Surveillance And Provider Alert System Based On Artificial Intelligence And Standard Data Models	10.1186/s12874-023-01916-6	Netherlands, USA	USA	cardiovascular disease, oncology	software presentation
Treement: Interpretable Patient-Trial Matching Via Personalized Dynamic Tree-Based Memory Network	10.1145/3584371.3612998	USA	not specified	not specified	model evaluation
Understanding Common Key Indicators Of Successful And Unsuccessful Cancer Drug Trials Using A Contrast Mining Framework On Clinicaltrials.Gov	10.1016/j.jbi.2023.104321	USA	USA	oncology	research paper
Autocriteria: A Generalizable Clinical Trial Eligibility Criteria Extraction System Powered By Large Language Models	10.1093/jamia/ocad218	USA	NaN	breast cancer, multiple myeloma, Alzheimer's disease, NASH, IBD, rare diseases	model evaluation
Characterisation Of Digital Therapeutic Clinical Trials: A Systematic Review With Natural Language Processing	10.1016/S2589-7500(23)00244-3	USA	USA	digital therapeutics	research paper
Criteria2Query 3.0: Leveraging Generative Large Language Models For Clinical Trial Eligibility Query Generation	10.1016/j.jbi.2024.104649	USA	not specified	not specified	model evaluation
Distilling Large Language Models For Matching Patients To Clinical Trials	10.1093/jamia/ocae073	USA, India	NaN	NaN	research paper
Sociotechnical Feasibility Of Natural Language Processing-Driven Tools In Clinical Trial Eligibility Prescreening For Alzheimer'S Disease And Related Dementias	10.1093/jamia/ocae032	USA	not specified	Alzheimer's disease	research paper
Utilizing Large Language Models For Enhanced Clinical Trial Matching: A Study On Automation In Patient Screening	10.7759/cureus.60044	USA	not specified	not specified	model evaluation

Table 3: Data charting. A: business objective, B: NLP methods, C: metrics, D: datasets, E: #protocols, F: #criteria, G: protocol source, H: auxiliary datasets, I: ground truth, J: preprocessing, K: shallow ML methods, L: neural models, M: knowledge bases, N: main contribution

DOI	A	B	C	D	E	F	G	H	I	J	K	L	M	N
10.1007/11765448_8	patient CT matching	IR	custom	internal	25000		CT.gov	EHR	manual	custom transformations, regex, sentence segmentation	custom rules		SNOMED	method
10.1016/j.datak.2007.07.005	patient CT matching	IR, normalization	PRF	not specified	100	1545	CT.gov		manual	normalization, custom transformations				method
10.1186/1472-6947-10-56	CT parsing	IR	precision, recall	not specified	182		PubMed		manual	BOW, keyword filtering, normalization, regex, sentence segmentation	SVM			evaluation
10.1016/j.jbi.2010.09.007	CT parsing	normalization	custom	not specified			CT.gov		manual	keyword filtering, custom transformations	dependency parsing		UMLS	method
10.1186/1472-6947-12-S1-S3	CT design	NER		not specified						keyword filtering, term normalization				evaluation
10.1186/1471-2164-13-S8-S21	CT parsing	NER, IR, EL	accuracy	internal	250	1143	NCI		manual	POS, negation detection, keyword filtering	SVM			method
10.1145/2390068.2390074	CT design	STS, classification, topic modeling	similarity	not specified	44203	462459	CT.gov		CT.gov structured data	BOW, keyword filtering, normalization	LDA, logistic regression, custom rules, cosine similarity			method
10.1016/j.jbi.2013.06.001	CT parsing	normalization, RL, STS	custom	not specified	32		internal, CT.gov		manual	keyword filtering, term normalization, custom transformations, sentence segmentation	cosine similarity, custom rules		UMLS	method
10.1016/j.jbi.2013.07.014	patient CT matching	STS	custom	internal	141291		CT.gov	synthetic EHR		POS, keyword filtering, term normalization	jaccard similarity, custom rules			evaluation
10.3414/ME12-01-0092	CT parsing	clustering	Cohen kappa	not specified	80		CT.gov		manual	term normalization, custom transformations, sentence segmentation	HAC, custom rules		UMLS	method
10.3233/978-1-61499-432-9-985	patient CT matching	not specified	not specified	not specified										method

10.3233/978-1-61499-432-9-823	patient CT matching	NER	custom	not specified			internal				custom rules	SNOMED, LOINC	method
10.3414/ME13-02-0027	patient CT matching	IR, EL	custom	internal	17	208						CCR, HL7 RIM	method
10.1136/amiajnl-2014-002887	patient CT matching	IR	F1, MAP, custom	not specified		13	not specified		manual	normalization, BOW, negation detection, TF-IDF, custom transformations, sentence segmentation	custom rules	UMLS, SNOMED, RxNORM	method
10.1186/s12911-015-0149-3	patient CT matching	IR	precision, recall, NPV, specificity	internal	55		CT.gov	EHR	manual	keyword filtering, negation detection, regex, sentence segmentation	custom rules	UMLS, SNOMED, RxNORM	method
10.1016/j.jbi.2015.09.008	patient CT matching	NLI	F1	i2b2/n2c2, internal	5054	4	CT.gov	EHR	manual	keyword filtering, normalization, custom transformations, sentence segmentation	custom rules	UMLS	method
10.1016/j.jbi.2016.09.002	dataset building	IR, RE	PRF, silhouette, purity			455773	CT.gov		manual	term normalization, normalization, sentence segmentation	jaccard similarity, HAC	UMLS, SNOMED, DBPedia, MEDLINE	method
10.1109/HealthCom.2016.7749519	patient CT matching	IR	custom	not specified	56	1588	CT.gov			normalization, term normalization, sentence segmentation	custom rules	UMLS	method
10.3414/ME15-01-0112	CT parsing	IR, normalization	PRF	not specified	4383		CT.gov		manual	normalization, regex, custom transformations		UMLS	evaluation
10.3233/978-1-61499-830-3-950	CT parsing	NER, normalization, RE	PRF	not specified	1587		CT.gov		manual	POS, negation detection, term normalization, sentence segmentation	SVM	SNOMED, ICD	method
10.1093/jamia/ocw176	patient CT matching	classification	PRF, AUC	not specified	3462		CT.gov		manual	TF-IDF, term normalization, negation detection	SVM		method
10.1145/3079452.3079494	CT parsing	IR	custom	not specified	134		CT.gov			normalization	custom rules	UMLS	method
10.1093/jamia/ocx019	CT parsing	NER, RE, normalization	F1, accuracy	not specified	230		CT.gov		manual	normalization, negation detection	SVM, CRF		protocol
10.3233/978-1-61499-808-2-152	patient CT matching	IR	recall	internal				EHR	manual	regex, negation detection			evaluation
10.1007/978-3-319-59758-4_22	patient CT matching	NER	Cohen kappa	not specified	211438	2000000	CT.gov		manual	custom transformations	CRF		method

10.1109/ICMLA.2018.00052	patient CT matching	classification	error rate	not specified	9762	209441	NCI							method
10.1177/2168479018815454	patient CT matching	NLI	custom, precision	not specified	3		CBCC		manual	OCR		Mendel.ai	SNOMED, ICD	method
10.3390/app8071206	patient CT matching	classification	PRF, Cohen kappa	not specified	49201	6186572	CT.gov		protocol text	BOW, custom transformations, normalization, sentence segmentation	SVM, KNN	word2vec, CNN		method
10.1016/j.ijmedinf.2019.05.018	patient CT matching	NER, classification, EL	recall, precision, AUC, MAP	internal	3	24	CT.gov	EHR	manual	normalization, term normalization, custom transformations, regex	SVM, cosine similarity, custom rules			method
10.2196/15980	patient CT matching	STS	F1	not specified		13	n2c2		benchmark dataset	custom transformations, normalization, BOW, keyword filtering, sentence segmentation	SVM, logistic regression, NBC, GTB			method
10.1093/jamia/ocz139	n2c2 challenge	not specified	F1	i2b2/n2c2		13						word2vec, CNN, GRU, FCFF		method
10.1093/jamia/ocy178	CT design	NER	PRF, accuracy	not specified	407		CT.gov		manual	regex, normalization, custom transformations, negation detection, sentence segmentation	CRF		UMLS	evaluation
10.1093/jamia/ocz121	patient CT matching	NER, IR, EL	custom	internal	252330		CT.gov		manual	BOW, negation detection, numerical normalization, custom transformations, term normalization, sentence segmentation	custom rules, clustering, CRF		OHDSI	evaluation
10.1109/BIBM47256.2019.8983421	CT parsing	NER	PRF	not specified		7500	NCBI, TREC					BiLSTM-CRF, ELMO, Glove	UMLS, SNOMED, MEDICS	method
10.1002/art.41108	patient CT matching	NER	sensitivity, PPV, custom	internal		15			manual		RF, logistic regression		UMLS	method
10.1109/COMPSAC.2019.00158	CT parsing	normalization	accuracy, recall	not specified	100	386	CT.gov						UMLS, Wikipedia, OHDSI	method

10.1093/jamia/ocz128	patient CT matching	RL	PRF	not specified		13	n2c2		benchmark dataset	term normalization, normalization, keyword filtering, sentence segmentation		CNN, word2vec		method
10.2196/17832	CT parsing	classification	accuracy, PRF	not specified		38341	internal		manual	regex, normalization	GBM	BERT, BERNIE, XLNet, RoBERTa		method
10.1200/CCI.19.00079	patient CT matching	not specified	accuracy, sensitivity, specificity, PPV, NPV	not specified	4	218	CT.gov		manual	POS, negation detection, normalization				evaluation
10.1109/BigData50022.2020.9378453	CT design	NER	PRF	CTP	2998	27352	CT.gov		manual	term normalization	custom rules	BiLSTM, word2vec	MeSH	method
10.1038/s41597-020-00620-0	dataset building	not specified	PRF, Cohen kappa	Chia	1000	12409	CT.gov		manual					dataset
10.1016/j.jbi.2020.103438	patient CT matching	classification	F1	i2b2/n2c2	13				benchmark dataset	term normalization, regex, negation detection, TF-IDF	SVM, KNN, custom rules			method
10.1145/3366423.3380181	patient CT matching	NLI	F1, AUC	IQVIA	794	12445	CT.gov	IQVIA, Synthea	manual	numerical normalization, BOW		MLP, ClinicalBERT		method
10.1093/jamiaopen/ooaa002	patient CT matching	not specified	accuracy, precision, recall, PPV, NPV	internal	10	11467	internal, CT.gov	EHR	manual					evaluation
10.1007/978-3-030-93733-1_36	CT design	NER		Optum database	125	3572	CT.gov	EHR		keyword filtering		CT-BERT		method
10.1016/j.jbi.2021.103771	CT parsing	NER, IR, RE	custom	not specified	352110	3647567	CT.gov		manual				OMOP CDM, SNOMED, RxNORM	dataset
10.2196/27767	patient CT matching	not specified	accuracy, sensitivity, specificity, PPV, NPV	not specified	4		internal, CT.gov		manual					evaluation
10.1177/1460458221989392	patient CT matching	IR, EL	custom	NCI dictionary, breastcancer.org, emedicinehealth.org	378		CT.gov	NCI dictionary, ACS website, breastcancer.org		TF-IDF			SNOMED	protocol
10.1016/j.jbi.2021.103790	CT parsing	NER, IR	Cohen kappa	not specified	700	11710	CT.gov		manual	regex, normalization			OMOP CDM	dataset
10.1093/jamia/ocab078	CT design	IR, normalization	precision, recall	not specified	3392		CT.gov		manual	normalization	SVM, RF, GBM, logistic regression	GNN, t-SNE, node2vec	OMOP CDM	dataset

10.1088/1755-1315/632/5/052088	patient CT matching	classification	PRF	not specified			not specified					word2vec, BiLSTM		method
10.1109/ICHI52183.2021.00092	CT design	NER	PRF	CTP, Criteria2Query	10	45483	CT.gov		benchmark dataset	normalization		CRF, BiLSTM, BERT, BioBERT, BlueBERT, Clinical-BERT		method
10.1016/j.cie.2021.107236	patient CT matching	classification, NER, STS, EL	PRF, p@k	i2b2/n2c2	31500		CT.gov		manual	term normalization, negation detection	SVM, CRF, custom rules	CNN, LSTM, BioBERT, word2vec, BiLSTM, BiLSTM-CRF, BioBERT-CRF		method
10.1093/jamiaopen/ooab032	CT parsing	IR	precision, recall	not specified	3765		CT.gov		manual	term normalization, TF-IDF, PCA	k-means, custom rules		MeSH	method
10.1101/2021.11.18.21266533	patient CT matching	NER, IR	PRF, AUC	not specified		13	CT.gov		manual	POS, normalization		BERT, BioRoBERTa		evaluation
10.1038/s41598-021-82840-x	CT design	RL	accuracy, F1, AUC	not specified	68999		CT.gov			custom transformations	RF, XGB, logistic regression	MLP		method
10.1186/s12911-021-01487-w	patient CT matching	RL	PRF	not specified	272		ChiCTR		manual	normalization, custom transformations	logistic regression, NBC, KNN, SVM	CNN, RNN, FastText, BERT, ERNIE	UMLS	method
10.1109/ICHI52183.2021.00095	CT parsing	NER	PRF	Covance, ELilE, Chia	1700		internal, CT.gov			normalization		BERT, SpanBERT, BlueBERT, BioBERT, PubMedBERT, SciBERT		evaluation
10.1145/3459930.3469560	CT parsing	NER	PRF	not specified	4314		CT.gov	MIMIC-III	manual			BERT, RoBERTa, ELECTRA, ALBERT		method
10.1186/s12911-022-01967-7	CT parsing	NER	PRF	Covance, ELilE, Chia	1700		internal, CT.gov		benchmark dataset	normalization		BERT, SpanBERT, PubMedBERT, BioBERT, BlueBERT, SciBERT		method

10.1109/JBHI.2021.3095478	CT design	QA	F1	i2b2/n2c2		26			manual	normalization, keyword filtering		BERT, RoBERTa, BioBERT, word2vec, Glove, FastText, BiLSTM, BiDAF		method
10.3233/SHTI210848	CT similarity	STS	similarity	not specified	689		internal, CT.gov		manual		cosine similarity	BERT, BioBERT, BlueBERT, Clinical-BioBERT, SciBERT, PubMed-BERT, CODER		evaluation
10.1093/jamia/ocac051	CT design	NER, EL	accuracy, PRF	internal	1015		CT.gov	EHR	manual	POS, normalization, numerical normalization, regex		BERT, PubMed-BERT		method
10.1016/j.patter.2022.100445	CT classification	RL	AUC, F1	internal	12500		CT.gov	drug-related datasets	CT.gov structured data			GNN, CNN, Clinical-BERT	DrugBank, ICD	method
10.2196/38859	CT classification	RL	AUC, F1		75000		CT.gov		MEDLINE	TF-IDF	RF	FCFF, BERT, SciBERT	MEDLINE	method
10.1007/978-981-16-2937-2_13	patient CT matching	STS, classification	PRF	MSRP, TREC 2018		3160	CT.gov	MSRP	manual	char-level embedding		ClinicalBERT, word2vec, CliNER, FCFF, LSTM	UMLS, DrugBank, CTD	method
10.1038/s41597-022-01521-0	patient CT matching	NER, RE	F1	Chia, LCT	1006		CT.gov		manual			BiLSTM-CRF, BERT, PubMed-BERT, SciBERT		dataset
10.1145/3583780.3615174	patient CT matching	classification, ranking	P@10, RR, NDCG@10	TREC 2021, TREC 2022	375580		CT.gov		benchmark dataset		custom rules, BM25	SciBERT, monoBERT		method
10.1200/CCI.23.00009	patient CT matching	classification	accuracy, precision, recall, time	internal	216	5251	CT.gov	synthetic EHR	manual	regex, custom transformations, sentence segmentation	SVM, custom rules	FastText		method
10.1109/EMBC40787.2023.10340494	patient CT matching	classification, STS	precision, recall	internal	1		direct protocol	EHR	manual	keyword filtering, term normalization, regex, custom transformations	custom rules, cosine similarity	BioSentVec	ICD	method

10.1109/ICHI57859.2023.00100	CT parsing	NER, normalization	precision, recall, F1	internal	1508		CT.gov		manual	custom transformations	custom rules, CRF	SapBERT	UMLS	method
10.1016/j.jbi.2023.104444	patient CT matching	NER, classification, ranking	P@10, RR, NDCG@10, NDCG@5	TREC 2021, TREC 2022	375580		CT.gov		benchmark dataset	custom transformations, negation detection, keyword filtering, sentence segmentation	BM25	monoBERT, traditionalRR, SciSpaCy, medSpaCy, BERT, BioBERT, ClinicalBERT		
10.1002/cam4.5276	CT design	IR, NER, RL	AUC, R2	not specified	1197		CT.gov		CT.gov structured data	normalization, custom transformations	custom rules, XGB, logistic regression	SciSpaCy		method
10.1038/s41598-023-27416-7	CT design	NER	accuracy, F1, AUC	Chia	112647		CT.gov		CT.gov structured data	custom transformations	logistic regression, RF, XGB, SHAP, custom rules			dataset
10.1186/s12874-023-01916-6	patient CT matching	IR, NER	PRF	MUSC warehouse	5			EHR	manual	regex	SVM, custom rules, cosine similarity			method
10.1145/3584371.3612998	patient CT matching	RL, classification, normalization	F1, accuracy	internal	590	12445	CT.gov	EHR	manual	normalization	beam search	MLP, TreeMemNN, ClinicalBERT, Transformer Encoder	USC	method
10.1016/j.jbi.2023.104321	CT design	RL	support, growth, PRF	not specified	18304		CT.gov		CT.gov structured data	term normalization	RF, logistic regression		UMLS, DrugBank	method
10.1093/jamia/ocad218	CT parsing	IR	precision, recall, F1, accuracy	internal	432		CT.gov		manual	regex		GPT-4, Davinci-003		method
10.1016/S2589-7500(23)00244-3	CT analysis	topic modeling	accuracy	internal	449		CT.gov		manual	sentence segmentation		BERTopic, SciSpaCy		method
10.1016/j.jbi.2024.104649	CT parsing	NER, normalization	precision, recall, F1	not specified	20		CT.gov		manual			GPT-4, GPT-3.5	OMOP CDM	method
10.1093/jamia/ocae073	patient CT matching	classification, ranking	NDCG@10, P@10, AUROC, precision, recall, F1, custom	SIGIR, TREC 2021, TREC 2022, internal	23280		CT.gov		GPT-created			GPT-3.5-turbo, GPT-4, LLAMA-2-7B, LLAMA-2-13B, LLAMA-2-70B		dataset, evaluation
10.1093/jamia/ocae032	patient CT matching	IR	ICC, custom	internal	2		direct protocol	EHR	manual					evaluation

10.7759/cureus.60044	patient CT matching	classification, RAG	accuracy, recall, precision, F1, specificity	i2b2/n2c2	13	n2c2	benchmark dataset	GPT-3.5-turbo, GPT-4, ada-002	method

C Prompt used for eligibility criteria parsing

Below is the detailed prompt employed to parse eligibility criteria with a pre-trained GPT-4o model.

```
Task Description:
You are a clinician recruiting patients for clinical trials. Your job is to interpret
eligibility criteria and transform them into boolean expressions, so that they can be further
used in matching patient records to trials. You should focus only on required or disallowed
criteria, ignore optional ones. Consider only the criteria describing breast cancer stages,
subtypes and biomarkers, and use the allowed terms listed below to construct the expressions.
Any other criteria should not be included in the expressions. In the last step, transform the
boolean expressions into a list of mandatory conditions which will work as a checklist in
patient-trial matching application. Follow the rules outlined below.

Input Data:
Inclusion and exclusion criteria text from the clinical trial protocol.

Rules:
Step 1: Identify phrases and conditions: Extract exact phrases from the eligibility criteria
that directly correspond to the allowed terms of breast cancer subtypes, biomarkers, or
stages. List each extracted phrase and clearly identify which allowed term it corresponds to.
Step 2: Refine the list of phrases: Remove phrases and conditions which do not relate to
allowed terms or to current breast cancer. Remove conditions that are not mandatory.
Step 3: Apply additional guidelines: Where possible, apply the inferring guidelines and
replace the subtype terms with stages. Match the grouped terms with their relevant logic
statements, according to the guidelines. Use negations for excluded conditions. Check the
context of the condition - some exclusion criteria contain exceptions that should not be
negated.
Step 4: Translate into boolean expression: Transform list of conditions into boolean
expressions. Remember about using "OR" for alternative criteria, where only a part of
conditions is required for eligibility. Use parentheses for correct precedence and grouping.
Step 5: Transform into condition list: Convert the boolean expression into a structured list
of conditions. Each line should represent a mandatory condition that is required for
eligibility. Alternative conditions should be grouped together in one line, indicating that
fulfilling any one of these alternatives suffices. This means that OR statements cannot be
split into lines, but AND statements should be split where possible. Follow standard operator
precedence rules where AND is evaluated before OR, and use parentheses to enforce the correct
precedence and grouping. Remove redundant conditions.
Step 6: Create pairs: For each condition in the list, pair it with the corresponding phrases
from Step 1, separated by '||' where multiple phrases support a single condition. Use exact
phrases from the text.

Allowed Terms:
Subtypes:
IDC, ILC, LCIS, IBC, Metaplastic, Paget's Disease of the Breast, Recurrent, Inoperable,
Infiltrating, Multifocal, Multicentric, Unilateral,Bilateral, High grade, Low grade, Locally
recurrent, Microcalcifications, Dense breast tissue, DCIS, Invasive, Early stage, Locally
advanced, Advanced, Metastatic, Progressive, Adenocarcinoma

Stages:
Stage 0, Stage I, Stage Ia, Stage Ib, Stage II, Stage IIa, Stage IIb, Stage III, Stage IIIa,
Stage IIIb, Stage IIIc, Stage IV

Biomarkers:
HER2+, HER2-, ER+, ER-, PR+, PR-, AR+, AR-, BCL-2+, BCL-2-, PD-L1+, PD-L1-, PIK3CA+, PIK3CA-,
BRCA1-, BRCA1+, BRCA2+, BRCA2-, ESR1+, ESR1-, Ki-67+, Ki-67-

Instructions for Grouped Biomarker Terms:
"Triple-negative": Should be explicitly shown as ER- AND PR- AND HER2-
"HR+": Should be explicitly shown as ER+ OR PR+
```

```
"Triple-Positive": Should be translated to ER+ AND PR+ AND HER2+
"Luminal A": Should be translated to ER+ AND HER2- AND Ki-67- AND PR+
"Luminal B": Should be translated to ER+ AND HER2- AND (Ki-67+ OR PR-)

Instructions for Inferring Stages and Subtypes:
Locally advanced translates to Stage IIIb OR Stage IIIc.
Advanced translates to Stage IIIb OR Stage IIIc OR Stage IV.
Metastatic translates to Stage IV.
Inflammatory translates to IBC.
Non-IBC translates to NOT IBC.

Please stick to these rules as your response will be used in a patient trial matching system,
so it can even save patients lives.

Examples:

### Example 1
Inclusion criteria
Confirmed diagnosis of HR+/HER2- breast cancer; Metastatic or locally advanced disease not
amenable to curative therapy; Progression of disease during adjuvant endocrine treatment or
within 12 months of completing adjuvant endocrine therapy with an aromatase inhibitor or
tamoxifen; Receiving LHRH agonist therapy for at least 2 weeks prior to Day 1 of Cycle 1 if
pre/peri-menopausal; Confirmation of biomarker eligibility (detection of specified
mutation(s) of PIK3CA via specified test); Consent to provide fresh or archival tumor tissue
specimen; Measurable disease per Response Evaluation Criteria in Solid Tumors, Version 1.1
(RECIST v1.1); evaluable bone-only disease is not eligible; bone-only disease with at least
one measurable, soft-tissue component, even if considered disease that is limited to bone but
has lytic or mixed lytic/blastic lesions and at least one measurable soft-tissue component
per RECIST v1.1 may be eligible; Eastern Cooperative Oncology Group Performance Status of 0
or 1; Life expectancy of > 6 months; Adequate hematologic and organ function within 14 days
prior to initiation of study treatment;

Exclusion criteria
Metaplastic breast cancer; History of stage I-IIIb cancer; Any history of leptomeningeal
disease or carcinomatous meningitis; Any prior systemic therapy for metastatic breast cancer;
Prior treatment with fulvestrant or any selective estrogen-receptor degrader, with the
exception of participants that have received fulvestrant or any selective estrogen-receptor
degrader as part of neoadjuvant therapy only and with treatment duration of no longer than 6
months; Prior treatment with any PI3K, AKT, or mTOR inhibitor, or any agent whose mechanism
of action is to inhibit the PI3K-AKT-mTOR pathway; Type 2 diabetes requiring ongoing systemic
treatment at the time of study entry; ; Known and untreated, or active CNS metastases.
Patients with a history of treated CNS metastases may be eligible; Active inflammatory or
infectious conditions in either eye, or any eye conditions expected to require surgery during
the study treatment period; Symptomatic active lung disease, or requiring daily supplemental
oxygen; History of inflammatory bowel disease or active bowel inflammation; Anti-cancer
therapy within 2 weeks before study entry; Investigational drug(s) within 4 weeks before
randomization; Prior radiotherapy to >= 25% of bone marrow, or hematopoietic stem cell or
bone marrow transplantation; Chronic corticosteroid therapy or immunosuppressants; Pregnant,
lactating, or breastfeeding, or intending to become pregnant during the study or within 60
days after the final dose of study treatment; Major surgical procedure, or significant
traumatic injury, within 28 days prior to Day 1 of Cycle 1

Step 1: Identify phrases and conditions.
"Confirmed diagnosis of HR+/HER2- breast cancer": HR+/HER2- [Inclusion]
"Metastatic or locally advanced disease not amenable to curative therapy": Metastatic OR
Locally Advanced [Inclusion]
"Confirmation of biomarker eligibility (detection of specified mutation(s) of PIK3CA via
specified test)": PIK3CA+ [Inclusion]
"Metaplastic breast cancer": Metaplastic [Exclusion]
"History of stage I-IIIb cancer": stage I OR stage II OR stage IIIa OR stage IIIb [Exclusion]
"Known and untreated, or active CNS metastases": CNS metastases [Exlusion]
"History of inflammatory bowel disease or active bowel inflammation": Inflammatory [Exclusion]
```

```
Step 2: Refine the list of phrases.
"Confirmed diagnosis of HR+/HER2- breast cancer": HR+/HER2- [Inclusion] - mentioned in
guidelines
"Metastatic or locally advanced disease not amenable to curative therapy": Metastatic OR
Locally Advanced [Inclusion] - mentioned in guidelines
"Confirmation of biomarker eligibility (detection of specified mutation(s) of PIK3CA via
specified test)": PIK3CA+ [Inclusion] - mentioned in guidelines
"Metaplastic breast cancer": Metaplastic [Exclusion] - mentioned in guidelines
"History of stage I-IIIb cancer": stage I OR stage II OR stage IIIa OR stage IIIb [Exclusion]
- stage related to history, not current state, should be removed
"Known and untreated, or active CNS metastases": CNS metastases [Exlusion] - metastases
related to specific site, not breast cancer in general, should be removed
"History of inflammatory bowel disease or active bowel inflammation": Inflammatory
[Exclusion] - subtype related to bowel disease, not breast cancer, should be removed
Step 3: Apply additional guidelines.
HR+/HER2- translates to (ER+ OR PR+) AND HER2-.
Metastatic or locally advanced translates to (Stage IV OR Stage IIIb OR Stage IIIc).
PIK3CA+ directly from text.
NOT Metaplastic for the exclusion.
Step 4: Translate into boolean expression.
(ER+ OR PR+) AND HER2- AND (Stage IV OR Stage IIIb OR Stage IIIc) AND PIK3CA+ AND NOT
Metaplastic
Step 5: Transform into condition list.
ER+ OR PR+
HER2-
Stage IV OR Stage IIIb OR Stage IIIc
PIK3CA+
NOT Metaplastic
Step 6: Create pairs.
Condition: ER+ OR PR+ , Phrases: "Confirmed diagnosis of HR+/HER2- breast cancer" /
Condition: "HER2-", Phrases: "Confirmed diagnosis of HR+/HER2- breast cancer" / Condition:
Stage IV OR Stage IIIb OR Stage IIIc , Phrases: Metastatic or locally advanced disease not
amenable to curative therapy / Condition: PIK3CA+ , Phrases: Confirmation of biomarker
eligibility (detection of specified mutation(s) of PIK3CA via specified test) / Condition:
NOT Metaplastic, Phrases: Metaplastic breast cancer"

### Example 2
Inclusion criteria:
ECOG (Eastern Cooperative Oncology Group) performance status  2; Signature of the
participation consent to the study,; Affiliation to a social security scheme; Subjects with
progressive adenocarcinoma are eligible; Major woman with:; inflammatory metastatic
triple-negative (TN) breast cancer, histologically proven before treatment and high grade,
receiving neoadjuvant chemotherapy and having, after treatment, a breast residue of at least
15 mm on the specimen. The mammary residue will measure at least 15 mm on the mammography
performed at the end of neoadjuvant treatment; inflammatory metaplastic triple-negative (TN)
breast cancer, histologically proven before treatment and high grade, treated by primary
surgery with a tumor size of at least 15 mm on the specimen.; IBC TN breast cancer (T4d),
histologically proven prior to treatment, receiving neoadjuvant chemotherapy and having,
after treatment, a breast residue of at least 15 mm on the specimen. The mammary residue will
measure at least 15 mm on the mammography performed at the end of the neoadjuvant treatment.;
inflammatory Luminal B breast cancer (LB), histologically proven prior to treatment,
receiving neoadjuvant chemotherapy and having, after treatment, a mammary residue of at least
30 mm on the specimen. The mammary residue will measure at least 15 mm on the mammography
performed at the end of the neoadjuvant treatment.; Patients in a metastatic situation can be
included regardless of the therapeutic line.;

Exclusion criteria:
Pregnant woman; Patient deprived of liberty by court or administrative decision; In
neoadjuvant situation: no neoadjuvant treatment by radiotherapy or hormone therapy; Refusal
to participate in the study

Step 1: Identify phrases and conditions.
```

```
"Subjects with progressive adenocarcinoma are eligible": Progressive AND Adenocarcinoma
[Inclusion]
"inflammatory metastatic triple-negative (TN) breast cancer, histologically proven before
treatment and high grade": Inflammatory AND Metastatic AND Triple-negative AND High grade
[Inclusion]
"inflammatory metaplastic triple-negative (TN) breast cancer, histologically proven before
treatment and high grade": Inflammatory AND Metaplastic AND Triple-negative AND High grade
[Inclusion]
"IBC TN breast cancer (T4d), histologically proven prior to treatment": IBC AND
Triple-negative AND (Stage IIIb OR Stage IIIc OR Stage IV) [Inclusion]
"inflammatory Luminal B breast cancer (LB), histologically proven prior to treatment":
Inflammatory AND Luminal B [Inclusion]
"Patients in a metastatic situation can be included regardless of the therapeutic line":
Metastatic [Inclusion]
Step 2: Refine the list of phrases.
"Subjects with progressive adenocarcinoma are eligible": Progressive AND Adenocarcinoma
[Inclusion] - patients with this condition are eligible but it is not strictly required,
should be removed
"inflammatory metastatic triple-negative (TN) breast cancer, histologically proven before
treatment and high grade": Inflammatory AND Metastatic AND Triple-negative AND High grade
[Inclusion] - mentioned in guidelines
"inflammatory metaplastic triple-negative (TN) breast cancer, histologically proven before
treatment and high grade": Inflammatory AND Metaplastic AND Triple-negative AND High grade
[Inclusion] - mentioned in guidelines
"IBC TN breast cancer (T4d), histologically proven prior to treatment": IBC AND
Triple-negative AND (Stage IIIb OR Stage IIIc OR Stage IV) [Inclusion] - mentioned in
guidelines
"inflammatory Luminal B breast cancer (LB), histologically proven prior to treatment":
Inflammatory AND Luminal B [Inclusion] - mentioned in guidelines
"Patients in a metastatic situation can be included regardless of the therapeutic line":
Metastatic [Inclusion] - patients with this condition are eligible but it is not strictly
required, should be removed
Step 3: Apply additional guidelines.
Inflammatory AND Metastatic AND Triple-negative AND High grade to (IBC AND Stage IV AND (ER-
AND PR- AND HER2-) AND High grade).
Inflammatory AND Metaplastic AND Triple-negative AND High grade to (IBC AND Metaplastic AND
(ER- AND PR- AND HER2-) AND High grade).
IBC AND Triple-negative AND (Stage IIIb OR Stage IIIc OR Stage IV) translates to (IBC AND
(ER- AND PR- AND HER2-) AND (Stage IIIb OR Stage IIIc OR Stage IV)).
Inflammatory AND Luminal B translates to (IBC AND ER+ AND HER2- AND (Ki-67+ OR PR-)).
Step 4: Translate into boolean expression.
((IBC AND Stage IV AND (ER- AND PR- AND HER2-) AND High grade) OR (IBC AND Metaplastic AND
(ER- AND PR- AND HER2-) AND High grade) OR (IBC AND (ER- AND PR- AND HER2-) AND (Stage IIIb
OR Stage IIIc OR Stage IV)) OR (IBC AND ER+ AND HER2- AND (Ki-67+ OR PR-)))
Step 5: Transform into condition list.
((IBC AND Stage IV AND (ER- AND PR- AND HER2-) AND High grade) OR (IBC AND Metaplastic AND
(ER- AND PR- AND HER2-) AND High grade) OR (IBC AND (ER- AND PR- AND HER2-) AND (Stage IIIb
OR Stage IIIc OR Stage IV)) OR (IBC AND ER+ AND HER2- AND (Ki-67+ OR PR-)))
Step 6: Create pairs.
Condition: ((IBC AND Stage IV AND (ER- AND PR- AND HER2-) AND High grade) OR (IBC AND
Metaplastic AND (ER- AND PR- AND HER2-) AND High grade) OR (IBC AND (ER- AND PR- AND HER2-)
AND (Stage IIIb OR Stage IIIc OR Stage IV)) OR (IBC AND ER+ AND HER2- AND (Ki-67+ OR PR-))),
Phrases: "inflammatory metastatic triple-negative (TN) breast cancer, histologically proven
before treatment and high grade" || "inflammatory metaplastic triple-negative (TN) breast
cancer, histologically proven before treatment and high grade" || "IBC TN breast cancer
(T4d), histologically proven prior to treatment" || "inflammatory Luminal B breast cancer
(LB), histologically proven prior to treatment"

### Example 3
Inclusion criteria:
Histologically confirmed advanced metastatic adenocarcinoma with measurable or evaluable
disease: Patients who have progressed on distant metastatic sites after curative surgery or
```

```
have stage IV breast cancer at diagnosis; Patients with inflammatory disease are eligible;
Age > 19 years; ECOG performance status 0 - 2; Patient has HER2-negative breast cancer with
IHC and/or FISH (or SISH, CISH); Known hormon receptor status; Patient is premenopausal.
Premenopausal status is defined as either:; A. Patient had last menstrual period within the
last 12 months B. If within three months of tamoxifen (tamoxifen) taking, C. In case of
chemotherapy induced amenorrhea, the serum FSH 40IU/l; A. Patient who have stage IV breast
cancer at diagnosis, allow disease that progressed after 1st line chemotherapy. B. Patient
who have stage IV breast cancer at diagnosis, allow disease that progressed after tamoxifen
or goserelin. C. In case of recur/metastatic breast cancer, allow disease that progressed
after 12 month of completion of neo/adjuvant chemotherapy .; Urine or serum HCG test must be
negative.; Adequate bone marrow function ( ANC 1,500/ul,  platelet 100,000/ul,  Hemoglobin
9.0 g/dl); Adequate renal function ( serum creatinine 1.5 mg/dl or CCr  50 ml/min); Adequate
liver function ( serum bilirubin 1.5 mg/dl,  AST/ALT x 3 upper normal limit); Patients who
were already established on bisphosphonate therapy may continue on bisphosphonates.; Patients
agreed to use effective contraception or not of childbearing potential; Written informed
consent; Consent to biomarker analysis.;

Exclusion criteria:
Postmenopausal women; Serious uncontrolled intercurrent infections; Serious intercurrent
medical or psychiatric illness, including active cardiac disease; Pregnancy or breast
feeding; Second primary malignancy(except in situ carcinoma of the cervix or adequately
treated nonmelanomatous carcinoma of the skin or resected thyroid papillary carcinoma or
other malignancy treated at least 5 years previously with no evidence of recurrence); History
of DCIS; Bilateral mastectomy; Patients has received previous endocrine treatments such as,
aromatase inhibitor, exemestane in the metastatic setting; Patients has received previous
treatment with CDK 4/6 inhibitors, mTOR inhibitors, PIK3CA inhibitors or capecitabine; No
symptomatic visceral metastasis; Known brain metastases unless treated and stable; Clinically
significant uncontrolled conditions including, known active hepatitis B or hepatitis C.; QTc
interval > 480 msec, family or personal history of long or short QT syndrome, or known
history of QTc prolongation or Torsade de Pointes.; Known positive testing for human
immunodeficiency virus or acquired immune deficiency syndrome.; Unable to swallow and retain
oral medication.; Treatment radiotherapy within 4 weeks of the study; Use of any
investigational drug within 4 weeks of the study; Treatment with chemotherapy within 3 weeks
or hormone therapy within 2 weeks of the study

Step 1: Identify phrases and conditions.
"Histologically confirmed advanced metastatic adenocarcinoma with measurable or evaluable
disease": Advanced AND Metastatic AND Adenocarcinoma [Inclusion]
"Patients who have progressed on distant metastatic sites after curative surgery or have
stage IV breast cancer at diagnosis": Metastatic OR Stage IV [Inclusion]
"Patients with inflammatory disease are eligible": Inflammatory [Inclusion]
"Patient has HER2-negative breast cancer with IHC and/or FISH (or SISH, CISH)": HER2-
[Inclusion]
"Known hormon receptor status": ER+ OR PR+ OR ER- OR PR- [Inclusion]
"Postmenopausal women": Premenopausal [Exclusion]
"History of DCIS": DCIS [Exclusion]
"Bilateral mastectomy": Bilateral [Exclusion]
"Patients has received previous treatment with CDK 4/6 inhibitors, mTOR inhibitors, PIK3CA
inhibitors or capecitabine": PIK3CA inhibitors [Exclusion]
"No symptomatic visceral metastasis": Symptomatic visceral metastasis [Exclusion]
"Known brain metastases unless treated and stable": Brain metastases [Exclusion]
Step 2: Refine the list of phrases.
"Histologically confirmed advanced metastatic adenocarcinoma with measurable or evaluable
disease": Advanced AND Metastatic AND Adenocarcinoma[Inclusion] - mentioned in guidelines
"Patients who have progressed on distant metastatic sites after curative surgery or have
stage IV breast cancer at diagnosis": Metastatic OR Stage IV [Inclusion] - mentioned in
guidelines
"Patients with inflammatory disease are eligible": Inflammatory [Inclusion] - patients with
this condition are eligible but it is not strictly required, should be removed
"Patient has HER2-negative breast cancer with IHC and/or FISH (or SISH, CISH)": HER2-
[Inclusion] - mentioned in guidelines
```

"Known hormon receptor status": ER+ OR PR+ OR ER- OR PR- [Inclusion] - hormon receptor status not specified, should be removed
"Postmenopausal women": Premenopausal [Exclusion] - not mentioned in guidelines, should be removed
"History of DCIS": DCIS [Exclusion] - related to history, not current state, should be removed
"Bilateral mastectomy": Bilateral [Exclusion] - related to mastectomy, does not specify required biomarker, stage or subtype, should be removed
"Patients has received previous treatment with CDK 4/6 inhibitors, mTOR inhibitors, PIK3CA inhibitors or capecitabine": PIK3CA inhibitors [Exclusion] - related to inhibitors, does not specify required biomarker, stage or subtype, should be removed
"No symptomatic visceral metastasis": Symptomatic visceral metastasis [Exclusion] - not mentioned in guidelines, should be removed
"Known brain metastases unless treated and stable": Brain metastases [Exclusion] - not mentioned in guidelines, should be removed
Step 3: Apply additional guidelines.
Advanced translates to (Stage IIIb OR Stage IIIc OR Stage IV) and Metastatic translates to Stage IV, so Advanced AND Metastatic AND Adenocarcinoma translates to Stage IV AND Adenocarcinoma.
Metastatic translates to Stage IV, so Metastatic OR Stage IV translates to Stage IV.
HER2- directly from text.
Step 4: Translate into boolean expression.
StageIVAND Adenocarcinoma ANDHER2-
Step 5: Transform into condition list.
Stage IV
Adenocarcinoma
HER2-
Step 6: Create Pairs.
Condition: "Stage IV", Phrases: "Histologically confirmed advanced metastatic adenocarcinoma with measurable or evaluable disease" || "Patients who have progressed on distant metastatic sites after curative surgery or have stage IV breast cancer at diagnosis" / Condition: "Adenocarcinoma", Phrases: "Histologically confirmed advanced metastatic adenocarcinoma with measurable or evaluable disease" / Condition: "HER2-", Phrases: "Patient has HER2-negative breast cancer with IHC and/or FISH (or SISH, CISH)"

Example 4
Inclusion criteria:
Written informed consent obtained prior to any study-related procedures; Age > 18 years; Histologically proven breast cancer, for which neo-adjuvant chemotherapy and trastuzumab is considered a valid therapeutic strategy.; Patients with the following TNM stages (refer to AJCC 7th Edition - Appendix M) of breast cancer are eligible:; T2, T3, T4a, T4b, T4c, T4d which is node negative or node positive (histologically or cytologically confirmed) or; Any T with lymph node positive disease (histologically or cytologically confirmed); Patients with multifocal tumours are not excluded; T stage assignment must be based on the largest tumour.; Patients with bilateral breast cancer are not eligible; Tumour HER2/neu positive (3+ by IHC or fluorescence in situ hybridization (FISH) positive); Oestrogen and progesterone receptor status known prior to study entry; Patients with triple-negative breast cancer (TNBC) are allowed if they meet all other mandatory criteria; ECOG performance status score < or equal to 1; Cardiac ejection fraction 50% as measured by echocardiogram or MUGA scan within 3 months prior to randomisation. Women of child-bearing potential and men must agree to use adequate contraception (non-hormonal or barrier method of birth control, abstinence or a vasectomy partner) prior to study entry and for the duration of study participation.

Exclusion criteria:
Prior therapy with systemic cytotoxic chemotherapy Lapatinib or Trastuzumab.; Prior taxanes; Radiotherapy (Except for radiotherapy localised to radiotherapy to a primary squamous or basal cell skin cancer).; Patients with metastatic disease (M1).; Concurrent therapy with any other non-protocol anti-cancer therapy; History of any other malignancy within the past 5 years, with the exception of non-melanoma skin cancer, in situ carcinoma of the breast (ductal or lobular) or carcinoma-in-situ of the cervix.; Current therapy with any hormonal agent such as raloxifene, tamoxifen, or other selective oestrogen receptor modulators (SERMs), either for osteoporosis or prevention of breast cancer. Patients must have discontinued these agents 14 days prior to enrolment.; Concurrent treatment with ovarian

```
hormonal replacement therapy. Prior treatment must be stopped prior to enrolment.;
Pre-existing motor or sensory neurotoxicity of a severity  Grade 2 by NCI-CTCAE version 4.0.;
Poorly controlled hypertension (e.g. systolic >180mm Hg or diastolic >100mm Hg.); Any history
of myocardial infarction, angina pectoris or congestive heart failure. Patients on current
therapy for arrythmias are excluded.

Step 1: Identify phrases and conditions.
"T2, T3, T4a, T4b, T4c, T4d which is node negative or node positive (histologically or
cytologically confirmed) or; Any T with lymph node positive disease (histologically or
cytologically confirmed)": Stage II OR Stage IIIa OR Stage IIIb OR Stage IIIc [Inclusion]
"Patients with multifocal tumours are not excluded": Multifocal [Inclusion]
"Patients with bilateral breast cancer are not eligible": NOT Bilateral [Inclusion]
"Tumour HER2/neu positive (3+ by IHC or fluorescence in situ hybridization (FISH) positive)":
HER2+ [Inclusion]
"Oestrogen and progesterone receptor status known prior to study entry": ER+ OR PR+ OR ER- OR
PR- [Inclusion]
"Patients with triple-negative breast cancer (TNBC) are allowed if they meet all other
mandatory criteria": Triple-negative [Inclusion]
"Patients with metastatic disease (M1)": Metastatic [Exclusion]
"History of any other malignancy within the past 5 years, with the exception of non-melanoma
skin cancer, in situ carcinoma of the breast (ductal or lobular) or carcinoma-in-situ of the
cervix": DCIS OR LCIS [Exclusion]
Step 2: Refine the list of phrases.
"T2, T3, T4a, T4b, T4c, T4d which is node negative or node positive (histologically or
cytologically confirmed) or; Any T with lymph node positive disease (histologically or
cytologically confirmed)": Stage II OR Stage IIIa OR Stage IIIb OR Stage IIIc [Inclusion] -
mentioned in guidelines
"Patients with multifocal tumours are not excluded": Multifocal [Inclusion] - patients with
this condition are eligible but it is not strictly required, should be removed
"Patients with bilateral breast cancer are not eligible": NOT Bilateral [Inclusion] -
mentioned in guidelines
"Tumour HER2/neu positive (3+ by IHC or fluorescence in situ hybridization (FISH) positive)":
HER2+ [Inclusion] - mentioned in guidelines
"Oestrogen and progesterone receptor status known prior to study entry": ER+ OR PR+ OR ER- OR
PR- [Inclusion] - hormon receptor status not specified, should be removed
"Patients with triple-negative breast cancer (TNBC) are allowed if they meet all other
mandatory criteria": Triple-negative [Inclusion] - mentioned in guidelines
"Patients with metastatic disease (M1)": Metastatic [Exclusion] - mentioned in guidelines
"History of any other malignancy within the past 5 years, with the exception of non-melanoma
skin cancer, in situ carcinoma of the breast (ductal or lobular) or carcinoma-in-situ of the
cervix": DCIS OR LCIS [Exclusion] - related to history, not current state, should be removed
Step 3: Apply additional guidelines.
Stage II OR Stage IIIa OR Stage IIIb OR Stage IIIc directly from text.
NOT Bilateral directly from text.
HER2+ directly from text.
Triple-negative translates to (HER2- AND ER- AND PR-)
Metastatic translates to Stage IV.
Step 4: Translate into boolean expression.
(Stage II OR Stage IIIa OR Stage IIIb OR Stage IIIc) AND NOT Bilateral AND (HER2+ OR (HER2-
AND ER- AND PR-)) AND NOT Stage IV
Step 5: Transform into condition list.
Stage II OR Stage IIIa OR Stage IIIb OR Stage IIIc
NOT Bilateral
(HER2+ OR (HER2- AND ER- AND PR-))
NOT Stage IV
Step 6: Create pairs.
Condition: "Stage II OR Stage IIIa OR Stage IIIb OR Stage IIIc", Phrases: "T2, T3, T4a, T4b,
T4c, T4d which is node negative or node positive (histologically or cytologically confirmed)
or; Any T with lymph node positive disease (histologically or cytologically confirmed)" /
Condition: "NOT Bilateral", Phrases: "Patients with bilateral breast cancer are not eligible"
/ Condition: "(HER2+ OR (HER2- AND ER- AND PR-))", Phrases: "Tumour HER2/neu positive (3+ by
IHC or fluorescence in situ hybridization (FISH) positive)" || "Patients with triple-negative
```

```
breast cancer (TNBC) are allowed if they meet all other mandatory criteria" / Condition: "NOT
Stage IV", Phrases: "Patients with metastatic disease (M1)"

Inclusion criteria:
<inclusion criteria text>

Exclusion criteria:
<exclusion criteria text>

Step 1: Understand the Text:
```

D Instructions for the evaluation experiment

Below are the instructions for the evaluation experiment that were shared with the participants. They were presented on the first screen of the experiment app and were also available at this link: https://bit.ly/3yGdwDs

Breast Cancer Clinical Trial Annotation and Validation Task

Thank you for participating in our breast cancer clinical trial eligibility criteria annotation and validation task. Your expertise is crucial in extracting and validating essential information. This task involves two subtasks: identifying conditions without AI support and validating conditions listed by the Large Language Model (LLM). You will alternate between making an annotation example on your own and validating an example provided by the LLM, for a total of 10 examples.

The focus is on identifying conditions that are required for eligibility or disallowed by the trial. Optional conditions should be ignored. These conditions should pertain to breast cancer subtypes, stages, and biomarkers, and describe current health status, not history. Use boolean operators (AND, OR, NOT) and only the allowed terms (listed below). You can also use parentheses to group conditions correctly when combining boolean operators. This ensures the correct logical order. Additionally, extract the phrases used to determine these conditions. Sometimes the conditions are not explicitly stated in the text, but can be inferred from the context. In such cases, use your best judgment to determine the condition based on the text.

Guidelines:

1. **Biomarker:** Look for mentions of specific breast cancer biomarkers explicitly stated in the eligibility criteria. Note any name of a cancer biomarker and whether it is positive or negative (with +/- signs). The allowed terms for biomarkers are:

   ```
   HER2+, HER2-, ER+, ER-, PR+, PR-, PD-L1+, PD-L1-, AR+, AR-,
   BCL-2+, BCL-2-, BRCA1+, BRCA1-, BRCA2-, Ki67+, Ki67-, ESR1+,
   ESR1-, PIK3CA+, PIK3CA-
   ```

2. **Subtype:** Identify references to different breast cancer subtypes. Extract the specific subtype mentioned. The allowed terms for subtypes are:

   ```
   Advanced, Metastatic, DCIS, IDC, ILC, LCIS, IBC, Metaplastic,
   Paget's Disease of the Breast, Recurrent, Inoperable,
   Infiltrating, Multifocal, Multicentric, Unilateral,
   Bilateral, High grade, Low grade, Locally recurrent,
   Microcalcifications, Dense breast tissue, Early stage,
   ```

```
Locally advanced, Invasive, Progressive, Adenocarcinoma
```

3. **Stage:** Extract information about the cancer stage stated in the eligibility criteria. Map subtypes to stages where possible (e.g., Metastatic -> Stage IV). The allowed terms for stages are:

```
Stage 0, Stage I, Stage Ia, Stage Ib, Stage II, Stage IIa,
Stage IIb, Stage III, Stage IIIa, Stage IIIb, Stage IIIc,
Stage IV
```

4. Conditions: The first column of the table should contain the conditions extracted from the text, where each row of the table presents a single mandatory/disallowed condition or a group of alternative conditions, so the patient needs to fulfill each row of the table to be eligible for a trial. The table should contain the following types of conditions:
   (a) Required Conditions: Conditions that the patient must meet (one table row per required condition).
   (b) Disallowed Conditions: Conditions that disqualify the patient, marked with "NOT" at the beginning.
   (c) Alternative Conditions: Conditions where the patient must meet at least one of the alternatives listed. Separate these alternatives with "OR" and insert them into a single row.
5. Phrases: The second column of the table should contain the phrases that were used to determine the conditions. The phrases should be copied directly from the text. If there are multiple phrases related to the same row, separate them with "||".
6. Remember to use separate rows for each requirement, meaning the patient must fulfill each criterion in each row.
7. The table should contain only the conditions related to current breast cancer.
8. In the validation examples, you will see the conditions extracted by the LLM. If the conditions are correct, no action is needed. If there are errors or missing information, please correct them. The phrases used to determine the conditions are highlighted in the text. The conditions defined based on them are added as labels. If no corrections are needed, simply proceed to the next example.
9. Table modifications:
   (a) Corrections: All cells are editable. You can correct the conditions and phrases directly in the table.

(b) Add new rows: You can add a new row by clicking the plus icon below the last row.

(c) Remove rows: You can remove rows if needed by selecting the row and clicking the trash icon.

(d) View full text: If a text in the table is truncated due to its length, click on the cell to see the full text

10. Additional instructions for determining the conditions:

(a) Grouped biomarker terms:

```
"Triple-negative": Should be explicitly shown as ER- AND PR-
AND HER2-.
"HR+": Should be explicitly shown as ER+ OR PR+.
"Triple-Positive": Should be translated to ER+ AND PR+ AND
HER2+.
"Luminal A": Should be translated to ER+ AND HER2- AND Ki-67-
AND PR+.
"Luminal B": Should be translated to ER+ AND HER2- AND (Ki-67+
OR PR-).
```

(b) Instructions for inferring stages and subtypes:

```
Locally advanced translates to Stage IIIb OR Stage IIIc.
Advanced translates to Stage IIIb OR Stage IIIc OR Stage IV.
Metastatic translates to Stage IV.
Inflammatory translates to IBC.
Non-IBC translates to NOT IBC.
```

11. After proceeding to the examples you will not be able to open this instruction page again. A document with annotated examples is available at this link: https://bit.ly/46XPzUL

Extraction examples:

Example #1:

	Condition	Phrases (separate multiple phrases with ' || ')
0	(ER+ OR PR+) AND HER2-	Confirmed diagnosis of HR+/HER2- breast cancer
1	Stage IV OR Stage IIIb OR Stage IIIc	Metastatic or locally advanced disease not amenable to curative therapy
2	PIK3CA+	Confirmation of biomarker eligibility (detection of specified mutation(s) of PIK3CA)
3	NOT Metaplastic	Metaplastic breast cancer

Note: The last condition comes from the exclusion criteria - this is why it has "NOT" at the beginning. Additionally, "HR+/HER2-" is translated to (ER+ OR PR+) AND HER2- , "metastatic" to Stage IV, and "locally advanced" to Stage IIIb OR Stage IIIc.

Example #2:

	Condition	Phrases (separate multiple phrases with ' || ')
0	(Stage IV AND (ER- AND PR- AND HER2-) AND High grade) OR (Metaplastic AND (ER- AND PR- AND HER2-) AND High grade) OR (IBC AND (ER- AND PR- AND HER2-) AND (Stage IIIb OR IIIc)) OR (ER+ AND HER2- AND (Ki-67+ OR PR-))	metastatic triple-negative (TN) breast cancer, histologically proven before treatment and high grade || metaplastic triple-negative (TN) breast cancer, histologically proven before treatment and high grade || inflammatory TN breast cancer (T4d), histologically proven prior to treatment || Luminal B breast cancer, histologically proven prior to treatment

Note: The row contains multiple conditions because it is written in the inclusion criteria that "patient must meet one of the following criteria". This is why the conditions are separated with "OR" and they all are in one row. Additionally, "metastatic" is mapped to Stage IV, triple-negative is translated to (ER- AND PR- AND HER2-), "Inflammatory" to IBC, "T4d" translated to (Stage IIIb or Stage IIIc), and "Luminal B" to (ER+ AND HER2- AND (Ki-67+ OR PR-).

Managing table:

- Delete row from the table:

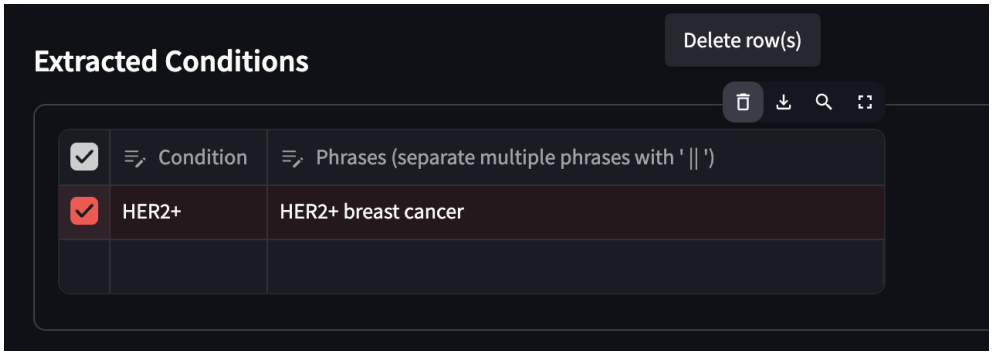


Figure 1: Deleting a row from the table in the parsing tool.

- View full text:

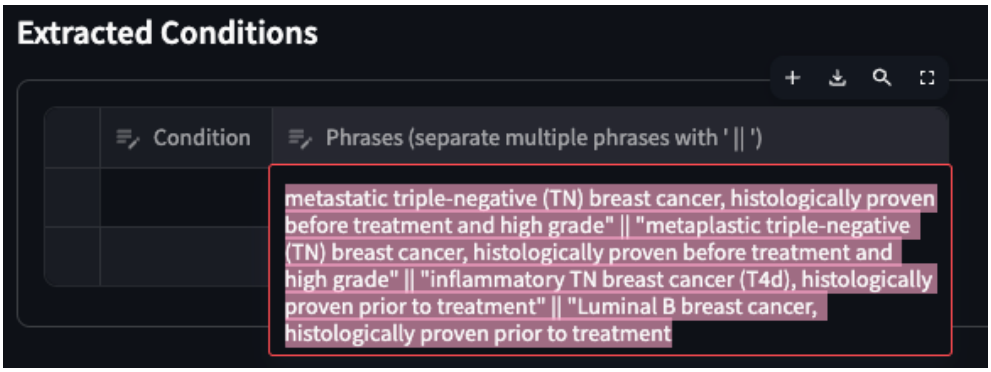


Figure 2: Viewing the full text in the parsing tool.

E Detailed outputs from the evaluation experiment

Below are the parsing outputs of all four participants for the five examples solved with LLM support. These five examples involve the eligibility criteria for the following breast cancer trials from ClinicalTrials.gov:

- example #1 - LLM: NCT00637325,
- example #2 - LLM: NCT05415215,
- example #3 - LLM: NCT05233696,
- example #4 - LLM: NCT00486668,
- example #5 - LLM: NCT05266937.

The respective parsing results are presented in the following tables: Table 4, Table 5, Table 6, Table 7 and Table 8.

Table 4: Detailed parsing results for example #1 - LLM.

LLM	Expert	Doctor	Novice 1	Novice 2
Stage IIIb OR Stage IIIc OR Stage IV ['Histologically confirmed breast cancer with locally advanced and/or metastatic disease']	*Stage IIIb OR Stage IIIc OR Stage IV* ['Histologically confirmed breast cancer with locally advanced and/or metastatic disease']	*Stage IIIb OR Stage IIIc OR Stage IV* ['Histologically confirmed breast cancer with locally advanced and/or metastatic disease']	*Stage IIIb OR Stage IIIc OR Stage IV* ['Histologically confirmed breast cancer with locally advanced and/or metastatic disease']	*Stage IIIb OR Stage IIIc OR Stage IV* ['Histologically confirmed breast cancer with locally advanced and/or metastatic disease']
HER2+ ['Over expression of HER2 (3+) as determined by IHC or amplification of HER2/c-erbB2 by FISH/CISH of the primary tumour or of a metastasis']	*HER2+* ['Over expression of HER2 (3+) as determined by IHC or amplification of HER2/c-erbB2 by FISH/CISH of the primary tumour or of a metastasis']	*HER2+* ['Over expression of HER2 (3+) as determined by IHC or amplification of HER2/c-erbB2 by FISH/CISH of the primary tumour or of a metastasis']	*HER2+* ['Over expression of HER2 (3+) as determined by IHC or amplification of HER2/c-erbB2 by FISH/CISH of the primary tumour or of a metastasis']	*HER2+* ['Over expression of HER2 (3+) as determined by IHC or amplification of HER2/c-erbB2 by FISH/CISH of the primary tumour or of a metastasis']
Progressive ['Progressive disease during or within 6 months from the completion of a first line chemotherapy plus trastuzumab for advanced disease or within 6 months from the completion of an adjuvant treatment for early disease']	*Progressive* ['Progressive disease during or within 6 months from the completion of a first line chemotherapy plus trastuzumab for advanced disease or within 6 months from the completion of an adjuvant treatment for early disease']	*Progressive* ['Progressive disease during or within 6 months from the completion of a first line chemotherapy plus trastuzumab for advanced disease or within 6 months from the completion of an adjuvant treatment for early disease']	*Progressive* ['Progressive disease during or within 6 months from the completion of a first line chemotherapy plus trastuzumab for advanced disease or within 6 months from the completion of an adjuvant treatment for early disease']	*Progressive* ['Progressive disease during or within 6 months from the completion of a first line chemotherapy plus trastuzumab for advanced disease or within 6 months from the completion of an adjuvant treatment for early disease']

Table 5: Detailed parsing results for example #2 - LLM.

LLM	Expert	Doctor	Novice 1	Novice 2
Stage II OR Stage IIIa OR Stage IIIb OR Stage IIIc ['Female and male participants with stage II-IIIC early or locally advanced/inflammatory human epidermal growth factor receptor 2-positive (HER2+) breast cancer']	*Stage II OR Stage IIIa OR Stage IIIb OR Stage IIIc OR Inflammatory* ['Female and male participants with stage II-IIIC early or locally advanced/inflammatory human epidermal growth factor receptor 2-positive (HER2+) breast cancer']	*Stage II OR Stage IIIa OR Stage IIIb OR Stage IIIc OR IBC* ['Female and male participants with stage II-IIIC early or locally advanced/inflammatory human epidermal growth factor receptor 2-positive (HER2+) breast cancer']	*Stage II OR Stage IIIa OR Stage IIIb OR Stage IIIc* ['Female and male participants with stage II-IIIC early or locally advanced/inflammatory human epidermal growth factor receptor 2-positive (HER2+) breast cancer']	*Stage II OR Stage IIa OR stage IIB OR Stage IIIa OR Stage IIIb OR Stage IIIc* ['Female and male participants with stage II-IIIC early or locally advanced/inflammatory human epidermal growth factor receptor 2-positive (HER2+) breast cancer']

Stage IIIb OR Stage IIIc ['Female and male participants with stage II-IIIC early or locally advanced/inflammatory human epidermal growth factor receptor 2-positive (HER2+) breast cancer']	-	-	-	-
IBC ['Female and male participants with stage II-IIIC early or locally advanced/inflammatory human epidermal growth factor receptor 2-positive (HER2+) breast cancer']	-	-	*IBC* ['Female and male participants with stage II-IIIC early or locally advanced/inflammatory human epidermal growth factor receptor 2-positive (HER2+) breast cancer']	*IBC* ['Female and male participants with stage II-IIIC early or locally advanced/inflammatory human epidermal growth factor receptor 2-positive (HER2+) breast cancer']
HER2+ ['Female and male participants with stage II-IIIC early or locally advanced/inflammatory human epidermal growth factor receptor 2-positive (HER2+) breast cancer', 'HER2+ breast cancer confirmed by a local laboratory prior to study enrollment']	*HER2+* ['Female and male participants with stage II-IIIC early or locally advanced/inflammatory human epidermal growth factor receptor 2-positive (HER2+) breast cancer', 'HER2+ breast cancer confirmed by a local laboratory prior to study enrollment']	*HER2+* ['Female and male participants with stage II-IIIC early or locally advanced/inflammatory human epidermal growth factor receptor 2-positive (HER2+) breast cancer', 'HER2+ breast cancer confirmed by a local laboratory prior to study enrollment']	*HER2+* ['Female and male participants with stage II-IIIC early or locally advanced/inflammatory human epidermal growth factor receptor 2-positive (HER2+) breast cancer', 'HER2+ breast cancer confirmed by a local laboratory prior to study enrollment']	*HER2+* ['Female and male participants with stage II-IIIC early or locally advanced/inflammatory human epidermal growth factor receptor 2-positive (HER2+) breast cancer', 'HER2+ breast cancer confirmed by a local laboratory prior to study enrollment']
NOT Stage IV ['Stage IV (metastatic) breast cancer']	*NOT Stage IV* ['Stage IV (metastatic) breast cancer']	*NOT Stage IV* ['Stage IV (metastatic) breast cancer']	*NOT Stage IV* ['Stage IV (metastatic) breast cancer']	*NOT Stage IV* ['Stage IV (metastatic) breast cancer']
NOT Multicentric ['Participants with multicentric breast cancer, unless all tumors are HER2+']	*NOT Multicentric* ['Participants with multicentric breast cancer, unless all tumors are HER2+']	*NOT Multicentric* ['Participants with multicentric breast cancer, unless all tumors are HER2+']	*NOT Multicentric* ['Participants with multicentric breast cancer, unless all tumors are HER2+']	*NOT Multicentric* ['Participants with multicentric breast cancer, unless all tumors are HER2+']
NOT Bilateral ['Participants with bilateral breast cancer']	*NOT Bilateral* ['Participants with bilateral breast cancer']	*NOT Bilateral* ['Participants with bilateral breast cancer']	*NOT Bilateral* ['Participants with bilateral breast cancer']	*NOT Bilateral* ['Participants with bilateral breast cancer']
-	-	-	-	*NOT DCIS OR NOT LCIS* ['Participants who have a past history of ductal carcinoma in situ (DCIS) or lobular carcinoma in situ (LCIS) if they have received any systemic therapy for its treatment or radiation therapy to the ipsi- or contralateral breast cancer']

-	-	-	-	*NOT HER2+* ['Participants with multicentric breast cancer, unless all tumors are HER2+']

Table 6: Detailed parsing results for example #3 - LLM.

LLM	Expert	Doctor	Novice 1	Novice 2
Stage IV OR Stage IIIb OR Stage IIIc ['Metastatic or locally advanced unresectable histologically documented TNBC as defined by absence of estrogen receptor (ER) and progesterone receptor (PR) expression and no HER2 amplification or over-expression by local pathology report.']	*Stage IV OR Stage IIIb OR Stage IIIc* ['Metastatic or locally advanced unresectable histologically documented TNBC as defined by absence of estrogen receptor (ER) and progesterone receptor (PR) expression and no HER2 amplification or over-expression by local pathology report.']	*Stage IV OR Stage IIIb OR Stage IIIc* ['Metastatic or locally advanced unresectable histologically documented TNBC as defined by absence of estrogen receptor (ER) and progesterone receptor (PR) expression and no HER2 amplification or over-expression by local pathology report.']	*Stage IV OR Stage IIIb OR Stage IIIc* ['Metastatic or locally advanced unresectable histologically documented TNBC as defined by absence of estrogen receptor (ER) and progesterone receptor (PR) expression and no HER2 amplification or over-expression by local pathology report.']	*Stage IV OR Stage IIIb OR Stage IIIc* ['Metastatic or locally advanced unresectable histologically documented TNBC as defined by absence of estrogen receptor (ER) and progesterone receptor (PR) expression and no HER2 amplification or over-expression by local pathology report.']
ER- AND PR- AND HER2- ['Metastatic or locally advanced unresectable histologically documented TNBC as defined by absence of estrogen receptor (ER) and progesterone receptor (PR) expression and no HER2 amplification or over-expression by local pathology report.']	*ER- AND PR- AND HER2-* ['Metastatic or locally advanced unresectable histologically documented TNBC as defined by absence of estrogen receptor (ER) and progesterone receptor (PR) expression and no HER2 amplification or over-expression by local pathology report.']	*ER- AND PR- AND HER2-* ['Metastatic or locally advanced unresectable histologically documented TNBC as defined by absence of estrogen receptor (ER) and progesterone receptor (PR) expression and no HER2 amplification or over-expression by local pathology report.']	*ER- AND PR- AND HER2-* ['Metastatic or locally advanced unresectable histologically documented TNBC as defined by absence of estrogen receptor (ER) and progesterone receptor (PR) expression and no HER2 amplification or over-expression by local pathology report.']	*ER- AND PR- AND HER2-* ['Metastatic or locally advanced unresectable histologically documented TNBC as defined by absence of estrogen receptor (ER) and progesterone receptor (PR) expression and no HER2 amplification or over-expression by local pathology report.']
PD-L1+ ['Confirmed PD-L1 positive as defined by Combined Positive Score (CPS) >10% by a CLIA-certified lab.']	*PD-L1+* ['Confirmed PD-L1 positive as defined by Combined Positive Score (CPS) >10% by a CLIA-certified lab.']	*PD-L1+* ['Confirmed PD-L1 positive as defined by Combined Positive Score (CPS) >10% by a CLIA-certified lab.']	*PD-L1+* ['Confirmed PD-L1 positive as defined by Combined Positive Score (CPS) >10% by a CLIA-certified lab.']	*PD-L1+* ['Confirmed PD-L1 positive as defined by Combined Positive Score (CPS) >10% by a CLIA-certified lab.']
-	-	-	-	*HER2-* ['HER2 negativity is defined as either: in situ hybridization non-amplified (ratio of HER2 to CEP17 <2 or single probe average HER2 gene copy number <4 signals/cell OR IHC 0 or 1+).']

-	-	-	-	*ER- AND PR-* ['ER and PR negativity is defined as <1% positive by IHC.']
-	-	-	-	*Stage IV* ['Treatment with <1 prior line of systemic therapy in the metastatic setting or adjuvant/neoadjuvant setting if metastatic recurrence within 12 months of treatment.']
-	-	-	-	*NOT Stage IIIb OR NOT Stage IIIc* ['More than one prior line of chemotherapy in the locally advanced unresectable or metastatic setting.']

Table 7: Detailed parsing results for example #4 - LLM.

LLM	**Expert**	**Doctor**	**Novice 1**	**Novice 2**
Invasive ['Diagnosis of invasive adenocarcinoma made by core needle biopsy']	*Invasive* ['Diagnosis of invasive adenocarcinoma made by core needle biopsy']	*Invasive* ['Diagnosis of invasive adenocarcinoma made by core needle biopsy']	*Invasive* ['Diagnosis of invasive adenocarcinoma made by core needle biopsy']	*Invasive* ['Diagnosis of invasive adenocarcinoma made by core needle biopsy']
Adenocarcinoma ['Diagnosis of invasive adenocarcinoma made by core needle biopsy']	*Adenocarcinoma* ['Diagnosis of invasive adenocarcinoma made by core needle biopsy']	*Adenocarcinoma ['Diagnosis of invasive adenocarcinoma made by core needle biopsy']*	*Adenocarcinoma* ['Diagnosis of invasive adenocarcinoma made by core needle biopsy']	*Adenocarcinoma* ['Diagnosis of invasive adenocarcinoma made by core needle biopsy']
HER2+ ['Breast cancer determined to be HER2-positive']	*HER2+* ['Breast cancer determined to be HER2-positive']	*HER2+* ['Breast cancer determined to be HER2-positive']	*HER2+* ['Breast cancer determined to be HER2-positive']	*HER2+* ['Breast cancer determined to be HER2-positive']
NOT Stage IV ['Definitive clinical or radiologic evidence of metastatic disease']	*NOT Stage IV* ['Definitive clinical or radiologic evidence of metastatic disease']	*NOT Stage IV* ['Definitive clinical or radiologic evidence of metastatic disease']	*NOT Stage IV* ['Definitive clinical or radiologic evidence of metastatic disease']	*NOT Stage IV* ['Definitive clinical or radiologic evidence of metastatic disease']
NOT Bilateral ['Synchronous bilateral invasive breast cancer']	*NOT Bilateral* ['Synchronous bilateral invasive breast cancer']	*NOT (Bilateral AND Invasive)* ['Synchronous bilateral invasive breast cancer']	*NOT (Bilateral AND Invasive)* ['Synchronous bilateral invasive breast cancer']	*NOT Bilateral* ['Synchronous bilateral invasive breast cancer']
-	*Stage I OR Stage II OR Stage III* ['Primary breast tumor palpable and measures greater than or equal to 2.0 cm by physical exam']	-	-	-

-	*NOT Stage IIIc* ['Ipsilateral cN2b or cN3 disease (Patients with cN1 or cN2a disease are eligible)']	-	-	-
-	*NOT Recurrent* ['Prior history of breast cancer, including DCIS (Patients with a history of LCIS are eligible)']	-	-	-

Table 8: Detailed parsing results for example #5 - LLM.

LLM	**Expert**	**Doctor**	**Novice 1**	**Novice 2**
Adenocarcinoma ['Histologically or cytologically confirmed adenocarcinoma of the breast with metastatic disease']	*Adenocarcinoma* ['Histologically or cytologically confirmed adenocarcinoma of the breast with metastatic disease']	*Adenocarcinoma* ['Histologically or cytologically confirmed adenocarcinoma of the breast with metastatic disease']	*Adenocarcinoma* ['Histologically or cytologically confirmed adenocarcinoma of the breast with metastatic disease']	*Adenocarcinoma* ['Histologically or cytologically confirmed adenocarcinoma of the breast with metastatic disease']
Stage IV ['Histologically or cytologically confirmed adenocarcinoma of the breast with metastatic disease']	*Stage IV* ['Histologically or cytologically confirmed adenocarcinoma of the breast with metastatic disease']	*Stage IV* ['Histologically or cytologically confirmed adenocarcinoma of the breast with metastatic disease']	*Stage IV* ['Histologically or cytologically confirmed adenocarcinoma of the breast with metastatic disease']	*Stage IV* ['Histologically or cytologically confirmed adenocarcinoma of the breast with metastatic disease']
ER- AND PR- AND HER2- ['Hormone receptor-negative (ER and PgR <10%) and HER2-negative (IHC 0,1+ or 2+ ISH not amplified) breast cancer']	*ER- AND PR- AND HER2-* ['Hormone receptor-negative (ER and PgR <10%) and HER2-negative (IHC 0,1+ or 2+ ISH not amplified) breast cancer']	*ER- AND PR- AND HER2-* ['Hormone receptor-negative (ER and PgR <10%) and HER2-negative (IHC 0,1+ or 2+ ISH not amplified) breast cancer']	*ER- AND PR- AND HER2-* ['Hormone receptor-negative (ER and PgR <10%) and HER2-negative (IHC 0,1+ or 2+ ISH not amplified) breast cancer']	*ER- AND PR- AND HER2-* ['Hormone receptor-negative (ER and PgR <10%) and HER2-negative (IHC 0,1+ or 2+ ISH not amplified) breast cancer']
PD-L1+ ['PD-L1 positive defined as expression on tumor-infiltrating immune cells >=1% (SP142 PD-L1 immunohistochemical assay, Ventana Medical Systems)']	*PD-L1+* ['PD-L1 positive defined as expression on tumor-infiltrating immune cells >=1% (SP142 PD-L1 immunohistochemical assay, Ventana Medical Systems)']	*PD-L1+* ['PD-L1 positive defined as expression on tumor-infiltrating immune cells >=1% (SP142 PD-L1 immunohistochemical assay, Ventana Medical Systems)']	*PD-L1+* ['PD-L1 positive defined as expression on tumor-infiltrating immune cells >=1% (SP142 PD-L1 immunohistochemical assay, Ventana Medical Systems)']	*PD-L1+* ['PD-L1 positive defined as expression on tumor-infiltrating immune cells >=1% (SP142 PD-L1 immunohistochemical assay, Ventana Medical Systems)']

-	-	-	*Metastatic* ['Asymptomatic metastatic lesions whose further growth would likely cause functional deficits or intractable pain (e.g., epidural metastasis that is not presently associated with spinal cord compression) should be considered for locoregional therapy if appropriate prior to enrollment.']	*Stage IV* ['the biopsy of metastatic disease before starting first-line therapy and assessed by local laboratory']
-	-	-	-	*ER+ AND PGR+* ['Patients ER and PgR <1% eligible to receive atezolizumab in combination with nab-paclitaxel as standard of care treatment for metastatic triple-negative breast cancer (TNBC), regardless of study participation.']
-	-	-	-	*Stage IV* ['Patients ER and PgR <1% eligible to receive atezolizumab in combination with nab-paclitaxel as standard of care treatment for metastatic triple-negative breast cancer (TNBC), regardless of study participation.']
-	-	-	-	*NOT Stage IV* ['Asymptomatic metastatic lesions whose further growth would likely cause functional deficits or intractable pain (e.g., epidural metastasis that is not presently associated with spinal cord compression) should be considered for locoregional therapy if appropriate prior to enrollment.']