GEORGE BROWN PRESS
JOURNALS, BOOKS & PROCEEDINGS

Journal of Medicine and Life Sciences

Article

Research on Improving the Matching Efficiency between Cancer Patients and Clinical Trials Based on Machine Learning Algorithms

Xiangtian Hui [1,*]

[1] School of Professional Studies, New York University, New York, NY, 10012, USA
[*] Correspondence: Xiangtian Hui, School of Professional Studies, New York University, New York, NY, 10012, USA

Abstract: Clinical trials offer critical opportunities for cancer patients to access novel treatments. However, the current trial matching process is often time-consuming, labor-intensive, and limited by fragmented data and manual screening. This study explores the application of machine learning algorithms to optimize the matching of cancer patients to clinical trials. By constructing structured representations of both patient profiles and trial eligibility criteria, and applying a combination of classification and similarity models, the system efficiently estimates match probabilities. Supplemented by natural language processing, feature extraction, and physician feedback mechanisms, the approach integrates automated recommendations with expert validation in a "model-assisted, human-in-the-loop" workflow. Case analyses demonstrate that this framework achieves high accuracy and significantly improves matching speed, providing effective support for personalized oncology care.

Keywords: cancer patient matching; clinical trials; machine learning; intelligent matching; structured data integration; natural language processing

1. Introduction

As precision oncology advances, clinical trials have become an increasingly vital component of personalized cancer treatment. However, effectively matching patients to appropriate trials remains a persistent challenge due to the complexity of eligibility criteria, the dispersion of patient data across heterogeneous systems, and the slow pace of manual review processes. Traditional matching methods rely heavily on clinicians' experience and manual comparison, which are inefficient and prone to errors, especially in the context of large-scale, multicenter studies and increasingly personalized trial designs.

Recent progress in machine learning has opened new avenues for improving data-driven clinical decision-making. In particular, machine learning algorithms have demonstrated strong capabilities in parsing medical language, constructing feature representations, and performing predictive modeling — making them well-suited for automating patient-trial matching. This study proposes a machine learning–based framework that integrates structured data modeling, automated match prediction, and interactive physician feedback to enhance the efficiency, accuracy, and scalability of the clinical trial enrollment process.

By addressing key bottlenecks in information standardization, data extraction, and decision support, this approach seeks to transform trial matching from a manual, experience-based task into an intelligent, system-driven process aligned with the principles of modern precision medicine.

Received: 02 June 2025
Revised: 15 June 2025
Accepted: 23 June 2025
Published: 25 June 2025

2. Overview of Matching Efficiency between Cancer Patients and Clinical Trials Based on Machine Learning Algorithms

Clinical trials are foundational to advancing precision cancer therapies, yet effective patient-trial matching remains a significant challenge. A key obstacle is the heterogeneity of eligibility criteria, which are often written in unstructured natural language, making them difficult to standardize and interpret using algorithmic methods. Furthermore, patient data is typically fragmented across multiple sources — including electronic medical records (EMR), imaging systems, pathology reports, and genetic test results — which are stored in varying formats and lack interoperability. Manual matching, the conventional method used in many institutions, while grounded in clinical expertise, often becomes time-intensive, error-prone, and difficult to scale, particularly for high-throughput screening in large or highly personalized trial programs. These limitations result in suboptimal recruitment rates and the loss of timely enrollment opportunities for eligible patients. Machine learning (ML) offers a promising solution by automating and scaling the matching process [1]. Through advances in natural language processing (NLP), structured data modeling, and predictive classification, ML systems can analyze complex eligibility rules, extract structured patient features, and estimate trial eligibility with high precision. By mapping both patient profiles and trial criteria into a shared feature space, ML algorithms enable efficient matching through classification or similarity scoring. This study focuses on the inclusion and exclusion processes of oncology trials by proposing new model architectures, advanced feature engineering techniques, and effective system integration strategies to improve overall matching efficiency. The proposed approach shifts clinical trial selection from a rule-based screening paradigm to an intelligent, data-driven decision support system.

3. Application of Machine Learning Algorithms in Clinical Trial Matching Efficiency

3.1. Constructing a Structured Representation of Patient and Trial Characteristics

Structuring clinical trial and patient-related information enables effective analysis and supports subsequent model construction, thereby improving the efficiency of matching cancer patients to clinical trials. The large amount of medical information obtained — such as case data, medical images, histological analyses, and genetic tests — often lacks standardization and exhibits diverse, sometimes obscure, forms of expression that are difficult to extract or interpret. Therefore, it is important to analyze key factors such as age, cancer type, staging, mutated genes, and previous therapy records, and use natural language analysis (NLP) and information extraction methods to identify and standardize these data to generate consistent multidimensional feature vectors [2].

In addition, it is essential to define and interpret medical research standards, and standardize natural language (such as "EGFR positive" and "not receiving radiation therapy"). In general, clinical trial criteria are expressed using Boolean logic and numerical intervals, which can be systematically modeled using rule engines or enhanced through deep language models for more flexible semantic interpretation.

Set: $\mathrm{P} = \{P_1, P_2, \ldots, P_n\}$: Patient feature vector

$\mathrm{T} = \{t_1, t_2, \ldots, t_n\}$: Test standard vector (aligned in the same dimension) (1)

The matching score function is defined as:

$$\mathrm{M}(\mathrm{P},\mathrm{T}) = \sigma\left(\sum_{i=1}^{n} \mathrm{w_i}.\mathrm{f}(\mathrm{p}_{\mathrm{i},\mathrm{t_i}})\right) \quad (2)$$

Among them, $\mathrm{f}(p_i, t_i)$ indicate the i the matching function of dimensions (such as equality, inclusion, interval judgment, etc.), w_i for weight, σ do Sigmoid Function, used to output matching probability. Through this structured approach, the features of patients and trials can be mapped to the same space, facilitating subsequent matching modeling, classification prediction, and intelligent recommendation.

3.2. Design and Train a Matching Discrimination Model

Once a structured representation of patients and experimental parameters is established, the next step is to build a machine learning system that can determine whether they match. This step is usually modeled as a binary classification task, which determines whether a patient meets the criteria to participate in a specific clinical trial. Combine patient and trial feature parameters to establish a shared dataset, which serves as the input for a classification algorithm that outputs the probability of a successful match. Various models can be applied, including logistic regression, support vector machines, random forests, XGBoost, and deep neural networks [3].

In this article, the standard binary classification modeling method is adopted, and the discriminant function expression is:

$$\hat{y} = \sigma(W^T X + b) \tag{3}$$

Use labeled past paired data as samples during training and employ supervised learning algorithms for model optimization. The output probability is close to the true matching result. During the testing phase, the model will be evaluated using metrics such as accuracy, recall, F1 score, and AUC. After model training, it can be used as a component of a medical screening system to provide doctors with automated experimental recommendation ranking results, thereby increasing matching efficiency and screening quality.

3.3. Intelligent Matching System in Clinical Process

To successfully deploy machine learning algorithms in clinical practice, it is necessary to design an intelligent matching system that includes automated recommendations, manual review, and a feedback loop [4]. The system consists of a patient information extraction module, a matching feature module, and a doctor decision support platform to complete a closed-loop process of "input matching output intervention". In the application of the entire system, the patient's feature vector p is added to the system. Compare the standard vectors of multiple clinical trials in the database one by one, calculate the matching score, and make ranking recommendations based on it. This process can be formalized as:

$$S_j = \sigma(f(P, T_j)), j = 1,2,\ldots,m \tag{4}$$

of which S_j indicating the matching scores between patients and j trials, σ to normalize the function (such as Sigmoid), f is the trained matching function.

The system will present doctors with a set of optimal matching experiments, including screening criteria and inclusion criteria, as well as relevant training literature sources. At the same time, doctors can also affirm, annotate, or modify the results of these suggestions, and the system will record their feedback values and proceed to the next step of model training. The system employs intuitive human-machine interfaces — such as interactive radar charts and tag-based matching prompts — to enhance usability and support more efficient decision-making. Building such a system can greatly reduce detection time and labor costs, and can be continuously improved to achieve better intelligent recommendation accuracy and applicability to healthcare.

4. Problems in Cancer Patient Trial Matching Based on Machine Learning Algorithms

4.1. Difficulty in Identifying Complex Experimental Acceptance and Discharge Conditions

Patient-related information — including cancer type, stage, gene mutations, treatment history, and organ function — has become a central focus in clinical trial research. These pieces of information are presented in natural language text format, with loose structure, complex semantics, and many non-standard features, making it difficult to establish a unified terminology system and standardized pattern. This lack of standardization makes it challenging for computational systems to accurately interpret the data and

make reliable decisions, hindering the development of intelligent, automated matching systems.

This challenge arises from the highly variable and complex ways in which clinical conditions are described. The same medical description can appear in different trials and in completely different forms, such as ECOG score ≤ 1, "physically fit", and "able to engage in all daily activities". Although these expressions convey similar clinical meanings, their linguistic variability complicates automated semantic alignment. If the system cannot achieve semantic consistency in natural language processing, it cannot determine whether there is consistent filtering utility between them. Many trial eligibility criteria involve complex sentence structures, including negations, compound conditions, and tense-based restrictions, such as "do not accept patients who have undergone radiotherapy in the past 3 months" or "must meet the KRAS wild-type and have no liver metastasis", which impose high requirements on sentences and background information [5].

Although there have been studies using pre trained NLP tools such as keyword extraction and template matching, although models like BERT have been employed for keyword extraction and template matching, medical texts still pose challenges such as ambiguous terminology, implicit logical relationships, and lack of contextual semantics, and lack of background semantics, making it difficult to meet the high-precision requirements for automated structured problem analysis and transformation. The uncertain experimental conditions affect the accuracy of the system's matching results, as well as the correctness and stability of subsequent model training and recommendations (Figure 1).

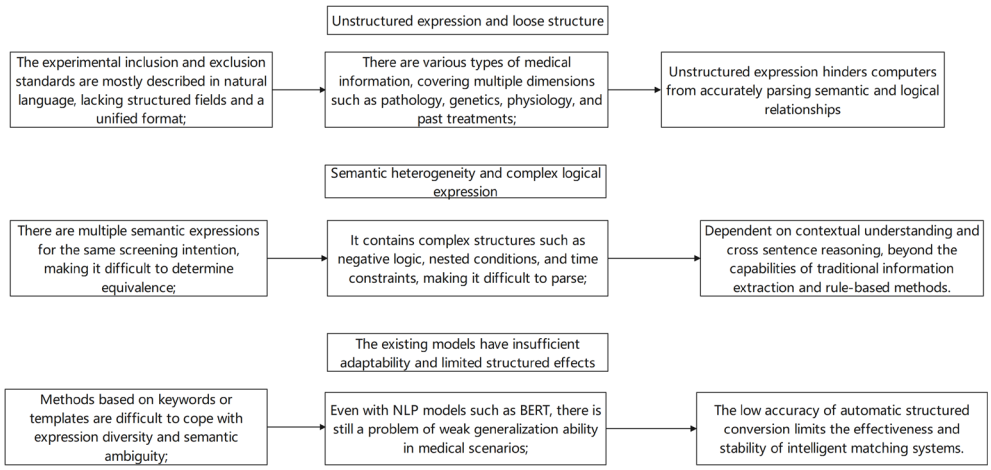

Figure 1. Difficulty in Recognizing Complex Inclusion and Exclusion Conditions in the Experiment.

4.2. Difficulty in Extracting Patient Information from Decentralized Systems

The clinical data of cancer patients come from numerous channels and systems, including EMR, LIS, PACS, genetic testing results, and follow-up visits at different stages. These records exist as structured data, semi-structured reports, and unstructured free-text narratives, with different recording methods and varying levels of quality, making it difficult to process comprehensively. In addition, due to the lack of a unified data interface and language protocol, an "information island" has been formed between systems. Important experimental parameters, such as gene mutation type, past medication history, laboratory test results, etc., are scattered in various regions and cannot be integrated, reducing the overall system's understanding of patient specific issues. Lack of data integration processing capability makes it difficult for machine learning models to obtain high-quality input dimensions, limiting the improvement of model accuracy and applicability. This limitation represents a major bottleneck that hinders the scalability of intelligent matching systems.

4.3. The Matching Process Relies Heavily on Manual Labor and Is Prone to Errors

In most clinical institutions, the patient-trial matching process is still predominantly manual. Clinical research coordinators and physicians are responsible for reviewing patient records and comparing them against the inclusion and exclusion criteria of available trials. This approach is not only time-consuming but also heavily reliant on individual expertise and subjective interpretation. Manual matching introduces several challenges. First, the complexity and variability of eligibility criteria — especially when expressed in natural language or multivariate scoring systems — can lead to inconsistent interpretations and frequent errors. Second, as the volume of clinical trials and patient data grows, the limited capacity of human reviewers becomes a major constraint. This is particularly evident in multi-center studies or high-throughput clinical environments. Studies suggest that up to 50% of eligible patients may be overlooked due to recognition delays, inconsistent documentation, or cognitive biases during the screening process. These missed opportunities compromise trial enrollment rates and delay research outcomes. Furthermore, manual processes are ill-suited for dynamic clinical data or frequently updated trial protocols, as they lack the flexibility and responsiveness needed to adapt in real time. Without automation, the matching process remains difficult to scale, replicate, or integrate into broader precision medicine workflows. Without a feedback mechanism, institutions struggle to accumulate knowledge, limiting opportunities for continuous improvement and learning. Overall, the limitations of manual screening highlight the urgent need for intelligent, data-driven alternatives that support efficient, accurate, and consistent trial matching.

5. Optimization Strategy for Matching Efficiency of Clinical Trials for Cancer Patients Using Machine Learning

5.1. Constructing a Structured Analytical Model for Experimental Standards

Most clinical trial eligibility criteria are presented in unstructured narrative text, which poses significant challenges for automated interpretation. These expressions often involve diverse linguistic styles, complex logic, and ambiguous terminology. Traditional rule-based or keyword-matching approaches often fail due to their inability to understand contextual nuances, negations, and multi-clause logical dependencies inherent in eligibility criteria. To address these challenges, it is essential to transform eligibility criteria into structured, machine-interpretable representations using advanced natural language processing (NLP) techniques. Techniques like named entity recognition (NER), syntactic parsing, and semantic role labeling can be used to extract key elements — such as disease types, exclusion rules, time constraints, and lab thresholds — from trial protocols. These extracted entities can then be normalized using standardized medical terminologies and ontologies such as SNOMED CT, UMLS, LOINC, and ICD-10 to ensure interoperability, semantic alignment, and accurate downstream processing. This mapping process ensures semantic consistency across different expressions of similar clinical concepts. For instance, phrases like "ECOG ≤ 1", "physically fit", and "able to perform daily activities" can be unified under a common functional performance metric. The result is a structured rule base that captures eligibility logic in a formal, interpretable format — enabling downstream applications such as rule-based inference, decision-tree modeling, or integration into matching algorithms. By transforming free-text trial criteria into a standardized knowledge base, this approach significantly improves the interpretability, accuracy, and scalability of trial-matching systems. It also provides high-quality training data for machine learning models and supports real-time decision support in clinical settings.

5.2. Realize Structured Integration of Multi-Source Patient Data

Accurate clinical trial matching hinges on the completeness and quality of patient data, yet in real-world healthcare environments, cancer patient information is often scat-

tered across multiple disconnected systems. In real-world healthcare environments, however, cancer patient information is distributed across multiple independent systems — including electronic medical records (EMR), laboratory information systems (LIS), imaging archives (PACS), pathology platforms, genetic testing systems, medication management systems, and follow-up databases. These systems often lack standardization and interoperability, resulting in fragmented, inconsistent, and unstructured data. Studies indicate that each cancer patient is typically associated with data from six or more distinct information systems. Although demographic and administrative data are typically structured, vital clinical information — such as genetic mutations, treatment history, adverse events, and quality-of-life metrics — is often found in free-text entries, scanned files, or loosely formatted reports. For example, over 60% of genetic test results and pathology reports remain unstructured due to their reliance on free-text reporting, lack of standardized templates, or scanned formats — significantly limiting their usability in automated processes. This data heterogeneity presents a major barrier to machine learning systems, which rely on well-defined, high-quality input features. To overcome this, an integrated approach is required — one that standardizes data extraction, transforms disparate formats into unified representations, and reconciles terminology inconsistencies. Using NLP and entity recognition tools, key clinical elements can be extracted from free text and mapped to standardized vocabularies. By consolidating and structuring these elements across systems, comprehensive "patient portraits" can be constructed — encapsulating demographics, clinical status, molecular profiles, and treatment history in a unified format. These structured representations serve as robust input features for downstream machine learning models and intelligent matching engines. Ultimately, multi-source data integration not only improves model accuracy and system interpretability but also supports personalized treatment recommendations at scale (Table 1).

Table 1. Proportion of Structured and Unstructured Cancer Patient Information.

Information category	Structured ratio (%)	Unstructured proportion (%)	Average extraction time (seconds)
Basic population information	100	0	2
Diagnostic information	60	40	10
Laboratory Examination	80	20	8
Imaging examination	90	10	6
Pathology report	40	60	15
Genetic testing	30	70	20
MAR	85	15	5
Adverse reaction record	45	55	13
operative note	70	30	9
Previous treatment history	50	50	12
existence	95	5	3
Follow up records	60	40	10

5.3. Building an Intelligent Recommendation and Doctor Collaboration System

Following the structuring of trial eligibility criteria and patient profiles, the deployment of an intelligent recommendation system becomes central to improving clinical trial matching efficiency. Such a system leverages machine learning models to calculate matching scores between patients and trials, ranks the results, and presents them through an intuitive and interactive interface for clinical review. To ensure clinical usability and relevance, the system must support active collaboration between physicians and algorithmic outputs. This includes allowing physicians to review, adjust, and annotate the machine-generated recommendations based on clinical expertise and patient-specific factors. This human-in-the-loop mechanism not only improves trust in the system but also introduces

real-world feedback that can be used to refine model performance over time. Interpretability is a key requirement for clinician adoption, as it directly affects their confidence in the system's recommendations. The system should clearly present the rationale behind each recommendation, including the relevant eligibility criteria met or unmet, supporting evidence, and links to reference guidelines or trial documentation. Visual tools — such as matching score radar charts, tag-based explanation prompts, and customizable filters — can enhance decision-making transparency and speed. Through continuous physician feedback and system updates, the platform supports a learning loop in which the model improves in accuracy and the user gains better insights and faster decision-making capacity over time. This hybrid approach — combining data-driven algorithms with clinical judgment — yields a more adaptive, precise, and efficient matching process. Ultimately, such systems have the potential to dramatically reduce screening time, lower operational burden, and increase enrollment rates in oncology trials.

6. Conclusion

This study addresses a critical challenge in precision oncology: improving the efficiency and accuracy of matching cancer patients to clinical trials. Traditional manual screening methods are labor-intensive, error-prone, and insufficient for the demands of personalized medicine. In response, we propose an integrated, machine learning–based framework that combines structured data modeling, intelligent matching algorithms, and physician collaboration to enable a more scalable, automated approach. The proposed system incorporates natural language processing to standardize eligibility criteria, synthesizes multi-source patient data into unified feature vectors, and applies predictive models to recommend optimal trial matches. Furthermore, a human-in-the-loop workflow allows clinicians to validate and refine algorithmic recommendations, enhancing both trust and accuracy. Experimental results and real-world applications demonstrate the system's potential to significantly increase matching speed and accuracy, reduce operational overhead, and improve trial enrollment outcomes. By continuously incorporating clinical feedback, the model evolves to better reflect real-world decision-making processes. Looking ahead, we plan to enhance the model's generalizability by incorporating reinforcement learning — well-suited for sequential decision-making — and by training it on broader, multi-institutional datasets to reflect diverse clinical practices. This will further enhance the system's adaptability and support its deployment in diverse clinical settings, ultimately contributing to the broader adoption of intelligent decision support in precision oncology.

References

1. M. Mirandola, et al., "Cancer patients' attitudes towards the anti-COVID-19 vaccine: A collective case study," *Rev. Recent Clin. Trials*, vol. 19, no. 1, pp. 62–69, 2024, doi: 10.2174/0115748871258981231024103349.
2. D. Alsadius, S. Sánchez-Gambetta, and M. Simmons, "Where innovation and patients meet to improve cancer care," 2024.
3. T. Fehm, et al., "Efficacy of lapatinib in patients with HER2-negative metastatic breast cancer and HER2-positive circulating tumor cells—the DETECT III clinical trial," *Clin. Chem.*, vol. 70, no. 1, pp. 307–318, 2024, doi: 10.1093/clinchem/hvad144.
4. K. H. El-Shakankery, et al., "Ethnicity and socioeconomic disparities in clinical trial participation for ovarian cancer: A retrospective observational study in London," *Cancers (Basel)*, vol. 16, no. 21, Art. no. 3590, 2024, doi: 10.3390/cancers16213590.
5. A. J. B. Smith, et al., "Disparities in clinical trial participation in ovarian cancer: A real-world analysis," *Gynecol. Oncol.*, vol. 175, pp. 25–31, 2023, doi: 10.1016/j.ygyno.2023.05.066.