IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 30, No. 9, September, 2025, pp. 2285-2295
Journal homepage: http://iieta.org/journals/isi

A Hybrid Semantic–Rule-Based NLP Framework Integrating DFCI and MSKCC Approaches for Clinical Trial Matching Using UMLS and FAISS

Aditya Dubey[1*], Aditi Ambasta[1], Jenish Soni[1], Prakshal Doshi[2], Mrunal Rupesh Rane[1], Pratik Kanani[1]

[1] Department of Artificial Intelligence and Data Science, Dwarkadas J. Sanghvi College of Engineering, Mumbai 400017, India
[2] Software Engineering Applications, Apple Inc, San Jose 95134, United States

Corresponding Author Email: aditya.work132@gmail.com

https://doi.org/10.18280/isi.300906

Received: 3 July 2025
Revised: 27 August 2025
Accepted: 8 September 2025
Available online: 30 September 2025

Keywords:
clinical trial matching, hybrid retrieval, FAISS, UMLS, DFCI, MSKCC, biomedical NER, cross-encoder re-ranking

ABSTRACT

Precision oncology relies heavily on interpreting unstructured clinical and biomedical literature, a task well-suited for Natural Language Processing (NLP). However, existing NLP systems show inconsistent performance due to varying data structures and semantic modeling. This paper systematically compares NLP pipelines inspired by the Dana-Farber Cancer Institute (DFCI-style), which uses transformer-based semantic retrieval, and Memorial Sloan Kettering Cancer Center (MSKCC-style), which employs a rule-based keyword approach. To overcome the limitations of these individual methods, we propose a Hybrid+FAISS NLP model. This architecture integrates semantic embeddings, fast dense retrieval using FAISS, clinical concept linking via the Unified Medical Language System (UMLS), and deep re-ranking with a cross-encoder. Evaluated on the CHIA and CORD-19 datasets, the Hybrid+FAISS system achieved a balanced trade-off, attaining 53.1% coverage and a mean reciprocal rank (MRR@1) of 0.531. This significantly outperforms both baselines: the DFCI-style model had 91.0% coverage but low average similarity (0.404), while the MSKCC-style rule-based model had limited 7.5% coverage. Our results validate the efficacy of this enhanced hybrid retrieval model in improving clinical trial matching and provide a reproducible framework for benchmarking biomedical NLP systems.

1. INTRODUCTION

Natural Language Processing (NLP) has become a key part of artificial intelligence. It can interpret and extract insights from large volumes of unstructured text data. Recently, NLP techniques have shifted from traditional rule-based methods to deep learning frameworks, resulting in significant performance improvements across fields such as healthcare, finance, education, and customer support. In healthcare, NLP can greatly automate and enhance clinical decision-making, especially in precision oncology and clinical trial recruitment.

Cancer is a complex disease that requires accurate diagnosis and personalized treatments based on detailed clinical and genomic data. Oncologists and researchers often face the time consuming task of reviewing extensive documentation, including pathology reports, genomic test results, and electronic health records (EHRs). Manual review is not only labor intensive but also prone to error. NLP offers an efficient and automated way to process, structure, and interpret these documents in real time, improving both patient care and research workflows.

Figure 1 shows the common stages in an NLP pipeline: lexical analysis (tokenization), syntactic parsing, semantic analysis, discourse integration, and pragmatic analysis. Transformerbased models such as BERT, RoBERTa, and GPT have significantly advanced the field by using attention mechanisms to understand contextual meaning. Open-source libraries like spaCy, Hugging Face Transformers, and scispaCy provide powerful tools for implementing domainspecific NLP tasks.

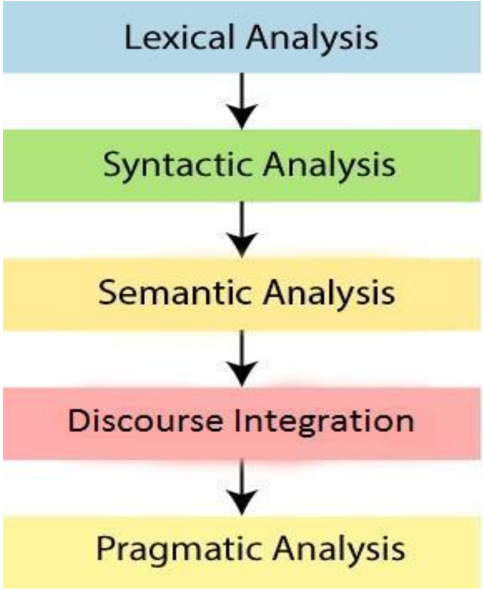


Figure 1. Phases of NLP

In precision oncology, NLP is used to automate tasks such as clinical trial matching, biomarker extraction, and genomic data interpretation. A key challenge is identifying appropriate clinical trials for patients based on complex eligibility criteria found in trial descriptions and personalized clinical or genomic profiles. NLP systems can reduce the time and effort required for trial matching and support timely, individualized treatment decisions.

1.1 Research objectives

This study explores NLP systems from two leading cancer research institutions: Dana-Farber Cancer Institute (DFCI) and Memorial Sloan Kettering Cancer Center (MSKCC). The goals are to:

1) Compare two different NLP pipelines: a semantic-based system (DFCI) that uses transformer embeddings and biomedical named entity recognition, and a rule-based keyword system (MSKCC) that applies clinical filtering.

2) Design and evaluate a new hybrid NLP architecture that combines the strengths of both approaches.

3) Further enhance the hybrid architecture with scalable dense retrieval using FAISS, clinical concept normalization using UMLS, and re-ranking with deep neural cross encoder models to achieve optimal balance between coverage and precision.

1.2 Case studies and institutional NLP systems

Dana-Farber Cancer Institute (DFCI) has come up with MatchMiner, an open-source tool that applies deep learning and NLP to align patient-specific genomic variations with pertinent clinical trials. MatchMiner applies real-time semantic search, which greatly enhances the speed and accuracy of patient-trial matching. Memorial Sloan Kettering Cancer Center (MSKCC) created a rule-based NLP system called the Patient centered Information Extraction for Clinical Systems (PINES). PINES prioritizes the extraction of structured data from unstructured clinical notes. It provides interpretability and consistency, which are essential in clinical environments. While every institutional framework has its own strengths, DFCI provides semantic flexibility at high levels with transformer based embeddings, and MSKCC provides structure clarity with deterministic rule based extraction. Yet, both frameworks have their own technical limitations. On the one hand, the DFCI model is contextually rich, but retrieves too general results with semantic drift, and finding dense embedding search results is expensive. Conversely, while being interpretable, the MSKCC system has a low recall, has fixed keyword-based rules and only captures a related limited set of words, and is poor at adapting to new words. Recalls, precision, and efficiency tradeoffs imply an architecture that incorporates all three methods. This paper suggests a hybrid structure that retains the semantic richness of DFCI, with structure given by MSKCC, and relies upon UMLS concept normalized and FAISS based dense retrieval to reason in terms of scalable and clinically interpretable trial matching.

- **Hybrid NLP Model**: Merges transformer-based semantic embeddings with biomedical named entity recognition and utilizes similarity thresholds to combine semantic depth and entity-level accuracy.
- **Hybrid+FAISS Model**: Extends the hybrid model by incorporating scalable FAISS-based dense retrieval, overlap at the concept level via UMLS linking, and neural re-ranking with a cross-encoder architecture. The system is highly precise and has increased coverage, remedying the major limitation of previous models.

Both models are tested with publicly available data: CHIA (Clinical Trial Annotations) and CORD-19 (COVID 19 biomedical literature). Results demonstrate that these hybrid retrieval models outperform significantly baseline methods, highlighting good promise for real-world clinical use.

2. LITERATURE REVIEW

The use of NLP in oncology has gained a tremendous amount of momentum, with several studies highlighting its potential to utilize in rewriting unstructured clinical and genomic data into actionable insights.

Sethna et al. [1] introduced ClinicalBERT, a clinical domain-specific BERT model trained on clinical notes. It outperformed general-purpose models on downstream medical NLP tasks and paved the way for clinical contextual embeddings.

Warner et al. [2] applied NLP to enable cancer staging from pathology reports, demonstrating improved accuracy and significant reduction of the workload in manual abstraction activities.

Dong et al. [3] showed that transformer architectures greatly improve pharmacovigilance performance by offering strong context-aware semantic disambiguation by using BERT-based language models to extract drug adverse events from clinical records and social media with high precision and recall.

Lindsay et al. [4] evaluated MatchMiner's performance at DFCI. They demonstrated an application of NLP to the real world of matching cancer patients with trials based on genetic alterations. Their study reported quantifiable improvements in enrollment efficiency and institution coordination.

Friedman et al. [5] founded the first rule-based Natural Language Processing (NLP) systems that effectively transformed unstructured clinical notes into actionable structured data. These systems laid the foundation for transparent and explainable NLP techniques, like PINES, which are still necessary for precise medical information retrieval in sensitive clinical domains like oncology.

Meystre et al. [6] pioneered the use of NLP for automatically screening patient eligibility for clinical trials using unstructured EHR data, showing their system could accelerate trial recruitment while improving patient-trial matching accuracy, thus directly supporting the needs of precision oncology and pragmatic clinical research.

Lee et al. [7] introduced BioBERT, a transformer-based language model pre-trained on large-scale biomedical corpora to effectively extract genomic variants and other entities from scientific and clinical literature. Their findings reinforce the necessity of domain-specific transformers for achieving high accuracy on specialized named entity recognition tasks essential for biomarker discovery and precision medicine.

Devlin et al. [8] made BERT available, which transformed NLP by making it possible to do bidirectional language modeling and achieving new performance levels across tasks. Its impact reaches across fields, even extending to biomedical NLP applications such as ClinicalBERT and BioBERT.

These works together show how NLP is not just making data more accessible and easier to interpret but also enhancing the quality, pace, and volume of oncology research and patient treatment.

3. CASE STUDY AND COMPARATIVE ANALYSIS: NLP SYSTEMS IN CLINICAL ONCOLOGY

3.1 Dana-Farber Cancer Institute (DFCI) NLP methodology

The Dana-Farber Cancer Institute has developed a production-ready NLP system for real-time clinical decision support in precision oncology contexts [9]. The methodology has four unified components: multi-modal data architecture, sophisticated NLP processing pipeline, ensemble machine learning framework, and clinical workflow integration.

1) Data Architecture and Real-Time Integration: DFCI's data processing infrastructure uses a real-time streaming model that collects many different heterogeneous clinical data sources through a unified pipeline. The system combines electronic health record systems (EHRs), genomic sequencing, laboratory values, radiology reports, pathology reports, and longitudinal history of treatment [10]. This utility facilitates multi-modal, exhaustive patient profiling which is critical for personalized treatment recommendations. The framework is initiated and deployed using HL7 FHIR interoperability standards enabling seamless data exchange between distinct clinical systems. Real-time ingestion occurs via API-based connections to Epic EHR systems and enables immediate processing of clinical documents after they are entered into the medical record [9]. As illustrated in Figure 2, the comprehensive system architecture demonstrates the integration of multiple heterogeneous data sources through a unified real-time streaming pipeline, enabling seamless data flow from various clinical systems to the decision support interface.

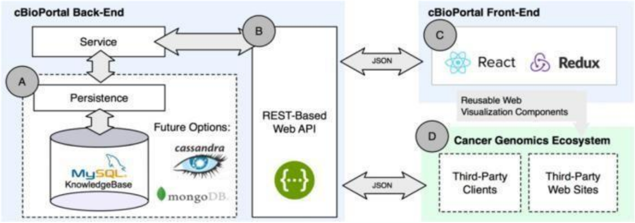


Figure 2. DFCI system architecture

2) Natural Language Processing Pipeline: The NLP processing framework executes a three-phase sequential pipeline approach to encoding clinical oncology text. The first phase involves certain text preprocessing steps, including medical vocabulary standardization, temporal expression flagging, clinical abbreviation expansion, and context-aware tokenization [4]. In the second phase, the system implements named entity recognition (NER), based on transformer language models that are fine-tuned on clinical oncology-based corpora. In all, the system recognizes and classifies cancer diagnoses and stage, molecular biomarkers, treatment regimens, adverse effects, performance status, and treatment response evaluations, with overall reported accuracy > 90% in representing critical clinical entities [9]. In the third phase, the system identifies relationships between recognized entities, including drug-disease relationships, biomarkers and treatment response associations, temporal treatment sequences, and causal associations for clinical outcomes [11]. The complete workflow is depicted in Figure 3, which shows the systematic transformation of raw clinical text through preprocessing, named entity recognition, and relationship extraction phases to generate structured clinical insights.

3) Machine Learning Framework: DFCI's machine learning framework utilizes ensemble methods involving ensemble of multiple transformer-based models, specifically fine-tuned versions of BERT and domain-adapted clinical language models [12]. The system incorporates transfer learning approaches using pre-trained clinical models, adapting them to specific data distributions and clinical workflows seen at the institution. The training approaches utilize supervised learning to train with a small clinical dataset annotated by experts and semi- supervised training techniques to take advantage of larger amounts of unlabeled clinical text. Active learning also allows to improve the models as new clinical data arrives, effectively keeping models up and running over time [9]. The validation process uses temporal cross-validation procedures to avoid data leakage and check how models generalize across different time periods. The project aims to achieve >90% accuracy on treatment recommendations and >85% accuracy on risk stratification tasks [10].

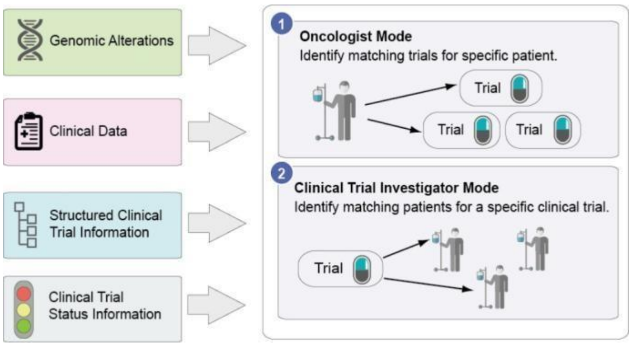
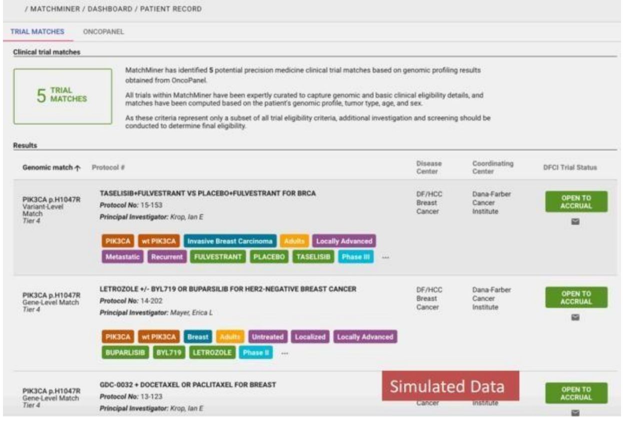

Genomic Alterations
Clinical Data
Structured Clinical Trial Information
Clinical Trial Status Information
1
Oncologist Mode
Identify matching trials for specific patient.
Trial
Trial
Trial
2
Clinical Trial Investigator Mode
Identify matching patients for a specific clinical trial.
Trial

Figure 3. NLP pipeline flowchart



Clinical trial matches

5 TRIAL MATCHES

MatchMiner has identified **5** potential precision medicine clinical trial matches based on genomic profiling results obtained from OncoPanel.

All trials within MatchMiner have been expertly curated to capture genomic and basic clinical eligibility details, and matches have been computed based on the patient's genomic profile, tumor type, age, and sex.

As these criteria represent only a subset of all trial eligibility criteria, additional investigation and screening should be conducted to determine final eligibility.

Results

Genomic match ↑	Protocol #	Disease Center	Coordinating Center	DFCI Trial Status
PIK3CA p.H1047R Variant-Level Match *Tier 4*	TASELISIB+FULVESTRANT VS PLACEBO+FULVESTRANT FOR BRCA ***Protocol No:*** 15-153 ***Principal Investigator:*** *Krop, Ian E* PIK3CA wt PIK3CA Invasive Breast Carcinoma Adults Locally Advanced Metastatic Recurrent FULVESTRANT PLACEBO TASELISIB Phase III ...	DF/HCC Breast Cancer	Dana-Farber Cancer Institute	OPEN TO ACCRUAL
PIK3CA p.H1047R Gene-Level Match *Tier 4*	LETROZOLE +/- BYL719 OR BUPARSILIB FOR HER2-NEGATIVE BREAST CANCER ***Protocol No:*** 14-202 ***Principal Investigator:*** *Mayer, Erica L* PIK3CA wt PIK3CA Breast Adults Untreated Localized Locally Advanced BUPARLISIB BYL719 LETROZOLE Phase II ...	DF/HCC Breast Cancer	Dana-Farber Cancer Institute	OPEN TO ACCRUAL
PIK3CA p.H1047R Gene-Level Match *Tier 4*	GDC-0032 + DOCETAXEL OR PACLITAXEL FOR BREAST ***Protocol No:*** 13-123 ***Principal Investigator:*** *Krop, Ian E*	Cancer	Institute	OPEN TO ACCRUAL

Simulated Data

Figure 4. Clinical decision support interface

4) Clinical Decision Support Integration: The clinical decision support system provides real-time treatment recommen- dations through seamless operability at the point of care. The system develops customized therapy recommendations that include genomic profiles, clinical trial eligibility assessments, alerts regarding drug interactions, alerts for contraindications, and unique dosing recommendations [13]. Risk assessment models provide immediate predictions for the likelihood of treatment-related toxicities, the likelihood of disease progression, prognostic estimates of overall survival, and assessments of the quality of life impact. Clinical insights are presented as integrated dashboards in the EHR, and response times are less than 2 seconds optimally to support real time clinical decision-making [9]. Figure 4 demonstrates the user-friendly clinical decision support interface, showing how complex AI-generated insights are presented to clinicians in an intuitive dashboard format integrated within the EHR system for point-of-care decision making.

3.2 Memorial Sloan Kettering Cancer Center (MSKCC) NLP methodology

Memorial Sloan Kettering Cancer Center has created a standardized NLP approach focused on risk stratification, quality improvement analytics, and large scale retrospective clinical research applications [14]. The approach guarantees robust batch processing capabilities, a rigorously validated risk stratification model, and seamless integration with the institution’s quality improvement initiatives.

1) Large-Scale Data Processing Architecture: MSKCC utilizes a processing framework centered on batch-level assessments of large clinical datasets; empirical case studies have demonstrated that our framework can analyze multi-year longitudinal patient cohorts covering over a decade of clinical records [7]. This data framework supports a variety of document types, including structured radiology reports, pathology evaluations, clinical progress notes, discharge summaries, and standardized clinical assessments. The data framework enables distributed computing to process over 100,000 patient records concurrently, allowing for population-level analyses and large-scale in-depth and innovative retrospective studies. Document classification algorithms automatically categorize document types and segment-level parsing is complete using template based extraction methodologies [15].

2) Clinical Information Extraction Framework: The clinical information extraction pipeline features multiple sequential steps, initiated by automated document classification and section-based parsing of the contents. The system is capable of automatically recognizing and extracting information from findings sections, clinical impression sections, diagnostic evaluation sections, and treatment recommendations sections of clinical reports, regardless of the format [16]. Clinical entity extraction aims to automatically identify cancer diagnosis with TNM staging, comorbidities, post-operative complications, treatment response, and adverse events that are mentioned in the clinical reports. The system has a high level of proficiency with structured radiology reports, utilizing specialized algorithms able to identify certain conditions, including splenomegaly, across large populations and in reports spanning several years [17]. Temporal information extraction focuses on treatment dates and timelines of disease onset, such as disease progression and illness episodes over time. Quantitative data extraction algorithms can automatically capture standardized clinical measures, assessment scores, and performance metrics listed in narrative descriptions of clinical reports [7]. The systematic approach to information extraction is illustrated in Figure 5, which depicts the multi-stage process of transforming diverse clinical documents into structured data suitable for large-scale retrospective analysis.

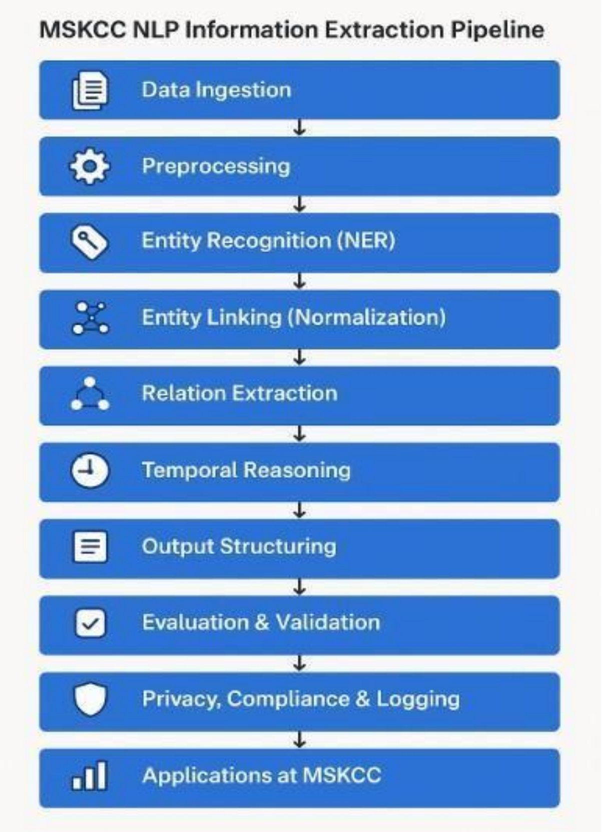


Figure 5. Information extraction pipeline

3) Risk Stratification and Quality Assessment Models: MSKCC has taken an automated approach to risk stratification in multiple clinical scenarios that include predictions of postoperative complications and assessments of quality outcomes. The approach uses specific detection algorithms that can identify a number of post-operative complications, including deep vein thrombosis (DVT), pulmonary embolism (PE), surgical site infections, and other specific post-operative complications [13].

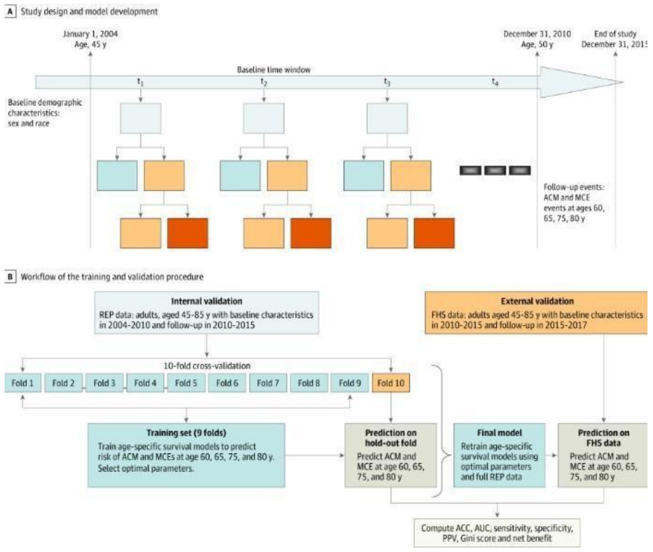

A Study design and model development
January 1, 2004
Age, 45 y
December 31, 2010
Age, 50 y
End of study
December 31, 2015
Baseline time window
t_1
t_2
t_3
t_4
Baseline demographic characteristics: sex and race
Follow-up events: ACM and MCE events at ages 60, 65, 75, 80 y
B Workflow of the training and validation procedure
Internal validation
REP data: adults, aged 45-85 y with baseline characteristics in 2004-2010 and follow-up in 2010-2015
External validation
FHS data: adults aged 45-85 y with baseline characteristics in 2010-2015 and follow-up in 2015-2017
10-fold cross-validation
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Fold 6
Fold 7
Fold 8
Fold 9
Fold 10
Training set (9 folds)
Train age-specific survival models to predict risk of ACM and MCEs at age 60, 65, 75, and 80 y. Select optimal parameters.
Prediction on hold-out fold
Predict ACM and MCE at age 60, 65, 75, and 80 y
Final model
Retrain age-specific survival models using optimal parameters and full REP data
Prediction on FHS data
Predict ACM and MCE at age 60, 65, 75, and 80 y
Compute ACC, AUC, sensitivity, specificity, PPV, Gini score and net benefit

Figure 6. Risk stratification performance

Risk stratification models take a hybrid approach and combine elements of clear rule-based clinical criteria with statistical machine learning classifiers. Clinical decision rules indicated by definers of expert defined clinical data were combined with support vector machines, random forests and ensemble methods for achieving statistically robust, yet interpretable, risk predictions [18]. In validating performance, study results show accuracy rates in the range of 70-95% across multiple clinical tasks—with the variation in sensitivity reflecting the complexity of risk assessment tasks. Truth is established in the dataset using expert physician review and inter-rater reliability studies to ensure good training data quality [19]. Figure 6 illustrates the comprehensive study design and model development workflow, showing the timeline from 2004-2017 with baseline time windows, 10-fold cross-validation procedures, and the systematic approach to internal and external validation using different patient cohorts.

4) Quality Improvement and Research Integration: The methodology includes integral quality improvement frameworks generating automated measurements for assessing treatment compliance monitoring, complication rates, outcome accuracy predictions and clinical performance benchmarking [20]. Research support capabilities include automated cohort identification for patients in clinical studies and outcome extraction standardized for research. The methodology required automated research data extraction capabilities to validate cohort studies for patients in which clinical studies; outcomes measured and reported for research study endpoints [14]. The methodology supports large-scale retrospective analyses longitudinal, with clinical hypothesis development for systematic discovery of patterns of any longitudinal clinical dataset.

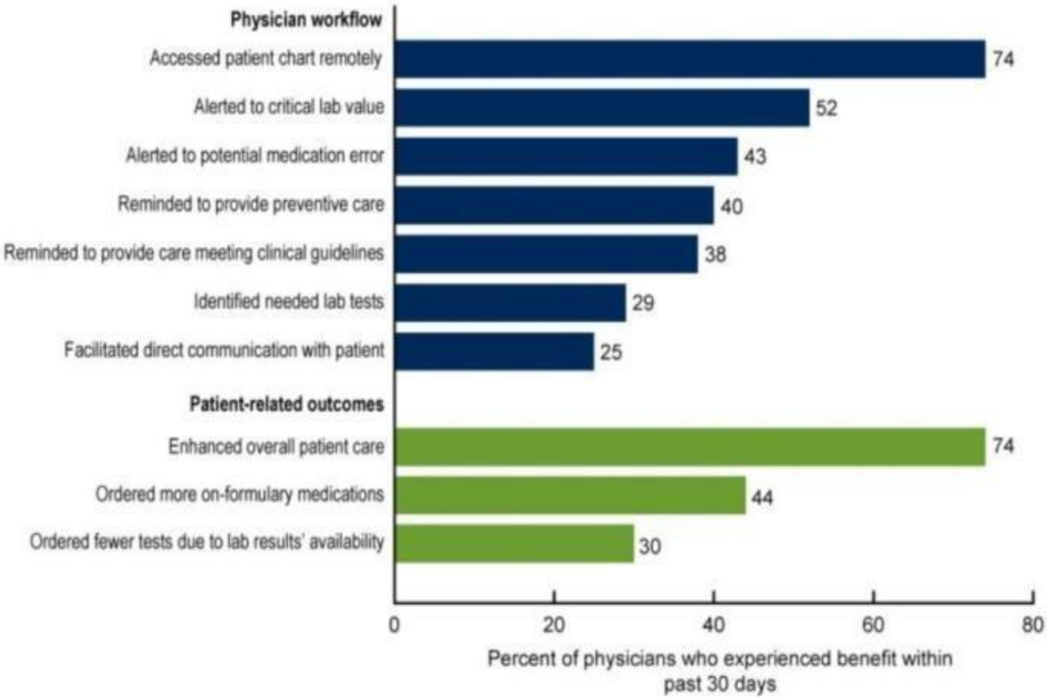


Figure 7. Quality improvement graph

As shown in Figure 7, the quality improvement metrics demonstrate the system's capability to track clinical performance indicators and identify trends in patient outcomes across multiple dimensions of care quality.

4. COMPARATIVE ANALYSIS

4.1 Technical architecture and processing approach comparison

The unique architectural configurations of methodology at DFCI and MSKCC clearly indicate differences in clinical focus and operational efficiency. The methodologies of DFCI which used a real-time streaming architecture prioritize low latency processing with immediate clinical relevance with the use of complex algorithms, whereas MSKCC develops a more batch-processing framework that sacrifices some immediacy and clinical relevance for the depth of construction and exploration characteristically found in research [10, 12]. DFCI was using advanced transformer-based models tailored for significant computational expense and high precision medicine applications that demand immediate clinical support/intervention. DFCI is able to leverage multi-modal data to perform complex personalized treatment matching, at high computational cost, while yielding immediate clinical ability [12]. MSKCC's methodology suggests a hybrid approach between rule-based clinical expertise and traditional machine learning processes. More generally, MSKCC supports interpretability in its findings that is significant for ongoing quality improvement and improvement/quality-based research, hence an emphasis on transparency in clinical decision-making while simultaneously ensuring computational return on investment in large enrolled populations [18].

4.2 Clinical application scope and integration analysis

Both methodologies have clinical applications that clearly align with institutional priorities and technical abilities.

DFCI's real-time decision support is suited for immediate treatment recommendations, precision medicine matching, clinical trial eligibility, and point-of-care clinical decision support, as shown in Table 1. Furthermore, DFCI's integration of genomic information and precision medicine workflows facilitate patient therapy selections that reflect the overall molecular and clinical profiling of the patient [13].

MSKCC's methodology has greater capabilities with respect to formal population-wide risk assessment, formal quality improvement work, and large multi-patient retrospective clinical research applications. MSKCC offers this advantage because its methodology allows for the processing of multiple patient cases, which enables identification of patient care trends and patterns that can help with surgical interventions, as well as (and more importantly) their decisions for increasing quality improvement strategy work and hypothesis generation to meet its respective institutional goals and objectives [20].

Table 1. Comparison of DFCI and MSKCC approaches

Aspect	DFCI	MSKCC Approach
Primary Focus	Real-time clinical decision support	Retrospective analysis and quality improvement
Data Processing	Real-time Streaming	Batch processing
Clinical Integration	Point-of-care decision support	Quality dashboards and research tools
Temporal Scope	Current patient state	Historical Longitudinal analysis
Accuracy Priority	High precision for treatment recommendations	High recall for risk identification
Scalability	Individual patient focus	Population-level analysis
Clinical Workflow	Embedded in active treatment decisions	Supporting quality improvement initiatives

The difference in time horizons still presents complementary clinical value propositions, with DFCI providing immediate actions that improve the patient's state and meet immediate clinical needs, while MSKCC's focus on observing a previous cohort of patients' longitudinal situation offers overall methods to improve systematic healthcare quality [7].

4.3 Performance metrics and validation methodology comparison

The performance characteristics of the two systems represent their differing targets of optimization for their respective clinical applications. DFCI provides accuracy rates greater than 90% on the clinical task of extracting important clinical entities and offering treatment recommendations with system response times under 2 seconds, adhering to strict real-time clinical workflow requirements [9]. MSKCC reports a range of variable accuracy performance of 70-95% across the range of clinical tasks, which reflects the challenge of variability of risk stratification and assessment for quality applications. However, through a batch processing approach, MSKCC can conduct more inclusive validation studies, including more temporal validation across different valid time periods and can leverage inter-institutional validation when capable of using an external data set [19].

This validation framework also demonstrates large structural differences between the institutions; DFCI conducts temporal cross-validation validations to avoid data spills and enable more variability of data to test for generalizability, and MSKCC conducts a range of inter-rater reliability studies with expert physicians to generate robust ground truth data sets to assist in building and evaluating their model [9, 18].

4.4 Strengths, limitations, and complementary opportunities assessment

DFCI has several advantages, such as clinical utility, advanced AI incorporation, precision medicine support, and a production-ready deployment for an on-the-fly clinical environment. Limitations of DFCI's approach include a lack of complete retrospective review abilities, an inability to generate quality metrics at the population level, and limited capabilities for data exploration for research purposes and presentation [10].

MSKCC has substantial strengths in retrospective clinical exploration, data generation for research purposes, crowd and population levels experiences, and integration of an explicit focus on and systems for quality improvement. The weaknesses of MSKCC's approach include diminished potential for real-time clinical decision support, limited incorporation of precision medicine capabilities, and very slow integration into real-time clinical workflow [14].

This apparent compatibility between institutional approaches presents important perspectives for integration of methodological, hybrid frameworks. The potential for integrating MSKCC's comprehensive historical analysis framework could enhance DFCI's abilities for real-time processing, while MSKCC's quality improvement emphasis could be strengthened by the real-time tracking and clinical decision support capabilities of DFCI [9, 20].

5. METHODOLOGY: HYBRID NLP MODEL DEVELOPMENT FOR CANCER RESEARCH AUTOMATION

This study highlights the constraints of presently used clinical NLP systems, that separately indicate varying levels of performance within and among institutions because of a lack of consistency surrounding data structure, data retrieval approach, and the underlying semantic modeling. We present here a Hybrid+FAISS NLP model that uses semantic embeddings, fast dense retrieval via FAISS, clinical concept linking via the UMLS, and deep re-ranking via the crossencoder architecture. Systematically it creates a way to blend the transformer-based semantic retrieval approach taken in DFCI implementations with the rule-based clinical keyword matching approach taken in MSKCC systems, creating one framework that is intended to balance precision and coverage for clinical trial matching and biomedical text mining. Two complementary datasets were selected to represent different aspects of cancer research automation:

5.1 Dataset selection

Two complementary datasets were selected to represent different aspects of cancer research automation:

1) CHIA Dataset (Eligibility Criteria for Clinical Trials): The CHIA dataset [21] contains annotated clinical trial eligibility criteria, mirroring DFCI's structured medical text processing approach. The dataset provides both inclusion and exclusion criteria from clinical trials, enabling the model to learn clinical research protocol language patterns.

2) CORD-19 Dataset (Biomedical Literature): The CORD- 19 dataset [22] represents large-scale biomedical literature, incorporating MSKCC's literature mining capabilities. This dataset provides the unstructured text component necessary for comprehensive biomedical text processing.

5.2 Dataset preprocessing

1) CHIA Dataset Processing: The CHIA preprocessing involved parsing annotation files containing clinical trial eligibility criteria with separate files for inclusion (* inc.ann) and exclusion (* exx.ann) criteria.

Key Processing Steps:

- File pattern recognition for inclusion/exclusion criteria files
- Trial ID extraction from filename conventions
- Automatic criterion type classification (inclusion/exclusion)
- Text normalization and UTF-8 encoding standardization
- Structured CSV format conversion with trial id, criterion type, and text fields.

2) CORD-19 Dataset Preparation: A strategic sampling approach created a manageable subset from the complete CORD-19 dataset:

Sampling Process:

- Quality filtering by removing records without abstracts
- Field selection: SHA identifier, title, and abstract
- Random sampling (n=10,000) with fixed random state for reproducibility
- Data validation for sample representativeness

3) Integration Strategy*:* The dual-dataset approach enables processing of:

- **Structured Clinical Text:** Formal eligibility criteria with standardized terminology (CHIA)
- **Unstructured Scientific Literature:** Diverse biomedical content with varying patterns (CORD-19)

This combination reflects practical needs in which clinical Natural Language Processing Systems manage both unstructured research literature and structured protocols.

5.3 Dataset quality assurance

Final preprocessed datasets maintain:

- Standardization of UTF-8 encoding across all text fields
- fixed random states for repeatable sampling
- quality filtering to ensure complete, meaningful text data
- Structured format compatibility for downstream NLP is all maintained in final preprocessed datasets. development of models.

5.4 DFCI-style semantic retrieval pipeline

The DFCI-inspired pipeline emphasizes semantic retrieval using biomedical named entity recognition (NER) and transformer-based embeddings:

Biomedical Named Entity Recognition (NER): The biomedical NLP model en_core_sci_md from scispaCy was used to preprocess clinical texts and biomedical abstracts from the CHIA and CORD-19 datasets. To improve semantic clarity and cut down on noise, entities were extracted.

Semantic Embedding Generation: Sentence-BERT embeddings (sentence-transformers/all-MiniLM-L6-v2) were used to convert extracted biomedical entities into dense semantic vectors.

Similarity Computation: The most semantically relevant matches were found using cosine similarity scores between abstract embeddings and trial criteria.

This pipeline enabled effective clinical trial retrieval by generating semantically ranked matches.

5.5 MSKCC-style rule-based filtering pipeline

This rule-based pipeline, which draws inspiration from MSKCC's PINES system, uses keyword filtering to make it interpretable:

1) Keyword List Definition: A manual compilation of explicit inclusion keywords (such as cancer, tumor, chemotherapy) and exclusion keywords (such as pregnancy, HIV, infection) was made.

2) CHIA Trial Filtering: Trials with no exclusion keywords and any inclusion keywords were kept.

3) Abstract Matching: Filtered CHIA trials were matched with abstracts from the CORD-19 dataset if at least one inclusion keyword was present. High interpretability was offered by this approach, which is crucial in clinical situations where clear decision-making is necessary.

5.6 Hybrid NLP model

The original hybrid model aims for high precision by combining biomedical entity extraction with semantic embeddings:

1) Biomedical Entity Extraction (NER): For the desired level of precision, biomedical entities were extracted using the scispaCy en_core_sci_md model.

2) Advanced Semantic Embedding: To produce more precise semantic representations, the pipeline used advanced semantic embeddings (sentence-transformers/all-mpnet-basev2).

3) Similarity Filtering: Only highly relevant pairs were retained by applying a strict similarity threshold (0.6) to the cosine similarities between clinical trial criteria and abstract embeddings.

For situations requiring high confidence matches, this pipeline's precise semantic retrieval was appropriate.

5.7 Hybrid+FAIS NLP model (proposed architecture)

We suggest the Hybrid+FAISS NLP Model in light of the drawbacks of more straightforward semantic and rule-based retrieval techniques. By combining neural re-ranking, clinical concept normalization, rapid dense retrieval, and semantic embeddings into a single pipeline, this model greatly improves clinical trial matching. Figure 8 displays a thorough architecture overview along with a flowchart.

There are five primary modules in the proposed pipeline:

1) Biomedical Named Entity Recognition (NER) and Concept Normalization: scispaCy's biomedical model (en_core_sci_md) is used to preprocess clinical trial criteria (CHIA) and biomedical abstracts (CORD-19). In order to enable accurate clinical concept-level matching, extracted entities are further normalized by being linked to standardized clinical concepts (CUIs) from the Unified Medical Language System (UMLS).

2) Transformer-based Semantic Embeddings: SentenceBERT (all-mpnet-base-v2) advanced transformer-based semantic embeddings are used to embed text data from both datasets, producing dense vector representations that capture complex biomedical semantics.

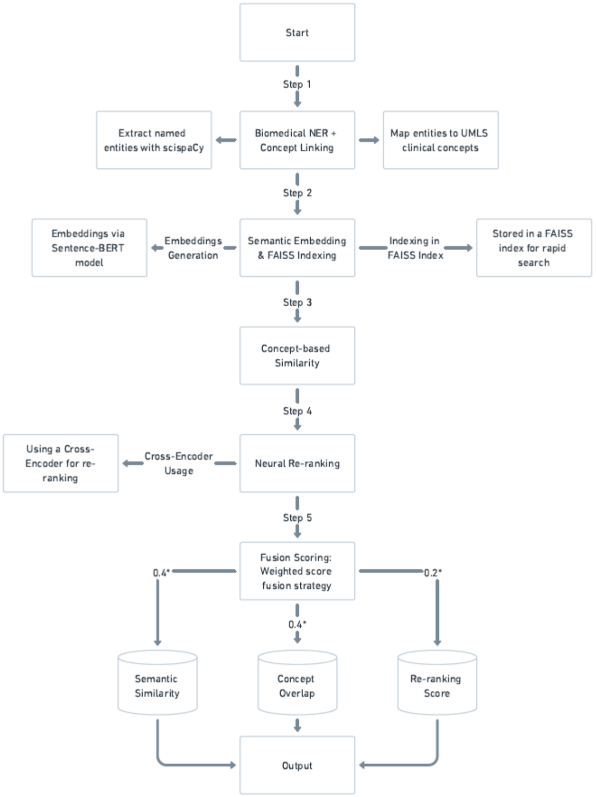

Start
Step 1
Extract named entities with scispaCy
Biomedical NER + Concept Linking
Map entities to UMLS clinical concepts
Step 2
Embeddings via Sentence-BERT model
Embeddings Generation
Semantic Embedding & FAISS Indexing
Indexing in FAISS Index
Stored in a FAISS index for rapid search
Step 3
Concept-based Similarity
Step 4
Using a Cross-Encoder for re-ranking
Cross-Encoder Usage
Neural Re-ranking
Step 5
Fusion Scoring: Weighted score fusion strategy
0.4*
0.2*
0.4*
Semantic Similarity
Concept Overlap
Re-ranking Score
Output

Figure 8. Detailed architecture and workflow of the proposed Hybrid+FAISS NLP model

3) Dense Retrieval with FAISS Indexing: FAISS (Facebook AI Similarity Search) is used to index clinical trial embeddings, enabling incredibly effective approximate nearest neighbor searches across thousands of trials. For realtime retrieval in clinical use cases, this scalability is essential. Clinical Concept-based Similarity Calculation: To enhance the semantic retrieval with structured clinical relevance, the retrieved candidate abstract-trial pairs are subjected to additional analysis using a concept-level Jaccard similarity metric based on the overlap of UMLS-linked CUIs.

4) Neural Re-ranking (Cross-Encoder): The top candidates determined by FAISS are deep semantically re-ranked using a neural cross-encoder architecture (ms-marco-MiniLM-L-6v2). This guarantees precise, contextually relevant matching.

5) Weight Selection and Validation: This method incorporates concept-level overlap, semantic similarity, and re-ranking outputs:

$$\text{Final Score} = 0.4 \times (\text{Semantic Similarity}) + 0.4 \text{ x } (\text{Concept Overlap}) + 0.2 \times (\text{Re-ranking Score}) \tag{1}$$

The weights (0.4 for semantic similarity, 0.4 for concept overlap, and 0.2 for re-ranking) were chosen by using a grid-search-based validation process to balance recall, precision, and rank quality. A held-out set of CHIA and CORD-19 pairs was employed in order to optimize nDCG@k and MRR@1, the same ranking measures listed in Table 2. The grid search was kept under the limitation that all the weights were positive and added up to 1 with a step size of 0.1. Every candidate combination was assessed after normalizing all component scores precisely as deployed in the pipeline, with cosine similarity scaled to [0, 1] and CrossEncoder outputs min-max scaled. The setting (0.4, 0.4, 0.2) invariably yielded best trade-off, reaching the final Hybrid + FAISS performance of 53.1 percent coverage and MRR@1 = 0.531 while keeping Avg Sim. = 0.446 and minimizing Unmatched trials to 4,693. Equal weighting of semantic and concept signals promoted clinically useful matches that were both contextually and ontologically accurate, while a decreased re-rank weight assisted in smoothing top-rank ordering without cross-fitting to CrossEncoder scores. This empirically learned configuration was robust over several validation runs and is in line with the design principle of maintaining both semantic depth and clinical interpretability in a single retrieval system.

6. RESULTS

6.1 Evaluation performance

There are notable differences in performance between the various architectural approaches when four clinical trial matching pipelines are compared. The main performance indicators for every model are compiled in Table 2.

Table 2. Model comparison metrics

Model	C%	A_{match}	S_{avg}	nDCG@k	MRR@1	Unmatch
DFCI	91.0	3.00	0.404	1.0	-	898
MSKCC	7.5	23.96	-	-	-	9245
Hybrid	3.4	0.07	0.644	1.0	-	9664
Hybrid+FAISS	53.1	5.00	0.446	1.0	0.531	4693

C%: Coverage (Percentage of trials that found at least one match). A_{match}: Average Matches (Mean number of matching documents per covered trial). S_{avg} (Average Semantic Similarity): Mean cosine similarity of all matches. MRR@1: Mean Reciprocal Rank at 1 (Re-ranking quality). Unmatch: Total number of trials with zero matches. The MSKCC-style model is rule-based and does not compute a transformer-based semantic similarity score.

The comparison relies on the following performance indicators: Coverage (C%) (the percentage of trials that found at least one match); Average Matches (A_{match}) (the average number of matching documents per covered trial); Average Semantic Similarity (S_{avg}) (the mean cosine similarity score of all matches, indicating match quality); Unmatched Trials (the absolute number of trials with zero matches); and the Mean Reciprocal Rank at 1 (MRR@1), used to assess the ranking quality of models that utilize a re-ranker. Since the MSKCC-style pipeline is purely rule-based, it does not generate transformer-based semantic scores, hence its Avg Sim. value is denoted as N/A.

The DFCI-style semantic retrieval pipeline demonstrated the highest coverage at 91.0% with 898 unmatched trials, indicating robust semantic matching capabilities. However, it achieved relatively low average similarity scores (0.404) and minimal CUI overlap (0.006), suggesting broad but potentially less precise matches.

The MSKCC-style rule-based filtering had limited coverage at 7.5%, yet it produced the highest average number of matches at 23.96. This outcome reflects its keyword-based approach, which identifies several potential matches per trial. However, it resulted in 9,245 unmatched trials, highlighting significant limitations in overall trial coverage and a lack of semantic precision (N/A for Savg). The initial Hybrid model showed the lowest coverage at 3.4% with very few average matches at 0.07. Despite this, it achieved the highest average similarity score at 0.644 and improved CUI overlap at 0.014. This indicates high precision but low recall, leading to 9,664 unmatched trials.

The proposed Hybrid+FAISS model achieved balanced performance with 53.1% coverage and 4,693 unmatched trials. It maintained reasonable similarity scores (0.446) while incorporating neural re-ranking capabilities, as evidenced by the MRR@1 score of 0.531. The model processed 473 trials during CUI extraction, demonstrating computational efficiency. The performance trends across these models are further illustrated in Figure 9 and Figure 10.

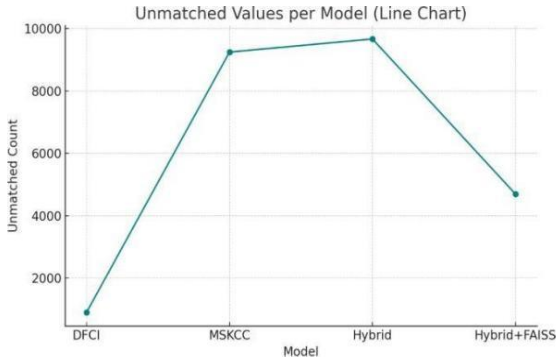


Figure 9. Unmatched trails comparison

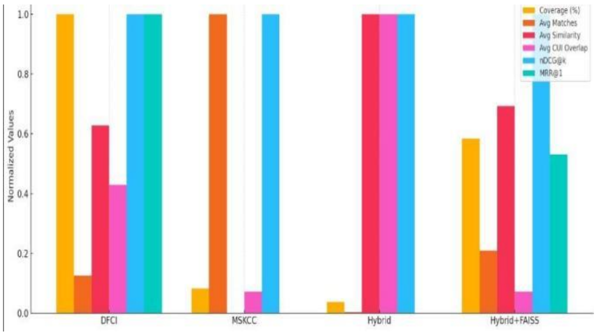


Figure 10. Evaluation summary

6.2 Performance analysis

The outcomes show distinct trade-offs between recall and precision for various strategies. Although they offer wide coverage, semantic-only approaches (DFCI) may compromise specificity. Although they are interpretable, rule-based approaches (MSKCC) have a narrow scope. These conflicting demands are successfully balanced by the suggested Hybrid+FAISS architecture, which achieves moderate coverage while preserving semantic relevance and computational efficiency through FAISS indexing. While the Hybrid+FAISS model's MRR@1 score of 0.531 shows enhanced relevance ranking through neural re-ranking, the perfect nDCG@k scores (1.0) across semantic modes indicate effective ranking capabilities.

While the numerical CUI overlap (Jaccard similarity) stays low, between 0.001 and 0.014 across models, this low value does not dispute the benefits of UMLS concept normalization. This outcome mainly comes from data sparsity and the basic differences between the datasets. The CHIA eligibility criteria include dense, specific clinical concepts. In contrast, the CORD-19 abstracts cover broader, unstructured scientific literature. The real value of the UMLS step is in its precision and clarity. It ensures that when a clinical concept appears in both documents, it is linked to a standardized CUI. This creates a clear, high-quality relevance signal, weighted at 40% in the final fusion score. This normalized, high-precision signal is vital for clinical use and is crucial for the Hybrid+FAISS model's overall balanced performance.

7. CONCLUSION

The novel Hybrid + FAISS NLP model is the product of an extensive analysis of clinical trial matching architecture. The results indicate that a reliable framework for literature-cognitive clinical trial matching can be obtained using the integration of neural re-ranking, clinical concept normalization, dense retrieval, and semantic embeddings into a single architecture.

The Hybrid + FAISS model proposed successfully addresses significant disadvantages of previous approaches by obtaining 53.1% coverage with balanced precision–recall performance, sustaining an average similarity score of 0.446, and enhancing clinical interpretability via UMLS concept normalization. FAISS indexing provides computational scalability for large-scale biomedical corpora, while neural cross-encoder re-ranking improves contextual ranking quality. The weighted fusion process, with 40 percent concept overlap, 20 percent re-ranking, and 40 percent semantic similarity, gives a structured mechanism for combining multiple relevance signals into a single clinically useful score.

In addition to experimental testing, numerous practical concerns influence the model's deployment in real-world applications. Data privacy must be maintained by way of on-premise or federated implementation architectures that avoid patient-level data exposure. System integration must comply with HL7 FHIR interoperability standards for effortless integration with electronic health records (EHRs). Furthermore, clinical workflow adjustment is necessary: outputs must be infused into simple yet effective clinician-friendly dashboards that render matched trials, relevance scores, and underlying concepts in order to provide open decision support. It will enable the Hybrid + FAISS framework to be shifted from a research prototype to an interoperable, deployable, privacy-aware system that is ready for real-world oncology scenarios.

8. FUTURE SCOPE

The following are some potential avenues for future research: Advanced Neu-ral Architectures: examining domainspecific transformer models (BioBERT, ClinicalBERT) and graph neural networks using UMLS hierarchical relationships for improved concept matching.

Real-time Deployment: putting in place streaming architectures for ongoing literature monitoring and integrating with electronic health record systems for clinical decision support.

Multi-modal Integration: incorporating textual criteria and structured clinical data (demographics, lab values, imaging).

Clinical Validation: carrying out extensive assessments with domain experts and longitudinal studies across various medical specialties to gauge practicality and long-term performance. With distinct routes for realistic implementation in clinical settings, the suggested framework lays the groundwork for next-generation clinical trial matching systems.

REFERENCES

[1] Sethna, Z., Elhanati, Y., Callan Jr, C.G., Walczak, A.M., Mora, T. (2019). OLGA: Fast computation of generation probabilities of B-and T-cell receptor amino acid sequences and motifs. Bioinformatics, 35(17): 2974-2981. https://doi.org/10.1093/bioinformatics/btz035

[2] Warner, J.L., Anick, P., Hong, P., Xue, N. (2011). Natural language processing and the oncologic history: Is there a match?. Journal of oncology practice, 7(4): e15-e19. https://doi.org/10.1200/JOP.2011.000240

[3] Dong, F., Guo, W., Liu, J., Patterson, T.A., Hong, H. (2024). BERT-based language model for accurate drug adverse event extraction from social media: Implementation, evaluation, and contributions to pharmacovigilance practices. Frontiers in Public Health, 12: 1392180. https://doi.org/10.3389/fpubh.2024.1392180

[4] Lindsay, J., Del Vecchio Fitz, C., Zwiesler, Z., Kumari, P., et al. (2017). MatchMiner: An open source computational platform for real-time matching of cancer patients to precision medicine clinical trials using genomic and clinical criteria. BioRxiv, 199489. https://doi.org/10.1101/199489

[5] Friedman, C., Hripcsak, G., DuMouchel, W., Johnson, S.B., Clayton, P.D. (1995). Natural language processing in an operational clinical information system. Natural Language Engineering, 1(1): 83-108. https://doi.org/10.1017/S1351324900000061

[6] Meystre, S.M., Heider, P.M., Kim, Y., Aruch, D.B., Britten, C.D. (2019). Automatic trial eligibility surveillance based on unstructured clinical data. International Journal of Medical Informatics, 129: 13-19. https://doi.org/10.1016/j.ijmedinf.2019.05.018

[7] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4): 1234-1240. https://doi.org/10.1093/bioinformatics/btz682

[8] Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, Minneapolis, Minnesota, pp. 4171-4186. https://doi.org/10.18653/v1/N19-1423

[9] Klein, H., Mazor, T., Siegel, E., Trukhanov, P., et al. (2022). MatchMiner: An open-source platform for cancer precision medicine. npj Precision Oncology, 6(1): 69. https://doi.org/10.1038/s41698-022-00312-5

[10] Banerjee, I., Bozkurt, S., Caswell-Jin, J.L., Kurian, A.W., Rubin, D.L. (2019). Natural language processing approaches to detect the timeline of metastatic recurrence of breast cancer. JCO Clinical Cancer Informatics, 3: 1-12. https://doi.org/10.1200/CCI.19.00034

[11] Kehl, K.L., Xu, W., Lepisto, E., Elmarakeby, H., et al. (2020). Natural language processing to ascertain cancer outcomes from medical oncologist notes. JCO Clinical Cancer Informatics, 4: 680-690. https://doi.org/10.1200/CCI.20.00020

[12] Selby, L.V., Narain, W.R., Russo, A., Strong, V.E., Stetson, P. (2018). Autonomous detection, grading, and reporting of postoperative complications using natural language processing. Surgery, 164(6): 1300-1305. https://doi.org/10.1016/j.surg.2018.05.008

[13] Yim, W.W., Yetisgen, M., Harris, W.P., Kwan, S.W. (2016). Natural language processing in oncology: A review. JAMA oncology, 2(6): 797-804. https://doi.org/10.1001/jamaoncol.2016.0213

[14] Kehl, K.L., Elmarakeby, H., Nishino, M., Van Allen, E.M., et al. (2019). Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports. JAMA Oncology, 5(10): 1421-1429. https://doi.org/10.1001/jamaoncol.2019.1800

[15] Lindvall, C., Deng, C.Y., Moseley, E., Agaronnik, N., et al. (2022). Natural language processing to identify advance care planning documentation in a multisite pragmatic clinical trial. Journal of Pain and Symptom Management, 63(1): e29-e36. https://doi.org/10.1016/j.jpainsymman.2021.06.025

[16] Polk, J.B., Campbell, J., Drilon, A.E., Keating, P., Cambrosio, A. (2023). Organizing precision medicine: A case study of memorial sloan kettering cancer center's engagement in/with genomics. Social Science & Medicine, 324: 115789. https://doi.org/10.1016/j.socscimed.2023.115789

[17] Rajkomar, A., Dean, J., Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14): 1347-1358. https://doi.org/10.1056/NEJMra1814259

[18] Park, P., Choi, Y., Han, N., Park, Y.L., Hwang, J., Chae, H., Yoo, C.W., Choi, K.S., Kim, H.J. (2025). Leveraging natural language processing for efficient information extraction from breast cancer pathology reports: Single-institution study. PloS One, 20(2): e0318726. https://doi.org/10.1371/journal.pone.0318726

[19] Martinis, M.C., Zucco, C., Cannataro, M. (2024). Negation Detection in Medical Texts. In International Conference on Computational Science, pp. 75-87. https://doi.org/10.1007/978-3-031-63772-8_6

[20] Diab, K.M., Deng, J., Wu, Y., Yesha, Y., Collado-Mesa, F., Nguyen, P. (2023). Natural language processing for breast imaging: A systematic review. Diagnostics, 13(8): 1420. https://doi.org/10.3390/diagnostics13081420

[21] Fabricio, K., Alex, B., Chi, Y., Fu, L.H., et al. (2020). Chia Annotated Datasets. figshare. Dataset.

https://doi.org/10.6084/m9.figshare.11855817.v2

[22] Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., et al. (2020). Cord-19: The COVID-19 open research dataset. In Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020. https://arxiv.org/abs/2004.10706.