Precision Mining of Potential Recruited Patients for Clinical Trials Based on Multimodal AI Algorithms

Nanyang Guo, Chen Gong, Peng Shen
Waterdrop Research Institute, Beijing, China

Abstract: The recruitment of clinical trial patients faces challenges such as information heterogeneity, semantic gap, and lack of temporal correlation. Traditional methods are difficult to achieve accurate and efficient potential subject mining. This article proposes a precision mining framework based on multimodal AI algorithms. Firstly, it elaborates on the medical informatics foundation of multimodal data fusion, semantic representation of inclusion and exclusion standards, and unified representation theory of heterogeneous patient data; Furthermore, key challenges such as difficulty in aligning multimodal data, mismatched information granularity, bottlenecks in text image fusion, and missing temporal window associations will be analyzed; Finally, a multimodal semantic alignment method for inclusion criteria, a joint representation learning architecture that integrates text and images, a dynamic matching model for temporal perception, and an interpretability mining framework are proposed. This framework provides a systematic technical path and theoretical support for intelligent clinical trial patient screening.

Keywords: Multimodal AI Algorithm; Clinical Trials; Accurate Patient Excavation; Semantic Alignment of Input and Output Standards

1. Introduction

The successful conduct of clinical trials highly relies on the efficiency and accuracy of patient recruitment, and traditional manual screening methods are time-consuming, labor-intensive, and prone to missing potential subjects. Multimodal AI algorithms can integrate structured examination data, unstructured text, and medical imaging information from electronic medical records, providing a new approach for accurately mining patients who meet inclusion and exclusion criteria. This article focuses on the theoretical basis of multimodal data fusion and clinical trial adaptation, systematically analyzes the key challenges faced by patient mining, and proposes a strategic framework for semantic alignment of inclusion and exclusion criteria, text image joint representation, temporal dynamic matching, and interpretability mining, aiming to provide theoretical support and technical reference for intelligent patient recruitment.

2. Multimodal AI Algorithm and Clinical Trial Adaptation Theory

2.1 Medical Informatics Fundamentals of Multimodal Data Fusion

The medical informatics foundation of multimodal data fusion lies in addressing the heterogeneity of clinical data in terms of format, granularity, and semantics. Structured fields (such as test indicators and diagnostic codes), semi-structured text (such as medical records), and unstructured images (such as CT and MRI) in electronic medical records naturally have spatiotemporal inconsistencies and information redundancy in the data generation process[1]. Medical informatics provides an interoperable underlying structure for multimodal data by establishing a unified data model (such as FHIR resource framework) and standardized terminology system (such as ICD-10, SNOMED CT). On this basis, fusion methods need to balance complementarity and redundancy: complementary fusion uses images to reveal anatomical structures and text contains functional states, while redundant fusion improves the robustness of features through multi-source cross validation. This lays a theoretical foundation for subsequent algorithms to convert clinical raw data into feature matrices that can be used for deep learning.

2.2 Semantic Representation Framework for Inclusion and Exclusion Criteria in Clinical

Trials

The inclusion and exclusion criteria for clinical trials are presented in natural language text form, including various constraints such as diagnosis, laboratory thresholds, past treatment history, and organ function status. Their semantic complexity is reflected in logical nesting ("and/or/not"), temporal dependence ("within the past 6 months"), and implicit medical knowledge ("severe liver injury" needs to be combined with Child Pugh score)[2]. The core of semantic representation framework is to transform free text input and output standards into computer understandable and executable logical units. This framework typically consists of three levels: entity extraction layer (identifying medical concepts such as diseases, drugs, and examination items), relationship construction layer (extracting comparison operators, time windows, and logical connectors), and knowledge enhancement layer (referencing external medical knowledge bases to complete concept standardization and threshold mapping). The final generation of structured rule expressions or query graphs such as "age ≥ 18 years old and have not received chemotherapy" provides a semantic bridge for automatic matching between patients and trials.

2.3 Unified Representation Model for Multi-Source Heterogeneous Data of Patients

The patient's multi-source heterogeneous data includes structured test indicators, unstructured text reports, time-series vital signs, and image sequences. Directly inputting them into classification or matching algorithms faces the challenges of exploding feature dimensions and semantic alignment. The goal of the unified representation model is to map the multi-source data mentioned above to the same embedding space, forming a comparable and computable patient vector representation. The existing methods include cross modal fusion based on Transformer architecture (such as aligning text reports with image features through attention mechanism), patient medical knowledge graph constructed by graph neural network (with entities as nodes and relationships as edges), and longitudinal data evolution trajectory processed by temporal encoder[3]. Unified representation requires two key properties: modality consistency (text descriptions and image features of the same patient are close in distance in the embedding space) and clinical task separability (patient representations with different inclusion states form separable clusters in space). This representation is directly used as input for downstream matching algorithms, determining the core performance of the mining system.

3. Exploring Potential Patients in Clinical Trials Faces Challenges

3.1 Heterogeneity and Alignment Difficulties of Multimodal Medical Data

Medical data is naturally heterogeneous in terms of source, format, and semantics: test indicators are continuous numerical values, disease records are unstructured text, images are pixel matrices, and gene sequencing is discrete variant sites. There is a lack of natural spatiotemporal alignment anchors between different modalities, for example, lung CT images and bronchoscopy pathology reports of the same patient may differ by several days in time, making it difficult to establish pixel level correspondence in space. This heterogeneity leads to problems such as high feature dimensions, missing modalities, and inconsistent sampling frequencies in multimodal fusion algorithms. Direct concatenation or simple weighted fusion often introduces noise rather than effective information.

3.2 Inconsistent Granularity between Inclusion and Exclusion Criteria and Electronic Medical Record Information

Clinical trial inclusion and exclusion criteria are often described using medical concepts (such as "refractory hypertension" or "mild renal insufficiency"), while electronic medical records record raw test values (such as creatinine 96 μ mol/L) or discrete diagnostic codes. Conceptual standards need to be combined with multi indicator joint judgment and domain knowledge thresholds to be transformed into executable conditions, but key indicators may be missing or incompletely recorded in medical records. The mismatch in granularity leads to a large number of patients being misjudged as unqualified due to incomplete information, or the algorithm cannot accurately determine their true status.

3.3 Bottleneck in the Fusion of Unstructured Medical Text and Image Features

There is a semantic gap between unstructured text (such as image reports and discharge summaries) and the original image: the text describes the radiologist's abstract interpretation

of the image ("ground glass like density shadow visible in the lower left lung"), while the image itself contains pixel level texture, edge, and density information. The text has lost the original details of the image, and the image lacks high-level semantic annotation in natural language. When directly fusing the two, it is easy to generate information redundancy or contradictions, and there is a lack of joint annotated training data, which limits the performance of end-to-end multimodal models.

3.4 Missing Association between Patient Time-Series Data and Dynamic Window of the Trial

The inclusion and exclusion criteria for clinical trials often include strict time constraints ("no use of biologics within 4 weeks prior to enrollment" or "first treatment must be received within 6 weeks after diagnosis"), and patients' electronic medical records are discretely recorded in an event driven manner, with state changes between two time points often being blind spots. In addition, the experiment itself has a dynamic window (such as a screening period lasting 14 days), during which the patient's test indicators may fluctuate. Most existing methods treat patient states as static snapshots and fail to model the dynamic correlation between temporal evolution and trial windows, resulting in a large number of patients who meet dynamic conditions being missed.

4. Precise Mining Strategy Based on Multimodal AI Algorithm

4.1 Multimodal Semantic Alignment Method for Inclusion and Exclusion Standards

To address the semantic gap between multimodal medical data and inclusion/exclusion criteria, a cross modal alignment mechanism centered on inclusion/exclusion criteria needs to be constructed. The core idea of this method is to treat each criterion in the inclusion and exclusion criteria as a semantic anchor, and map patient data of different modalities to the semantic space defined by the anchor[4]. Specifically, the inclusion and exclusion criteria text is first finely parsed into logical units that include constraint types (such as inclusion or exclusion), medical attributes (such as "serum creatinine"), comparison operators (such as "≥"), thresholds (such as "1.2 mg/dL"), and time windows (such as "within the first 7 days of the screening period"). Subsequently, for each semantic unit, design corresponding modal specific feature extractors: for structured inspection data, directly extract the values of corresponding indicators and normalize them; For unstructured text such as medical records and imaging reports, pre trained medical language models (such as BioBERT and ClinicalBERT) are used to extract contextual representations related to the target medical attributes; For medical images such as CT and MRI, 3D convolutional neural networks or visual transformers are used to extract image biomarkers (such as lesion size and density features) that are semantically related to the criteria. On this basis, a cross modal attention mechanism is introduced to enable the model to dynamically focus on the most relevant modal information based on the current criteria, and through contrastive learning, multiple modal representations of the same patient are brought closer to the inclusion and exclusion criteria, while different patients or irrelevant modal representations are pushed further away. This method not only solves the problem of modal heterogeneity alignment, but also provides a semantically consistent and interpretable feature basis for subsequent patient trial matching.

4.2 Joint Representation Learning Architecture Integrating Text and Image

In the face of the semantic gap between unstructured text and raw images, it is necessary to construct a joint representation learning architecture that can simultaneously understand information from both modalities[5]. This architecture is centered around a multimodal Transformer, with separate text encoders and image encoders set up. The text encoder uses pre trained medical language models (such as BioBERT) to convert clinical descriptions in image reports or discharge summaries into high-dimensional semantic vectors; Image encoders use 3D convolutional neural networks or visual transformers to extract spatial structure and texture features from raw medical images. To achieve effective fusion of two modalities, a cross modal attention module is introduced in the deep layers of Transformer, allowing text features to query spatial regions related to descriptions in images through attention mechanisms, while allowing image features to reverse query the corresponding diagnostic semantics in the text. This module adopts a

gating mechanism to dynamically control the information flow intensity between different modalities, avoiding the interference of noise modalities on the fusion results. In addition, a modal alignment loss function is designed in the architecture, which requires the text representation and image representation of the same patient to maintain a close distance in the joint embedding space, while different patients or mismatched text image pairs are kept away from each other. By pre training on large-scale unlabeled or weakly labeled medical data, this architecture can learn cross modal joint representations with clinical discriminative ability, providing rich feature representations that integrate morphological and semantic information for subsequent patient trial matching.

4.3 Patient Trial Dynamic Matching Model with Temporal Perception

To address the issue of missing associations between patient time-series data and experimental dynamic windows, a time-series aware dynamic matching model needs to be constructed. This model regards the patient's historical electronic medical record sequence as a multimodal event stream at discrete time points, and uses a time aware Transformer encoder to capture the time interval information and evolutionary trends between events. Unlike traditional static matching methods, the model introduces the concept of a time sliding window during the trial screening period, dynamically generating multiple candidate matching segments on the patient timeline, with each segment corresponding to a time interval of the specified screening days for the trial. The model calculates the aggregated representation of patient modal data for each candidate segment during that time period, including the trend change slope of the test indicators, the occurrence and disappearance time of key events, and the temporal consistency of imaging features. At the same time, the time constraints in the inclusion criteria are encoded as differentiable temporal logical predicates, and criteria such as "whether an event has occurred in the past N days" or "whether M consecutive measurements meet a certain condition" are simulated through neural networks. The matching score is calculated using a bidirectional attention mechanism: on the one hand, it evaluates the temporal characteristics of patient segments and their compliance with the time constraints of inclusion and exclusion criteria; on the other hand, it reverse weights the time periods that do not meet the constraints to provide interpretable attribution of failure. The final output includes the probability of patients meeting the trial conditions within future time windows and recommendations for the optimal enrollment time.

4.4 Heterogeneous Data-Driven Interpretability Mining Framework

In clinical trial patient mining, the interpretability of algorithm decisions directly affects the trust of clinical doctors and the pass rate of ethical review. To address the "black box" problem caused by multimodal heterogeneous data, it is necessary to construct a mining framework with interpretability as the core. This framework adopts a dual layer strategy combining post hoc interpretation and self explanatory models. The first layer is the feature attribution layer, which uses Shapley value based feature contribution decomposition for structured test data to quantify the marginal contribution of each assay indicator to the matching results; For unstructured text, attention heatmap and token attribution techniques are introduced to highlight keywords and phrases that trigger inclusion criteria in disease records or imaging reports (such as "metastasis" and "new lesion"); For medical imaging, a gradient weighted class activation mapping is used to generate a spatial heatmap of the lesion area, visually displaying the anatomical locations that affect matching decisions in the image. The second layer is the rule self interpretation layer, which distills the internal decisions of the multimodal fusion model into a set of readable clinical decision rules (such as "if the image shows a lesion diameter $\geq$ 3cm and CA19-9 levels rise twice in a row, it is judged to meet the inclusion criteria"), with confidence and supported patient sample size for each rule. The framework also features an interactive visual interface that allows doctors to trace back the patient's various modal data and model judgment criteria along a timeline, and supports counterfactual inference of a single criterion ("how the matching results will change if the patient's creatinine level drops to the normal range"). By presenting the decision-making process driven by heterogeneous data in a clinically understandable way, this framework significantly improves the transparency and

clinical usability of the system while maintaining high mining accuracy.

5. Conclusion

This article systematically explores the problem of precise exploration of potential patients in clinical trials based on multimodal AI algorithms. Firstly, starting from the foundation of medical informatics, a theoretical framework was constructed for the semantic representation of inclusion and exclusion standards and the unified representation of heterogeneous patient data; Furthermore, the core challenges of multimodal data heterogeneity, granularity mismatch, text image fusion bottleneck, and lack of temporal correlation were analyzed. On this basis, four strategies were proposed: multimodal semantic alignment for input and output standards, text image joint representation learning, dynamic matching model for temporal perception, and interpretability mining framework driven by heterogeneous data. Research has shown that multimodal AI algorithms can effectively integrate structured and unstructured clinical data, bridge the semantic and granularity gap between inclusion and exclusion criteria and real medical records, and achieve accurate and dynamic matching between patients and trials. Future work will focus on the prospective validation of federated learning paradigms and algorithms for multi center heterogeneous data in real clinical trials.

References

[1] Liu H Y, Yang Y, Li G L, et al. Application of Python-based Apriori Algorithm in Information Mining of Patients with Extremely Long Hospital Stay. Chinese Medical Record, 2022, 23(06): 68-71.

[2] Wang Q H. Research on Risk Factor Mining of Plateau-Related Cholelithiasis Patients Based on Machine Learning Intelligent Algorithms. Qinghai Province: Qinghai Normal University, 2021-06-18.

[3] Jin H P, Tao Y Q, Li Z H, et al. Survival Prediction Algorithm for Gastric Cancer Patients Based on Multimodal Multi-Instance Learning. Journal of Computer-Aided Design & Computer Graphics, 2025, 37(02): 349-360.

[4] Park E S, Wollstein Y, Hutchison L A, et al. Recruitment of Minority Patients in Metabolic Dysfunction-Associated Steatotic Liver Disease Clinical Trials. Digestive Diseases and Sciences, 2026 (prepublish): 1-7.

[5] Tan Y H, Peng J Y, Fang C, et al. Design and Application of an Intelligent Recruitment and Screening Information System for Trial Participants. Tumor Pharmacy, 2025, 15(04): 475-479.

