# Research on the Key Technologies of Medical Data Mining and Its Application in Clinical Decision-Making under the Background of Smart Healthcare

**Guangyu He**

Macau University of Science and Technology, Taipa, Macau, China

heguangyu09@163.com

**Keywords:** Smart healthcare; Medical data mining; Clinical decision-making; Machine learning

**Abstract:** Medical data mining in the context of smart healthcare (abbreviated as medical data mining in the following text) refers to disciplines that focus on medical informatics, biostatistics, artificial intelligence, and other related knowledge related to medical data. It is a general term for information science and medical science and has distinct interdisciplinary characteristics and application orientation. Compared with a single discipline, it presents more complex attributes. It is impossible to quantify and apply traditional statistical models to achieve deeper exploration in the research, which limits the diversity of research methods. The academic community's attention to research methods for medical data mining is increasing with the arrival of big data and artificial intelligence. In recent years, scholars have emphasized the combination of innovation and interdisciplinary approaches, with machine learning technology construction and application in medical data mining being a typical representative. However, due to the complexity of the research objects and issues involved in this field, the academic community still needs to solve the problem of effectively integrating machine learning technology with medical data mining. This study attempts to explore the practical application of the above theories to provide a theoretical basis and practical reference for promoting the intersection and innovation of internal and external elements in the discipline.

## 1. Introduction

The relevant policy documents on smart healthcare clearly state the need to strengthen the connection between medical services and health management, and the service process should reflect the continuity and advancement of personalization and precision. Medical workers need to understand the characteristics of patients and the characteristics of medical data to prepare patients for further treatment in advance. Data mining skills are an crucial component of healthcare workers forming the ideas and methods of smart healthcare, as well as the foundation for learning clinical knowledge and forming precision healthcare concepts [1].

This study adopts quantitative analysis, model construction, case study, and other methods to conduct comparative analysis on the collection, processing, and analysis of clinical data, based on data mining skills. Analysis shows that clinical data from different sources have great similarity and connectivity, and the correlation between the data is high. The content learned in the primary stage is involved in the advanced stage, but the content learned in the advanced stage is more complex. This article collects data on the quality and application status of medical data through data collection, selects different levels of medical institutions as control groups for field research and analysis, and analyses existing problems, finally concludes. This article presents the following strategies to cultivate the skills of healthcare workers: emphasizing the learning of data mining theory, emphasizing the cultivation of data analysis ability, combining with electronic medical record platforms to form intuitive feelings, emphasizing clinical guidance, and selecting data preprocessing, feature selection, and model application based on existing medical data and clinical decision-making needs for systematic design to enable the application of strategies.

Finally, this article provides a summary and outlook on the aspects of smart healthcare, including healthcare workers, medical institutions, medical management, and strengthening patient data privacy protection.

  DOI: 10.25236/semihs.2024.023

## 2. Basic Concepts of Medical Data Mining for Smart Healthcare

### 2.1. Definition and Scope of Medical Data Mining

Medical data mining research involves two fields: data preprocessing and data modelling. Data preprocessing focuses on the cleaning, standardization, and transformation of raw medical data, and usually uses data cleaning techniques and data transformation techniques to deeply explore individual clinical, physiological, genetic, and behavioral information, covering fields such as diagnosis, treatment, pharmacy, epidemiology, and psychology [2]. Data modelling generally focuses on the structure and patterns of medical data and utilizes statistical analysis and machine learning techniques to analyze macro medical phenomena in fields such as disease prediction, treatment effectiveness evaluation, and health management. Although there is a theoretical difference between data preprocessing and data modelling, in reality, "raw data" and "processed data" are closely connected and inseparable. Medical phenomena are deeply influenced by individual characteristics, and medical data mining inevitably involves multidimensional data analysis. Therefore, this study aims to provide an overview of the broad content of medical data mining, without specific classification or differentiation.

### 2.2. Medical Data Mining Information and Its Characteristics Classification

The data types involved in medical data mining research are complex, with various classification standards. This study only defines and distinguishes two aspects: data sources and data structures.

#### 2.2.1. Medical Data Source Perspective

From this perspective, the data involved in medical data mining research can be divided into raw and secondary data. The former refers to the data collected and processed by researchers for the first time through electronic medical record systems, wearable devices, gene sequencing, and more, while the latter refers to data sourced from research results of others and public medical databases. The original data source reflects the enormous value of medical data mining data. It adds practical difficulties to the standardization and large-scale utilization of subsequent data standards [3].

#### 2.2.2. Medical Data Structure Perspective

From the broad perspective of data types involved in scientific research activities, medical data can be divided into three categories: structured, semi-structured, and unstructured data. Structured data refers to digital health record data, such as electronic medical record data directly generated by hospital information systems, with strict standardization characteristics, and use relational databases to store and manage data. Semi-structured data refers to data with irregular or incomplete structures, without predefined data models. Common examples include medical imaging, clinical text records, and more Unstructured data refers to no fixed structure data, such as handwritten notes by doctors, patient narratives, etc, with mixed content and data structure characteristics. In medical data mining, data is mainly composed of semi-structured and unstructured data, which means that a single piece of data contains a large amount of information, a weak structure, and an unstable data format, making it impossible for researchers to analyse and utilize it directly. Standardization and data formatting are the main characteristics that distinguish it from other data fields, and they are also the prerequisites for the two key processes of data preprocessing and data modelling mentioned later [4].

### 2.3. Knowledge Graph Technology and Some Unique Features of Knowledge Graph Technology in Medical Data Mining

Knowledge graph technology, also known as graph technology or knowledge visualization technology, is an interdisciplinary technology that combines theoretical methods from disciplines such as medical informatics and data science with corresponding methods from computer science. Its basic application logic is to present diseases, symptoms, drugs, and more in a structured knowledge manner through visual graphics to achieve systematic and structured medical knowledge. From the perspective of graph application scale, knowledge graph technology can be divided into general knowledge graphs and professional domain knowledge graphs [5]. Among them, the General

Knowledge Graph is a structured knowledge base that contains a wide range of domain information, covering comprehensive knowledge from multiple disciplines and thematic areas, but the graph accuracy is low. The professional domain knowledge graph is a structured knowledge base that focuses on specific fields or topics. Compared to the former, the professional domain knowledge graph is generally constructed by experts in the field, focusing on deep mining of key information in the field. Therefore, the graph has higher accuracy and can efficiently assist in deep research and decision support. The specific architecture is shown in Figure 1.

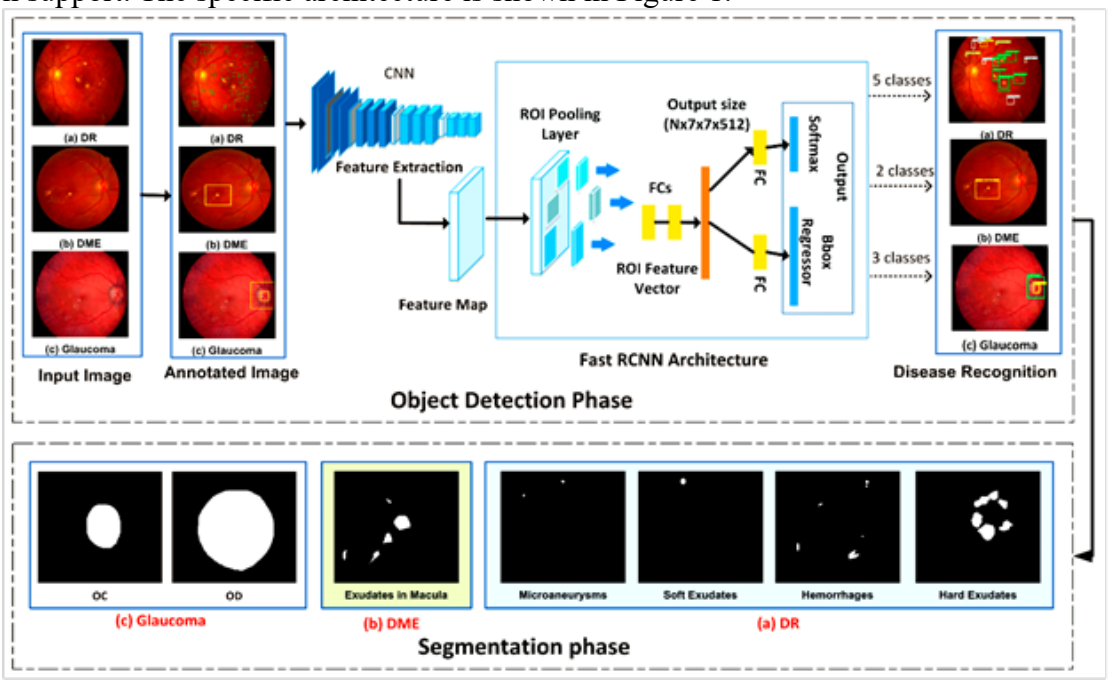


Figure 1: Medical data mining information and its characteristics classification architecture diagram

Unlike general knowledge graphs, medical data has different requirements in terms of accuracy, professionalism, and timeliness, and is required to solve specific problems. Taking the construction of knowledge graphs in medical data mining as an example, its particularity lies in firstly, in terms of data foundation, it is necessary to overcome the challenge of unstructured data. As mentioned earlier, most medical data is difficult to represent in a structured way. Therefore, when constructing knowledge graphs, natural language processing and other technologies need to be used for information extraction and basic database construction. Secondly, in terms of modelling logic, considering the subjectivity and complexity of medical knowledge, constructing a knowledge graph based on a single dimension of modelling logic cannot achieve good results. It is necessary to comprehensively reflect the knowledge architecture through comprehensive strategies and overlapping parameters in practice. Thirdly, in terms of interdisciplinary expression, the construction process needs to comprehensively consider the weights of different disciplines and the unity of concepts between disciplines, to achieve comprehensive expression from the perspective of big health disciplines. Fourthly, in terms of the unity of single decision-making and group intelligence, the construction of knowledge graphs in the field of medical data mining is a deep integration of knowledge graph technology into the field of medical data mining. This is mainly reflected in the introduction of the will of "domain experts" in the knowledge graph construction process. At the same time, there is a prominent contradiction that "group intelligence algorithms focus on how to make group intelligence emerge and surpass individual intelligence, but lack the mechanism to evolve individual intelligence. Therefore, they cannot become a self-evolving knowledge graph body without significant expansion." Therefore, their effective construction depends on the prior knowledge of domain experts in the knowledge graph stage for the overall domain's control, coverage, and accuracy. Only by achieving unity between individuals and groups can the practicality of knowledge graphs be better improved. Professionalism and accuracy [6].

## 3. Technical Process of Knowledge Graph Construction in Medical Data Mining

### 3.1. Construction of Medical Data Mining Knowledge Framework

In the process of constructing a knowledge graph, the construction of a knowledge framework is the first and most crucial step, commonly referred to as the "ontology construction" stage. Its ontology is a model that describes the structure and organization of a knowledge graph, defining the types of entities, concepts, and relationships in the knowledge graph and their relationships, typically expressed in ontology language. In this process, it is necessary to use a unified form to represent the pattern layer data of the upper-level knowledge system of the data entity, requiring the builder to have a certain level of medical domain abstraction ability and overall control of the entire domain knowledge, to maximize the restoration of known semantic relationships between existing knowledge. Taking the well-known disease ontology as an example, with disease as the narrative core, the narrative contains a lot of information such as disease, symptoms, examination, treatment, and medication, which can abstract narrative into a simple ontology structure, as shown in Figure 1, based on which the corresponding disease knowledge structure can be fully presented [7].

### 3.2. Medical Data Mining Information Extraction Mode

In knowledge graph construction (ontology construction), information extraction refers to the process of automatically extracting structured and meaningful information and knowledge from text, images, and sound. This process is of great significance for the field of medical data mining, which mainly focuses on unstructured and semi-structured data. This process aims to extract information from a large amount of text for further storage, organization, or analysis. It usually involves the recognition of entities, relationships, and attributes, as well as the mining of their relationships. It includes several key steps such as entity recognition, relationship extraction, event extraction, and attribute extraction, and the specific selection is based on the structure of the knowledge graph itself.

In this complex task, researchers can use various methods such as rule-based traditional methods or machine learning. For example, rule-based methods utilize predefined rules and patterns to extract information by matching entities, relationships, or events in text. In addition, there are various methods based on statistics, semantic analysis, or feature extraction processes using deep learning techniques, and natural language processing (NLP). In practical applications, these methods often combine to form a comprehensive medical data mining system [8].

### 3.3. Medical Data Aggregation Method

#### 3.3.1. Medical Data Entity Pairing Strategy

In the process of medical data mining, the first step is to determine the source of medical data. Based on research findings, data resources in the medical field mostly exist in massive medical information systems, divided into static medical data and real-time medical data. These resources include hospital databases, public health platforms, online health communities, and more. Therefore, research is conducted to obtain corresponding medical data from these systems, and an automatic acquisition model is constructed based on a knowledge graph framework. The acquisition framework diagram specifically includes the following modules: firstly, the request scheduling module, which is mainly used to receive link acquisition requests. Secondly, the data acquisition module primarily processes all responses and extracts the requested links that need to be retrieved. Thirdly, the data download module's main function is to download the obtained content. Fourthly, the medical data processing module is primarily utilized to process the medical data acquired during acquisition and handle corresponding requests. The main workflow is as follows: the engine receives the send request sent by the acquisition module and passes it to the scheduler. The scheduling module accepts the request and then passes it to the downloader. The download module retrieves and acquires the corresponding medical data from the Internet. After filtering, the medical data is returned to the engine and then transferred to the data acquisition module. This module needs to extract the necessary medical data, transmit it to the medical data processing module, and then make the processed medical data available.

### 3.3.2. Medical Data Entity Association Technology

When constructing a knowledge graph, there are usually two types of relationships to start with: explicit relationships and implicit relationships. Explicit relationships can be directly extracted from raw data, while implicit relationships require complex calculations on existing data to obtain. These calculations can reveal deeper connections between entities. The construction process includes constructing a single-layer knowledge graph at the disease entity level, constructing a single-layer knowledge graph at the drug entity level, constructing a single-layer knowledge graph at the treatment entity level, and creating a multi-level knowledge graph for the medical field. After constructing a multi-level knowledge graph, it is necessary to make correct judgments and quality evaluations on the domain entities that have been incorporated into the graph to ensure that the constructed knowledge graph maintains high accuracy. Compared to a single-level knowledge graph, a multi-level knowledge graph is structurally more complex and can store more entities, resulting in richer entity relationships. Therefore, this graph can be validated using a graph embedding model, which is a shared variable graph embedding model with smaller parameter sizes and greater application advantages compared to other models. In the application stage, this model can combine entities and relationships in the knowledge graph to complete the representation from entities to relationships. In addition, the model can score entities using the graph embedding function, select the knowledge graph with the highest score as the accurate graph, and multiple link predictions can enhance the accuracy of the knowledge graph, thereby confirming that the constructed knowledge graph is highly accurate [9].

### 3.4. Medical Data Acquisition Method

The acquisition of medical data is the main component of medical data mining, which emphasizes the quality and completeness of data, and directly reflects medical services and patient health status through data collection. As some constituent elements of medical data acquisition are gradually taking shape, the standardization of medical data and data privacy protection is also gradually receiving attention. . However, from a practical perspective, the collection practice of some medical data is still in its early stages, and there are still inconsistencies with the logical framework and generation mechanism of medical data mining, which leads to data quality issues.

## 4. Medical Data Mining and Clinical Decision-Making Integration Practice to the Electronic Medical Record System as a Sample

### 4.1. Electronic Medical Record Information Collection

From the data mining perspective, electronic medical records are the fundamental link of medical data mining and the core of clinical decision-making. Therefore, data mining focuses on generating electronic medical records as the main logic. Electronic medical records are the main data source for medical data mining and the data subject for clinical decision-making. Currently, medical data mining from the electronic medical records perspective, with three main forms to strengthen data control: first, data integration, which clarifies the integration of patients' medical records at different medical institutions and time points. Second, is data standardization, which achieves standardized control of medical data by establishing unified data terminology standards and disclosing these standards to medical institutions. Third, the internal process reengineering of electronic medical records. In recent years, electronic medical record systems and other methods have been used to improve medical services and decision-making efficiency through data mining. However, compared to the ideal state, the data quality of current medical data mining still needs improvement [10].

### 4.2. Electronic Medical Record Information Knowledge Extraction

The difference between electronic medical records and medical decision-making lies in their information attributes. The data standards and clinical guidelines for electronic medical records aim to improve the quality of medical services, and the development of medical decision-making mainly reflects personalization and precision. In the framework of medical data mining, accurate diagnostic

information, treatment plans, patient feedback, and efficacy evaluation are the core values and highest criteria for the development of medical decision-making. The diversity of electronic medical records and the differences in medical needs have led to a complex situation in medical decision-making. Despite technological advancements, electronic medical record systems are still not well-established, and medical decision-making itself lacks effective data support mechanisms. Therefore, this creates an information gap in medical decision-making, which affects the quality and efficiency of decision-making.

### 4.3. Summary of Electronic Medical Record Information Knowledge

#### 4.3.1. Medical Data Mining Structure Design Logic

From a clinical decision-making perspective, existing electronic medical record systems cannot accurately provide the key information needed for decision-making. The main form of clinical decision-making for electronic medical records is satisfaction evaluation, but electronic medical records lack relevant information and intelligent recommendation mechanisms for decision support. The core of this problem is the deep mining of information. In electronic medical records, patient information is often described as "structured data", which directly reflects the availability of information in support of clinical decision-making. However, electronic medical records are mostly about basic information, diagnostic information, and more, lacking in-depth analytical information. Usually, deep-level analysis information is difficult to obtain or measure. Asymmetric information and imperfect data mining directly lead to obstacles in clinical decision-making.

#### 4.3.2. Arrangement of Medical Data Mining Functional Components

From a technical implementation perspective, the long-standing data silos have constrained the ability of medical data mining. Since the information age, electronic medical record systems integrating data storage and processing have reshaped medical services through information technology. However, the drawbacks of traditional medical information systems still constrain data mining. Due to data standardization issues and the data privacy protection impact, medical data mining still needs improving. Under the premise of data-driven, knowledge graph is a direct way of data mining. However, the practical role of data mining based on knowledge graphs in clinical decision-making remains debated. Meanwhile, difficulties in data quality have led to a lack of practicality in knowledge graphs. Therefore, knowledge graphs may not always achieve the goals of clinical decision-making. It is evident that knowledge graphs are a technical challenge and face the challenge of application implementation [11].

### 4.4. Building a Knowledge Graph for Electronic Medical Record Systems

Undoubtedly, medical data mining is unavoidable being a "core tool" for information processing in clinical decision-making. In clinical decision-making mechanisms, knowledge graphs are standard and effective data mining tools that play a crucial role in medical decision-making, making knowledge graphs a data organization and a decision support concept. Therefore, data mining based on "knowledge graphs" has become an intelligent mechanism for clinical decision-making. The practical interpretation of knowledge graphs is generally a decision support path gradually formed on the basis of electronic medical record systems, although this path involves attempts at technological implementation. The knowledge graph revolves closely around clinical decision-making, from data collection to knowledge discovery. Knowledge graphs should be dedicated to providing decision support to meet the requirements of precision medicine, however, the magnified complexity of data poses a dilemma of information overload. Overall, knowledge graphs in data mining and their decision support functions also need to be further improved, which are crucial tasks in medical data mining.

## 5. Conclusion

We can systematically analyze the research hotspots and development trends in the medical field by applying medical data mining techniques. This analysis deepens the research, and provides guiding

suggestions for clinical decision-making, helping healthcare workers accurately grasp the direction and innovation of subsequent treatments. Therefore, it is crucial to explore how to integrate medical data mining technology into clinical decision-making effectively, revealing the potential value of this technology in new fields and promoting its wider application and development.

## References


[1] Pramanik M I, Lau R Y K, Demirkan H, et al. Smart health: Big data enabled health paradigm within smart cities[J]. Expert Systems with Applications, 2017, 87: 370-383.

[2] Sun W, Cai Z, Li Y, et al. Data processing and text mining technologies on electronic medical records: a review[J]. Journal of healthcare engineering, 2018, 2018(1): 4302425.

[3] Islam M S, Hasan M M, Wang X, et al. A systematic review on healthcare analytics: application and theoretical perspective of data mining[C]//Healthcare. MDPI, 2018, 6(2): 54.

[4] Singh A K, Anand A, Lv Z, et al. A survey on healthcare data: a security perspective[J]. ACM Transactions on Multimidia Computing Communications and Applications, 2021, 17(2s): 1-26.

[5] Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine[J]. Scientific Data, 2023, 10(1): 67.

[6] Rotmensch M, Halpern Y, Tlimat A, et al. Learning a health knowledge graph from electronic medical records[J]. Scientific reports, 2017, 7(1): 5994.

[7] AbdulAmeer D A H. Medical data mining: Health care knowledge discovery framework based on clinical big data analysis[J]. International Journal of Scientific and Research Publications, 2015, 5(7): 6.

[8] Xiao W, Jing L, Xu Y, et al. Different data mining approaches based medical text data[J]. Journal of Healthcare Engineering, 2021, 2021(1): 1285167.

[9] Sun W, Cai Z, Li Y, et al. Data processing and text mining technologies on electronic medical records: a review[J]. Journal of healthcare engineering, 2018, 2018(1): 4302425.

[10] Brundin-Mather R, Soo A, Zuege D J, et al. Secondary EMR data for quality improvement and research: a comparison of manual and electronic data collection from an integrated critical care electronic medical record system[J]. Journal of critical care, 2018, 47: 295-301.

[11] Kovalchuk S V, Funkner A A, Metsker O G, et al. Simulation of patient flow in multiple healthcare units using process and data mining techniques for model identification[J]. Journal of biomedical informatics, 2018, 82: 128-142.