Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 214 (2022) 495–502

Procedia Computer Science

www.elsevier.com/locate/procedia

9th International Conference on Information Technology and Quantitative Management

Evaluation System Framework of Artificial Intelligence Applications in Medical Diagnosis and Treatment

Xueqing Tian[a] , Haocheng Tang[b], Long Cheng[c], Zirui Liao[d], Yuxiao Li[e], Jing He[f], Ping Ren[a], Mao You[a], Zhen Pang[g]

[a]*China National Health Development Research Center, Beijing 100044, China*
[b]*Institute of Automation, Chinese Academy of Science, Beijing 100190, China*
[c]*Chinese Medical Information and Big Data Association, Beijing 100037, China*
[d]*Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100730, China*
[e]*Rensselaer Polytechnic Institute, NY 12180, US*
[f]*Nanjing University of Posts and Communications, Nanjing 210049, China*
[g]*Xiyuan Hospital CACMS*

Abstract

Current medical artificial intelligence applications and products are confronted the dilemma of lacking standardized practical evaluation guidelines and management methods. This research adopts the Donabedian medical quality management classic model and the DeLone & Mclean benefit evaluation model to develop a basic framework of the medical artificial intelligence application evaluation system for auxiliary diagnosis and treatment and to design the corresponding evaluation procedures. This paper illustrates the overall project framework, followed by a detailed structure and construction process of the model proposed for fast and accurate medical evaluations, NHCKG. NHCKG provides a holistic view of diseases, medical regulations and evaluations. The proposed NHCKG can provide partial solutions to significant challenges faced by knowledge graphs and natural language processing (NLP). Based on the knowledge graph, the structure of a related medical question answering system is elaborated along with its advantages and disadvantages. The evaluation process can take this procedure as a blueprint for exploring standardized and practical evaluation of medical artificial intelligence applications. The process aims to build a concrete measurement basis for medical artificial intelligence products and to promote the healthy and stable development of the medical artificial intelligence industry.

Keywords: Diagnosis and treatment; Medical artificial intelligence; Evaluation system applications; Pharmaceutical agent for gray matter volume damage

10.1016/j.procs.2022.11.204

1. Introduction

Following the pace of rapid technology development, artificial intelligence technology has been widely applied in various fields. The intelligent evolution of the medical field is also considered a significant subject. Still, it is an inevitable tendency to turn information technology systems into intelligent systems. A core issue in transforming medical healthcare with artificial intelligence is whether the medical artificial intelligence applications can be effectively reviewed and supervised. The basic structure of the medical artificial intelligence application evaluation system for auxiliary diagnosis and treatment proposed by this research provides thinking and reference for the supervision and approval of medical artificial intelligence in China. It accelerates the process of standardizing artificial intelligence-assisted diagnosis and treatment in real-world clinics and promotes artificial intelligence applications, pricing and reimbursement. Besides, this system improves medical resource allocation and service quality, leading to the healthy and steady development of the medical artificial intelligence industry.

1.1. Scope and Performance of artificial intelligence applications in auxiliary diagnosis and clinical decision making

"Artificial intelligence-assisted diagnosis and treatment" refers to the use of a new generation of artificial intelligence technology to assist medical staff in decision-making, for example, diagnosis criteria and treatment execution, efficacy evaluation, dynamic optimization and process management. The "new generation of artificial intelligence technology" refers to the technology that employs data-driven methods to train algorithms such as deep learning and neural networks. The application evaluation system provides evaluation bases for old technologies that have already used by medical institutions for medical insurance reimbursement catalogs and post-marketing evaluation for new technologies. It also makes a forward-looking evaluation of future technologies. The evaluation goal varies based on purposes. It can continuously evaluate the application process to guide the application research and development of medical artificial intelligence products. It can also evaluate the impacts of specific applications on the health system. The evaluation results can guide the products' use process and quality improvement by figuring out the risks and providing potential solutions.

1.2. Significance of developing the artificial intelligence application evaluation system

Noticing there is a lack of legislation on medical artificial intelligence applications in China [1], the evaluation system can promote the establishment of laws and regulations. It also helps convert technology research results to commercial products and contributes to the establishment of a full-lifecycle medical artificial intelligence supervision system. It can guide and promote effective and healthy research and application development.

2. The framework of the application evaluation system

The theory of the evaluation framework relies on the classic model of medical quality management proposed by Donabedian - the "structure-process-result" model, and the benefit evaluation framework proposed by DeLone & Mclean (D&M) [2]. Traditional medical software can be divided into medical device-embedded software and medical-device-independent software. The medical artificial intelligence for auxiliary diagnosis and treatment should be categorized as a medical information technology evaluation system because it shares more similarities with the medical device-independent software. However, according to the essential design principle of evaluation system development, the evaluation content and critical criteria should be redesigned regarding their uniqueness.

Knowledge graphs (KG) and natural language processing (NLP) serve to build this evaluation system. KG integrate knowledge and data with entities and relationships. It has been widely applied in intelligent question answering systems, intelligent recommendation systems, structured search, exploratory search, digital assistants, etc. NLP is frequently employed to automatically extract unstructured texts in knowledge base construction [3].

2.1. Dimensions of the evaluation system

The framework of the evaluation system includes four levels and nine dimensions. Subjects are evaluated using specific indicators correspondingly during the implementation of the evaluation process.

- Institution/facility/equipment layer

Technical effectiveness evaluates the use effect of artificial intelligence applications in real-world clinics. Technical accessibility emphasizes its accessibility to patients, medical staff, and medical institutions. Technology affordability stresses the personal burden and group/institution burden.

- Application and plan evaluation (policy evaluation)

Technical effectiveness evaluates the use effect of artificial intelligence applications in real-world clinics. Technical accessibility emphasizes its accessibility to patients, medical staff, and medical institutions. Technology affordability stresses the personal burden and group/institution burden.

- Process layer

Operational effectiveness focuses on scenario fit, ease of use, and operation management. User satisfaction describes user acceptance, doctor-patient experience, user habits, and the doctor-patient role.

- Result layer

The individual effect includes patient/clinical effectiveness and medical staff/service effectiveness. Group effect mainly focuses on the effectiveness of the service system, economic benefits, and social benefits.

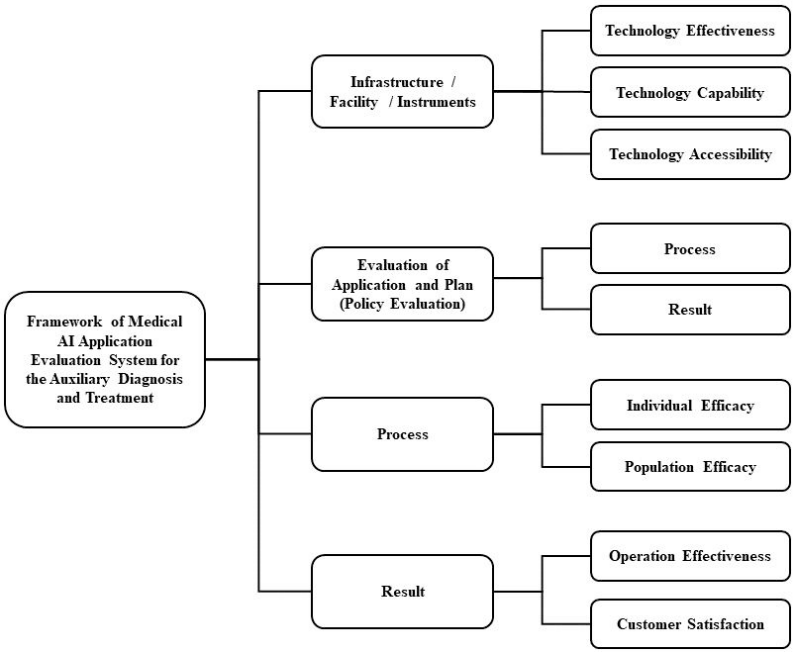

Fig. 1. Framework of medical artificial intelligence application evaluation system for auxiliary diagnosis and treatment.

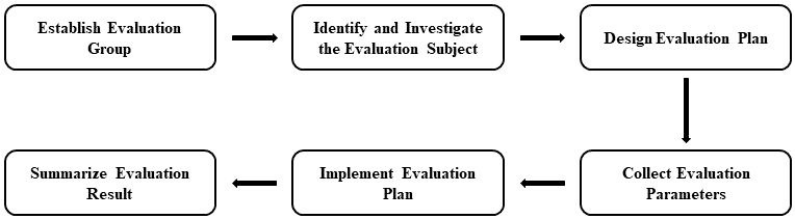

Fig. 2. Evaluation process.

2.2. Evaluation process

The process of the application evaluation system adopts the Delphi method (Delphi) design [4].

An evaluation team of professionals is formed first to collect descriptive documents of the evaluation subjects. After generating an initial description report, an evaluation plan including the purposes of the evaluation, selection of evaluation indicators, users, using methods, and protocols of the evaluation results is designed. To improve the quality of the evaluation, the team further collects evidence focusing on acceptability, suitability and patient preference through literature reviews, questionnaires, and expert consultation. Then, three methods - examination evaluation method, expert scoring method, and system review method (?) - are employed to systematically evaluate the clinical effect, cost-effectiveness, fairness, and impact on the implementation of the artificial intelligence applications for diagnosis and treatment to form the first draft of the evaluation report. Finally, the evaluation team organizes meetings with multiple parties to explain and discuss the evaluation results, form final recommendations, generate a final formal evaluation report, and ensure the effective sharing and use of the evaluation results.

3. NHCKG: building knowledge graphs for precise medical evaluation

3.1. Knowledge graphs

The term "knowledge graph" was firstly proposed by Google in 2012 and has been widely applied in various fields. It is a form of graph organization that associates various entities through semantic correlations. It extracts and fuses structured and unstructured data together, embodying the idea of data governance and semantic connection. This is conducive to the utilization and migration of large-scale data. Knowledge graph changes the traditional way of data storage in traditional drug research and development. The construction of the knowledge graph proposed by this research is based on the dynamic ontology theory. A complete set of data ontology proposed in this project can concatenate and convert different data forms and integrate them into a comprehensive data system. The data gene system is a complete ontology framework of drug development data (in simple words, a data catalog).

In addition to helping the integration of multi-source heterogeneous data, knowledge graphs are also conducive to the integration of multi-modal (text, video, image) data. The multi-modal learning task proposed in this project is to take multiple modalities of text, pictures, and videos together as input. These multiple modalities can be converted through deep learning and connected using semantics. The feature engineering in the proposed knowledge graph is different from machine learning, mainly reflected in use in stages, feature splicing, and fusion based on semantic information.

3.2. NHCKG introduction

NHCKG is a knowledge graph for government medical device licensing and food and drug evaluation and regulation. It provides a holistic view of the diseases and regulations and evaluations of medical devices, food, and drug. Twenty high-quality resources are integrated by NHCKG to describe 17,080 diseases with 4,050,249 relationships. These relationships represent ten major biological scales, including disease-associated protein perturbations, biological processes and pathways, anatomical and phenotypic scales, and all of the approved and experimental drugs and their therapeutic effects. The graph structure of NHCKG is combined with textual descriptions of drugs and clinical practice guidelines to enable multimodal analysis.

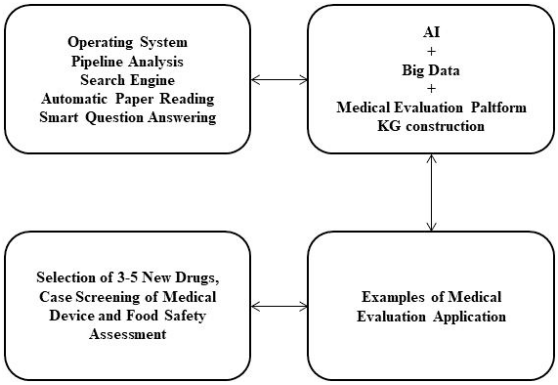

Fig. 3. Development and application of the evaluation system based on artificial intelligence big data and reliable safe calculation.

In this study, the reported information on medical devices, food, and drugs is firstly translated by NLP to complete automatic reading, pipeline analysis, search engine, and the framework of automatic question answering. It helps integrate raw data resources of diseases and assessment information into a comprehensive, disease-rich, and functional knowledge graph. Noticing the major challenges faced by NLP, NHCKG realizes higher precision in pipeline analysis, automatic reading, and automatic question-answering systems. It also extends previous work on disease-based knowledge graph creation, adds indications, contradictions, and off-label uses of edges, and enables multimodal analysis.

3.3. NHCKG construction

NHCKG is designed as a heterogeneous network with 100,000 types of nodes and 300,000 types of undirected edges. The drug and disease nodes in this network are augmented with textual descriptions by retrieving and organizing the resources shown in Figure 4a, and the relationships among the resources depicted in Figures 4b and 4c and Figure 4d with textual descriptions.

The major data resources include 200,000 resources, for instance, Mayo Clinics, Orphanet, DisGeNET, UMLS, DrugBank, etc. The selected 200,000 resources are standardized and coordinated by defining node types and selecting a common ontology, coordinating external data resources, and resolving overlaps between phenotypes and disease nodes. The unified raw data resources are merged into a single graph, and its maximum connected component is extracted, as shown in Fig 4c. Finally, the drug nodes and disease nodes are supplemented with clinical information. Take olanzapine, a drug that possibly increases the gray matter volumes in the caudate nucleus in schizophrenia patients, as an example [5]. Textual and numerical features of the drug nodes from the knowledge graphs of DrugBank and Drug Central are extracted and mapped directly to the knowledge graph. Features of disease nodes from Mondo Disease Ontology, Orphanet, Mayo Clinics, and UMLS are extracted and mapped as well.

3.4. Test result verification

In the example below, NHCKG contains 129,375 nodes and 8,100,498 edges with ten types of nodes and 30 types of edges. Figure 5a shows the graph structure, and Figure 5b demonstrates that the disease nodes are closely related to other node types in the knowledge graph. Disease characteristics include information on disease prevalence, symptoms, etiology, risk factors, epidemiology, clinical description, management and treatment, complications, prevention, and when to seek medical attention. Drug characteristics include molecular weight information, indications, mechanism of action, pharmacodynamics, protein binding events, and pathway information of compounds. This extensive clinical information describing the entire spectrum of drugs and diseases is a unique feature of NHCKG and sets NHCKG apart from its peer knowledge graph. Figure 5c provides an example of the supporting information available in these characterizations.

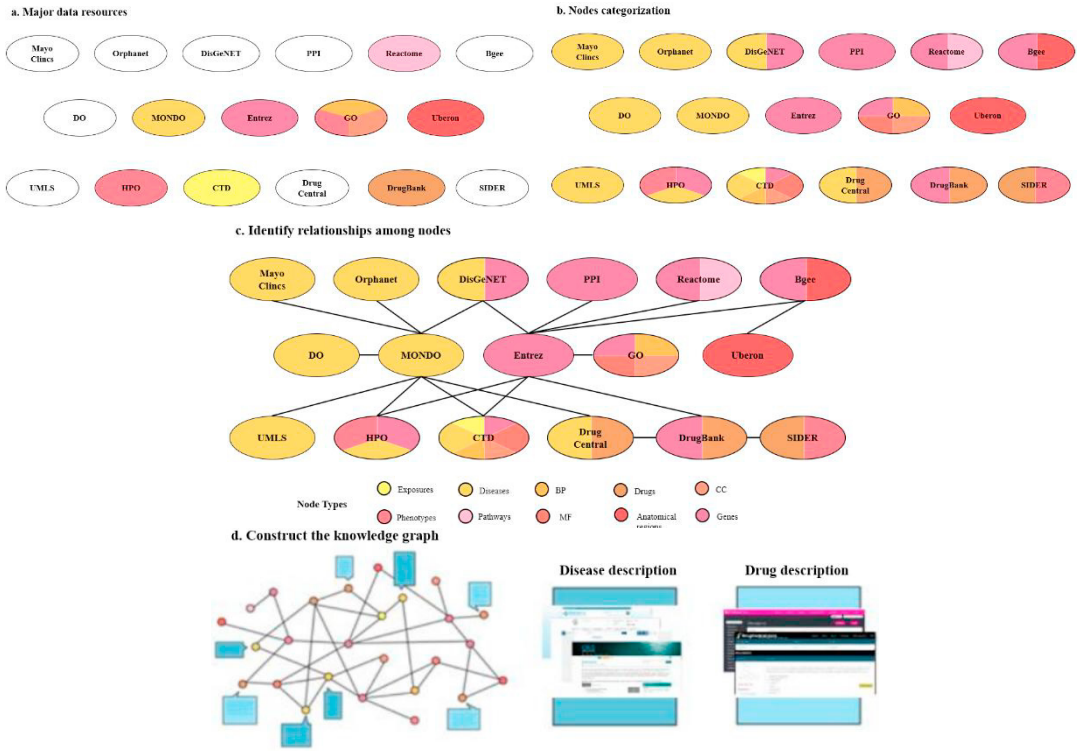

Fig. 4. NHCKG construction.

For example, a company reports a new drug for gray matter damage treatment. We first perform NLP on the reported materials. Analysis of the correlation between the disease manifestations of NHCKG and its clinical manifestations is carried out in two steps by conducting case studies of brain gray matter damage: by performing entity resolution of gray matter damage concepts in all relevant raw data sources; and by examining the relationship between these gray matter damage concepts and clinical subtypes of it. It is possible that the disease concepts in Mondo do not correlate well with medical subtypes. Since Mondo contains a lot of duplicated disease entities with

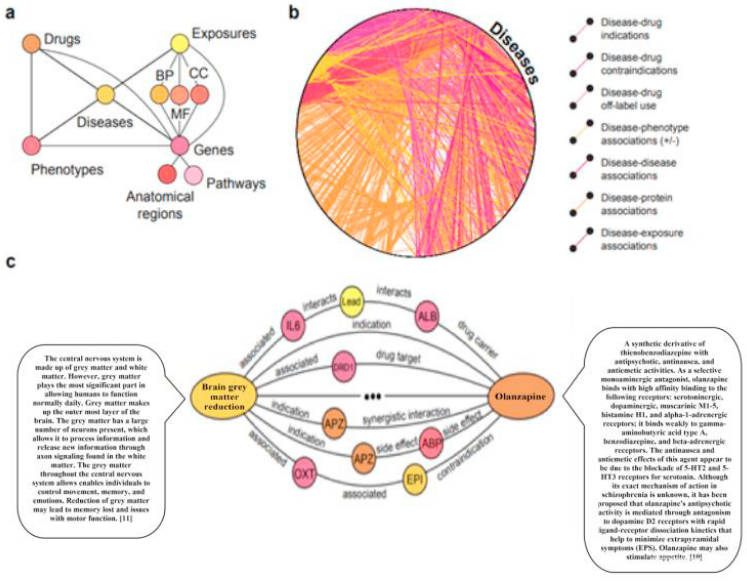

Fig. 5. NHCKG characterization.

ambiguous clinical relevance, diseases in Mondo are grouped into medically related entities. A semi-automated unsupervised approach is adopted to group disease concepts in PrimeKG, identify disease groups with a string matching strategy across disease names, and explore word embedding similarities between disease names. This further enhanced the grouping by string matching.

4. Medical evaluation question answering system

The intelligent applications of knowledge graphs can also be used for intelligent question answering, medical services, and library information services. In question answering and search applications, the knowledge graph presents search with accurate results instead of returning a bunch of similar pages for users to filter by themselves, providing expected answers for the questions. It includes key steps such as problem decomposition, hypothesis generation, and evidence-based fusion ranking. A real understanding of questions is achieved through in-depth analysis. In this project's non-real-time tasks, automatic recommendation results returned by the machine and further manual editing and review are combined to ensure user experience and reduce the disadvantages of pure manual working, such as low efficiency and large workload. The unique human-machine integration in this project gives artificial intelligence a new connotation: artificial intelligence + Human Intelligence = Augmented intelligence.

4.1. Capabilities required

According to the artificial intelligence application capability framework, the capability requirements for medical evaluation search engines can be divided into three parts: basic capabilities, task accumulation, and intelligent technologies. Among them, the basic capabilities are mainly to solve the general and basic capabilities in intelligent question answering, such as sensitive word filtering, multiple question method identification, etc.; task accumulation mainly refers to specific question and answer fields, which need to be accumulated in tasks. For example, how many categories the questions can be roughly divided into, what is contained in each category, and what kind of answering methods are generally adopted, etc.; intelligent technology is mainly a technology for more advanced applications, such as how to realize multiple rounds of dialogue, identify multiple models' state, dynamic loading, etc. The project mainly adopts three major technologies: retrieval technology, knowledge network, and deep learning.

4.2. Process of question answering matching

Search engines and question-answering systems for medical evaluation can be divided into task-oriented, knowledge-oriented, and chat-oriented. From the perspective of key technologies, they can also be divided into question-answering systems based on search technology, collaboration, and knowledge bases. The process of question answering matching for knowledge graph processing can generally be summarized into eight processes, as shown in figure 6.

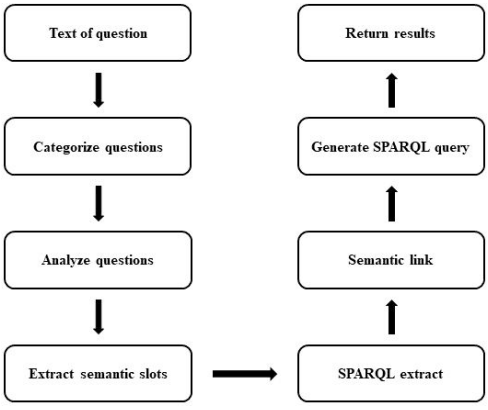

Fig. 6. Process of question answering matching.

Conclusion

The framework for artificial intelligence-assisted diagnosis and treatment application evaluation provides comprehensive evaluation criteria for medical artificial intelligence applications. By dividing the evaluation into different layers, it measures the effectiveness and capability of the applications thoroughly. The process of the evaluation system using D&M model further specifies the procedures of evaluation, thus avoiding unreasonable results that are difficult to recognize. The development of the evaluation system will bring about the establishment of laws and regulations, flourishing commercial applications and healthy steady development of the medical artificial intelligence industry and research field.

NHCKG characterizes drugs at a deeper biological level and disease at a deeper clinical level. It can be paired with machine learning to discover new disease biomarkers, describe disease processes, refine disease classifications, identify phenotypic features, predict biological mechanisms, and repurpose drugs. With the realization of machine learning capabilities, it is expected that NHCKG and similar knowledge graphs will become a key tool in advancing the rapid and accurate safety assessment of medical devices, food and medicine.

References

[1] Wang, C. & Qin, J. (2008) “A Study of FDA Medical Devices Software Guild.” *Soudu Yixue* **25** (**2**).

[2] Donabedian, A. (1980) “Explorations in Quality Assessment and Monitoring: The Definition of Quality and Approaches to its Assessment." *Ache Management*.

[3] M. Nickel, K. Murphy, V. Tresp and E. Gabrilovich. (2015) “A Review of Relational Machine Learning for Knowledge Graphs." arXiv:1503.00759 [stat.ML].

[4] Zhang, D. M. “A Study of Delphi Application Based on Two Cases-USA and Belgium cases.” *Information Studies: Theory & Application* **41** (**5**).

[5] Okugawa G, Nobuhara K, Takase K, Saito Y, Yoshimura M, Kinoshita T. (2007) “Olanzapine increases grey and white matter volumes in the caudate nucleus of patients with schizophrenia.” *Neuropsychobiology* **55** (**1**): 43-6.