Journal of Big Data

RESEARCH Open Access

Risk and UCON-based access control model for healthcare big data

Rong Jiang[1,2], Xue Chen[1,2], Yimin Yu[1,2], Ying Zhang[4,5*] and Weiping Ding[3*]

*Correspondence:
angie17@qq.com;
dwp9988@163.com

[1] Institute of Intelligence Applications, Yunnan University of Finance and Economics, Kunming, China
[2] Yunnan Key Laboratory of Service Computing, Kunming, China
[3] School of Information Science and Technology, Nantong University, Nantong, China
[4] School of International Business, Yunnan University of Finance and Economics, Kunming, China
[5] School of Management and Marketing, Charles Sturt University, NSW, Wagga Wagga, Australia

Abstract

The rapid development of healthcare big data has brought certain convenience to medical research and health management, but privacy protection of healthcare big data is an issue that must be considered in the process of data application. Access control is one of the methods for privacy protection, but traditional access control models cannot adapt to the dynamic, continuous, and real-time characteristics of healthcare big data scenarios. In this paper, we propose an access control model based on risk quantification and usage control (RQ-UCON). The model adds a risk quantification module to the traditional UCON model to achieve privacy protection of medical data. This module classifies risks into direct and indirect risks and quantifies them based on the physician's visit history. The model stores the quantified risk values as subject attributes. The RQ-UCON model uses an improved Exponentially Weighted Moving Average (EWMA) and penalty factors to predict risk value and to update the risk values of the subject attributes in real-time. The RQ-UCON model uses agglomerative hierarchical clustering to cluster the risk values of physicians within the department, resulting in risk intervals for each physician's operational behavior. Each risk interval is stored as a condition in the RQ-UCON model. Finally, according to the model whether the subject attributes meet the model conditions to determine whether the subject has the corresponding access rights, and according to the risk interval to grant the subject the corresponding access rights. Through the final experiment, it can be seen that the access control model proposed in this paper has a certain control on the excessive access behavior of doctors and has a certain limitation on the privacy leakage of healthcare big data.

Keywords: Healthcare Big Data, Risk Quantification, UCON, Agglomerative Hierarchical Clustering, Access Control

Introduction

With the rapid development of the Internet industry, the development of various industries has entered the era of big data. The continuous integration and development of medical technology and information technology have provided a constant impetus for the generation of medical big data and laid a stable cornerstone for the application and development of big data technology in the medical field [1]. In 2014, the U.S. Office of the National Coordinator for Health Information Technology released the U.S. Federal Government Healthcare Information Technology Strategic Plan 2015–2020, which specifies the goal of achieving healthcare data sharing and proposes three application goals, including enhancing healthcare services, improving the health of the public and the community, and promoting medical knowledge research and innovation. In 2016, the General Office of the State Council of the People's Republic of China issued [2016] No. 47 document "Guidance of the General Office of the State Council on Promoting and Regulating the Development of Health Care Big Data Applications". It states "health care big data is an important basic strategic resource for the country and we should vigorously promote the interconnection and integration of government health care information systems and public health care data, open Share, eliminate information silos, and promote the development of safe and standardized health care big data and innovative applications". Big Data is now one of the UK's major strategic areas of development, with the UK spending £5.5 billion on a national integrated healthcare information storage service that collects and stores data from over 23,000 healthcare information systems and already serves 1.3 million healthcare professionals.

Medical big data refers to the healthcare-related data generated by people in the process of disease prevention and treatment, health management, etc.; it is a kind of big data. Volume velocity variety value Medical big data has the 4 "V" characteristics of big data, namely: large volume, fast velocity, wide variety, and high value [2]. Healthcare big data has high value, standardized and reasonable use of healthcare big data, the government can promote health care clinical and scientific research big data applications and public health big data applications, etc., to form a new industry of health care big data applications [3]. By accessing medical data through the hospital's medical information system, doctors can quickly retrieve a patient's medical history and provide a more accurate treatment plan for the patient. Researchers are pushing the boundaries of smart device research and health management by conducting visual analytics research on medical big data.

In using the information in medical big data to provide new ideas for health analytics [4], medical applications [5], etc., we also face the problem of inevitable privacy breaches [6]. For example, in early 2018, employees at a maternal and child health hospital in China downloaded the personal information of newborns and mothers for over-authorized access, totaling more than 89,000. In February 2020, during a critical period in the fight against the Corona Virus Disease 2019 (COVID-19) in China, the Indian APT hacking group launched an attack on Chinese medical institutions and government departments, using the topic of "COVID-19" to lure victims into executing phishing commands. The 2020 U.S. Healthcare Industry Data Breach Incident Report mentions that the total number of medical records compromised in the U.S. for the year 2020 exceeded 29 million, with unauthorized access/disclosure incidents accounting for 22.7% of annual healthcare breach incidents, involving 787,015 compromised records; and that U.S. healthcare data breaches caused $13 billion in losses in 2020. The leakage of medical information comes partly from hackers exploiting the security loopholes of medical information systems, and partly from excessive data access caused by the unreasonable granting of job privileges by hospital staff. Leaked medical records include detailed personal information and medical details, and this data can be extremely troubling to patients if used by unscrupulous individuals. The leakage of medical privacy information causes difficulties in protecting patients' privacy, resulting in the utilization of medical big data is not high. At present, the privacy protection of medical big data has become an urgent issue to be solved.

Although medical big data has high research value, medical big data is vulnerable to external attacks and insider over-authorization resulting in privacy data leakage leading to the underutilization of medical big data. For big data security and privacy protection problems some scholars have carried out a lot of research work on big data security and privacy protection problems, which mainly focuses on a differential method for private data leakage, encryption algorithm to prevent data intrusion, anonymization of data release, user identity authentication and access control model [7], etc. To solve the problem of privacy data leakage caused by excessive authorization of insiders, various access control models have been proposed to enable accessing subjects to use the object resources within the scope of legitimate authorization. Traditional access controls such as autonomous access control [8] 、mandatory access control [9] and role-based access control [10] have static and explicit access authorization, which cannot be well adapted to the dynamic and real-time nature of big data scenarios. If the traditional access control is used to authorize all-access subjects separately, the managers of the information system have to refine the scope of authority to authorize separately while completing their daily work. This method cannot adapt to the real-time and dynamic characteristics [11] of the big data environment, and it also leads to a heavy workload for managers. It is a privacy protection method that adds risk [12] to the access control model to achieve dynamic adjustment of access rights of the access subject by calculating the risk value for the historical access behavior of the access subject. There are more studies on risk-based access control for big data, but fewer studies on risk-based access control [13] in the context of big data in healthcare, so it is of theoretical interest to study access control related to big data in healthcare [14, 15].

In this paper, we investigate the risk-adaptive access control model for healthcare big data and propose a RQ-UCON model that takes dynamicity and real-time into account and implements the model for continuous access.

The innovations of this paper are the following:

- Using a combined risk and UCON access control model to achieve privacy protection of medical data in the context of big data in healthcare. We add a risk quantification module to the traditional UCON model and quantify the risk value of user history access records through a risk control component.
- To improve the accuracy of risk quantification by dividing it into direct risk quantification and indirect risk quantification in the risk quantification stage, and we use the EWMA algorithm and penalty factors to realize dynamic updates of risk values.
- The user clustering stage introduces an agglomerated hierarchical clustering algorithm to cluster doctors into four classes, and the risk intervals of corresponding types of doctors appear as conditions in the UCON model.

The rest of the paper is organized as follows. "Related work" presents the current state of research in access control, and "Access control model" details the modules of the model proposed in this paper. "Simulation experiment" conducts simulation experiments to verify the feasibility and superiority of the RQ-UCON model proposed in this paper. "Conclusion" concludes the paper.

Related work

With the development of medical information technology, more and more hospitals are using medical information systems to carry out medical services [16]. As a national development strategy, big data in health care is going to be applied and developed rapidly. Although the development of health care big data has just started, the large amount of personal privacy involved in health care big data is facing unprecedented threats and challenges. Currently, the issue of privacy breach of personal data has become the core concern of big data in health care [17, 18]. Soceanu et al. [19] proposed a new privacy protection approach for healthcare big data to address the privacy and security issues of clinical data. The approach achieves layered privacy protection of eHealth data by using an advanced encryption scheme ARCANA and an attribute-based access control authorization framework with partial visibility of the authorized part. Wu et al. [20] researched the key issues of privacy protection, studied medical data from three aspects of data collection, data transmission, and data sharing, and proposed a medical big data privacy protection sharing platform based on the Internet of Things to achieve separation of users and data to ensure medical data security. Aiming at the problem of medical data abuse, Jiang et al. [21] proposed an access control model based on the credibility of requesting users, quantified user access records, and introduced the historical behavior trend of users into the trust evaluation model. This model improves the overall behavior of users in the system and realizes the protection of medical data. Gan Lin et al. [22] proposed a three-tier electronic medical record sharing scheme of a private chain of the patient electronic medical records(EMR), a private chain of medical institution electronic medical records, and a public chain of the electronic medical record to solve the problem of privacy protection and shared use of health electronic record. In this scheme, the electronic medical record data is stored in a distributed manner with attribute encryption in the patient chain's super ledger, and the owner of the electronic medical record identifies the applicant to share the data through authorization. Lee et al. [23] studied medical privacy data and proposed a medical big data privacy protection system based on the Diffie-Hellmann protocol. Through this system, access rights are assigned to authorized physicians, who can access and share patients' private information, thus achieving the purpose of protecting the privacy and confidentiality of medical big data. Hossain et al. [24] conducted a study on electronic healthcare data sharing to provide a secure model for cloud data, where the model can exchange information between healthcare providers and healthcare professionals, retrieve a patient's complete prior medical history without the violating privacy, and reduce the probability of involuntary disclosure of personal information affecting the patient's life. Zabar et al. [25] studied the security and privacy of electronic health records and proposed a new framework that uses distributed databases to avoid centralized storage problems. The new framework uses Hyperledger composer functions to store hashes of data and control access when retrieving data, promoting the robustness of the healthcare management system.

Currently, the development of privacy protection for medical big data is mostly focused on encryption technology, which can reduce the privacy leakage of medical big data due to external attacks to a certain extent, but cannot solve the privacy leakage problem caused by insiders. Access control techniques have evolved from traditional access control techniques such as mandatory access control and autonomous access control to widely used role-based access control techniques, attribute-variable and decision-continuous UCON access control, and risk-based and trust-based access control techniques. To address the privacy data leakage caused by over-authorization of hospital internal personnel and to adapt to the characteristics of dynamic and real-time authorization in big data scenarios, Researchers take risk into account in access control [26]. Hui et al. [27] a risk access control model applicable to medical context. This model quantifies risk values based on physician access records using information entropy and EM algorithms, controls physician access to medical records using quantified risks, and adjusts physician access rights based on risk values. Wang et al. [28] proposed a practical access control approach to protect patient privacy in health care information systems, allowing physicians to make access decisions while still being able to detect and control excessive physician access to patient medical data by quantifying the risks associated with physician data access activities, in the context of practical healthcare considerations. Li et al. [29] proposed a comprehensive risk management model for the privacy leakage problem, which divides users into two categories, namely access subjects and objects, and uses information entropy techniques to estimate the risk value of each category of users, thus enabling dynamic tracking, dynamic assignment of thresholds, and regulation of the user's access range. Daoud et al. [30] proposed the use of dynamic risk- and role-based access control to prevent intrusions against cloud computing, which combines eXtensible access control markup language decisions, risk analysis, and vulnerability reviews to obtain the appropriate access decisions. Rajani Kanth Aluvalu et al. [31] proposed a risk-aware model based on dynamic attributes. It is combined with a common access control model, where risk calculation and attributes used for access control will be combined. Shi et al. [32] proposed a risk quantification method using fuzzy wavelet neural network, which uses fuzzy theory to evaluate the attribute information of subject and object comprehensively and calculate the final risk value by wavelet neural network. This method can effectively reduce the influence of human factors on risk and thus better control the risk. Zakaria et al. [33] proposed an IoT security risk management model for security practices in healthcare environments, which includes healthcare IoT risk management, hospital accountability metrics, and implementation phases, for the risk of IoT compromise in healthcare environments.

To accommodate dynamic and continuous access in big data access control, experts propose the next-generation access control model of usage control [34], which is characterized by the continuity of decisions and variability of attributes and allows real-time monitoring of access requests during the use of resources by the subject. Liu et al. [35] introduced role elements and divided roles into provider roles and consumer roles based on the UCON model, using both direct use and the need for authorization to use, to manage the permissions in the UCON model more effectively while ensuring the reliability of the attributes. Fan et al. [36] proposed a UCON-based multi-use control protocol model based on the inability of access control to adapt to changes in a multi-tenant model in a cloud environment, which features flexible authorization, feature binding, and offline control. Based on the idea of temporary authorization, Li et al. [37] introduced the $\mathrm{UCON}_{\mathrm{ABC}}$ based information resource usage control strategy and subject reliability into the $\mathrm{UCON}_{\mathrm{ABC}}$ model as an authorization decision factor in information resource usage control, designed a subject reliability evaluation method based on subject relationship and subject attributes, and constructed a new resource usage control scheme. Arcetales et al. [38] proposed architecture for data usage and access control issues in data sharing, and this framework is based on the UCON model and an extended extensible access control markup language reference architecture. Wang et al. [39] analyzed the drawbacks of traditional access control in the current electronic medical record system to propose an improved electronic medical record system and gave a specific access control strategy through the UCON model.

Previous research by scholars on access control in the medical big data environment eventually gives two results of allowing access and denying access, and cannot dynamically adjust the user's access scope and perform operations based on the user's risk value. This paper takes a risk as the entry point to study the risk-based and UCON access control model in the medical big data environment. The model investigates the privacy leakage that may occur in the access control stage of physicians, quantifies direct and indirect risks and magically adjusts the attribute values of physicians according to their risk values. The model realizes the division of operational behavior risk intervals through agglomerative hierarchical clustering. According to the user's risk value, the model adjusts the user's access range and the operations performed in real time to reduce the probability of privacy leakage to a certain extent. Compared with the existing work, the model better adapts to the real-time and dynamic nature of the medical big data environment, solves the internal leakage problem to a certain extent, addresses the problem of continuous access in the current environment, realizes the dynamization of risk values, and can effectively protect the system privacy data.

Access control model

In this section, the complete architecture and working principle of the RQ-UCON model proposed in this paper are presented. The model architecture consists of several modules, each with different functions. This section is divided into four parts: the basic framework, the risk quantification module, the user clustering module, the user attribute update module, and the access policy module.

Basic framework

In this paper, we combine risk and the UCON access control model with variable attributes and continuous decision making and propose an access control model based on risk quantification and UCON. Based on the UCON model, a risk quantification component is added to quantify the risk of information leakage that may be caused during subject access by the risk quantification component, and the risk value is used as an attribute of the subject to achieve variable attributes, while the subject is subdivided into an access subject and a production subject. The attribute update of the UCON model is divided into the pre-access update, during-access update, and

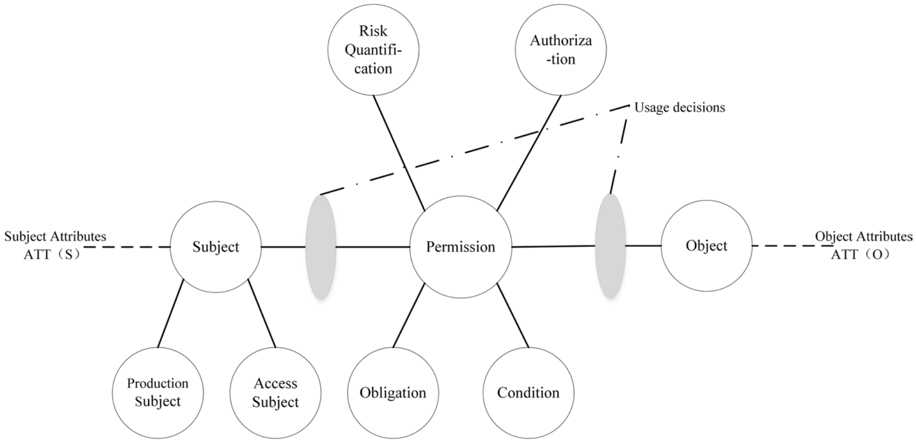


Fig. 1 RQ-UCON model

post-access update. In this paper, the risk value of this visit is predicted based on the risk value derived from the historical records, so the pre-access update is need. Figure 1 is a schematic diagram of the model.

The relevant definitions are as follows:

- Subject *S* is divided into production subject *SP* and access subject *SA*. Production subject *SP*: internal hospital data producers, patients, etc.; access subject *SA*: internal hospital medical staff, patients, the general public, etc. In the model of this paper, we use physicians as access subjects.
- Subject attribute *ATT(S)*: regular attributes include doctor's duty, doctor's department, doctor's number, etc.; variable attributes are subject risk value, access log, etc.
- Object *O*: Medical big data, which includes patients' personal information, examination records, patients' electronic medical records, medical research data, etc.
- Object attributes *ATT(O)*: patient's department, disease category, confidentiality level, admission time, etc., attributes usually are immutable.
- Permission *P*: The right of the subject to operate on the object resources. Access to the operations of the subject *SA* to view, modify, delete, etc. of the object *O*, such as the attending physician to query the patient's medical record, add treatment records, administer medication, etc.
- Authorization *A*: In the UCON model, there are two types of authorization: pre-authorization and process authorization; in this paper, we propose to use the pre-authorization method; each doctor is given the appropriate access rights according to his or her role, and the authorization range is dynamically adjusted by the doctor's risk value.
- Obligation *B*: The access subject *SA* needs to complete the corresponding operation on the object after obtaining access rights.
- Condition *C*: The conditions that need to be satisfied when the access request is sent by the access subject *SA*, e.g., whether the risk value of the subject is in the risk interval that allows access, whether the access time is within the office hours, etc. In this paper, we propose to use agglomerated hierarchical clustering for risk interval classification.

- Risk quantification component *RQ*: quantify the access request behavior of the access subject *SA* according to the risk quantification component, and the quantification result will update the subject' s risk attributes.

The risk and UCON based access request model proposed in this paper consists of four main modules, which are:

- Risk quantification module: When a user sends an access request, the risk quantification module calculates the risk value (hereinafter referred to as "direct risk") of the doctor's behavior by using the doctor's work target, operation behavior, access time, and sensitivity of the accessed information. It also calculates the risk value (hereinafter referred to as "indirect risk") of the entire department's physician's access behavior through the history of the entire department's access records for the same work target. The module finally calculates the total risk caused by the physician's visit history based on the direct risk and indirect risk.
- Doctor clustering module: The risk values derived from the doctor's historical visit records are clustered into four classes using an agglomerated hierarchical clustering algorithm, and the risk value intervals corresponding to the four classes of doctors are obtained.
- Subject property update module: The EWMA algorithm is used to predict the risk value of a physician's current visit based on the risk value derived from the physician's visit history, and to update the physician's risk value.
- Access Control Policy Module: The access control policy is based on the risk value of the doctor's visit and the corresponding risk value interval of the four categories of doctors.

Risk quantification module

When doctors may need to access other patients' EMR to help them determine a patient's condition during a patient visit, the process of accessing other patients' EMR may result in privacy breaches for other patients. Therefore, before a physician visits another patient's EMR, the risk quantification model calculates the risk value that could cause a patient information breach during the physician's historical visit by using the physician's and the entire department's historical visit for the previous month.

In this paper, to avoid the inaccuracy of considering only the physician's access behavior that may cause the risk of patient information disclosure, we divide the physician's risk into two parts: direct risk and indirect risk. Direct risk is calculated based on the physician's visit behavior, reflecting whether the physician's visit is within the scope of a normal visit. In this paper, we introduce indirect risk, which is obtained by averaging the risk values obtained by the authority of the physician's role in the entire department. The risk value of the current physician's historical record is obtained by the weighted average of direct and indirect risks [40] to avoid high-risk values due to multiple emergency visits to the medical records of other departments, which prevent physicians from performing their daily work properly.

Direct risk

(1) Identify direct risk influencing factors. When analyzing the factors influencing the direct risk, the main factors are the historical visit behavior of the physician in the visit log in *ATT(S)*, the operational behavior of the physician, the visit time of the physician, and the sensitivity of the patient information in *ATT(O)*. In this paper, four factors are used as influencing factors to calculate the risk of physical access control, and the risk factor set $U = \{U_1, U_2, U_3, U_4\}$ is constructed. Since physicians' operation behaviors on medical records can be classified as viewing, copying, adding, and deleting, the first level evaluation factor U_2 can be divided into $U_2 = \{U_{21}, U_{22}, U_{23}, U_{24}\}$. The doctor's visit time can be divided into on-time, off-time, and emergency time, so the first level evaluation factor U_3 can be divided into $U_3 = \{U_{31}, U_{32}, U_{33}\}$ The sensitivity of patient information can be classified as top secret, private and general according to the patient's influence and the type of illness suffered, so the first level evaluation factor U_4 can be classified as $U_4 = \{U_{41}, U_{42}, U_{43}\}$ where $U = U_1 \cup U_2 \cup U_3 \cup U_4$, and $U_i \cap U_j = \emptyset, \forall i \neq j, i, j = 1, 2, 3, 4$

(2) Determining the set of direct risk weights. Different risk-influencing factors have different degrees of influence on physician risk when conducting direct risk quantification of physicians, and this paper uses the coefficient of variation weighting method [41] to assign different weights [42] to different risk factors. The original indicator data matrix was first constructed based on the number of physicians m in the entire department and four factors affecting physician risk.

$$X = \begin{pmatrix} x_{11} & \cdots & x_{14} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{m4} \end{pmatrix}$$

where X_{ij} denotes the value of the j^{th} risk evaluation factor for the i^{th} physician. The mean and standard deviation of the j^{th} risk evaluation factor was calculated by Eq. 1 and Eq. 2. Then, the calculated mean and standard deviation are used to calculate the corresponding coefficient of variation by Eq. 3.

$$\overline{x_j} = \frac{1}{n} \sum\nolimits_{i=1}^{n} x_{ij} \tag{1}$$

$$S_j = \sqrt{\frac{\sum_{i=1}^{n} \left(x_{ij} - \overline{x_{ij}}\right)}{n-1}} \tag{2}$$

$$v_j = \frac{S_j}{\overline{x_j}}, j = 1, 2, \ldots, 4 \tag{3}$$

The coefficient of variation is normalized, and then the weights of each risk influencing factor are obtained. Finally, the last weights $\lambda = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ are calculated by Eq. 4.

$$\lambda_j = \frac{v_j}{\sum_{j=1}^{4} v_j} \tag{4}$$

(3) The risk value that may result from each risk factor is calculated and the direct risk to the physician is calculated. The direct risk consists of four main factors that affect the risk value: the physician's historical visit behavior, the physician's operational behavior, the physician's visit time, and the sensitivity of the patient's information. We calculate the risk value of the first three factors based on the access log in *ATT(S)* and the last factor based on the confidentiality level in *ATT(O)*.

Firstly, the risk value of the physician's historical visit behavior is calculated. The ICD codes in the electronic medical record of a physician's visit for a certain work objective are abstracted into a 7*26 matrix according to the coding rules. For example, B01.901 (chickenpox) is abstracted as matrix Q. The first column of the matrix represents the first letter B of the code, and each column of the matrix corresponds to each bit of the ICD code, in turn, yielding the matrix Q of chickenpox:

$$Q = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & 1 & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & 1 & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & 1 & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \vdots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Therefore, the n records visited by doctor i for a certain work target Q can be abstracted as n matrices: $Q^i = \left[Q_1^i, Q_2^i, \cdots, Q_n^i\right]$. All the medical records visited by the m doctors of the whole department for the work objective Q can be abstracted as.

The abstracted matrix was calculated by Euclidean distance [43] to find the deviation between doctor i's job goal Q and visit behavior, the deviation of each doctor in the whole department under job goal Q from visit behavior, and the deviation of doctor's visit behavior under this job goal from the visit behavior of doctors in this department.

The Euclidean distance represents the distance between the point (x_2, y_2) and the point (x_1, y_1) in two dimensions, and $|X|$ denotes the Euclidean distance from the origin to the point (x_2, y_2), calculated as:

$$\rho = \sqrt{(x_2 - x_1)^2 + \left(y_2 - y_1\right)^2}, |X| = \sqrt{x_2^2 + y_2^2} \tag{5}$$

Let Q_{1j}^i be the j^{th} row of the matrix Q_1^i, and let Q_k be the k^{th} row of matrix Q. We use Eq. 6 to calculate the deviation $dev[j][k]$ between and Q_1^i:

$$
\begin{aligned}
& dev[j][k] = Q = \left[Q^1, Q^2, \cdots, Q^m\right] \\
& \sqrt{\left(Q_{11}^i - Q_1\right)^2 + \left(Q_{12}^i - Q_2\right)^2 + \cdots + \left(Q_{1m}^i - Q_m\right)^2 + \cdots + \left(Q_{126}^i - Q_{26}\right)^2} \\
& = \sqrt{\left\| Q_{1j}^{i2} \right\| + \left\| Q_k^2 \right\| - 2 * Q_{1j}^i Q_k^T}
\end{aligned} \tag{6}
$$

Then Eq. 6 is extended to the formula of the i^{th} row of the matrix. Finally, Eq. 6 is extended to the whole matrix to calculate the deviation dev of the matrix Q_1^i and the matrix Q.

$$
dev = \sqrt{\begin{pmatrix} Q_{11}^{i2} & \cdots & Q_{11}^{i2} \\ \vdots & \ddots & \vdots \\ Q_{126}^{i2} & \cdots & Q_{126}^{i2} \end{pmatrix} + \begin{pmatrix} Q_1^2 & \cdots & Q_1^2 \\ \vdots & \ddots & \vdots \\ Q_{26}^2 & \cdots & Q_{26}^2 \end{pmatrix} - 2 * Q_1^i Q^T} \tag{7}
$$

According to Eq. 7, dev_i^k the deviation between a certain work target k of doctor i and the visiting behavior, the deviation $dev_{sum_1}^k$、$dev_{sum_2}^k$、…、$dev_{sum_m}^k$ of each doctor in the whole department under the work target k and the visiting behavior $dev_{i\_sum}^k$ of doctor i under the work target k and the visiting behavior of doctors in this department are calculated.

Therefore, the deviation of doctor i's work to target and visit behavior in a certain period is the weighted average of the deviation between the doctor's work target and visit behavior and the deviation of doctor's visit behavior under the work target and the visit behavior of doctors in this department. The deviation degree dev_i is the risk value r_{i1} calculated by doctor i based on the historical visit behavior. i.e. $r_{i1} = dev_i$.

$$
r_{i1} = dev_i = \frac{\sum_{j=1}^{k} dev_i^j + \sum_{j=1}^{k} dev_{i\_sum}^j}{2} \tag{8}
$$

Doctors' manipulation of medical records mainly includes four types of operations: viewing patients' medical records, copying electronic medical records, adding medical records, and deleting related operations. The viewing, copying, adding, and deleting of a doctor's previous medical records or medical records of other doctors through work objectives may lead to the leakage of patients' privacy. However, the probability that a physician's copying, adding, and deleting medical records of patients under the supervision of other physicians generates risk is greater than the probability that a physician's adding medical records of his or her patients generates risk. So, the risk of a physician's manipulation of medical records is expressed as a ratio between the number of times a physician copy, adds, and deletes medical records of other physicians and the total number of times a physician copies, adds, and deletes medical records.

$$
r_{i2} = \frac{\sum_{i \neq j} C_{Oi}^j + \sum_{i \neq j} A_{D_i}^j + \sum_{i \neq j} D_{E_i}^j}{\sum_{j=1}^{k} C_{Oi}^j + \sum_{j=1}^{k} A_{D_i}^j + \sum_{j=1}^{k} D_{E_i}^j} \tag{9}
$$

where $\sum_{i \neq j} C_{Oi}^j$ indicates the total number of times that doctor i copy medical records of other doctors, $\sum_{i \neq j} A_{D_i}^j$ indicates the total number of times that doctor i add medical records of other doctors, and $\sum_{i \neq j} D_{E_i}^j$ indicates the total number of times that doctor i deletes medical records of other doctors; $\sum_{j=1}^{k} C_{O_i}^j$ indicates the total number of times that doctor i copies medical records, $\sum_{j=1}^{k} A_{D_i}^{j}$ indicates the total number of times that doctor i adds medical records, and $\sum_{j=1}^{k} D_{E_i}^{j}$ indicates the total number of times that doctor i deletes medical records.

The doctor's visit time can be divided into on-time, off-time, and emergency time, so the first level evaluation factor U_3 can be divided into $U_3 = \{U_{31}, U_{32}, U_{33}\}$. The sensitivity of patient information can be classified as top secret, private and general according to the patient's influence and the type of illness suffered, so the first level evaluation factor U_4 can be classified as $U_4 = \{U_{41}, U_{42}, U_{43}\}$.

Physicians' access time can be divided into office hours, off-duty hours, and emergency hours. Physicians' access behavior during off-duty hours largely causes the risk of patient privacy leakage, so the risk of physician-to-medical access time is expressed by the ratio of the number of times physicians access patient data during off-duty hours to the number of times physicians access patient data during compound office hours, off-duty hours and emergency hours.

$$r_{i3} = \frac{\sum_{j=1}^{k} offduty_i^j}{\sum_{j=1}^{k} onfduty_i^j + \sum_{j=1}^{k} offduty_i^j + \sum_{j=1}^{k} urgency_i^j} \tag{10}$$

where $\sum_{j=1}^{k} offduty_i^j$ indicates the number of times that doctor i visits patient information during off-duty hours, $\sum_{j=1}^{k} onduty_i^j$ indicates the number of times that doctor i visits patient information during office hours, and $\sum_{j=1}^{k} urgency_i^j$ indicates the number of times that doctor i visits patient information during emergencies.

The level of confidentiality of patient information is stored as *ATT(O)* at the time of admission and the level of sensitivity of patient information accessed by physicians may lead to a risk of patient privacy disclosure. Therefore, this risk value is calculated by the number of patient information accessed by physicians with sensitivity levels in top secret and private as a percentage of all patient information accessed by physicians.

$$r_{i4} = \frac{\sum_{j=1}^{k} TS_i^j + \sum_{j=1}^{k} S_i^j}{\sum_{j=1}^{k} TS_i^j + \sum_{j=1}^{k} S_i^j + \sum_{j=1}^{k} G_i^j} \tag{11}$$

where $\sum_{j=1}^{k} TS_i^j$ and $\sum_{j=1}^{k} S_i^j$ denote the number of times physician i accessed top secret and private patient information and $\sum_{j=1}^{k} G_i^j$ denotes the number of times physician i accessed ordinary patient information.

(4) Calculating direct risk. The direct risk r_d of doctor i is calculated by the already calculated risk value $r = \{r_{i1}, r_{i2}, r_{i3}, r_{i4}\}$ and the weight $\lambda = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ derived from the coefficient of variation weighting method.

$$r_d = \lambda_1 * r_{i1} + \lambda_2 * r_{i2} + \lambda_3 * r_{i3} + \lambda_4 * r_{i4} \tag{12}$$

Indirect risk

The concept of indirect risk is introduced to reconcile the risk value of a physician with the direct risk calculated from the physician's historical visit behavior, which lacks a certain degree of accuracy, and the fact that unexpected circumstances of a physician may lead to a higher risk value for some physicians than the risk value generated by his normal work. The risk value for indirect risk is calculated from the deviation between the work objectives of all physicians in the entire department and the visit records, i.e.

$$r_{id} = \frac{\sum_{i=1}^{m} r_{i1}}{m} \tag{13}$$

Thus the physician's risk value is the weighted average of the risk values of the physician's direct and indirect risks. At this point, the physician's risk value is calculated based on the physician's visit history, and the risk value is stored in the *ATT(S)* of *SA* according to the RQ-UCON model. Where r_d is the direct risk,r_{id} is the indirect risk, and ξ is the weighting parameter, i.e.

$$R = \xi * r_d + (1 - \xi) * r_{id} \tag{14}$$

Physician clustering module

The operation behaviors of doctors accessing the hospital information system(HIS) are mainly viewing patient records, copying electronic medical records, adding treatment records, and deleting related operations. In this paper, we need to classify doctors according to their different risk values and divide them into different risk intervals and operational behaviors; based on the advantages that the cohesive hierarchical clustering algorithm can pre-specify the number of clusters and discover hierarchical relationships between classes, this paper uses bottom-up cohesive hierarchical clustering to cluster doctors within departments. Therefore, we use agglomerated hierarchical clustering to classify the corresponding risk value intervals for the above four operational behaviors of physicians in this department according to the risk values in the *ATT(S)* of physicians in the department. When the risk value of the doctor is not in the risk range of the operation, the HIS does not authorize the doctor to operate, which protects the patient's privacy to a certain extent.

According to the bottom-up agglomerated hierarchical clustering [44] algorithm each physician's risk value is first considered as a cluster, and then the distance between each cluster is calculated, and the two clusters with the closest or most similar distance are merged. Until the number of clusters stops at 4. Finally, the risk interval corresponding to each class of doctors is output. In this paper, the average distance is chosen as the metric to measure the distance between two clusters among many methods to calculate the distance. It defines the distance between two clusters [45] as the average distance between the data points in the first cluster and the data points in the second cluster, as shown in Eq. (15).

$$dist(c_1, c_2) = \frac{1}{c_1 c_2} \sum_{p_1 \in c_1, p_2 \in c_2} |p_1 - p_2| \tag{15}$$

where c_1 *and* c_2 denote the number of samples in the clusters, and p_1 *and* p_2 are different clusters.

The detailed algorithmic procedure of agglomerated hierarchical clustering is as follows.

Algorithm 1: Agglomerated Hierarchical Clustering Algorithm

```
Input: Sample set composed of risk values visited by department doctors $R=\{r_1,r_2,\cdots,r_n\}$ ; Clustering cluster distance metric function dist; Numbers of cluster 4
Output: Cluster division results $D=\{D_1,D_2,D_3,D_4\}$
for ($i$ = 1, 2, … , $n$){
    $D_i$={$r_i$};
}
for ($i$ = 1, 2, … , $n$){
    for ($j$ = 1, 2, … , $n$) {
        M[$i$][$j$] = dist($D_i$ , $D_j$) ;
        M[$j$][$i$] = M[$i$][$j$];
    }
}
Set the current number of clusters: $q$=$n$ ;
while ($q$ > 4) {
    Find the two closest clusters: $D_i$ and $D_j$ ;
    Merge $D_i$ and $D_j$:  $D_i$=$D_i \cup D_j$ ;
    for ($j$ = $i$+1, $i$+2, …, $q$-1) {
        $D_j$=$D_{j-1}$ ;
    }
    Delete row i and column j of matrix M[i][j] = dist($D_i$ , $D_j$) ;
    for ($j$=1, 2, …, $q$-1) {
        M[$i$][$j$] = dist($D_i$ , $D_j$) ;
    }
    $q$ = $q$-1;
}
```

Four categories of doctors are obtained by Algorithm 1, and the corresponding risk intervals are obtained from the risk values of the four categories of doctors, but it is not possible to determine the correspondence between each risk interval and the access operation behavior, and category determination is required. According to the previous description, it is known that the more operation behaviors of access authority, the lower the risk value of this doctor; when the risk value of a doctor exceeds a certain risk interval, the access request of this doctor will be denied. Therefore, according to this feature, the correspondence between risk interval and access operation behavior can be determined for each category. The risk intervals of operational behaviors obtained by agglomerated hierarchical clustering algorithm are used as conditions *C* in the RQ-UCON model to implement the access control policy together with other modules. i.e.

$$P=\begin{cases} L_O,C_O,A_D,D_E.....T_1 \le R_{t+1} < T_2 \\ L_O,C_O,A_D...........T_2 \le R_{t+1} < T_3 \\ L_O,C_O................T_3 \le R_{t+1} < T_4 \\ NoAuthority.....T_4 \le R_{t+1} < T_5 \end{cases}$$

where *P* is the doctor's permission, L_O indicates that the doctor views the patient's medical record, C_O indicates that the doctor copies the electronic medical record, A_D indicates that the doctor adds a consultation record, and D_E indicates that the doctor deletes the related operation. $T_1...T_5$ denotes the breakpoint value of the risk interval. Therefore, the permissions *P* of the RQ-UCON model are divided into the following four types: $P_1=[L_O,C_O,A_D,D_E]$, $P_2=[L_O,C_O,A_D]$, $P_3=[L_O,C_O]$, $P_4=[No\ Authority]$; the corresponding conditions *C* of the RQ-UCON model are: $C_1=[T_1,T_2)$, $C_2=[T_2,T_3)$, $C_3=[T_3,T_4)$, $C_4=[T_4,T_5)$.

Subject property update

The risk value of the doctor's visit history is calculated based on the doctor's history and the risk value of the doctor's current visit is predicted based on this risk value, and the doctor's risk value needs to be dynamized, and the EWMA algorithm is used for the dynamization.

The EWMA algorithm is an evolution of the Weighted Moving Average (WMA) method, which assigns different weights to the known data and derives the moving average based on these weights, and uses this value as the basis for determining the predicted value. EWMA, on the other hand, refers to the exponential decrease of the weighting coefficients of the numbers over time and has the main advantage that it does not need to save the past values, and only a very small amount of memory is used [46]. It can predict the value at moment $t + 1$ based on the value at moment t. Therefore, this paper combines the EWMA algorithm with the penalty factor to propose an improved EWMA algorithm to predict the risk value of doctor i for this visit, where the penalty factor is related to the number of times doctor i is denied access. The specific form is shown below.

$$EWMA_t = \omega EWMA_{t-1} + (1 - \omega)R_t \tag{16}$$

$$R_{t+1} = \omega PT_i + (1 - \omega)EWMA_t \tag{17}$$

$$PT_i = \sqrt{\frac{n(n+1)}{2}} \tag{18}$$

where R_{t+1} denotes the risk value predicted for doctor i for this visit;PT_i denotes the penalty factor for doctor i, where n is the number of denied visits for doctor i;R_t indicates the risk value of the doctor i at the moment of t; $EWMA_t$ and $EWMA_{t\text{-}1}$ represents the exponentially weighted moving average for time t and time at time t-1; $\omega(0 < \omega < 1)$ denotes for the historical measurement weight coefficient, also called exponential weighting, the weighting coefficient ω is exponentially decreasing, i.e., each index decreases with time and decreases exponentially. Using this method, the risk value of the current visit can be predicted from the risk value calculated from the physician's historical behavior for one month. And during access, the *ATT(S)* in the RQ-UCON access control model is updated in real-time by this algorithm.

Access control policies

In this section, this paper proposes a specific policy for controlling the authorization behavior in the RQ-UCON model. The RQ-UCON model works according to the access policy library mandatory access subjects are related according to the access control policy. To better control the access authorization in the RQ-UCON model, this paper classifies the user access scenarios into two types: normal access and emergency access.

Normal access: This scenario is a normal access scenario where the access subject is accessing the resources of the HIS and has legal identity rights as well as the risk value of allowed access.

Emergency access: In this access scenario, the HIS system has a higher tolerance for the access rights of the access subject. In case of emergency, the HIS system will grant the highest access rights to the access subject and record this access to avoid delaying patient treatment due to the access rights of the HIS system.

This access control policy framework diagram implements the control of physicians' access requests based on the risk value of the access subject's access, which limits the excessive access of the access subject to a certain extent and has a certain protective effect on patient information. The overall access control policy framework diagram is shown in Fig. 2.

As can be seen from Fig. 2, the overall access control process of the model is as follows: the physician as the access subject *SA* initiates a request to the HIS system to access the object *O*. The system determines whether the *SA* is an urgent access and if the *SA* is an urgent access, the authorization module *O* of the RQ-UCON model grants the *SA* full access rights; if the *SA* is not urgent access, the RQ component of the RQ-UCON model will perform the risk value calculation as well as permission *P* and condition *C* determination, and the real-time update of the calculated risk value R_{t+1} in the *ATT(S)*. The model matches the risk value R_{t+1} in the *ATT(S)* of *SA* with the condition *C*. If the match is successful, the authorization module *O* grants the corresponding permission *P* to *SA*; if the match fails, the authorization module *O* will deny this access to *SA*; finally, the access result is fed back to *SA*.

Simulation experiments

Data source

This paper relies on the National Natural Science Foundation of China project, and the data used in the experiments are obtained from a tertiary hospital in Kunming and a county people's hospital in Yunnan Province, which are the collaborators of the author's

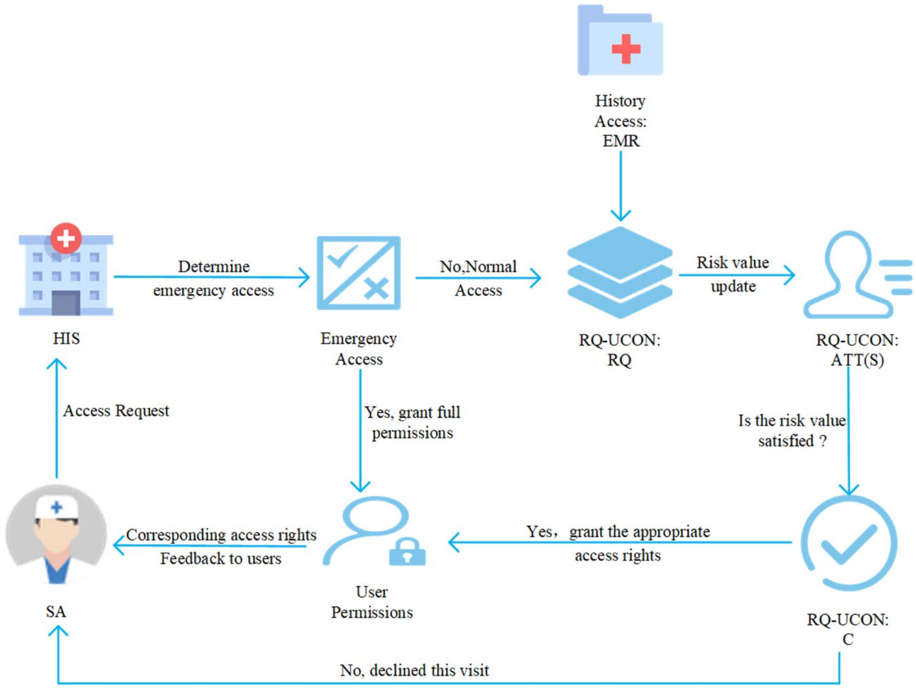


Fig. 2 Access control policy framework diagram

project team. The HIS system of a tertiary hospital in Kunming is deployed in Windows operating environment, using Oracle 11G database, and this team owns 1200 GB of its medical data, with 1360 tables and 2139373 records in the database; the HIS system of a people's hospital in Yunnan Province uses MySQL Server database, and the system is based on Internet/Intranet, and this project team has all its medical data. This experiment extracted some medical data from the two HIS system databases to quantify the risk value and simulate the access of doctors in the process of diagnosis and treatment.

Purpose of the experiment

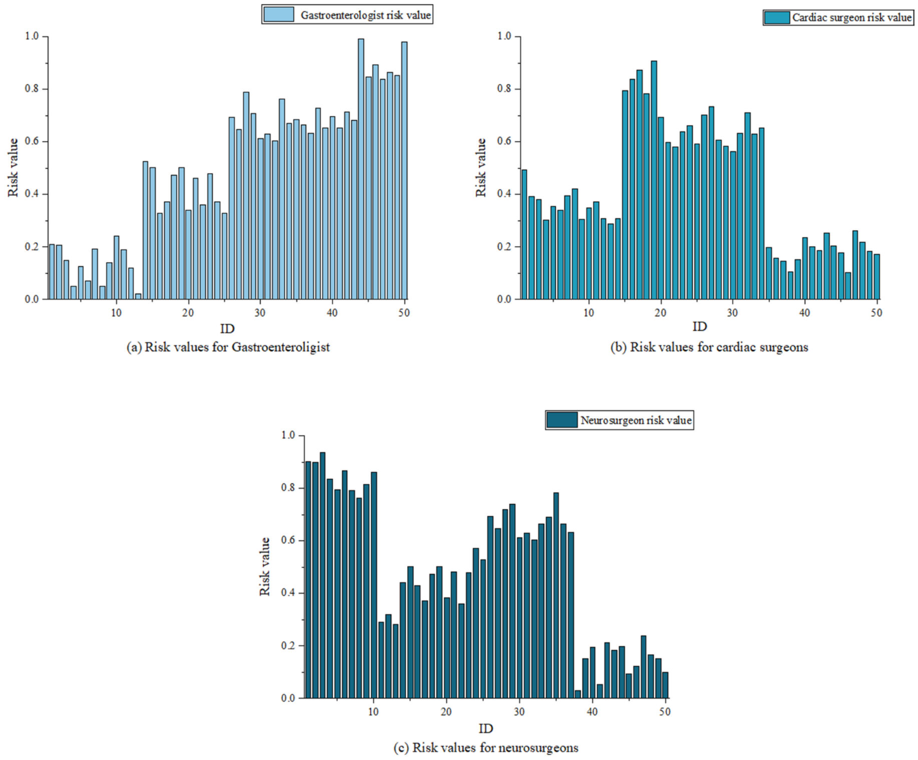
The purpose of the experimental setup is to verify the validity of the RQ-UCON model proposed in this paper, i.e., whether the model proposed in this paper can calculate the risk value of physicians based on historical access records, and verify whether the model proposed in this paper can control physician over-access. In this part, doctors who may access electronic medical records outside their work scope will be called over-access doctors, and doctors who only access electronic medical records within their work scope will be called normal-access doctors. The experimental procedure will be set up to observe the effect of different physicians in different departments on the risk value calculation and risk interval delineation to check the feasibility of the model. In the access control module, the success rate of interception of the proposed model in this paper is observed by the difference in risk value and risk interval of three different department doctors. To test the overall performance of the model, 50 doctors from each of the three different departments are selected for the experiment and compared with the data using 800 doctors with Huizhen model.

Risk quantification and access control experiments

Risk quantification for doctors in different department

The purpose of this experiment is to test whether the risk quantification module proposed in this model can quantify the risk value of physician access control based on the medical records of physicians in different departments. This module extracts one month of historical visit records of 50 physicians from each of three different departments, namely gastroenterology, cardiac surgery, and neurosurgery, and calculates the risk value of each physician through the risk quantification module, and the experimental results are shown in Fig. 3.

As can be seen from Fig. 3, the quantification of risk values based on different historical visit records of physicians in different departments shows that there are some differences in the risk values of physicians between different departments, and it is known that the risk quantification module proposed in this paper can quantify risk values based on different historical records. From Fig. 3, it can be seen that doctors with risk values bigger than 0.8 in gastroenterology are about 14% of the number of doctors in gastroenterology; doctors with risk values bigger than 0.8 in cardiac surgery are about 6% of the number of cardiac surgeons; doctors with risk values bigger than 0.8 in neurosurgery are about 14% of the number of neurosurgeons; the number of doctors with higher risk values in each department is small, which also indicates that most doctors do not make excessive visits.



(a) Risk values for Gastroentroligist

(b) Risk values for cardiac surgeons

(c) Risk values for neurosurgeons

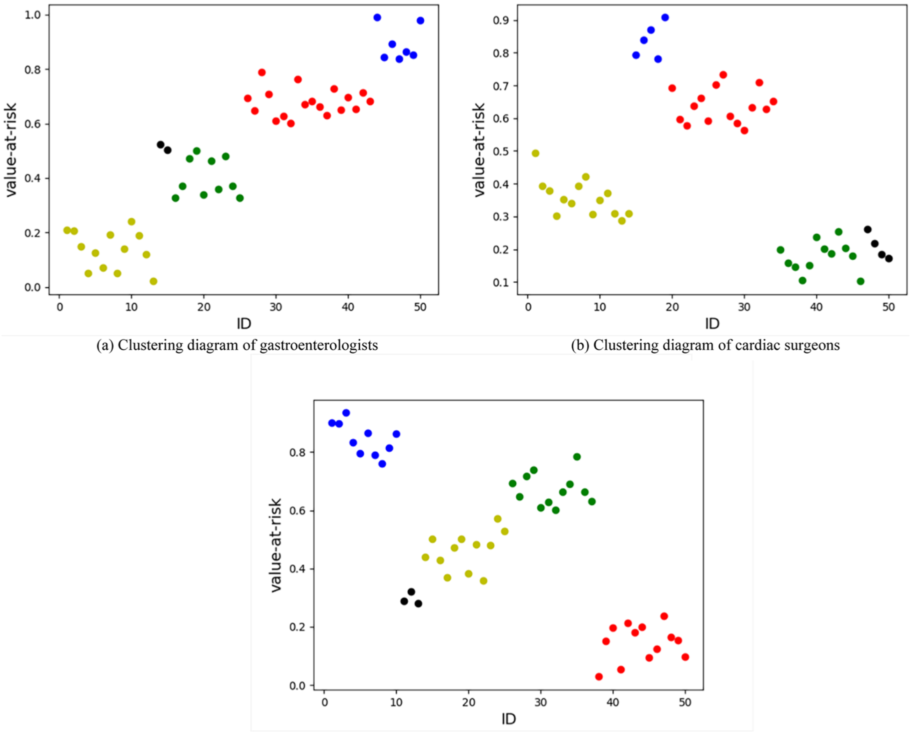
Fig. 3 Risk values for physicians in the department

After that, we performed agglomerated hierarchical clustering based on the risk values of each of the three different sections of physicians to classify different risk intervals for the operational behaviors of the different sections of physician visits. The clustering results are shown in Fig. 4.

As can be seen from Fig. 4, four different clustering results were obtained according to the risk values of doctors in different departments, where the black points are noise points, which are excluded. Based on the clustering results in Fig. 4, the risk value interval of different access control actions of doctors is obtained and the risk value interval of no access rights is obtained. According to Table 1, it can be seen that there are some differences in the risk intervals of different departments' operational behaviors, but overall, the differences are not large, which also indicates that most doctors only visit the electronic medical records related to their work, and the doctors who over-visit only account for a small proportion of all doctors. The model proposed in this paper has some validity and can identify a small percentage of doctors.

Access control blocking success rate and recall for doctors in different departments

To verify that the access control model proposed in this paper can deny access requests from over-accessing physicians, we conducted experiments on the accuracy of access control using the risk values and risk intervals of access operation behaviors of physicians in different departments derived from the experiments in the previous module. We took visit histories of 50 physicians from each of three different



(a) Clustering diagram of gastroenterologists

(b) Clustering diagram of cardiac surgeons

(c) Cluster diagram of neurosurgeons

Fig. 4 Clustering results of departmental doctors

Table 1 Risk intervals for operational behavior

	No authority	View, Copy	View, Copy, Add	View, Copy, Add, Delete
Gastroenterology	0.786–0.992	0.52–0.785	0.25–0.51	0.014–0.249
Cardiac Surgery	0.778–0.916	0.557–0.777	0.279–0.556	0.098–0.278
Neurosurgery	0.785–0.935	0.59–0.784	0.353–0.59	0.027–0.243

departments over one month (30 days), with more than 30,000 visit controls in each department for the experiment. Where the interception success rate and recall are calculated as follows.

$$f = \frac{D_{PA}}{D_P} \tag{19}$$

$$recall = \frac{D_{PA}}{D_A} \tag{20}$$

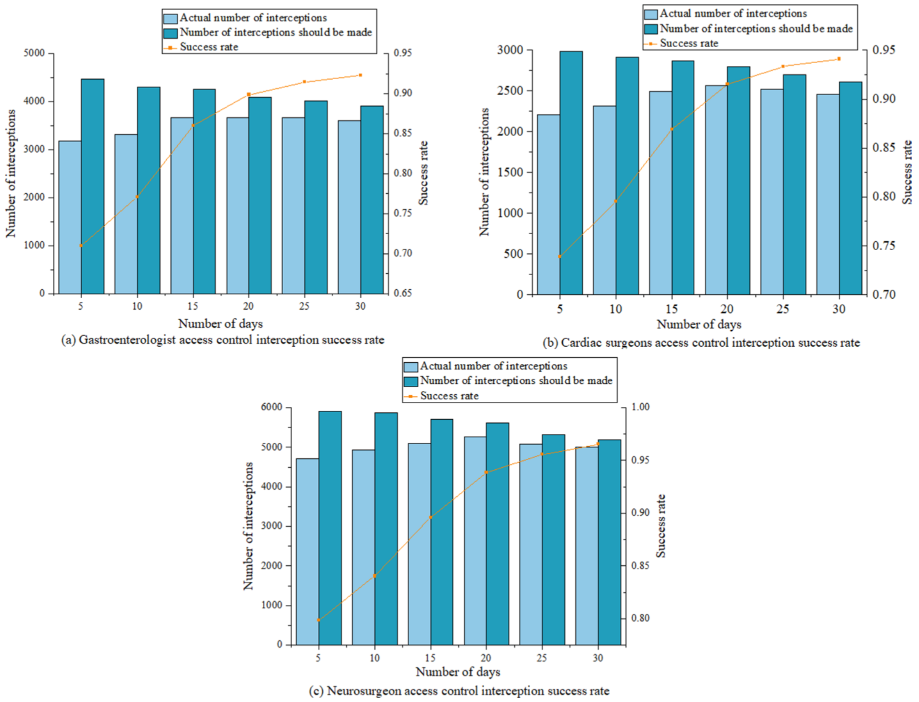
where f denotes the success rate of the RQ-UCON model proposed in this paper in blocking excessive visits by doctors, *recall* denotes recall of the RQ-UCON model proposed in this paper in blocking excessive visits by doctors, D_{PA} denotes the number of correct predictions and interceptions D_A denotes the number of actual interceptions, and D_P denotes the number of predicted interceptions.

From the analysis of the experimental results in Fig. 5, it can be seen that the success rate of access control denial for the three sections is gradually increasing with the increase of access days, which can show that the model proposed in this paper is effective. And from (a) (b) (c) of Fig. 5, it can be seen that the number of predicted interceptions is gradually decreasing with the increase of time, and the number of correct predictions and interceptions is gradually increasing. This indicates that the over-access of doctors is gradually decreasing with the increase of time, which can prove that the performance stability of the model proposed in this paper is good and the interception success rate can reach more than 90%.

From the analysis of the experimental results in Fig. 6, it can be seen that the recall rate of access control interceptions in the three departments is gradually increasing with the increase of access days, which can be seen that the model proposed in this paper is more accurate in identifying the excessive access behavior of doctors. From (a) (b) (c) of Fig. 6, it can be seen that the number of doctor's actual visit interceptions is gradually decreasing and recall is gradually increasing as time increases. This indicates that the model can better predict the over-visiting behavior of doctors and intercept the over-visiting doctors. When the experiment was conducted to 30 days, the recall of all three departments could reach more than 90%.

Comparison experiments

In this paper, we will conduct a comparison experiment with the model proposed by Huizhen [27], in which 800 physicians' historical visit information is selected for two comparison tests, and the experiment tests the superiority of the model approach through three metrics: accuracy rate, recall rate, and F1 score. Where, the accuracy rate

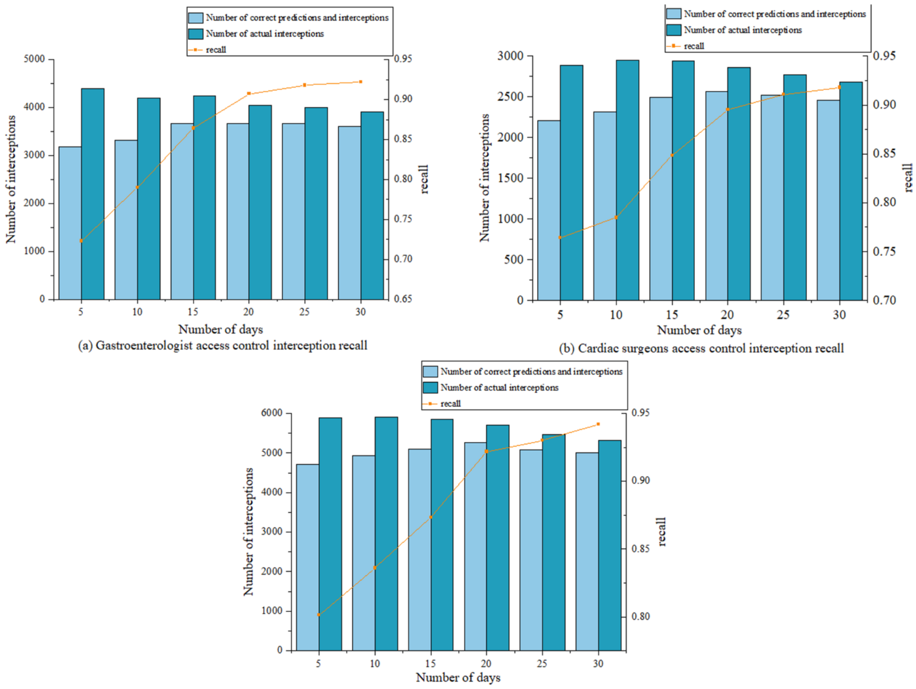


(a) Gastroenterologist access control interception success rate

(b) Cardiac surgeons access control interception success rate

(c) Neurosurgeon access control interception success rate

Fig. 5 Departmental physician access control interception success rate



(a) Gastroenterologist access control interception recall

(b) Cardiac surgeons access control interception recall

(c) Neurosurgeon access control interception recall

Fig. 6 Departmental physician access control interception recall

indicates the proportion of over-visiting physicians among the top X physicians with the highest risk; the recall rate indicates the proportion of curious physicians among the top X physicians with the highest risk to all curious physicians; F1 Score indicates the geometric mean of the accuracy and recall rates.

Experiment 1 sets different proportions of over-visiting physicians for 800 physicians, sets the number of visit requests to 10, and calculates the accuracy, recall, and F1 score to verify the effectiveness of the model proposed in this paper.

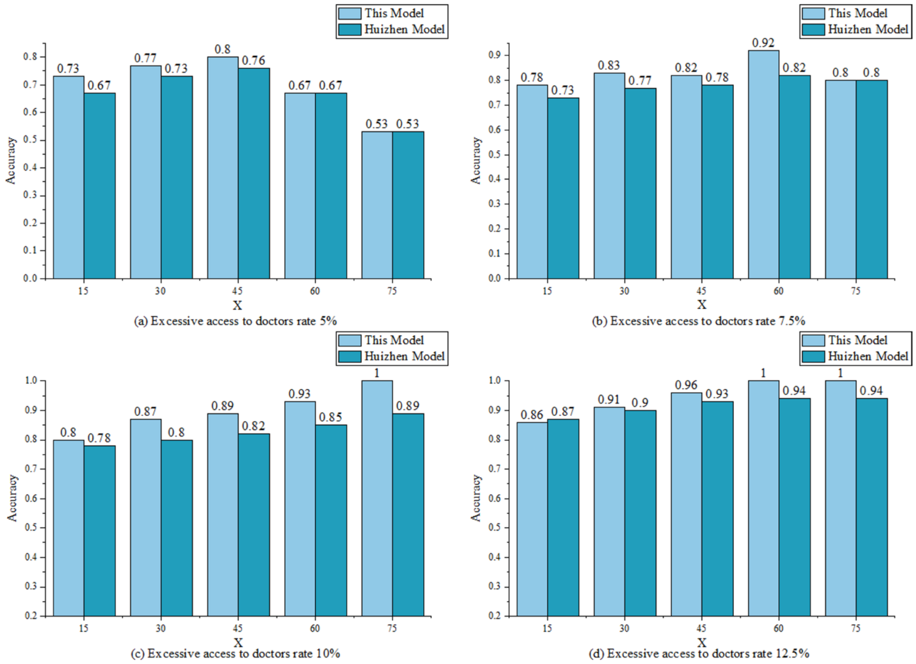
As can be seen from Table 2 and the figure above, the performance of the experiment improves as the proportion of over-visiting physicians increases. As can be seen from the Figs. 7, 8 and 9, the values of the three metrics proposed in this paper show an increasing trend, which indicates that the performance of the model proposed in this paper gradually improves as the proportion of over-visiting doctors to all doctors rises. At the same time, the performance indexes of the model proposed in this paper have some improvement over the Huizhen model under different proportions of over-visiting doctors.

Experiment 2 sets the proportion of over-visiting doctors among 800 doctors to 10%, sets the number of access requests to 10 and conducts experiments by controlling the probability of over-visiting by over-visiting doctors.

From the analysis in the comparison in Table 3, Figs. 10, 11, and 12, it is obtained that when the excess visit ratio is 6% or higher, the accuracy of the model proposed in this paper reaches more than 85% and the recall rate is somewhat improved than that of the Huizhen model, and the three performance indicators keep improving with the increase of the over-visit ratio of excess visit doctors, thus indicating that the performance of the model proposed in this paper keeps improving with the increase of the excess visit ratio of excess visit doctors.

Table 2 Performance metrics with different ratios of over-access doctors

Excessive access to doctors rate	X	Accuracy		Recall		F1 Score	
		This Model	Huizhen Model	This Model	Huizhen Model	This Model	Huizhen Model
5%	15	0.73	0.67	0.21	0.25	0.50	0.50
	30	0.77	0.73	0.53	0.55	0.63	0.64
	45	0.80	0.76	0.90	0.87	0.85	0.81
	60	0.67	0.67	1.00	1.00	0.83	0.83
	75	0.53	0.53	1.00	1.00	0.77	0.77
7.5%	15	0.78	0.73	0.19	0.22	0.50	0.46
	30	0.83	0.77	0.42	0.38	0.63	0.58
	45	0.82	0.78	0.58	0.60	0.72	0.73
	60	0.92	0.82	0.88	0.82	0.87	0.82
	75	0.80	0.80	1.00	1.00	0.90	0.90
10%	15	0.80	0.78	0.15	0.15	0.45	0.48
	30	0.87	0.80	0.33	0.30	0.57	0.55
	45	0.89	0.82	0.49	0.51	0.69	0.69
	60	0.93	0.85	0.70	0.64	0.82	0.78
	75	1.00	0.89	0.94	0.89	0.95	0.90
12.5%	15	0.86	0.87	0.14	0.13	0.54	0.50
	30	0.91	0.90	0.25	0.28	0.61	0.54
	45	0.96	0. 93	0.43	0.40	0.69	0.64
	60	1.00	0.94	0.60	0.55	0.80	0.73
	75	1.00	0.94	0.75	0.70	0.88	0.82



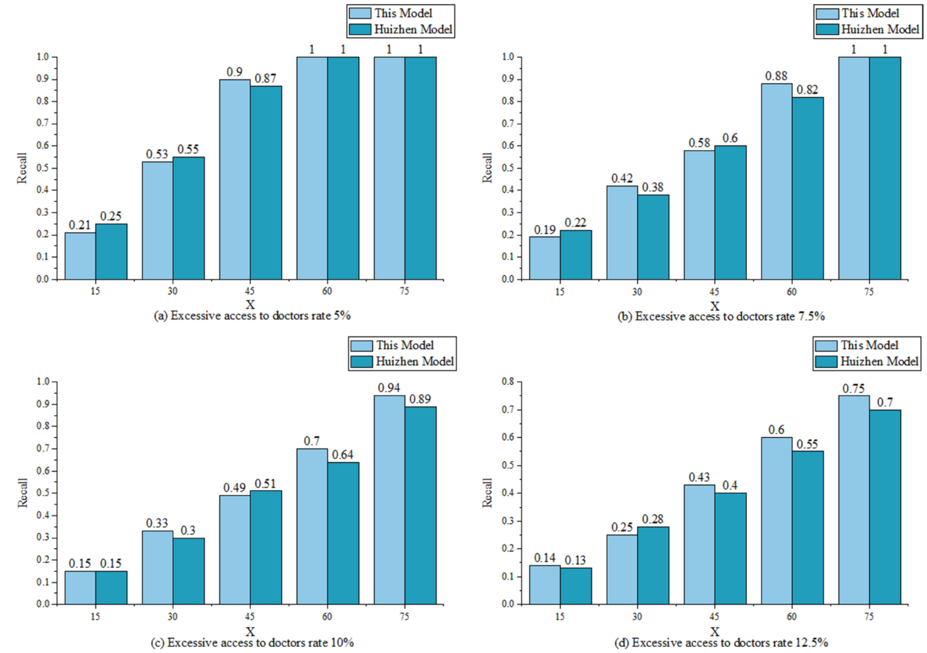
(a) Excessive access to doctors rate 5%

(b) Excessive access to doctors rate 7.5%

(c) Excessive access to doctors rate 10%

(d) Excessive access to doctors rate 12.5%

Fig. 7 Comparison of accuracy under different excessive access to doctors rate



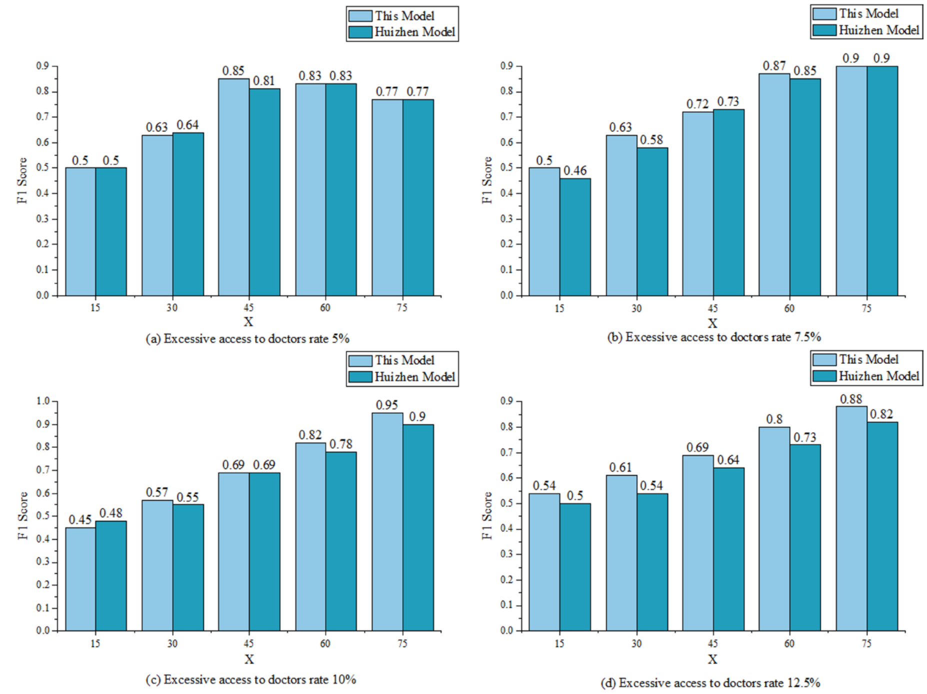
(a) Excessive access to doctors rate 5%

(b) Excessive access to doctors rate 7.5%

(c) Excessive access to doctors rate 10%

(d) Excessive access to doctors rate 12.5%

Fig. 8 Comparison of recall under different excessive access to doctors rate



(a) Excessive access to doctors rate 5%

(b) Excessive access to doctors rate 7.5%

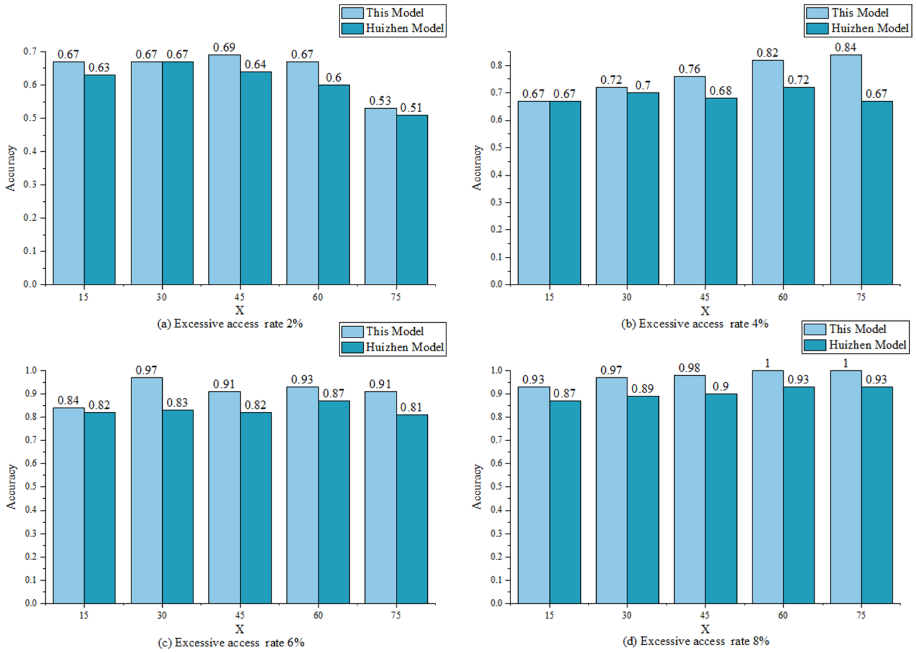
(c) Excessive access to doctors rate 10%

(d) Excessive access to doctors rate 12.5%

Fig. 9 Comparison of F1 score under different excessive access to doctors rate

Table 3 Performance metrics of over-access doctors with different over-access probabilities

Proportion	X	Accuracy		Recall		F1 Score	
		This Model	Huizhen Model	This Model	Huizhen Model	This Model	Huizhen Model
2%	15	0.67	0.63	0.13	0.14	0.40	0.36
	30	0.67	0.67	0.25	0.25	0.46	0.47
	45	0.69	0.64	0.39	0.38	0.54	0.50
	60	0.67	0.60	0.50	0.45	0.58	0.53
	75	0.53	0.51	0.50	0.48	0.52	0.49
4%	15	0.67	0.67	0.13	0.13	0.40	0.40
	30	0.72	0.70	0.28	0.26	0.50	0.49
	45	0.76	0.68	0.43	0.38	0.59	0.52
	60	0.82	0.72	0.61	0.54	0.71	0.63
	75	0.84	0.67	0.79	0.63	0.81	0.65
6%	15	0.84	0.82	0.16	0.15	0.50	0.53
	30	0.97	0.83	0.36	0.37	0.66	0.59
	45	0.91	0.82	0.51	0.46	0.71	0.64
	60	0.93	0.87	0.70	0.65	0.82	0.76
	75	0.91	0.81	0.85	0.76	0.88	0.79
8%	15	0.93	0.87	0.18	0.16	0.55	0.57
	30	0.97	0.89	0.36	0.33	0.62	0.62
	45	0.98	0.90	0.55	0.58	0.76	0.69
	60	1.00	0.93	0.75	0.68	0.88	0.79
	75	1.00	0.93	0.94	0.86	0.97	0.89



(a) Excessive access rate 2%

(b) Excessive access rate 4%

(c) Excessive access rate 6%

(d) Excessive access rate 8%

Fig. 10 Comparison of accuracy score under different excessive access rate

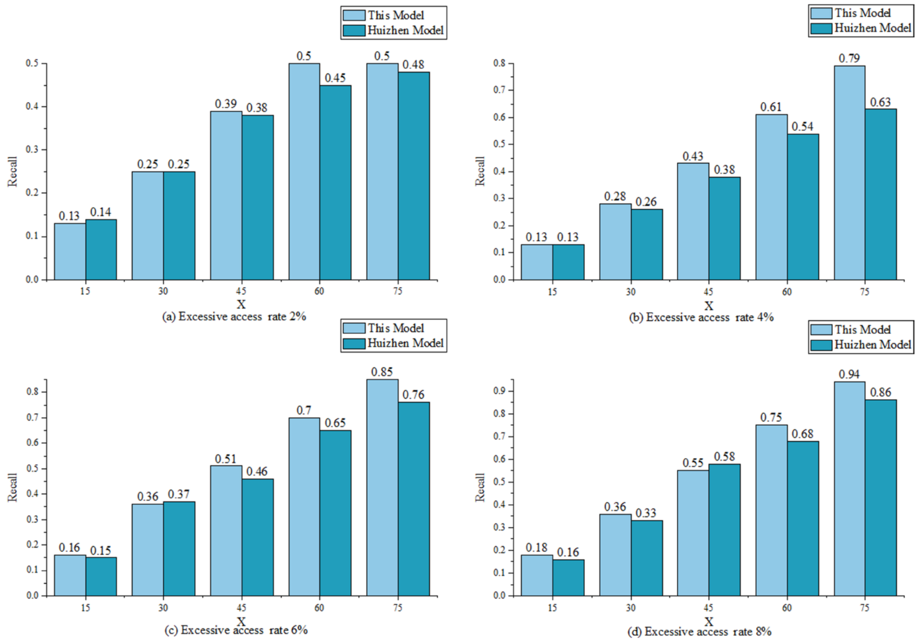


Fig. 11 Comparison of recall under different excessive access rate

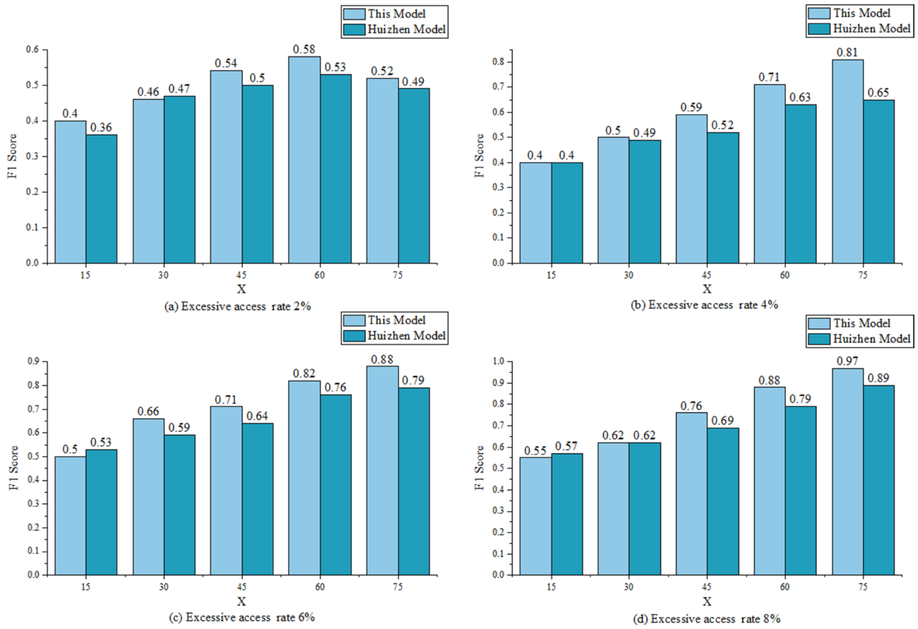


Fig. 12 Comparison of F1 score under different excessive access rate

Conclusion

Health care big data is a relatively important strategic resource in China's health care industry, but doctors accessing medical data in the process of accessing data unrelated to their work causes the risk of internal leakage of medical data. This paper proposes an access control model based on risk and UCON, which quantifies the risk value of doctors based on their historical access behaviors, updates the risk value of doctors for this access in real-time by calculating the risk value from historical records, calculates the risk interval of each operation behavior of doctors by clustering from cohesive levels, and grants doctors by matching the risk value of this doctor's access with the risk interval of the operation behavior. By matching the risk value of the doctor's access with the risk range of the operation, the doctor is granted the corresponding operation privileges. The final experiment shows that the access control model proposed in this paper has a certain control on the excessive access behavior of doctors and has a certain limitation on the privacy leakage of medical big data.

Abbreviations

RQ-UCON	Risk Quantification and Usage Control Model
EWMA	Exponentially Weighted Moving Average
COVID-19	The Corona Virus Disease
EMR	Electronic Medical Records
S	Subject
SA	Access Subject
SP	Production Subject
ATT(S)	Subject Attribute
O	Object
ATT(O)	Object Attributes
P	Permission
A	Authorization
B	Obligation
C	Condition
HIS	Hospital Information System
WMA	Weighted Moving Average

Acknowledgements
This work was supported by National Natural Science Foundation of China (Nos. 71972165, 61763048), Science and Technology Foundation of Yunnan Province (No. 202001AS070031).

Author contributions
RJ proposed the idea of this paper and designed the study. XC designed the study and wrote the manuscript. YY participated in experimental analysis. WD participated in experimental analysis. YZ reviewed and edited the paper. All authors have read and approved the final manuscript.

Funding
This work was supported by National Natural Science Foundation of China (Nos. 71972165,61763048,72271214,7176 2033), Science and Technology Foundation of Yunnan Province (No. 202001AS070031), Special Foundation for Local Science and Technology Development Guided by the Central Government of Yunnan Province Science and Technology Department (No.202307AB110009), Science and Technology Foundation of Yunnan Province Education Department (No.2023J0657), and Yunnan Social Science Foundation (No.ZD202213).

Availability of data and materials
Not applicable. For any collaboration, please contact the authors.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 12 July 2022 Accepted: 17 May 2023
Published online: 17 June 2023

References

1. Guo ZJ, Luo YC, Cai ZP, Zheng TF. Overview of privacy protection technology of big data in healthcare. Comput Sci Explor. 2021;15(03):389–402.
2. Shang JW, Jiang R, Hu XH, Shi MY. Medical big data and privacy disclosure. Comput Modernization. 2019;07:111–5.

3. Li HQ, Yin CQ, Fan JY. National strategic development study on china's health care Big Data. Lib. 2019;11:30–7.
4. Gao H, Zhou L, Kim JY, Li Y, Huang W. Applying probabilistic model checking to the behavior guidance and abnormality detection for A-MCI patients under wireless sensor network. ACM Trans Sen Netw. 2023;19(3):48.
5. Chen JT, Ying HC, Liu XC, Gu JJ, Feng RW, Chen TT, et al. A transfer learning based super-resolution microscopy for biopsy slice images: the joint methods perspective. IEEE-ACM Trans Comput Biol Bioinform. 2021;18(1):103–13.
6. Yin YP, Lin YD, Zhu FH. Hierarchical security model design of medical big data in cloud storage based on segmentation. Comput Netw. 2021;47(03):65–7.
7. Yang HY, Ning YG. Cloud platform dynamic risk access control model. J Xidian Univ. 2018;45(05):80–8.
8. Bugliesi M, Colazzo D, Crafa S, Macedonio D. A type system for discretionary access control. Math Struct Comput Sci. 2009;19(4):839–75.
9. McCune JM, Jaeger T, Berger S, Caceres R, Sailer R, editors. Shamon: a system for distributed mandatory access control. 2006 22nd annual computer security applications Conference (ACSAC'06); 2006 11–15 Dec. 2006.
10. Uddin M, Islam S, Al-Nemrat A. A dynamic access control model using authorising workflow and task-role-based access control. IEEE Access. 2019;7:166676–89.
11. Ma X, Xu H, Gao H, Bian M, Hussain W. Real-Time virtual machine scheduling in industry iot network: a reinforcement learning method. IEEE Trans Industr Inf. 2023;19(2):2129–39.
12. Akhuseyinoglu NB, Joshi J, editors. A risk-aware access control framework for cyber-physical systems. 2017 IEEE 3rd International Conference on collaboration and internet computing (CIC); 2017 15–17.
13. dos Santos DR, Marinho R, Schmitt GR, Westphall CM, Westphall CB. A framework and risk assessment approaches for risk-based access control in the cloud. J Network Comput Appl. 2016;74:86–97.
14. Lv ZH, Qiao L. Analysis of healthcare big data. Future Gener Comp Sy. 2020;109:103–10.
15. Li J, Zhang SH, Wang Y. Research on data security management mechanism of regional health and medical Big Data center. Chin Digital Med. 2020;15(12):1–4.
16. Hou MW, Lan X, Xing L, Na T, Lu L. Study on the application of privacy protection technology in the publication of health care Big Data. Chin Digital Med. 2020;15(02):92–4.
17. Xiao L, Li D, Sun Y, Shu Q, Xu XB, Xu SR, et al. Protection of personal privacy in the health and medical big data environment. China Med Record. 2019;20(12):48–50.
18. Wang Y, Jiang ZY, Pu C. Research on the current situation, problems and countermeasures of health and medical big data information security in China. Modern Med Health. 2021;37(17):3036–9.
19. Soceanu A, Vasylenko M, Egner A, Muntean T, editors. Managing the privacy and security of eHealth Data. 2015 20th International Conference on control systems and computer science; 2015 27–29.
20. Wu X, Zhang YT, Wang AM, Shi MY, Wang HH, Liu L. MNSSp3: medical big data privacy protection platform based on Internet of things. Neural Comp Appl. 2020. https://doi.org/10.1007/s00521-020-04873-z.
21. Jiang R, Xin Y, Chen Z, Zhang Y. A medical Big Data access control model based on fuzzy trust prediction and regression analysis. Appl Soft Comput. 2022;117: 108423.
22. Gan L, Yang JH, Lu SF. Research on electronic medical record sharing technology based on block chain. Chin Digital Med. 2019;14(12):11–3.
23. Lee NY, Wu BH, editors. Privacy protection technology and access control mechanism for medical Big Data. 6th IIAI International Congress on advanced applied informatics (IIAI-AAI); 2017 Jul 09–13; Hamamatsu, JAPAN2017.
24. Hossain A, Ferdous SMS, Islam S, Maalouf N, editors. Rapid Cloud data processing with healthcare information protection. IEEE World Congress on Services (SERVICES); 2014 Jun 27-Jul 02; Anchorage, AK2014.
25. Zaabar B, Cheikhrouhou O, Jamil F, Ammi M, Abid M. HealthBlock: a secure blockchain-based healthcare data management system. Comput Netw. 2021;200: 108500.
26. Jiang R, Han SS, Yu YM, Ding WP. An access control model for medical big data based on clustering and risk. Inf Sci. 2023;621:691–707.
27. Hui Z, Li H, Zhang M, Feng DG. Risk-adaptive access control model for big data in healthcare. J Comm. 2015;36(12):190–9.
28. Wang Q, Jin H, editors. Quantified risk-adaptive access control for patient privacy protection in health information systems. Proceedings of the 6th International symposium on information, computer and communications security, ASIACCS 2011; 2011: Association for Computing Machinery.
29. Li JS, Peng CG, Zhu YJ. Risk access control model for Hadoop. China J Network Inf Secur. 2016;2(01):46–52.
30. Daoud WB, Meddeb-Makhlouf A, Zarai F, editors. A Model of role-risk based intrusion prevention for cloud environment. 2018 14th International Wireless communications & mobile computing Conference (IWCMC); 2018 25–29 June 2018.
31. Aluvalu R, Muddana L, editors. A dynamic attribute-based risk aware access control model (DA-RAAC) for cloud computing. 7th IEEE International Conference on Computational Intelligence and Computing Research (ICCIC); 2016 Dec 15–17; Agni Coll Technol, Chennai, INDIA2016.
32. Shi XJ, Yu WH. Access control risk quantification method based on fuzzy neural network. Intell Comput Appl. 2018;8(01):31–5.
33. Zakaria H, Abu Bakar NA, Hassan NH, Yaacob S. IoT security risk management model for secured practice in healthcare environment. Procedia Comput Sci. 2019;161:1241–8.
34. Martinelli F, Mori P. On usage control for GRID systems. Future Gener Comp Sy. 2010;26(7):1032–42.
35. Liu ZF, Mao ZL. A UCON management model based on RBAC. Comput Sci. 2016;43(10):150–3.
36. Fan KF, Yao XZ, Fan XH, Wang Y, Chen MJ. A new usage control protocol for data protection of cloud environment. Eurasip J Inf Secur. 2016. https://doi.org/10.1186/s13635-016-0031-6.
37. Li YP, Zhou WL, Li ET, Inc DEP, editors. Usage Control Strategy Based on UCONABC Model and Subject Reliability. International Conference on Information Engineering and Communications Technology (IECT); 2016 Jun 25–26; Shanghai, PEOPLES R CHINA. LANCASTER: Destech Publications, Inc; 2016.
38. Munoz-Arcentales A, López-Pernas S, Pozo A, Alonso Á, Salvachúa J, Huecas G. An architecture for providing data usage and access control in data sharing ecosystems. Procedia Comput Sci. 2019;160:590–7.

39. Wang Y, Chen WH, Ju SG. An application control model for electronic medical record system. Comput Sci. 2010;37(11):190–3.
40. Bai X. The research of dynamic control based on risk and role in cloud computing environment [Master]. Chennai: Beijing Technology University; 2017.
41. Li Y, Liu J, Jiang AM. Research on safety ability evaluation of construction workers based on variation coefficient method and fuzzy theory. J Railw Sci Eng. 2020;17(08):2162–70.
42. Jiang R, Kang Y, Liu Y, Liang Z, Duan Y, Sun Y, et al. A trust transitivity model of small and medium-sized manufacturing enterprises under blockchain-based supply chain finance. Int J Product Econ. 2022;247: 108469.
43. Wang XL, Wang MJ, Zhou YT. Euclidean distance feature extraction and analysis of license plate characters. Comput Simul. 2014;31(04):184–7.
44. Liu SY, Song W, Ying MX, Sun WY, Wang R. Seismic phase analysis based on condensed hierarchical clustering of waveform eigenvectors. Geophys Geochem Explor. 2020;44(02):339–49.
45. Gao HH, Xu KL, Cao M, Xiao JS, Xu Q, Yin YY. The deep features and attention mechanism-based method to dish healthcare under social IoT systems: an empirical study with a hand-deep local-global net. IEEE Trans Comput Soc Syst. 2022;9(1):336–47.
46. Mei K, Liu XK, Mu C, Qin XQ. Fast defogging algorithm based on adaptive exponentially weighted moving average filtering. Chin J Lasers. 2020;47(01):250–9.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.