Research on Corpus Construction for Hospital Data Security

Authors: Yu Lizhang, Yu Lizhang

Date: 2025-11-06T00:00:00+00:00

Abstract

With the rapid development of smart healthcare, medical data is experiencing explosive growth, and data security has become a core challenge in hospital information system construction. This paper analyzes the current status of hospital data security, the internal management issues and external threats faced, and evaluates existing data security measures. It thoroughly investigates the medical data security corpus as a specialized resource repository, identifying its core characteristics of scalability, diversity, high quality, and security compliance. It elaborates on its pivotal role in supporting security policy formulation and advancing security technology development. The paper proposes a systematic corpus construction workflow encompassing corpus collection, data preprocessing, annotation, and management maintenance, emphasizing that the construction process must strictly adhere to core principles including legality, minimal necessity, security and confidentiality, and clear accountability to ensure the standardization and security of corpus construction. This paper aims to provide theoretical foundations and practical guidance for constructing efficient and compliant hospital data security corpora, employing systematic approaches to address increasingly complex medical data security risks.

Full Text

Preamble

Research on Corpus Construction for Hospital Data Security

Yu Lizhang[1]
(1. Shaoxing Second Hospital Medical Community General Hospital, Shaoxing, 312000)

Abstract

With the rapid development of smart healthcare, medical data has experienced explosive growth, making data security a core challenge in hospital information system construction. This paper analyzes the current status of hospital data security and evaluates existing security measures in the face of internal management issues and external threats. It delves into the specialized resource repository of medical data security corpora, clarifying their core characteristics of scalability, diversity, high quality, and security compliance, and expounds their critical role in supporting security strategy formulation and advancing security technology development. The paper proposes a systematic corpus construction process covering data collection, preprocessing, annotation, and management maintenance, emphasizing that the construction process must strictly adhere to core principles such as legality, minimal necessity, security confidentiality, and clear accountability to ensure the standardization and security of corpus construction. This study aims to provide theoretical foundations and practical guidance for building efficient and compliant hospital data security corpora, systematically addressing increasingly complex medical data security risks.

Keywords: Medical Data Security, Corpus, Smart Healthcare, Blockchain, Artificial Intelligence

1.1 Research Background

Driven by the wave of smart healthcare, the medical industry is undergoing an unprecedented digital transformation. The widespread application of emerging technologies such as big data, artificial intelligence, and the Internet of Things in healthcare has revolutionized traditional medical service systems. Smart healthcare breaks through the temporal and spatial limitations of conventional medicine through information technology, enabling the sharing and integration of medical resources and promoting personalized, precise, and intelligent medical services. However, this transformation has also brought enormous data security challenges. The volume of data generated daily by medical institutions grows exponentially, far exceeding the processing capacity of traditional databases. This data contains large amounts of sensitive personal information and medical records, and any leakage would cause irreversible damage to patient privacy and institutional reputation.

The sensitivity and particularity of medical data make its security protection especially critical. According to China' s Personal Information Protection Law, medical information is classified as sensitive personal information requiring special protection. In 2025, the National Health Commission, the National Administration of Traditional Chinese Medicine, and the National Disease Control and Prevention Administration jointly issued the Notice on Further Strengthening the Management of Electronic Medical Record Information Use in Medical Institutions, which further clarified the primary responsibility of medical institutions in managing electronic medical record information use. The notice requires

strict protection of patient privacy in accordance with laws and regulations, and prohibits the disclosure of patient medical information for non-medical, teaching, or research purposes. These regulatory policies impose higher requirements on data security management in medical institutions.

1.2 Research Significance

The sensitivity and privacy requirements of medical data demand extremely high security protection measures. Medical data involves patients' personal health information, genetic information, etc. Once leaked, it not only infringes on patient privacy but may also cause severe psychological and economic burdens [**?**]. Therefore, building a secure and reliable data corpus is fundamental to ensuring medical data security [**?**]. By collecting, organizing, and annotating medical security data, we can provide rich training resources for security large models, enabling them to more accurately identify and respond to various security threats, thereby effectively protecting medical data security.

The application of security large models in hospitals can significantly enhance information security management levels. As the primary producers and users of medical data, hospitals face multiple security threats from both internal and external sources, such as malware attacks, data breaches, and unauthorized internal operations. By deploying security large models, hospitals can achieve real-time monitoring and intelligent analysis of multi-source data including network traffic, system logs, and user behavior, enabling timely detection of potential security risks and implementation of appropriate protective measures. This effectively reduces the probability of security incidents and ensures the stable operation of hospital information systems.

Security large models can also promote data sharing and collaborative innovation in the healthcare industry. Currently, the "data silo"phenomenon in medical data severely restricts the mining and utilization of data value [**?**]. By building a unified security data corpus, we can break down data barriers and promote data sharing among different medical institutions, providing more comprehensive and accurate data support for medical research, clinical decision support, and public health management [**?**]. Meanwhile, the intelligent analysis capabilities of security large models can provide new ideas and methods for the deep mining and application of medical data, driving innovation and development in the healthcare industry.

From a theoretical perspective, corpus construction helps deepen understanding of hospital data security issues. By building corpora containing rich medical information, researchers can conduct comprehensive analyses of the generation, storage, transmission, and use processes of hospital data, revealing the causes and impact mechanisms of data security risks. Corpus construction also has long-term significance for promoting sustainable development in healthcare. With the application of emerging technologies such as big data, cloud computing, and artificial intelligence, the value of medical data is increasingly prominent. How

to fully mine and utilize this data while ensuring data security has become a new challenge for hospitals.

Building a corpus for medical industry security data and exploring the application of security large models in hospitals can not only effectively enhance the level of medical data security protection but also promote the improvement of hospital information security management levels and facilitate data sharing and collaborative innovation in the healthcare industry. This endeavor has important theoretical and practical significance [**?**]. It not only aligns with national strategic requirements for strengthening medical information security protection but also provides strong technical support for achieving high-quality development in the healthcare industry.

2.1 Internal Management Issues and External Threats in Medical Institutions

Medical institutions have numerous weak links in internal data security management. Incomplete management systems are a common problem. Chaotic permission management and non-standard operations are also prominent internal issues. Insufficient security awareness among medical staff likewise constitutes a security risk.

External data security threats faced by medical institutions are increasingly complex and diverse. Cyberattack methods continue to escalate, with medical institutions becoming prime targets for hackers. Technical vulnerabilities and system defects provide opportunities for external attacks.

2.2 Data Silo Phenomenon

In the era of big data, the medical industry faces unprecedented challenges in data management, with the "data silo"phenomenon being particularly prominent [**?**]. The causes of data silos are multifaceted. Non-uniform technical standards are an important reason for this phenomenon. Different medical institutions have adopted different technical standards and system architectures during informatization construction, making data standardization and interoperability difficult to achieve [**?**]. Legal and regulatory restrictions also contribute significantly to data silos. The sensitivity of medical data requires strict compliance with relevant laws and regulations during sharing and use, which to some extent restricts the free flow of data [**?**].

Conflicts of interest among medical institutions are another important cause of data silos. Different institutions may have conflicting interests during data sharing, such as concerns that data sharing might affect their competitive advantages, making them reluctant to actively share data. The data silo phenomenon is a significant issue facing the medical industry that requires comprehensive measures from technical, legal, cooperative, and security perspectives to effectively resolve and promote healthy development of the industry.

2.3 Privacy Protection Challenges

The application of large language models in clinical medicine provides new possibilities for medical informatization and intelligence but also brings privacy protection challenges [**?**]. Privacy protection is one of the most critical issues in medical data processing, especially when dealing with personal health information whose sensitivity and importance are self-evident. When processing and analyzing medical data, large language models must not only ensure data accuracy and effectiveness but also strictly comply with relevant laws and regulations to ensure patient privacy is not compromised.

Large language model training relies on massive amounts of data, which often contains patients'personal information such as names, addresses, phone numbers, and more sensitive health information including disease diagnoses, treatment plans, and medication usage [**?**]. Without effective privacy protection measures during data collection and use, personal information leakage can easily occur, causing unpredictable harm to patients.

3 Research Methods

This study aims to construct a specialized corpus for hospital data security to support in-depth analysis and research in the field of medical information security. The research methodology is divided into several stages including data collection, data preprocessing, data annotation, and management and maintenance, with rigorous methodologies employed at each stage to ensure the scientific nature and reliability of the corpus.

During the data collection stage, raw data covering the field of hospital data security was obtained through multiple channels. These data sources include but are not limited to internal hospital security reports, academic papers on medical information security, government-issued medical information security policy documents, and relevant news reports. To ensure data diversity and representativeness, particular attention was paid to data security practice cases from hospitals of different regions and scales, as well as data security incidents across different time periods, to comprehensively reflect the current status and challenges of hospital data security.

In the data preprocessing stage, the collected raw data was cleaned and organized. This process included removing duplicate data, correcting erroneous information, and performing preliminary structuring of unstructured data.

The data annotation stage is a critical step in ensuring corpus quality. A combination of manual and automatic annotation was adopted to provide detailed annotations for data in the corpus. For text data, annotation content includes data security incident types, occurrence times, involved medical equipment or systems, and impact scope. For image data, Google Cloud Vision tools were used to automatically generate image annotations, which were then manually added to corresponding positions in text files where images appeared and incor-

porated into subsequent multimodal analysis [**?**].

Through these methods, this study successfully constructed a high-quality hospital data security corpus, providing important data support for research in the field of medical information security.

4.1.1 Definition of Corpus A medical data security corpus is a specialized resource repository that systematically collects, organizes, and annotates various types of data related to medical data security. It contains not only traditional medical data such as electronic medical records, medical images, and gene sequences but also security threat intelligence, attack patterns, anomaly detection samples, and privacy protection strategies closely related to data security. In the context of smart healthcare, the legal definition of patient personal information covers three dimensions: formally emphasizing the identifiable connection between information and specific natural persons; content-wise encompassing physiological, psychological, and health status data generated throughout the entire medical activity process; and in terms of value attributes, such information has sensitive characteristics and is classified as sensitive personal information requiring special protection.

4.1.2 Characteristics of Corpus Medical data security corpora possess scalability and diversity. Medical data volumes are enormous and types are rich, including structured data (such as patient basic information and test results), unstructured data (such as medical images and doctor notes), logs from various security devices, and laws and regulations. Security and compliance are another important feature of medical data security corpora. Corpus construction must comply with relevant laws, regulations, and standards to ensure adequate protection of patient privacy.

4.2.1 Supporting Security Strategy Formulation Medical data security corpora provide data support and decision-making basis for medical institutions to scientifically formulate security strategies. By systematically collecting and analyzing historical security incident data, threat intelligence, and attack patterns, corpora can help security management personnel identify major risks and weak links in medical data security, thereby developing more targeted protection strategies. Corpora also support risk-based security decision-making.

4.2.2 Promoting Security Technology Development Medical data security corpora provide basic resources for security technology research and innovation. High-quality security corpora can accelerate the research and development of new-generation security technologies. Corpora also support the development and application of privacy-enhancing technologies.

5.1 Data Sources and Types

Medical data comes from diverse sources, including but not limited to hospital information systems, electronic medical record systems, medical imaging systems, laboratory information systems, and public health information systems. Data stored in these systems covers patients' personal health information, diagnosis and treatment records, medical images, laboratory test results, and medication usage records. With the popularization of wearable devices and mobile health applications, health data generated by patients in daily life has gradually become an important source of medical data. The collection of this data must follow strict privacy protection and data security standards to ensure the confidentiality and integrity of patient information.

Building high-quality datasets that are mounted to large models in the form of local vector databases can significantly enhance model affinity for specific business scenarios. This approach enables models to better understand and master industry-specific and private knowledge, thereby strengthening their application effectiveness in particular domains. In the process of constructing a medical industry security data corpus, the selection of data sources and types is a crucial step. This data must not only cover a wide range of medical scenarios but also ensure data quality, security, and compliance to support effective training and application of security large models.

In terms of data types, medical data has multimodal characteristics, including structured and unstructured data [**?**]. Structured data typically refers to data that can be stored in tabular form, such as patient personal information, diagnosis codes, and treatment plans. This data is easy to process and analyze and forms the foundation for building security large models. Unstructured data includes medical images, pathology reports, and doctors' clinical notes. Although difficult to process directly, this data contains rich clinical information that is crucial for improving model accuracy and interpretability [**?**].

The construction of medical industry security data corpora requires comprehensive consideration of data sources, types, quality, security, and compliance.

5.2 Construction Principles

Building a hospital security corpus requires adherence to a series of core principles. The principle of legality is paramount: corpus construction must comply with relevant laws and regulations such as the Cybersecurity Law of the People' s Republic of China, the Data Security Law of the People' s Republic of China, and the Personal Information Protection Law of the People' s Republic of China. Particularly when collecting and using patient personal information, legal authorization must be obtained to ensure the legality of data sources.

The principle of minimal necessity is a basic requirement for privacy protection. The corpus should only collect the minimum amount of data directly related to established objectives, avoiding excessive collection. For medical data security

corpora, only data essential for supporting security research and applications should be included, rather than indiscriminately collecting all medical data.

The principle of security and confidentiality is key to ensuring corpus credibility. A strict security protection system must be established to prevent unauthorized access, tampering, or leakage of data during storage, transmission, and use. The Luohu Medical Corpus Center' s adoption of blockchain and privacy computing technologies to enhance data security exemplifies this principle.

The principle of clear responsibility requires clarifying the responsibilities and obligations of all parties involved in corpus construction (medical institutions, technical personnel, management personnel, etc.) and establishing a sound responsibility system. The National Health Commission' s notice requiring medical institutions to assume primary responsibility for the use and management of electronic medical record information within their units also applies to corpus construction.

5.3.1 Data Collection Data collection is the primary step in building a hospital security corpus, requiring systematic acquisition and aggregation of various types of medical data and related security information. Data sources include but are not limited to business systems such as hospital information systems, electronic medical record systems, medical imaging systems, laboratory information systems, and public health information systems, as well as logs from various security devices and virus signature databases. They also include laws and regulations such as the Cybersecurity Law of the People' s Republic of China, the Data Security Law of the People' s Republic of China, the Personal Information Protection Law of the People' s Republic of China, and the Guiding Opinions on Information Security Level Protection Work in the Health Industry.

Regarding the construction of security management datasets, it covers the construction and continuous operation of regulatory documents including laws and regulations, management systems, and emergency response regulations. In the medical industry, data security is a critical link in building and applying security large models [**?**]. With the rapid development of medical large model technology, how to ensure these models operate compliantly within the legal and regulatory framework has become a focus of current research and practice.

These datasets are transformed into vector databases and integrated into security large models, thereby achieving high-quality security management knowledge Q&A functions. Based on management requirements and combined with real-time data inference and statistics, the model can provide compliance supervision-related answers, such as identifying servers that do not meet host security management requirements according to management systems. Additionally, for Q&A on single management knowledge points, users can make inquiries during specific operations, combining local management system knowledge bases or internet knowledge (for example: How should a host infected with computer viruses be handled? What are the regulatory systems

of industry supervisory units regarding internet assets?). These Q&A functions can be integrated into employee instant messaging (IM) systems and made available to internal staff.

Regarding the construction of security asset network datasets, it covers the construction and continuous operation of asset network-related data including asset inventories, network security devices, log systems, and threat intelligence platforms. By importing asset lists and transforming them into vector database content and integrating them into large model Q&A systems, asset sorting and generation of asset topology and access relationships based on large models can be achieved. This includes the investigation and management of special asset types (such as DNS servers, domain controllers, security devices, network devices) and special business assets (such as AI platforms, business interface APIs, big data platforms, OA systems).

Regarding the construction of security policy datasets, it covers the construction and continuous operation of security policy-related data including true/false positive annotations, key prevention and control strategies, and business preference knowledge. It supports the construction of data for personalized alert analysis ideas and methods, monitoring and analysis of key assets, special weak password and vulnerability analysis, emergency response strategies and contact person analysis, business knowledge bases, and personalized queries. Through the continuous operation of these datasets, security large models can provide customers with more precise and efficient security policy support.

During the data collection process, special attention must be paid to data quality and legality. The collected data should be sufficiently representative to reflect the diversity and complexity of medical institution data. At the same time, the legality of data collection must be ensured, with necessary authorization and consent obtained. Particularly for data containing personal sensitive information, strict desensitization processing should be conducted or only anonymized data should be collected.

5.3.2 Data Preprocessing Data preprocessing involves strict quality control of training data, including data cleaning, noise reduction, and anomaly detection, to ensure data accuracy and completeness. Through data processing, the impact of malicious samples on model training can be effectively reduced.

5.3.3 Data Annotation Data annotation is the process of adding labels and annotations to raw data and is the core link in building high-quality corpora. Accurate annotation can significantly enhance the value of corpora, enabling them to effectively support security research and applications. Annotation content for medical data security corpora includes data classification labels (such as data type and sensitivity level), security labels (such as threat type and risk level), and semantic labels (such as medical terminology and clinical concepts).

Data annotation and standardization are key links in building high-quality med-

ical industry corpora and implementing security large models in hospitals. In the medical field, the complexity and sensitivity of data require data annotation and standardization work to have high professionalism and rigor to ensure data accuracy and security, thereby supporting effective training and application of large models.

Data annotation standard formulation: Based on security requirements and the task objectives of security large models, determine the categories for data annotation. Detailed specifications should be made for the definition, characteristics, and annotation methods of each annotation category to ensure consistency and accuracy in annotation. Provide annotation personnel with rich annotation examples, including both correct and incorrect annotations, to help them better understand annotation standards and requirements. At the same time, regularly train annotation personnel to familiarize them with annotation norms and the latest security threat characteristics, improving annotation quality.

Annotation personnel training and management: Provide annotation personnel with training in security knowledge and related technologies to help them understand basic security concepts, common security threat types and characteristics, and the importance of data annotation for security large models. Training content can include basic network security knowledge, working principles of security devices, and analysis methods for security incidents.

Through practical operations and case analysis, train annotation personnel to master the use of annotation tools and annotation techniques, improving annotation efficiency and accuracy. For example, how to use annotation software for data annotation, how to quickly and accurately identify characteristics of security threats, and how to handle ambiguous or uncertain data. Establish a quality assessment mechanism for annotation personnel, regularly conduct spot checks and evaluations of annotation results, reward personnel with high annotation quality, and provide retraining or job adjustments for those with substandard quality. At the same time, strengthen supervision of the annotation process to ensure annotation personnel follow annotation norms and avoid arbitrary or subjective errors.

Data annotation is the process of converting raw data into machine-understandable formats, which is particularly important for medical data. Data annotation and standardization should also consider the dynamic and real-time nature of data. Medical data is highly dynamic, with patients' conditions changing over time. Therefore, corpora need regular updates to reflect the latest medical knowledge and clinical practice. During the data update process, annotation and standardization work must continue to ensure the quality and consistency of new data.

Annotation quality review and validation: Adopt a multi-person cross-review approach to repeatedly check annotated data. Different annotation personnel independently annotate the same batch of data, then compare annotation results. Inconsistent annotations are discussed and re-annotated to ensure accu-

racy and consistency. Invite security domain experts to review and validate annotated data. With their rich experience and professional knowledge, experts can identify potential problems and errors in annotations and provide improvement suggestions. Expert review can serve as the final quality check to ensure the annotation quality of datasets reaches a high level. Preliminarily train and test security large models using a portion of annotated data to verify the quality and effectiveness of annotated data. If the model performs poorly in testing, there may be problems with the annotated data that require further inspection and correction. Some evaluation metrics can be used to measure model performance on validation data, thereby indirectly assessing annotation quality.

5.3.4.1 Data Security and Privacy Protection In the process of building medical industry security data corpora and applying security large models, data security and privacy protection are crucial links [**?**]. Due to its high sensitivity, medical data involves not only personal privacy but may also affect patients' life safety [**?**]. Therefore, ensuring data security and privacy is not only a legal and ethical requirement but also key to improving medical service quality and enhancing patient trust.

In 2011, the Ministry of Health issued the Guiding Opinions on Information Security Level Protection Work in the Health Industry (hereinafter referred to as the Guiding Opinions). Based on the Recommendations of the CPC Central Committee on Formulating the 12th Five-Year Plan for National Economic and Social Development, this document proposed specific implementation measures, forming a regulatory system for information security in the healthcare industry [**?**]. The Guiding Opinions provide specific guidance for data security in the medical industry, requiring medical institutions to strengthen data security protection according to information security level protection requirements [**?**].

Security during the data collection phase is crucial. When building corpora, the legality and compliance of data sources must be ensured. Medical institutions should follow national standards such as the Information Security Technology - Basic Requirements for Information System Security Level Protection (GB/T22239-2008) and the Information Security Technology - Technical Requirements for Security Design of Information System Level Protection (GB/T 25070-2010) to ensure privacy protection during data collection. For example, anonymization techniques should be used to remove or encrypt personal identification information to prevent data from being traceable to individuals after leakage.

Security measures during the data storage phase are equally important. A medical information security storage model based on blockchain technology provides an effective method. Through private key signatures and access control protocols, it ensures data immutability and strict management of access permissions [**?**]. The decentralized nature of blockchain technology makes data storage more secure and reduces the risk of single points of failure. At the same time, encrypted data storage can further enhance data security, as even if data is

illegally accessed, its content cannot be directly read.

During the data processing and analysis phase, technologies such as security alignment, inference guidance, content filtering, penetration testing, and vulnerability scanning are key technical means for evaluating corpus security and can effectively prevent data leakage and misuse. For example, security alignment technology ensures that sensitive information is not leaked during model training; inference guidance technology can filter or blur information that may involve privacy during model output to avoid directly exposing patient information. Content filtering technology can identify and block the spread of harmful information, protecting users from potential threats. Penetration testing and vulnerability scanning should complement each other to build a complete data security assessment system. Penetration testing can discover unknown and complex vulnerabilities, while vulnerability scanning can efficiently detect known and common security issues. Combined use of both can more comprehensively assess the security status of corpora.

Building medical industry security data corpora and applying security large models requires comprehensive security measures from multiple stages including data collection, storage, processing, and application to ensure data security and privacy [**?**]. This not only improves medical service quality but also enhances patient trust in medical services and promotes healthy development of the medical industry.

5.3.4.2 Corpus Update and Dynamic Management In the process of building medical industry security data corpora and applying security large models, data update and dynamic management are key links to ensure system effectiveness and security. The sensitivity and importance of medical data require data management to ensure not only data accuracy and integrity but also data timeliness and security. Therefore, the design and implementation of data update and dynamic management mechanisms are particularly important.

Corpus data updates should follow strict standards and processes. This includes screening, annotating, and integrating new data. In the medical field, data accuracy and integrity are particularly important. Various virus signature databases and knowledge bases require regular maintenance and updates to ensure consistency with the latest regulations, rules, and standards. This means corpus updates involve not only data addition but may also involve correction or deletion of existing data. This process must be completed by personnel with professional knowledge to ensure data quality.

The frequency of data updates is also an important consideration. In the rapidly developing medical field, policy and practice changes can be very frequent. Therefore, corpus update frequency should be determined based on specific application requirements and data change speed. For some critical datasets, such as virus signature databases, updates may be needed monthly or even weekly to ensure timeliness.

At the technical level, the use of automated tools can greatly improve the efficiency and accuracy of data updates. For example, natural language processing can be used to automatically extract and annotate new corpora from medical literature and medical records. These tools can not only reduce errors from manual operations but also handle large amounts of data and speed up updates. However, the use of automated tools also requires caution to ensure their output quality meets standards. Typically, data after automated processing still requires manual review to ensure accuracy and reliability.

Dynamic management mechanisms should focus on data security and privacy protection [**?**]. The sensitivity of medical data requires strict compliance with relevant laws and regulations such as the Cybersecurity Law of the People' s Republic of China and the Personal Information Protection Law of the People' s Republic of China during data management. Dynamic management mechanisms should include security measures such as data encryption, access control, and audit tracking to ensure data security during transmission and storage [**?**]. For example, using Advanced Encryption Standard (AES) to encrypt sensitive data for storage, using Role-Based Access Control (RBAC) mechanisms to restrict data access permissions, and monitoring data access behavior through logging and audit tracking functions to timely detect and handle potential security threats.

Data update and dynamic management mechanisms should support data version control and rollback. The complexity and importance of medical data require systems to record historical versions of data so that when data errors occur or rollback is needed, the system can quickly restore to a historical state. Version control should record the time, content, and operator of each data update and support data version comparison and rollback operations.

Data update and dynamic management mechanisms should have flexible scalability and adaptability. As medical technology develops and medical data continues to grow, data management needs are constantly changing. Therefore, data update and dynamic management mechanisms should have good scalability and adaptability to flexibly adjust and optimize according to actual needs.

Data update and dynamic management are important links in medical industry security data corpus construction and security large model application. By establishing efficient data update mechanisms, strict data security management mechanisms, reliable data version control mechanisms, and flexible system expansion mechanisms, we can ensure the security, timeliness, and reliability of medical data, providing a solid data foundation for the digital transformation and intelligent development of the medical industry.

5.3.5.1 Simulating Potential Security Threats to Test Corpus Attack Resistance In the medical field, especially in hospital environments involving large amounts of sensitive patient information, corpus construction and maintenance require not only attention to data quality and usability but also high

emphasis on data security. With the development of information technology, various new attack methods emerge endlessly, posing serious threats to medical data security. Therefore, simulating potential security threats and testing corpus attack resistance has become an important link in ensuring medical information security.

For corpora, common security threats include but are not limited to data leakage, data tampering, Denial of Service (DoS) attacks, and improper operations by internal personnel. These threats may originate from malicious attacks by external hackers or from unintentional mistakes by internal employees. To effectively address these threats, security factors must be fully considered during the corpus design phase, adopting multi-layered security protection measures such as data encryption, access control, and audit tracking.

Simulating potential security threats is a key step in evaluating corpus security. This can be achieved by constructing virtual attack scenarios, such as simulating SQL injection attacks, XSS cross-site scripting attacks, and DDoS distributed denial-of-service attacks. Through these simulated attacks, we can detect corpus performance when facing real attacks and discover security vulnerabilities in the system. Additionally, automated tools can be used for penetration testing. These tools can simulate multiple attack methods, helping security teams more comprehensively understand system vulnerabilities.

Testing corpus attack resistance is not only technical work but also requires integration with management measures. For example, regular security training should be conducted to improve employee security awareness, ensuring they can properly handle sensitive information and avoid data leakage caused by human errors. At the same time, sound security management systems should be established, such as formulating strict data access permission strategies and implementing the principle of least privilege to ensure only authorized users can access specific data resources.

Continuous security monitoring is also indispensable. By deploying technical means such as Intrusion Detection Systems (IDS) and Intrusion Prevention Systems (IPS), network traffic and system logs can be monitored in real time to timely detect abnormal behavior. Once potential attack activities are detected, emergency plans should be immediately activated to take necessary measures to prevent further attack development and restore normal system operation as soon as possible.

Simulating potential security threats and testing corpus attack resistance is a complex but crucial task. It requires not only advanced technical means but also a complete security management system. Through these measures, corpus security can be significantly improved, protecting medical data from various threats and providing reliable information support for medical services.

5.3.5.2 Evaluating Data Processing Efficiency and Optimizing Corpus Structure and Access Control Strategies As an important tool for linguis-

tic research, corpus construction and application depend not only on large-scale authentic language materials but also involve multiple aspects such as data processing efficiency, corpus structure optimization, and access control strategies. Especially when dealing with sensitive information such as medical data, how to improve corpus usage efficiency while ensuring data security has become an urgent problem to be solved.

Evaluating data processing efficiency is the foundation for corpus optimization. Data processing efficiency directly affects corpus construction speed and the convenience of subsequent research. Efficient preprocessing workflows can significantly reduce time consumption in data cleaning, format conversion, and other links, thereby accelerating corpus update frequency and enabling researchers to obtain the latest language materials faster. To this end, advanced data processing technologies such as Natural Language Processing (NLP) algorithms can be used to achieve automated annotation, deduplication, classification, and other functions, improving data processing accuracy and speed. At the same time, using parallel computing and distributed storage technologies can further enhance processing capabilities under large data volumes, ensuring the real-time and dynamic nature of corpora.

Optimizing corpus structure is key to improving access efficiency. Reasonable corpus structure design can not only simplify the query process but also enhance data retrievability and usability. During the design phase, full consideration should be given to classification standards, indexing mechanisms, and metadata annotation methods. Additionally, using standardized metadata descriptions can increase corpus interoperability and promote cross-disciplinary and cross-domain data sharing.

Strengthening access control strategies is an important measure to ensure data security. In the medical field, personal information protection is particularly important. Therefore, a sound access permission management system must be established to ensure only authorized users can access specific data resources. This includes but is not limited to identity verification, role assignment, and operation logging mechanisms. By implementing fine-grained permission control, unauthorized access and data leakage can be effectively prevented. At the same time, regular security audits and technical updates are indispensable to address constantly changing security threats.

By evaluating data processing efficiency, optimizing corpus structure, and strengthening access control strategies, we can significantly improve corpus usage efficiency while ensuring data security, providing richer and more reliable data support for linguistic research. This process requires not only technological innovation but also institutional guarantees, with the two complementing each other to jointly promote the development of corpus construction.

In the process of building and maintaining hospital data security corpora, evaluating data processing efficiency is a key link to ensure corpus practicality and efficiency. Data processing efficiency concerns not only the speed of data process-

ing but also the quality of data processing, i.e., model performance in security detection and classification tasks. To comprehensively evaluate the performance of different models on the corpus, this paper adopts a series of recognized performance metrics, including Accuracy, Precision, Recall, F1-Score, and AUC value (Area Under the ROC Curve). These metrics reflect model capabilities in classification tasks from different perspectives: Accuracy measures overall classification correctness, Precision focuses on the accuracy of positive class predictions, Recall reflects the model' s ability to find all positive examples, F1-Score is the harmonic mean of Precision and Recall, and AUC value evaluates the model' s comprehensive classification performance under different thresholds.

By comparing and analyzing the performance of different models (such as BERT, LSTM, Random Forest, SVM, and CNN) on these metrics (as shown in Table 3), we can identify which models perform best on specific tasks, providing data support for corpus construction and optimization. For example, the BERT model performs excellently in Accuracy, Precision, Recall, and AUC value but has relatively slow processing speed. The Random Forest model has the fastest processing speed but slightly lower Accuracy. This analysis helps us select appropriate models based on specific needs (such as the trade-off between real-time requirements and accuracy requirements) in practical applications and further optimize corpus structure and access control strategies.

Table 3 shows detailed performance data. Test results indicate that different models demonstrate their own advantageous characteristics in medical data security tasks. The BERT model is most prominent in classification accuracy and AUC value, with Accuracy reaching 92.5% and AUC value as high as 0.968, showing excellent overall classification capability. The LSTM model achieves 89.1% Recall in anomaly detection tasks, demonstrating good threat discovery capability. The Random Forest model performs best in processing speed, reaching 3500 items/second, making it suitable for scenarios requiring fast response.

This paper uses trend charts, scatter plots, ROC curves, and radar charts to display model performance from multiple dimensions, intuitively revealing the advantages and disadvantages of each model.

[Figure 1: see original paper]

Overall Trend Pattern: Observing the performance of all models across various metrics reveals clear patterns. Deep learning models (BERT, CNN, LSTM) overall outperform traditional machine learning models (SVM, Random Forest) in classification performance metrics but are relatively slower in processing speed. This reflects the typical trade-off relationship between model complexity and performance/efficiency.

Metric Correlation: Accuracy, Precision, Recall, and F1-Score show high correlation, with relatively consistent ranking orders across models. AUC value, as a comprehensive performance metric, also maintains good consistency with these classification metrics but provides more detailed performance differentiation.

Speed-Performance Trade-off: Processing speed and classification performance show a clear negative correlation. The fastest Random Forest (3500 items/second) ranks last in classification performance, while the best-performing BERT (Accuracy 0.925) is the slowest (1200 items/second). This trade-off has important guiding significance for model selection in practical applications.

Comprehensive Analysis Conclusions: BERT-Security Classification performs best across all classification metrics, forming a clear high-value trend, making it particularly suitable for security-critical applications with extremely high accuracy requirements. CNN-Text Classification approaches BERT in classification performance but is significantly faster, forming a good balance between performance and efficiency on the trend chart. SVM-Multi-Classification shows balanced performance across all metrics with a smooth trend curve, making it a stable and reliable choice, especially suitable for resource-constrained environments. LSTM-Anomaly Detection performs well in anomaly detection tasks with relatively high processing speed, achieving a good balance between speed and performance on the trend chart. Random Forest leads significantly in processing speed but has relatively lower classification performance, with the trend chart clearly showing its advantage in efficiency. All models have identical F1-Scores and Accuracy, indicating that Precision and Recall maintain good balance across all models. Deep learning models have more obvious advantages in AUC values, indicating better comprehensive performance under different thresholds. The trend chart intuitively shows that no single model is optimal across all metrics, and practical applications require trade-off selection based on specific needs.

Model Characteristics Comparison: BERT has the highest Accuracy (0.925) but the slowest processing speed (1200 items/second), suitable for scenarios with extremely high security requirements. LSTM has good Accuracy (0.883) and faster processing speed (2800 items/second). CNN has the second-highest Accuracy (0.918) with medium processing speed (2100 items/second). Random Forest has the fastest processing speed (3500 items/second) but relatively lower Accuracy (0.857). SVM has moderate Accuracy (0.901) and moderate processing speed (1800 items/second).

Application Scenario Recommendations: For medical data encryption and desensitization, BERT or CNN are recommended to ensure the highest security standards. For real-time monitoring and early warning, Random Forest or LSTM-Anomaly Detection are recommended to guarantee processing efficiency. For balanced scenarios, CNN-Text Classification provides a good balance between speed and accuracy.

Performance Trade-off Analysis: Deep learning models (BERT, CNN, LSTM) typically have higher accuracy but relatively lower processing speed. Traditional machine learning models (Random Forest, SVM) have faster processing speed but slightly lower accuracy. Model performance shows a clear linear negative correlation, with a correlation coefficient of approximately -0.99.

Data Analysis Conclusions: BERT-Security Classification performs best with an AUC value of 0.968, showing excellent classification capability, especially suitable for security scenarios with extremely high accuracy requirements. CNN-Text Classification follows closely with an AUC value of 0.957, providing good processing speed while maintaining high accuracy. SVM-Multi-Classification balances performance and efficiency with an AUC value of 0.945, suitable for resource-constrained environments. LSTM-Anomaly Detection performs well in anomaly detection scenarios with an AUC value of 0.934 and relatively high processing speed. Although Random Forest has a lower AUC value (0.912), it has the fastest processing speed (3500 items/second), suitable for scenarios with high real-time requirements. All models have AUC values exceeding 0.9, indicating excellent performance in classification tasks. There is a clear accuracy-speed trade-off: models with higher accuracy usually have slower processing speed, and vice versa.

Comprehensive Analysis Conclusions: BERT-Security Classification performs best in AUC value (0.968), with all evaluation metrics in leading positions, showing excellent classification capability, but its processing speed is slower (1200 items/second), suitable for application scenarios with extremely high accuracy requirements but not high real-time requirements. CNN-Text Classification has comprehensive performance close to BERT, with an AUC value of 0.957 and significantly faster processing speed (2100 items/second), achieving a good balance between performance and efficiency. SVM-Multi-Classification has balanced metrics across the board, with an AUC value of 0.945 and moderate processing speed (1800 items/second), making it a stable and reliable choice, especially suitable for resource-constrained environments. LSTM-Anomaly Detection performs well in anomaly detection tasks with an AUC value of 0.934 and high processing speed (2800 items/second), suitable for application scenarios requiring certain real-time performance. Although Random Forest has slightly lower classification performance metrics than deep learning models, it has the fastest processing speed (3500 items/second), showing obvious advantages in scenarios with extremely high real-time requirements. All models perform relatively balanced across metrics, without any particularly outstanding or particularly poor metrics. Deep learning models (BERT, CNN, LSTM) overall outperform traditional machine learning models (SVM, Random Forest) in classification performance, but usually have slower processing speed.

In practical applications, trade-off selection should be made between performance and speed based on specific needs and resource constraints. Based on in-depth analysis of these performance metrics, we conducted targeted optimization of corpus structure and access control strategies. For sensitive data access control scenarios requiring high accuracy, the BERT model is prioritized for fine-grained permission judgment. For real-time monitoring and anomaly detection needs, the LSTM model' s good performance in anomaly detection tasks is combined to build dynamic protection systems. When processing large-scale batch data, the efficient processing capability of the Random Forest model is utilized to improve overall system throughput.

By establishing a multi-model collaborative working mechanism, corpus processing efficiency has been significantly improved. Specifically, we restructured the data index structure, clustering and storing data with similar security levels to reduce data retrieval overhead during model inference. At the same time, we designed a hierarchical access control strategy that dynamically selects the most suitable processing model based on data sensitivity and user permissions, maximizing processing efficiency while ensuring security.

This performance metric-driven optimization approach has improved overall corpus processing efficiency by approximately 40% while maintaining high-quality data services, providing more efficient and reliable technical support for hospital data security management.

6.1.1 Basic Principles of Large Models Large models, especially deep learning models, have demonstrated significant application potential in the field of hospital data security in recent years. These models can process and analyze large-scale medical data by simulating the structure and function of the human brain' s neural network. The basic principles of large models involve multiple aspects including data processing, model training, model optimization, and model application, which are discussed in detail below.

Data processing is the foundation of large model application. In the field of hospital data security, data processing includes not only cleaning, standardizing, and formatting raw data but also involves data annotation and feature extraction. During data processing, the diversity and complexity of data, as well as data timeliness and quality, must be considered to ensure the effectiveness and accuracy of model training.

Model training is the core link of large models. Large models are typically based on deep neural networks and learn complex features of data through multi-layer nonlinear transformations. During training, models continuously adjust weights through backpropagation algorithms to minimize the error between predicted output and actual output.

Model optimization is a key step to improve large model performance. Optimization methods include but are not limited to model structure optimization, parameter optimization, and training strategy optimization. Model structure optimization involves adjusting network layers, node numbers, activation functions, etc., to adapt to different task requirements. Parameter optimization prevents overfitting and improves model generalization ability through techniques such as regularization and early stopping. Training strategy optimization includes learning rate adjustment and batch size selection to accelerate training speed and improve training effectiveness.

Model application is the ultimate embodiment of large model value. In the field of medical data security, large models can be applied to security detection and identification scenarios, security analysis scenarios, security decision-making and command scenarios, and other aspects.

The application of large models in the medical health field involves multiple aspects including data processing, model training, model optimization, and model application. Through continuous optimization and innovation, large models are expected to play greater roles in the medical health field, providing higher-quality medical services for patients.

6.1.2 Characteristics of Security Large Models The application of security large models in the medical industry, especially in the field of hospital information security, demonstrates their advantages in handling complex data environments, ensuring data security, and improving medical service efficiency. These characteristics originate not only from the technical architecture of large models but are also closely related to their flexibility and adaptability in practical applications.

Security large models have high flexibility and adaptability. Information security needs in the medical industry constantly change with new security threats emerging. Security large models can continuously update their knowledge bases through continuous learning to address new security challenges. For example, when new viruses or attack methods appear, security large models can quickly identify these threats by analyzing the latest security reports and data and generate corresponding protection strategies. This dynamic adjustment capability enables security large models to maintain efficient security protection capabilities in constantly changing network environments.

Security large models can achieve intelligent security management. In traditional hospital information security systems, security management often relies on manual operations, which is not only inefficient but also prone to omissions. Security large models can achieve automated monitoring and management of hospital information systems by integrating advanced algorithms.

Security large models can provide personalized security services. In the medical industry, different hospitals, departments, and even patients have different information security needs. Security large models can provide customized security solutions for hospitals by analyzing their specific business processes, data types, and security requirements.

Security large models can promote knowledge innovation in the medical industry. In the field of medical information security, knowledge updates are also very rapid. Security large models can provide the latest knowledge support for medical industry information security by continuously learning the latest security technologies and research results.

The application of security large models in the medical industry can not only improve hospital information security levels but also promote knowledge innovation in the medical industry, providing strong support for the efficient and secure operation of medical services.

6.2.1 Application in Detection and Identification Scenarios In the construction of hospital data security corpora, the application in detection and identification scenarios is an important component involving data accuracy and integrity. This paper explores how to support detection and identification scenario applications in hospital data security by building specific corpora.

Focusing on multi-dimensional collaboration between security large models and products, a deep threat detection architecture based on a trinity of rule/intelligence engine, behavior engine, and generative AI large model is constructed. Security large models enhance the detection and identification capabilities of advanced threats, detecting high-bypass unknown threats such as web 0day vulnerability exploitation and obfuscated/bypass attacks, achieving a web application layer detection rate of $ $90% for network traffic with a false positive rate of $ $5%. Simultaneously, reinforcement learning based on manual feedback of attack markings is implemented to precipitate expert experience in analyzing traffic-based attacks and optimize automatic detection models.

By building domain-specific corpora, hospital data security and usability can be improved. Future research should further explore how to combine the latest information technologies, such as artificial intelligence and big data analysis, to optimize corpus construction and application to better serve hospital data security management needs.

6.2.2 Application in Analysis and Assessment Scenarios The application in analysis and assessment scenarios also requires corpora to have powerful data retrieval and analysis capabilities. By building efficient indexing mechanisms, fast retrieval of large-scale corpora can be achieved to meet the needs of different users.

Large models can conduct security operations through natural language interaction, providing security operations personnel with quick retrieval and effective response suggestions from different dimensions, greatly shortening MTTR and improving security operations efficiency. Large models can also achieve automated on-duty analysis of alerts, enabling comprehensive analysis of all alerts and high-precision classification based on built-in knowledge to identify real attacks, business false positives, and suspected attacks, preventing rule omissions. At the same time, explainable analysis conclusions are output through natural language, solving the security operations challenges of insufficient personnel, heavy protection tasks, and numerous security alerts. Based on Retrieval-Augmented Generation (RAG) technology, continuous false positive feedback and annotations from security operations personnel can continuously adjust and optimize the model' s security sensitivity to meet personalized use case requirements.

The application in analysis and assessment scenarios also requires continuous optimization and improvement of corpus construction. As medical data continues to grow and technology advances, corpus construction should maintain dynamic

updates to timely incorporate new data and research results. At the same time, corpus maintenance and management should be strengthened to ensure data integrity and consistency. Through continuous technological innovation and application practice, the ability and level of analysis and assessment can be continuously improved to provide strong support for medical service improvement and development.

6.2.3 Application in Decision-Making and Command Scenarios Utilizing the powerful data analysis capabilities of security large models, various security events over a period of time can be deeply analyzed in terms of quantity, type, impact scope, frequency, and trend changes. By comparing security conditions across different time periods, regions, or business systems, security risk hotspots and potential threats can be identified. Based on historical trend analysis results, security large models can automatically generate detailed security operation summary reports. These reports include not only security knowledge encyclopedias, event statistics, and trend charts but also expert interpretations and recommendations, providing decision-makers with intuitive and comprehensive security situation overviews. Automated summary reports can effectively enhance organizational security awareness, promote communication and collaboration across departments, and jointly drive the continuous optimization and upgrading of the government extranet security system. Through Retrieval-Augmented Generation (RAG) technology, exclusive process systems, business knowledge, and other data can be embedded, enabling the model to achieve personalized knowledge Q&A based on uploaded knowledge, experience, and system documents.

In the field of hospital security management, the application of decision-making and command scenarios is an important component of data-driven decision models, with its core lying in how to efficiently and accurately utilize the large amount of internal data resources accumulated by hospitals to provide scientific and reasonable decision support for management [**?**]. Data security and privacy protection are important aspects that cannot be ignored in the data-driven decision-making process. As a concentration of sensitive information, hospitals must take effective measures to ensure data security and patient privacy [**?**]. This includes but is not limited to establishing strict data access control mechanisms, implementing encrypted data transmission, conducting regular security audits, and establishing sound data leakage emergency response plans. At the same time, hospitals should strengthen employee data security awareness training to ensure all operations comply with laws and regulations and protect patient rights and interests.

The application of data-driven decision-making in hospital management decision-making and command scenarios can not only improve decision scientificity and accuracy but also promote the comprehensive improvement of hospital management levels [**?**]. However, the premise for achieving this goal is to establish a complete data processing process, ensure the rigor and

effectiveness of the decision-making process, and strengthen data security and privacy protection to provide solid guarantees for the sustainable development of hospitals.

6.4.1 Enhancing Medical Intelligence Level In recent years, with the rapid development of information technology, the intelligence level of the medical industry has been significantly improved. The application of intelligent technologies has not only improved the quality and efficiency of medical services but also provided new solutions for optimizing the allocation of medical resources.

In the construction of smart healthcare, data is the core resource. The collection, processing, analysis, and application of medical data are key to achieving medical intelligence. However, the sensitivity and importance of medical data determine that strict security standards must be followed during collection and use. Therefore, building a comprehensive, secure, and efficient medical data corpus has become an important prerequisite for enhancing medical intelligence levels.

Based on practical experience in scenario-based operations, work capabilities in asset sorting, reinforcement and prevention, monitoring and assessment, investigation and disposal, collaborative disposal, intelligence query, and traceability summary are formalized into processes, enabling security personnel to interact with security large models through natural language. Through rapid Q&A, corresponding tools, personnel, and processes can be mobilized to complete assisted driving of security operations, improving security operations efficiency and compressing the disposal speed of single security events to within one minute, enhancing cyberspace security confrontation response and disposal efficiency. Security large models can provide chain-of-thought-based operation on-duty and analysis and disposal capabilities, performing managed on-duty services, supporting 7x24 real-time online automatic analysis, and replacing security operations personnel in asset and vulnerability investigation and management work, achieving autonomous operation on-duty and analysis and disposal.

In the construction of smart hospitals, the application of security large models is not limited to data security and device management but can also be used to improve the intelligence level of medical services.

Building medical security data corpora and applying security large models are important ways to enhance medical intelligence levels. By ensuring data security and reliability, we can provide a solid foundation for the development of smart healthcare, drive innovation and optimization of medical services, and ultimately achieve efficient utilization of medical resources and improve the quality and efficiency of medical services.

6.4.2 Security and Ethical Issues In the process of building medical industry security data corpora and applying security large models in hospitals,

security and ethical issues are important topics that cannot be ignored.

There are numerous security risks in data collection, processing, and use. Due to its high sensitivity, medical data contains not only patients' personal identity information but also detailed health status, disease history, treatment plans, etc. [**?**]. Once this data is leaked, it will not only cause direct physical and psychological harm to patients but may also trigger social discrimination and employment barriers [**?**]. Improper use of data may also lead to infringement of patients' privacy rights, such as unauthorized data sharing and commercial use of data analysis results [**?**]. Therefore, establishing sound data security protection mechanisms, including but not limited to technical means such as data encryption, access permission management, and data desensitization, as well as strict legal and regulatory frameworks, is the foundation for ensuring data security.

In the application of large models in the medical field, the design and training process of the models themselves also faces significant ethical challenges. On the one hand, the training of large language models relies on huge datasets that may contain biases in gender, race, age, etc., causing models to amplify these biases when generating content, thereby affecting the fairness and impartiality of medical decisions.

The interpretability issue of large models is also a major obstacle to their application in the medical field. Medical decision-making requires high transparency and traceability, but large models are often regarded as "black boxes" due to their complex internal structures, making it difficult to intuitively explain their decision basis to medical workers.

Building medical industry security data corpora and applying security large models in hospitals requires not only technological innovation and breakthroughs but also sufficient attention at the ethical level to ensure that technological development can truly benefit patients and promote fairness, impartiality, and security in the medical industry.

7.1.1 Integration of Artificial Intelligence and Big Data As a technology that simulates human intelligence, artificial intelligence has demonstrated enormous potential and value in the medical field through the combination of algorithms, data, and computing power [**?**]. In today' s accelerating digital transformation, the health industry is undergoing unprecedented changes. Especially driven by artificial intelligence technology, multimodal learning has become a key technology for promoting innovation in the medical health field [**?**]. Multimodal learning refers to building large models that can understand and process complex medical information by integrating data from different sources and forms (such as text, images, audio, video, etc.). This technology can not only improve model accuracy and robustness but also promote precision and personalization of medical decision-making.

The deep integration of artificial intelligence and big data will significantly en-

hance the intelligence level of hospital data security corpora. On the one hand, the application of multimodal large model technology in the field of medical data security will enable corpora to have stronger semantic understanding and knowledge reasoning capabilities. These models can process and analyze different types of medical data such as electronic medical records, medical images, and genomics, identifying complex security threat patterns to support proactive protection. Intelligent data governance technology will also be widely applied. Through AI technologies such as natural language processing and knowledge graphs, corpora can achieve automated data classification, annotation, and quality assessment, greatly improving data governance efficiency. A health medical data platform solution mentions that through AI, NLP, and other artificial intelligence technologies, the top-level architecture of hospital data can be reconstructed, and through data governance, high-quality corpora can be provided for clinical decision support. This intelligent data governance model is also applicable to the construction of security corpora.

7.1.2 Application of Blockchain Technology The application of blockchain technology in medical data security corpora will become increasingly deep and extensive. The combination of blockchain and privacy computing will build a more secure and reliable data sharing mechanism. Blockchain provides an immutable audit trail, while privacy computing ensures that original data is not exposed. The combination of the two protects privacy while providing verifiability. Smart contracts will enable automated management of data access and use. Based on preset rules, smart contracts can automatically execute data access authorization, usage billing, compliance checking, and other operations, improving corpus management efficiency. Cross-chain technology will support interconnection between multiple corpora. As medical data security corpora from different institutions and regions increase, cross-chain technology can enable secure interaction and data collaboration between these corpora, forming a larger-scale medical data security collaboration network. This will greatly promote the sharing of medical security knowledge and the improvement of collaborative protection capabilities.

7.2 Application Prospects

In the construction and development process of smart healthcare, the construction of an information security system is particularly important. Given the high sensitivity and importance of medical data, ensuring the security of this data throughout its entire lifecycle has become the cornerstone of smart healthcare development [**?**]. An information security system based on dynamic network security models can effectively address various security threats from both internal and external sources by building a three-dimensional framework with information security organizations at the core and information security strategies, management, and technology as dimensions.

With the arrival of the big data era, legislative protection of personal health

medical information has become an indispensable part of smart healthcare development. The protection of personal information in personal information law should be a dynamic and continuously improvable process. By introducing a scenario- and risk-oriented dynamic protection model, privacy risks can be effectively controlled throughout the entire data processing process.

The development of smart healthcare also relies on the deep mining and utilization of data [**?**]. By building medical industry security data corpora and combining them with the application of security large models, efficient processing and analysis of medical data can be achieved, providing strong data support for clinical decision-making [**?**].

The future development of smart healthcare depends not only on technological innovation and service model optimization but also on establishing sound information security systems and effective data protection mechanisms. Through continuous exploration and improvement in these areas, smart healthcare is expected to achieve wider applications in the future and make greater contributions to human health [**?**].

8.1 Research Summary

This paper systematically studied the theoretical foundations, practical methods, and future directions of hospital data security corpus construction. By analyzing the current status of hospital data security, it revealed the internal and external challenges facing medical data security: internal management issues include incomplete systems, chaotic permission management, and insufficient personnel awareness; external threats mainly come from network attacks, technical vulnerabilities, and advances in quantum computing. Although existing data security measures have been implemented at both technical and management levels, systematic improvement is still needed.

The study examined the basic concepts and roles of corpora in hospital data security, clarifying that medical data security corpora, as specialized resource repositories, have characteristics of scalability, diversity, high quality, and security compliance. Corpora play increasingly important roles in the field of medical data security by supporting security strategy formulation and promoting security technology development.

The paper deeply explored the construction methods of hospital security corpora, proposing a systematic process including data collection, preprocessing, annotation, and management and maintenance. It emphasized that core principles such as legality, minimal necessity, security confidentiality, and clear accountability must be followed during construction to ensure standardized and orderly corpus construction.

8.2 Future Outlook

Looking ahead, hospital data security corpus construction will continue to develop along two dimensions: technological innovation and management optimization. At the technical level, the deep integration of artificial intelligence and big data will enhance the intelligence level of corpora, and the wide application of blockchain technology will enhance the security and credibility of corpora.

At the management level, regulatory compliance will become more dynamic and refined, and standardized data classification and grading systems will provide clear guidance for corpus management. Personnel training and awareness enhancement will receive more attention, and a data security culture in medical institutions will be built through the cultivation of compound talents and comprehensive security education.

Overall, hospital data security corpora will become key infrastructure for ensuring medical data security in the context of smart healthcare. By building a comprehensive protection framework of "legal regulation - technical guarantee - collaborative governance" and forming a multi-party participatory medical data governance ecosystem, hospital data security corpora can not only effectively address current data security challenges but also lay a security foundation for future innovative applications of medical data, ultimately achieving the win-win goal of protecting patient privacy and maximizing the value of medical data.

[1] Jin Meng, Sun Kexin, Hu Yonghua. Prospects for the Development of Medical Informatics in the Big Data Era[J]. Modern Preventive Medicine, 2016, 43(20):3831-3836.

[2] Chen Xiaohong, Liu Liu, Yuan Yige, et al. Research on Medical Large Model Technology and Application Development[J]. Chinese Engineering Science, 2024, 26(06):77-88.

[3] Ye Shaofang, Liu Chanjuan. Ethical Approaches to Digital Healthcare in the Intelligent Era—Reflections Based on "Moral Materialization" Theory[J]. Zhejiang Social Sciences, 2023(09):114-120+160. 10.14167/j.zjss.2023.09.006.

[4] Chen Shaomin, Chen Aimin, Liang Liping. Research on Constraints and Governance of Medical Big Data Sharing[J]. Health Economics Research, 2021, 38(09):18-20+24. 10.14055/j.cnki.33-1056/f.2021.09.004.

[5] Chen Jianfeng. Discussion on the Applicability of Large Language Models in Clinical Medicine[J]. Medicine and Philosophy, 2023, 44(21):1-6.

[6] Li Xintong, Ma Sufen, Zhang Fengcong, et al. Research Progress and Application Prospects of Large Language Models in Traditional Chinese Medicine[J]. Journal, 2024, 40(12):1393-1403. 10.14148/j.issn.1672-0482.2024.1393.

[7] Paul Baker, Luke Collins, Cheng Xiao, et al. Development and Analysis of Multimodal Corpora Based on Google Cloud Vision Automatic Image Annota-

tion Technology[J]. Foreign Language Teaching Theory and Practice, 2024(06):3-19.

[8] Hu Yaolin, Yu Donglei, Wang Jian. Development of Health and Medical Big Data Under the Background of "Healthy China"[J]. Social Scientist, 2022(03):79-87.

[9] Ma Lin, Bao Chenlu, Li Qing, et al. Design and Application of Tumor Clinical Big Data Management System[J]. Chinese Engineering Science, 2022, 24(06):127-136.

[10] Bi Dexu, Chang Liping. Research on Patient Privacy Protection Mechanisms in Medical Data Analysis Based on Artificial Intelligence[J]. Chinese Medical Ethics, 2025, 38(09):1184-1190.

[11] Liu Hui, Cong Yali. Preliminary Exploration of Ethical Issues in Clinical Medical Big Data[J]. Medicine and Philosophy (A), 2016, 37(10):32-36.

[12] Li Zhongmin, Wang Sihui, Chen Xianlai, et al. Analysis of the Construction of Laws and Regulations for Medical Data Security Governance in China[J]. Library, 2022(03):70-76.

[13] Wang Hui, Zhou Mingming. Medical Information Security Storage Model Based on Blockchain[J]. Computer Science, 2019, 46(12):174-179.

[14] Tong Feng, Zhang Xiaohong, Liu Jinhua. Legislative Protection of Personal Health and Medical Information in the Big Data Era[J]. Information and Documentation Services, 2020, 41(03):105-112.

[15] Zhou Xue, Wang Siwen, Li Xuemei, et al. Analysis of the Era Background, Necessity and Key Measures of Hospital Data Asset Management in China[J]. Chinese Hospital Management, 2025, 45(10):39-44.

[16] Liu Ziang, Huang Yuanyuan, Ma Jiali, et al. Design and Implementation of Medical Data Abuse Monitoring Platform Based on Blockchain[J]. Information Network Security, 2021, 21(05):58-66.

[17] Ye Qing, Liu Xun, Zhou Xiaomei, et al. Discussion on Problems and Countermeasures in the Application of Health and Medical Big Data[J]. Chinese Hospital Management, 2022, 42(01):83-85.

[18] Sun Haishuang, Yang Xiaoyan, Liu Min, et al. Advances in the Application of Artificial Intelligence in the Evaluation of Interstitial Lung Disease[J]. Chinese Journal of Medical Imaging, 2022, 30(05):509-513.

[19] Song Yuanming. Exploration of "Artificial Intelligence + Medicine" New Medical Talents Training—Taking Practices of Some Universities as Examples[J]. Chinese University Technology, 2020(08):65-68. 10.16209/j.cnki.cust.2020.08.015.

[20] Liu Yanfei, Wang Zhen. Research on New Business Forms of Health Service Industry Under "Internet Plus" Conditions[J]. Reform and Strategy, 2016, 32(11):151-154. 10.16331/j.cnki.issn1002-736x.2016.11.035.