Medical Big Data Analysis in Hospital Information System

RESEARCH-ARTICLE

Jing-Song Li*, Yi-Fan Zhang and Yu Tian
Show details

Abstract

The rapidly increasing medical data generated from hospital information system (HIS) signifies the era of Big Data in the healthcare domain. These data hold great value to the workflow management, patient care and treatment, scientific research, and education in the healthcare industry. However, the complex, distributed, and highly interdisciplinary nature of medical data has underscored the limitations of traditional data analysis capabilities of data accessing, storage, processing, analyzing, distributing, and sharing. New and efficient technologies are becoming necessary to obtain the wealth of information and knowledge underlying medical Big Data. This chapter discusses medical Big Data analysis in HIS, including an introduction to the fundamental concepts, related platforms and technologies of medical Big Data processing, and advanced Big Data processing technologies.

Keywords: medical Big Data analysis, hospital information system, cloud computing, data mining, Semantic Web technologies

1. Introduction

With the deepening of hospital information construction, the medical data generated from hospital information system (HIS) have been growing at an unprecedentedly rapid rate, which signifies the era of Big Data in the healthcare domain. These data hold great value to the workflow management, patient care and treatment, scientific research, and education in the healthcare industry. As a domain-specific form of Big Data, medical Big Data include features of volume, variety, velocity, validity, veracity, value, and volatility, commonly dubbed as the seven Vs of Big Data [1]. These characteristics of healthcare data, if exploited timely and appropriately, can bring enormous benefits in the form of cost savings, improved healthcare quality, and better productivity.

However, the complex, distributed, and highly interdisciplinary nature of medical data has underscored the limitations of traditional data analysis capabilities of data accessing, storage, processing, analysing, distributing, and sharing. New and efficient technologies, such as cloud computing, data mining, and Semantic Web technologies, are becoming necessary to obtain, utilize, and share the wealth of information and knowledge underlying these medical Big Data.

This chapter discusses medical Big Data analysis in HIS, including an introduction to the fundamental concepts, related platforms and technologies of medical Big Data processing (Section 2), and advanced Big Data processing technologies (Section 3, Section 4, and Section 5). In order to help readers understand more intuitively and intensively, two case studies are given to demonstrate the method and application of Big Data processing technologies (**Section 6**), including one for medical cloud platform construction for medical Big Data processing and one for semantic framework development to provide clinical decision support based on medical Big Data.

2. Medical Big Data in HIS

In the field of medical and health care, due to the diversity of the medical records, the heterogeneity of healthcare information systems and the widespread application of HIS, the capacity of medical data is constantly growing. Major data resources include: (1) life sciences data, (2) clinical data, (3) administrative data, and (4) social network data. These data resources are invaluable for disease prediction, management and control, medical research, and medical informatization construction.

Currently, there are two directions for designing Big Data processing systems, i.e., centralized computation and distributed computation. Centralized computation relies on mainframes, which are very expensive to implement. Besides, there still exists a bottleneck for scalable data processing using a single computer system; distributed computation relies on clusters of cheap commercial computers. Due to the scalability of cluster scale, the data processing ability of distributed computing systems is also scalable. Currently, Hadoop, Spark, and Storm are the most commonly used distributed Big Data processing platforms, which are all open source and free of charge.

Hadoop [2] is the core project of Apache foundation now; its development until now has already gone through many versions. Due to its open-source character, Hadoop becomes the de facto international standard for distributed computing system, and its technical ecosystem becomes larger and larger and more and more perfect, which covers all aspects of Big Data processing. The most fundamental Hadoop platform comes from the three technical articles from Google, including three parts, first the MapReduce distributed computing framework [3], second, the distributed file system (Hadoop distributed file system, HDFS) based on Google File System (GFS) [4], and third, the HBase data storage system based on Big Table [5].

Spark [6], another open-source project of the Apache foundation developed by a lab of the University of California, Berkeley, is another important distributed computing system. Spark achieves architecture improvement on the basis of Hadoop. The most essential difference between Hadoop and Spark is that Hadoop uses hard disk for saving original data, intermediate results, and final results, while Spark uses memory directly for saving these data. Thus, the computing speed of Spark could be 100 times than Hadoop in theory. However, since memory data will be missing after power failure, Spark is not suitable for processing data with long-term storage demand.

Storm [7], a free and open-source real-time distributed computing system, developed by BackType team of Twitter, is an incubated project of the Apache foundation. Storm offers real-time computation for implementing Big Data stream processing on the basis of Hadoop. Different from the above two processing platforms, Storm itself does not have the function of collecting and saving data; it uses the Internet to receive and process stream data online directly and post back analysis results directly through the network online.

Up to now, Hadoop, Spark, and Storm are the most popular and significant distributed cloud computing technologies in Big Data field. All the three systems have their own advantage for processing different types of Big Data; both Hadoop and Spark are off-line, but Hadoop is more complex, while Spark owns higher processing speed. Storm is online and available for real-time tasks. In medical industry, the data are more and have different application scenarios. We can build specific medical Big Data processing platform and develop and deploy related Big Data applications according to characters of the three different platforms while processing different types of medical Big Data with different demands.

A complete data processing workflow includes data acquisition, storage and management, analysis, and application. The technologies of each data processing step are as follows:

Big Data acquisition, as the basic step of Big Data process, aims to collect a large amount of data both in size and type by a variety of ways. To confirm data timeliness and reliability, implementing distributed platform-based high-speed and high-reliable data fetching or acquisition (extract) collection technologies are required to realize the high-speed data integration technology for data parsing, transforming and loading. In addition, data security technology is developed to ensure data consistency and security.

Big Data storage and management technology need to solve both physical and logical level issues. At the physical level, it is necessary to build reliable distributed file system, such as the HDFS, to provide highly available, fault-tolerant, configurable, efficient, and low-cost Big Data storage technology. At the logical level, it is essential to develop Big Data modelling technology to provide distributed non-relational data management and processing ability and heterogeneous data integration and organization ability.

Big Data analysis, as the core of the Big Data processing part, aims to mine the values hidden in the data. Big Data analysis follows three principles, namely processing all the data, not the random data; focusing on the mixture, not the accuracy; getting the association relationship, not the causal relationship. These principles are different from traditional data processing in data analysis requirements, direction, and technical requirements. With huge amounts of data, simply relying on a single server computing capacity does not satisfy the timeliness requirement of Big Data processing parallel processing technology. For example, MapReduce can improve the data processing speed as well as make the system facilitate high extensibility and high availability.

Big Data analysis result interpretation and presentation to users are the ultimate goal of data processing. The traditional way of data visualization, such as bar chart, histogram, scatter plot, etc., cannot meet the complexity of Big Data analysis results. Therefore, Big Data visualization technology, such as three-dimensional scatter plot, network, stream-graph, and multi-dimensional heat map, has been introduced to this field for more powerfully and visually explaining the Big Data analysis results.

3. Cloud computing and medical Big Data analysis

3.1. OVERVIEW OF CLOUD COMPUTING

According to the national institute of standards and technology (NIST), cloud computing is a model for enabling ubiquitous, convenient, and on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

Cloud computing has five essential characteristics [8]:

- On-demand service: Users do not need human interaction service provider, such as a server to automatically obtain time, network storage, and other computing resources according to their needs.

- Broad network access: Users can end on any heterogeneous access to resources through the network according to standard mechanisms, such as smart phones, tablet PCs, notebooks, workstations, and thin terminals.

- Pooling resource: All computing resources (computing, networking, storage, and application resources) are 'pooled' and fully dynamically reallocated based on user needs. Different physical and virtual resources are in possession for a plurality of service users. Based on this, high level of abstraction concept, even if the user has no concept of actual physical resources or control, can also be obtained as usual computing services.

- Rapid elasticity: All computing resources can quickly and flexibly configure publishing, to provide users with an unlimited supply capacity. For users, they can ask for computing resources acquired automatically increase or decrease with distribution according to their needs.

- Managed services: Cloud computing providers need to realize the measurement and control of resources and services in order to achieve the optimal allocation of resources.

According to different resource categories, the cloud services are divided into three service models, i.e., Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS).

- SaaS: It is a new software application and delivery model. Mode applications running on a cloud infrastructure that it will be application software and services delivered over the network to the user. Applications can access through a variety of end, and the user does not manage or control the underlying software required to run their own cloud infrastructure and software maintenance.

- PaaS: It is a kind of brand new software hosting service mode, users can interface with providers and own applications hosted on the cloud infrastructure.

- IaaS: It is a new infrastructure outsourcing mode, the user can obtain basic computing resources (CPU, memory, network, etc.) according to their needs. For users, it can be deployed on the service, operation, and control of the operating system and associated application software without the need to care or realize the underlying cloud infrastructure.

To meet the different needs of users, according to the cloud infrastructure deployment pattern difference, there are basically four deployment models, namely private cloud, public cloud, community cloud, and hybrid cloud, under different requirements for the deployment of the cloud computing infrastructure.

- Private cloud: Cloud platform is designed specifically for a particular unit of service and provides the most direct and effective control of data security and quality of service. In this mode, the unit needs to invest, construct, manage, and maintain the entire cloud infrastructure, platform, and software and owns risk.

- Public cloud: Cloud service providers provide free or low-cost computing, storage, and application services. The core attributes are to a shared resource service via the Internet such as Baidu cloud and Amazon Web Service.

- Community cloud: Multiple units share using the same cloud infrastructure for they have common goals or needs. Interest, costs, and risks are assumed jointly.

- Hybrid cloud: The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public).

3.2. TECHNOLOGIES OF CLOUD COMPUTING

Cloud computing is an emerging computing model, and its development depends on its own unique technology with a series of other traditional technique supports:

- Rapid deployment. Since the birth of data centre, rapid deployment is an important functional requirement. Data centre administrators and users have been in the pursuit of faster, more efficient, and more flexible deployment scheme. Cloud computing environment for rapid deployment requirements is even higher. First of all, in cloud environment, resources and application not only change in large range but also in high dynamics. The required services for users mainly adopt the on-demand deployment method. Secondly, different levels of cloud computing environment service deployment pattern are different. In addition, the deployment process supported by various forms of software system and the system structure is different; therefore the deployment tool should be able to adapt to the change of the object being deployed.

- Resource dispatching: In certain circumstances, according to certain regulations regarding the use of the resources, resource dispatching can adjust resources between different resource users. These resource users correspond to different computing tasks, and each computing tasks in the operating system corresponds to one or more processes. The emergence of virtual machine makes all computing tasks encapsulated within a virtual machine. The core technology of the virtual machine is the hypervisor. It builds an abstraction layer between the virtual machine and the underlying hardware; operating system calls to hardware interception down and provides the operating system virtual resources such as memory and CPU. At present, The Vmware ESX and Citrix XenServer can run directly on the hardware. Due to the isolation of virtual machine, it is feasible to use the virtual machine live migration technology to complete the migration of computing tasks.

- Massive data processing: With a platform of Internet, cloud computing will be more widely involved in large-scale data processing tasks. Due to the frequent operations of massive data processing, many researchers are working in support of mass data processing programming model. The world's most popular mass data processing programming model is MapReduce designed by Google. MapReduce programming model divides a task into many more granular subtasks, and these subtasks can schedule between free processing nodes making acute nodes process more tasks, which avoids slow processing speed of nodes to extend the task completion time.

- Massive message communication: A core concept of cloud computing is the resources, and software functions are released in the form of services, and it is often needed to communicate the message collaboration between different services. Therefore, reliable, safe, and high-performance communication infrastructure is vital for the success of cloud computing. Asynchronous message communication mechanism can make the internal components in each level of cloud computing and different layers decoupling and ensure high availability of cloud computing services. At present, the cloud computing environment of large-scale data communication technology is still in the stage of development.

- Massive distributed storage: Distributed storage needs storage resources to be abstract representations and unified management and be able to guarantee the safety of data read and write operations, the reliability, performance, etc. Distributed file system allows the user to access the remote server's file system like a visit to a local file system, and users can take the data stored in multiple remote servers. Mostly, distributed file system has redundant backup mechanism and the fault-tolerant mechanism to ensure the correctness of the data reading and writing. Based on distributed file system and according to the characteristics of cloud storage, cloud storage service makes the corresponding configuration and improvement.

3.3. APPLICATION OF CLOUD COMPUTING IN MEDICAL DATA ANALYSIS

With the continuous development of medical industry, expanding the scale of medical data and the increasing value, the concept of medical Big Data has become the target of many experts and scholars. In the face of the sheer scale of medical Big Data, the traditional storage architecture cannot meet the needs, and the emergence of cloud computing provides a perfect solution for the medical treatment of large data storage and call.

According to different functions, medical cloud platform is divided into five parts: cloud storage data acquisition layer, data storage layer, data mining layer, enterprise database, and application layer. Every part can form an independent child cloud. Data mining layer and application layer share using data storage layer. Medical cloud deployment is shown in **Figure 1**. The figure also illustrates the medical cloud data flow direction.

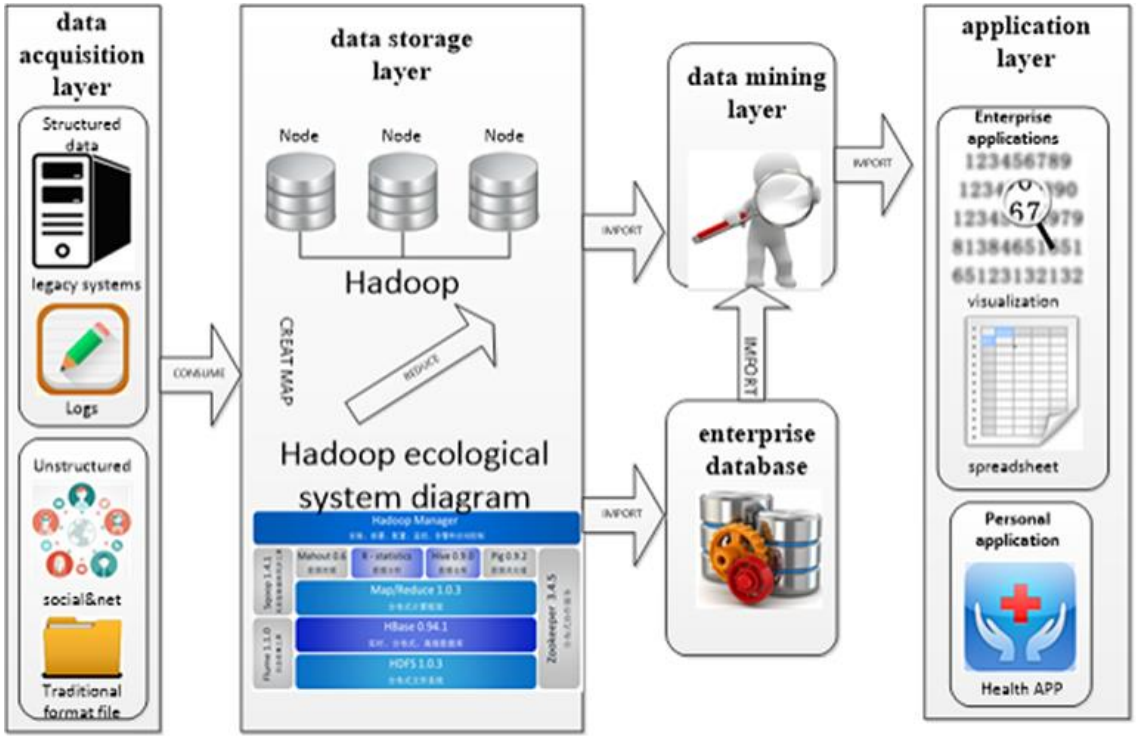

FIGURE 1.

Medical cloud deployment.

All the parts of the medical cloud platform are specific as follows:

- Data acquisition layer: The storage format of medical large data is diverse, including the structured and unstructured or semi-structured data. So data acquisition layer needs to collect data in a variety of formats. Also, medical cloud platform and various medical systems are needed for docking and reading data from the corresponding interface. Due to the current social software and network rapid development, combining medical and social networking is the trend of the future. So it is essential to collect these data. Finally, data acquisition layer will adopt sets of different formats of data processing, in order to focus on storage.

- Data storage layer: The data storage layer stores all data of the medical cloud platform resources. Cloud storage layer data will adopt platform model for architecture and merge the data collected from data acquisition layer and block for storage.

- Data mining layer: Data mining is the most important part of medical cloud platform which complete the data mining and analysis work through the computer cluster architecture. Using the corresponding data mining algorithms, data mining layer finds knowledge from the data in data storage layer and enterprise database and store the result in data storage layer. Data mining layer can also affect application layer using its digging rules and knowledge via methods of visualization.

- Enterprise database. Medical institutions require not only convenient, large capacity of cloud storage but also high real-time and high confidentiality to local storage of data. These would require the enterprise database. Enterprise

database needs interaction with data cloud storage layer and the data mining layer in data, and it will give the data to the application layer for display.

- Application layer: The application layer is mainly geared to the needs of users and displays data either original or derived through data mining.

4. Data mining and medical Big Data analysis

4.1. OVERVIEW OF DATA MINING

Cross Industry Standard Process for Data mining (CRISP-DM) is a general-purpose methodology which is industry independent, technology neutral, and the most referenced and used in practice DM methodology.

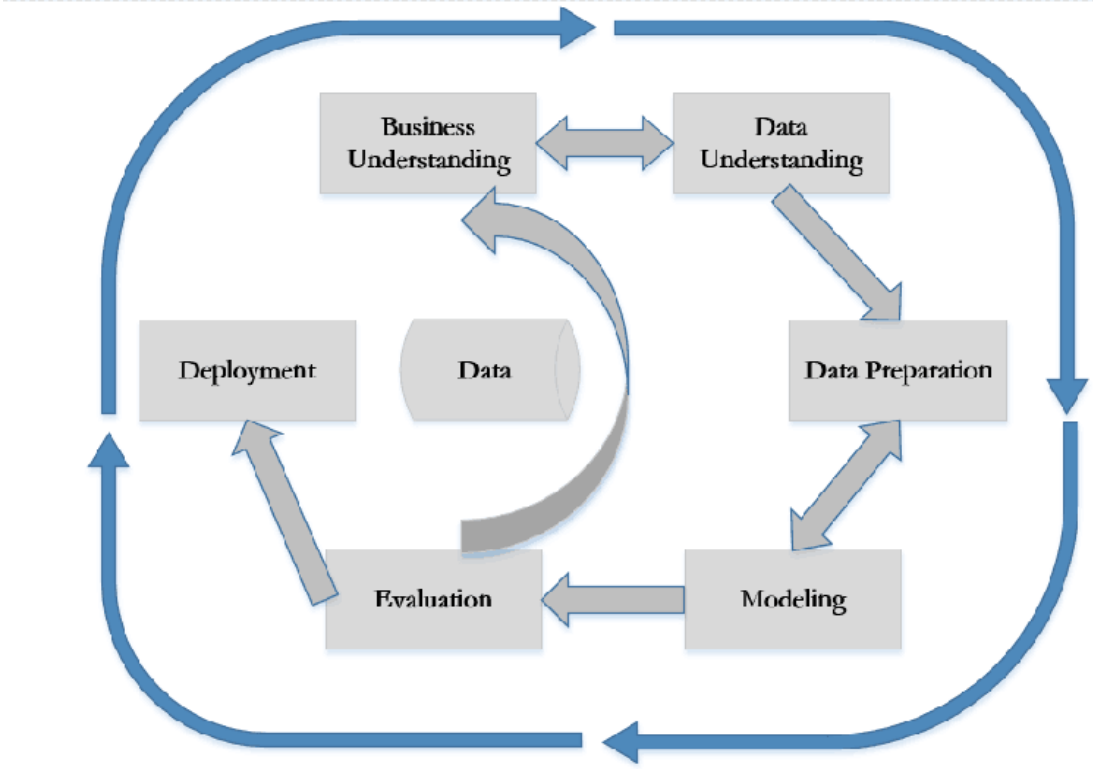

FIGURE 2.

Phases of the original CRISP-DM reference model.

As shown in **Figure 2**, CRISP-DM proposes an iterative process flow, with non-strictly defined loops between phases and overall iterative cyclical nature of DM project itself. The outcome of each phase determines which phase has to be performed next. The six phases of CRISP-DM are as follows: business understanding, data understanding, data preparation, modelling, evaluation, and deployment.

There are a few known attempts to provide a specialized DM methodology or process model for applications in the medical domain. Spečkauskienė and Lukoševičius [9] proposed a generic workflow of handling medical DM applications. However, the authors do not cover some important aspects of practical DM application, such as data understanding, data preparation, mining non-structured data, and deployment of the modelling results.

Catley et al. [10] proposed a CRISP-DM extension for mining temporal medical data of multidimensional streaming data of intensive care unit (ICU) equipment. The results of the work will benefit the researchers of ICU temporal data but not directly applicable for other medical data types or DM application goals.

Olegas Niaksu et al. [11] proposed a novel methodology, called CRISP-MED-DM, based on the CRISP-DM reference model and aimed to resolve the challenges of medical domain such as variety of data formats and representations, heterogeneous data, patient data privacy, and clinical data quality and completeness.

4.2. TECHNOLOGIES OF DATA MINING

There are five approaches for data mining tasks: classification, regression, clustering, association, and hybrid. Classification refers to supervised methods that determine target class value of unseen data. The process of classification is shown in **Figure 3**. In classification, the data are divided into training and test sets used for learning and validation, respectively. We have described most popular algorithms in medical data mining in **Table 1**. These algorithms are the most used in literatures and are also popular. Performance evaluation of classifiers can be measured by hold-out, random sub-sampling, cross-validation, and bootstrap. Among these, cross-validation is the most common.

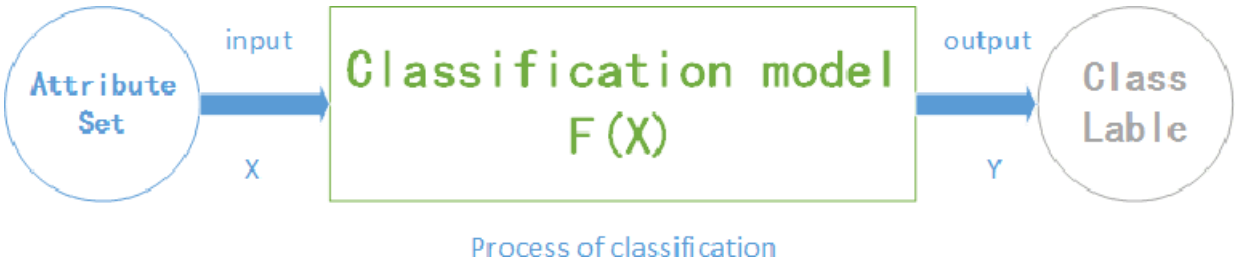

FIGURE 3.

Process of classification.

Regression analysis is a statistical technique that estimates and predicts relations between variables. Instances of regression algorithms are simple linear, multiple linear, fuzzy, and logistic. In data mining, regression is used to predict unseen data based on continuous training data. In this approach, the behaviour of dependent variable *y* is explored by independent variables *x*.

Algorithm	Advantage	Disadvantage	Characteristic
DT	Non-parametric, interpretable, resistant to noise and replication	Separation line parallel to axis *x*, *y*, sensitive to the inconsistent data	Eager approach, greedy, recursive, partitioning, stable
ANN	Diagonal separation line, popular in the other fields, ability to complex relation, resistant to replication	Black box, parametric, sensitive to the noise and missing value, increase time by increase hidden layers	Eager approach, multi-layer network with at least one hidden layer

Algorithm	Advantage	Disadvantage	Characteristic
Rule based	Interpretable, resistant to noise and imbalance data	Separation line parallel to axis *x, y*	Eager approach, produce if…then rules, partitioning
SVM	Diagonal separation line, appropriate for high-dimensional data and little training data	Black box, parametric	Eager approach, mathematics based, unstable, optimization, global minimum
NB	Resistant to noise, missing value, irrelevant features	Accuracy degraded by correlated attribute, required to determine initial probability	Eager approach, statistics based, nondeterministic
KNN	Simple, flexible, arbitrary decision boundaries	Sensitive to noise and replication, parametric	Lazy approach, instance based, required similarity measurement, prediction based on local data

TABLE 1.

Most popular classification algorithms in medical data mining.

Algorithm	Advantage	Disadvantage	Characteristic
K-means	Simple, fast, popular	Parametric, susceptible initial value, inappropriate for data different in size and density, different results in each run, sensitive to noise	Optimization problem, rototype based, partitioning problem, centre based
Hierarchical	Non-parametric, less susceptible to initial value	Time and space complexity, sensitive to noise	Graph based, prototype based, bottom-up
DBSCAN	Resistant to noise, handle arbitrary density and size	Time and space complexity	Density-based, non-complete, partitioning problem
Fuzzy c-means	Same as K-means	Same as K-means	Same as K-means, determining membership of each object to the clusters

TABLE 2.

Most popular data clustering methods.

Data clustering consists of grouping and collecting a set of objects into similar classes. In data clustering process, objects in the same cluster are similar to each other, while objects in different clusters are dissimilar. Data clustering can be seen as grouping or compression problem. Most popular data clustering methods are described in **Table 2**.

Association rule mining is a method for exploring sequential data to discover relationships between large transactional data. The result of this analysis is in the form of association rules or frequent items. In **Table 3**, most popular association algorithms are shown. Performance evaluation of discovered rules was done considering various criteria such as support and confidence.

Algorithm	**Advantage**	**Disadvantage**	**Characteristic**
Apriori	Popular, simple	Time and I/O complexity, reviewing entire database at each stage, searching in all variables	Using prior knowledge, iterative approach
DIC	Decrease I/O complexity	Sensitive to data homogeneity	Dynamic, retrieving lost patterns by moving forward, investigating the specified distance of transactions
DHP	Reducing the number of candidate patterns	Relation between runtime and database size, collision problem in the hash table	Using hash table
Eclat	Decreasing I/O complexity, exploring large length patterns, and discovering all sequential objects	Space complexity, inappropriate for large data	Bottom-up approach, uses lattice-theoretic
D-CLUB	Removing the empty bits, reduce time and space complexity, self-adaptive	–	Appropriate for parallel process and distributed database, dynamic, differential optimization

TABLE 3.

Most popular association rule methods.

Among the five data mining approaches, classification is known as the most important [12]. Interpretability of model is the key factor to select the best algorithm for extracting knowledge. It is important for the expert to understand extracted knowledge. Therefore, decision tree is the most popular method in medical data mining. SVM (Support Vector Machine) and artificial neural network are proved efficient but less popular compared with decision tree, due to the incomprehensibility.

4.3. APPLICATION OF DATA MINING IN MEDICAL BIG DATA ANALYSIS

The electronic medical record (EMR) system has been widely used around the world and has stored lots of data till today. With the data mining technologies, we can, in turn, use the data to improve the EMR system's performance, reduce medication errors, avoid adverse drug events, forecast patient outcomes, improve clinical documentation accuracy and completeness, increase clinician adherence to clinical guidelines, and contain costs and medical researches. However, the

highest functional level of the electronic health record (EHR) is process automation and clinical decision support (CDS), which are expected to enhance patient health and healthcare.

4.3.1. DATA MINING FOR BETTER SYSTEM USER EXPERIENCE

Tao et al. developed a closed-loop control scheme of electronic medical record (EMR) based on a business intelligence (BI) system to enhance the performance of hospital information system (HIS), which provides a new idea to improve the interaction design of EMR. The ranking of drugs in EMR for certain doctor is optimized and personalized based on his/her real-time pharmacy ranking. This illustrates the important applications of a BI system to automatically control the EMR. In addition, the applicability of drug ranking is verified. The system workflow is displayed in **Figure 4**.

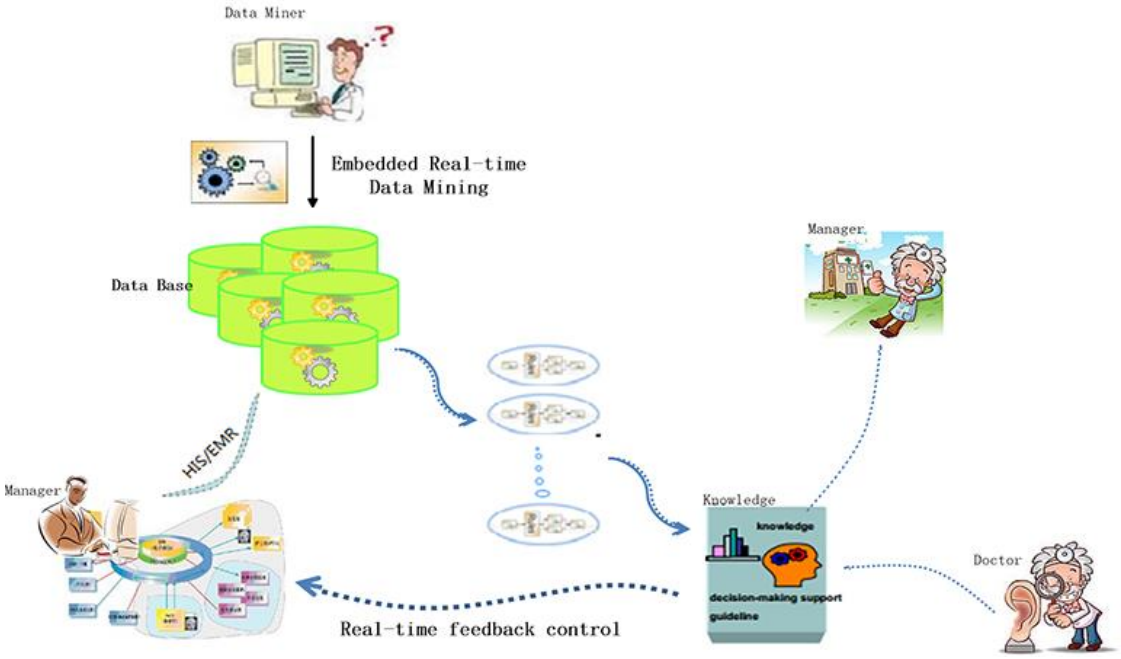

FIGURE 4.

Closed-loop HIS.

Using this EMR system, the ranking of drugs in the EMR is optimized with the real-time ranking of the doctor's pharmacies. With automated drug order in EMR, it realizes a personalized function for doctors, making doctors more convenient to make prescriptions compared to an irregular drug order. In addition, doctors can make orders faster with the help of personalized EMR.

4.3.2. DATA MINING FOR CLINICAL DECISION SUPPORT

Michael J. Donovan et al. [13] developed a predictive model for prostate cancer progression after radical prostatectomy. They collected 971 patients treated with radical prostatectomy at Memorial Sloan-Kettering Cancer Centre (MSKCC) between 1985 and 2003 for localized and locally advanced prostate cancer and for whom tissue samples were available. Although the patient number is relatively small, the dimension is high that they included clinicopathologic, morphometric,

molecular data, and outcome information to implement a systemic pathology approach. The complex relationships between predictors and outcomes were modelled by support vector regression (SVR) for censored data (SVRc), which is a machine learning way rather than the conventional statistical way, to take advantage of the ability of SVR to handle high dimensional data. The SVRc algorithm [14] can be summarized to minimize the following function:

$$\min 12\|W\|2+\sum i=1n(Ci\xi i+C*i\xi*i)\min 12\|W\|2+\sum i=1n(Ci\xi i+Ci*\xi i*)$$

given the constraints:

$$yi-(W\cdot\Phi(xi)+b)\leq\varepsilon i+\xi iyi-(W\cdot\Phi(xi)+b)\leq\varepsilon i+\xi i$$

$$(W\cdot\Phi(xi)+b)-yi\leq\varepsilon*i+\xi*i(W\cdot\Phi(xi)+b)-yi\leq\varepsilon i*+\xi i*$$

$$\xi(*)i\geq 0,i=1\ldots n\xi i(*)\geq 0,i=1\ldots n$$

The model performance was validated by a testing data set, and it was proved to be a highly accurate tool for predicting clinical failure within 5 years after prostatectomy using the integration of clinicopathologic variables with imaging and biomarker data.

5. Semantic Web technologies and medical Big Data analysis

5.1. OVERVIEW OF SEMANTIC WEB TECHNOLOGIES

First put forward by Tim Berners-Lee, the inventor of the World Wide Web and director of the World Wide Web Consortium (W3C), the Semantic web refers to ‘an extension of the current Web in which information is given a well-defined meaning, better enabling computers and people to work in cooperation’ [15]. According to W3C’s vision, the core mission of Semantic Web technologies is to convert the current Web, dominated by unstructured and semi-structured documents into a meaningful ‘Web of data’. The ultimate goal of the Web of data is to enable computers to do more useful work and to develop systems that can support trusted interactions over the network. To support this vision, the W3C has developed a set of standards and tools to enable human readable and computer interpretable representation of the concepts, terms, and relationships within a given knowledge domain, which can be illustrated by the Semantic Web Stack. As shown in **Figure 5**, it is a layered specification of increasingly expressive languages for metadata, where each layer exploits and uses capabilities of the layers below.

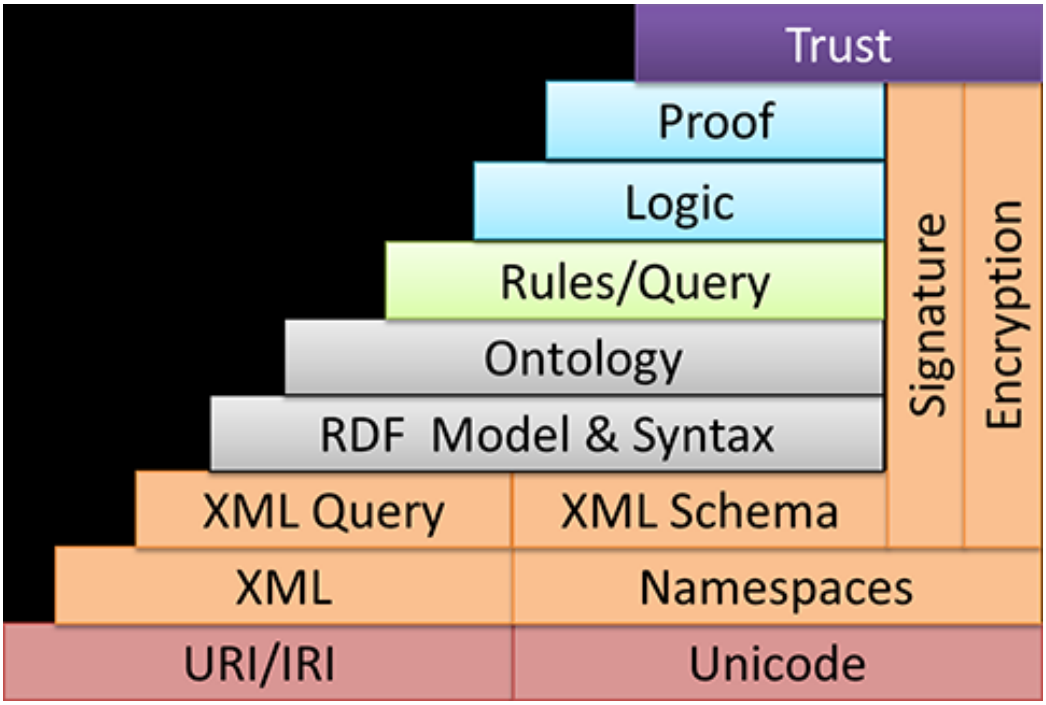

FIGURE 5.

Semantic web stack.

All layers of the stack need to be implemented to achieve full visions of the Semantic Web. The functions and relationships of each layer can be summarized as follows:

1. Hypertext Web technologies: The well-known hypertext web technologies constitute the basic layer of the Semantic Web.

    - Internationalized resource identifier (IRI), the generalized form of the uniform resource identifier (URI), is used to uniquely identify resources on the Semantic Web with Unicode, which serves to uniformly represent and manipulate text in many languages

    - Extendable mark-up language (XML) is a mark-up language that enables the creation of documents composed of structured data. XML namespaces are used for providing uniquely named elements and attributes in an XML document so that the ambiguity among more sources can be resolved to connect data together. XML schema is a description of a type of XML document, typically expressed in terms of constraints on the structure and content of documents of that type, above and beyond the basic syntactical constraints imposed by XML itself. XML query is to provide flexible query facilities to extract data from XML files.

2. Standardized Semantic Web technologies: Middle layers contain technologies standardized by W3C to enable building Semantic Web applications.

    - Resource description framework (RDF) is a framework for creating statements about Semantic Web resources in a form of ‘subject-predicate-object’ triples. A collection of RDF statements intrinsically represents a labelled,

directed multi-graph. As such, an RDF-based data model is more suited for lightweight, flexible, and efficient knowledge representation than relational models. RDF Schema (RDFS) is intended to structure RDF resources by providing basic vocabulary for RDF.

- Ontology is at the core of the Semantic Web Stack. It is originally defined as 'a formal, explicit specification of a shared conceptualization' [16]. By formally defining terms, relations, and constraints of commonly agreed concepts in a particular domain, ontology facilitates knowledge sharing and reuse in a declarative and computational formalism. Combined with rules and query languages, the static knowledge in the ontology can be dynamically utilized for semantic interoperation between systems.
- Logic consists of rules that enable advanced ontology-based inferences. These rules extended the expressivity of ontology with formal rule representation languages.

3. Unrealized Semantic Web technologies: Some technologies are proposed to realize a 'safer' Semantic Web, yet most of which have not come into a standard.

- Encryption is used to verify the reliability of data sources supporting the Semantic Web, typically using digital signature of RDF statements.

- Proof has been conceived to allow the explanation of given answers generated by automated agents. This will require the translation of Semantic Web reasoning mechanisms into some unifying proof representation language.

- Trust is supported by verifying that the premises come from trusted source and by relying on formal logic during deriving new information.

5.2. SEMANTIC WEB MODELLING LANGUAGES AND APPLICATION FRAMEWORK

The OWL Web Ontology Language (OWL) is a W3C recommended mark-up language for representing ontologies [17]. Compared with XML, RDF, and RDFS, OWL has more facilities for expressing semantics and thus goes beyond these languages in its ability to represent machine interpretable content on the Web. OWL is built upon the description logic (DL), which is a family of formal knowledge representation languages used in artificial intelligence to describe and reason about the relevant concepts of an application domain. Major constructs of OWL include individuals, classes, properties, and operations. The W3C-endorsed OWL specification includes three variants of OWL, with different levels of expressiveness. These are OWL Lite, OWL DL, and OWL Full, ordered by increasing expressiveness. Each of these sublanguages is a syntactic extension of its simpler predecessor. They are designed for use by different communities of implementers and users with varying requirements for knowledge representation.

SWRL, the Semantic Web Rule Language, is a W3C recommended encoding language for representing logic rules in the Semantic Web. It extends the expressivity of OWL ontologies with the Unary/Binary Datalog RuleML sublanguages of

the rule markup language. SWRL rules are represented as 'antecedent→consequent', indicating a derivation relationship from the antecedent conditions to the consequent conditions. Both the antecedent and consequent consist of zero or more atoms, written as '$a_1 \wedge a_2 \ldots \wedge a_n$'. Atoms can be of the form C(x) or P(x,y) where C is an OWL description, P is an OWL property, and x,y are either variables, OWL individuals, or OWL data values. Variables are prefixed with a question mark (e.g., ?x). Besides these basic atoms, SWRL provides modular, extensible, and reusable built-in atoms (identified using the http://www.w3.org/2003/11/swrlb namespace) as the flexible and robust infrastructure for specialized logical operations, such as swrlb:equal, swrlb:lessThan, and swrlb:greaterThanOrEqual for numeric comparisons; swrlb:add, swrlb:subtract, and swrlb:multiply for math operations; and swrlb:stringConcat, swrlb:uppercase, and swrlb:replace for string operations. A complete specification of SWRL built-in atoms can be found in [18].

Apache Jena (or Jena in short) is a free and open-source Java framework for building Semantic Web and linked data applications [19]. The framework is composed of different APIs (Application Programming Interface, API) interacting together to process RDF data. Providing various APIs for the development of inference engines and storage models, Jena is widely used in the development of systems or tools related with Web ontology management.

Jena has the following main features:

1. RDF API: Interacting with the core API, users can create and read resource description framework (RDF) graphs. The API can be used to serialize triples using popular formats such as RDF/XML and Turtle.

2. ARQ (SPARQL): It's a SPARQL 1.1 compliant engine which can be used to query RDF data. ARQ supports remote-federated queries and free text search.

3. TDB: It has a native high performance triple store and can be used to persist data. TDB supports the full range of Jena APIs.

4. Fuseki: It can be used to expose the triples as a SPARQL end-point accessible over HTTP. Fuseki provides REST-style interaction with RDF data.

5. Ontology API. It can be used to work with models, RDFS, and the Web Ontology Language (OWL) to add extra semantics to RDF data.

6. Inference API: It can be used to reason over the data to expand and check the content of the triple store. Users can use it to configure their own inference rules or use the built-in OWL and RDFS reasoners.

The interaction between the different APIs is shown in **Figure 6**.

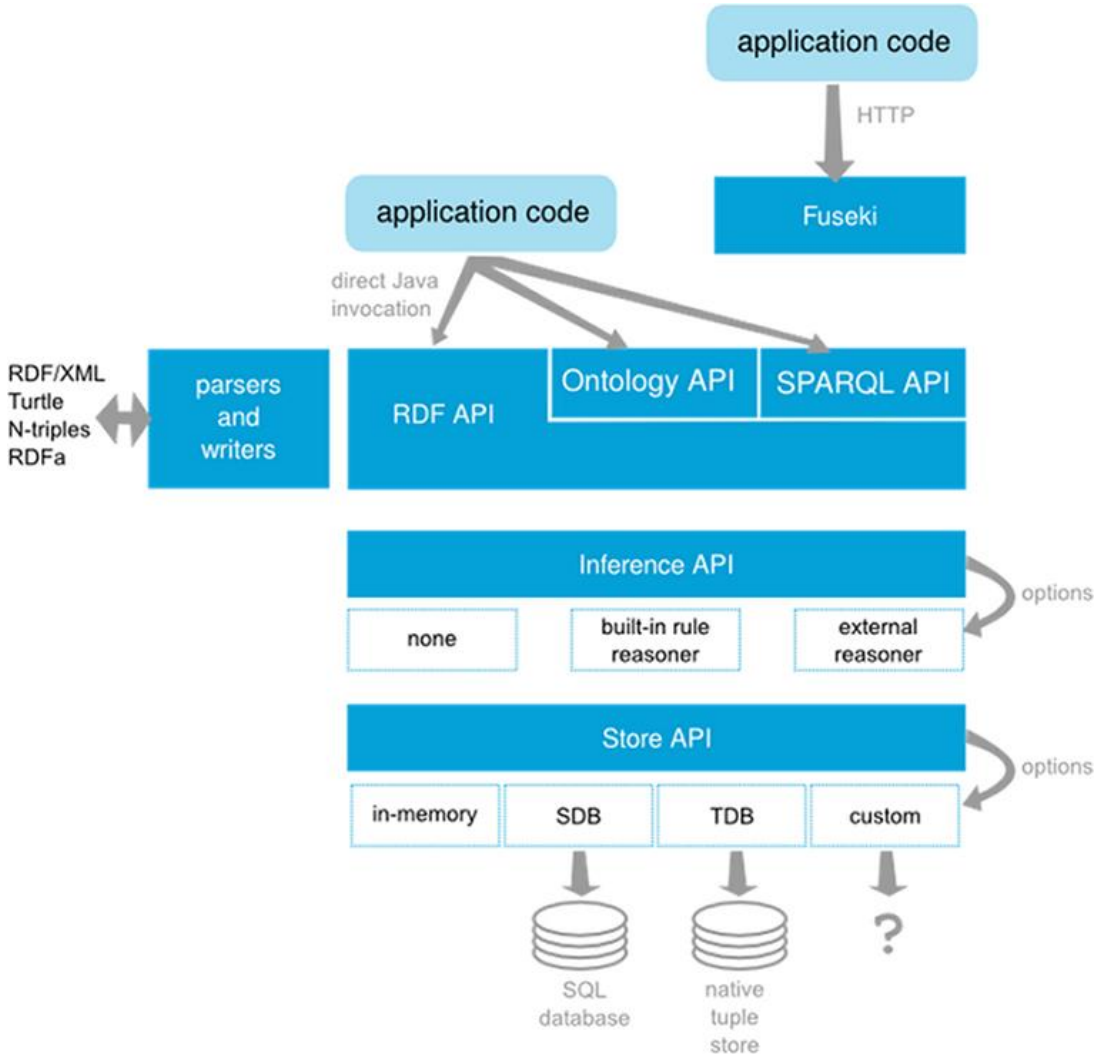

FIGURE 6.

Interaction between the different APIs of Jena.

5.3. THE APPLICATIONS OF SEMANTIC TECHNOLOGY IN THE ANALYSIS OF MEDICAL BIG DATA

The volume, velocity, and variety of medical data, which is being generated exponentially from biomedical research and electronic patient records, require special techniques and technologies [20]. Semantic Web technologies are meant to deal with these issues.

The Semantic Web is a collaborative movement, which promoted standard for the annotation and integration of data. Its aim is to convert the current web, dominated by unstructured and semi-structured documents, into a web of data, by encouraging the inclusion of semantic content in data accessible through the Internet.

The development of ontology on the basis of Semantic Web standards can be seen as a promising approach for a semantic-based integration of medical information. Many resources have ontology support, due to its consistency and expressivity. Important ontologies include UMLS [21], GO [22], UniProt [23], and so on.

The following diagram in **Figure 7** is an example showing the application of ontology in the big picture of Big Data analysis [20].

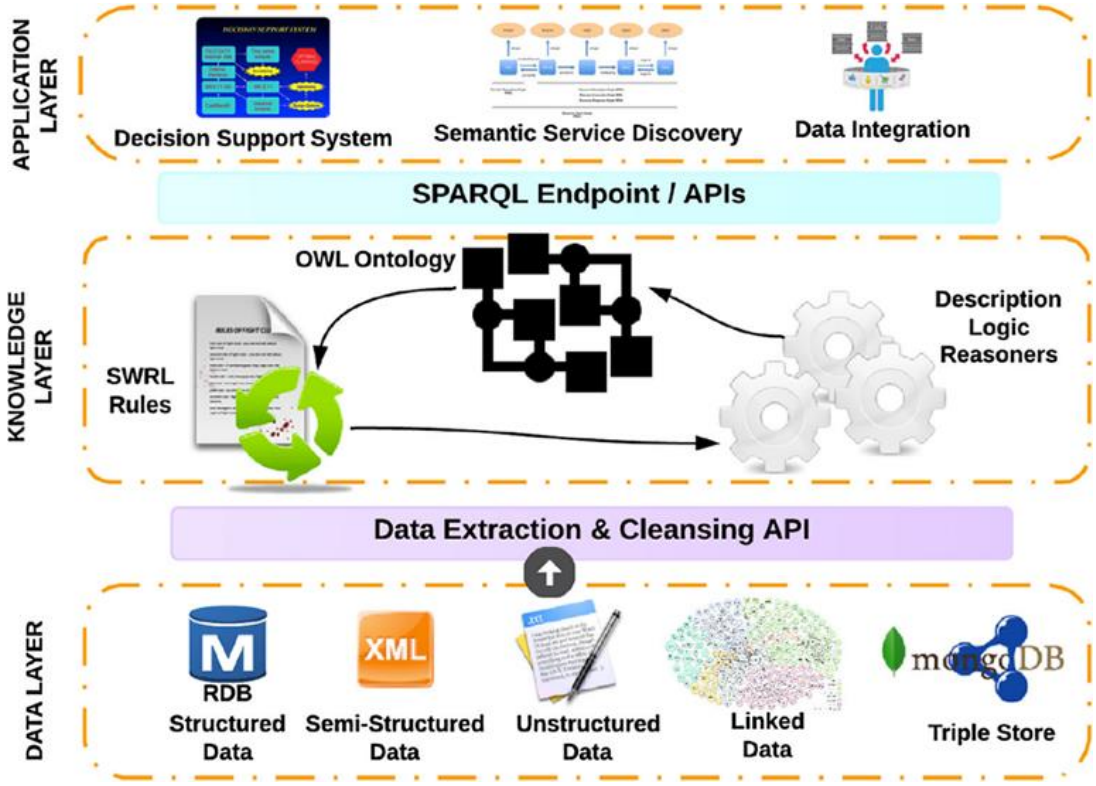

FIGURE 7.

Ontology and rules in the big picture of Big Data analysis.

The picture includes three layers: the data layer, knowledge layer, and the application layer. The data layer consists of a wide variety of heterogeneous and complex data including structured, semi-structured, and unstructured. In the knowledge layer, ontology can be used to access Big Data, which can be processed and analysed by the ontology, rules, and reasoners to derive inferences and obtain new knowledge from it. Then in the application layer, there are several applications that can use the new knowledge such as decision support, semantic service discovery, and data integration.

6. Two case studies of medical Big Data analysis in HIS

6.1. MEDICAL CLOUD PLATFORM CONSTRUCTION FOR MEDICAL BIG DATA PROCESSING

The medical cloud platform for Big Data processing is mainly divided into three levels wherein the first level achieves a hospital private cloud, which serves as the basis of the three-tier application model. It's the IaaS service model that achieves the infrastructure of a medical cloud and also reflects the core concept of 'maximization of resource utilization' in cloud computing. The second level achieves the medical community cloud, which is an upgrade based on the first level and achieves a medical cloud service. It is software-as-a-service (SaaS) service model that reflects the core concept of 'on-demand services' in cloud computing. The third level achieves the applications of medical Big Data. It builds a medical Big Data processing system based on a distributed computing platform named Hadoop.

6.1.1. HOSPITAL PRIVATE CLOUD BASED ON VIRTUALIZATION TECHNOLOGY

The overall architecture of hospital private cloud is shown in **Figure 8**. It is based on the concept of 'pool', and five standard IT resource pools (virtual computing pool, virtual storage pool, virtual network pool, virtual desktop pool, and virtual security pool) are built by highly integrating and fully making use of hospital information resources using virtualization, loading balancing, and high-availability technology. Besides, the dynamic data centre based on cloud computing technology and hospital information cloud service platform consisting of five business function clouds (production cloud, testing cloud, desktop cloud, security cloud, and disaster backup cloud) are also built in the hospital private cloud. All of the above realize unified deployment of systems, assignment on demand of resources, and security sharing of data in the platform, causing the improvement of overall utilization of IT resource and the full use of the performance of information systems, which comprehensively solve the problems existing in the traditional hospital IT structure.

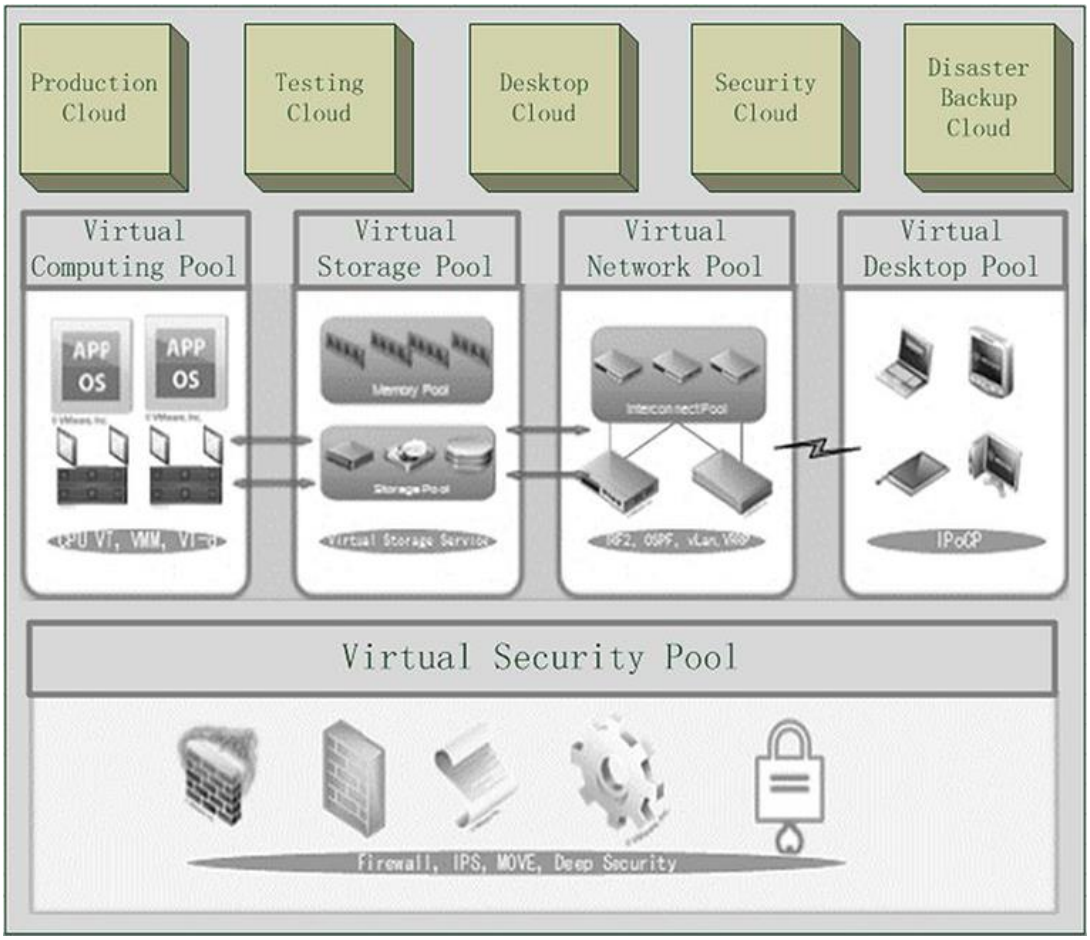

FIGURE 8.

The overall architecture of hospital private cloud.

In five virtual IT resource pools, virtual computing pool realizes the abstract of physical hardware resources by multiple types of virtualization technologies, making the computing resources to be assigned, dispatched, and managed; the role of virtual storage pool is mainly for storage integration; virtual network pool uses network virtualization technology to solve the problems of the interaction of data from the clinical data centre in the medical information system, the real-time backup

of medical, the transfer of virtual machine of medical information system, and other problems with large network flows; virtual desktop pool provides desktop system containing various packaged hospital information system application software; and virtual security pool divides the physical firewall into several independent logical firewall with different defence and security rules by virtualization division of the firewall device, making it easier to manage the firewall device and improve the utilization of the firewall device.

In the five business function clouds, production cloud is designed to maintain the hospital daily medical business under normal circumstances; testing cloud is designed for debugging the hospital's newly developed business systems; desktop cloud is designed to be used to provide virtual desktop delivery containing hospital information system applications; disaster backup cloud is designed for the backup of production cloud and providing the continuity of medical business under abnormal circumstances; and security cloud is designed for providing security services and user authority management.

6.1.2. MEDICAL CLOUD SERVICES BASED ON MEDICAL COMMUNITY CLOUD

Medical cloud services can provide access services everywhere in any time, regardless of the system's installation and implementation details of these services; secondly, medical cloud services can remain online forever. For the occasional unexpected problems, maintenance staff of medical cloud background can find and solve the problem at the first time, ensuring high availability and reliability of medical cloud services, providing normal medical information services; moreover, medical cloud service also supports a very large user base. By 'multi-tenancy' mode, a medical cloud platform provides tenancies of medical cloud services to multiple grassroots medical institutions. The platform can withstand the pressure of the mass of medical information system applications and data access, supporting large user base accessing the medical cloud services.

Community cloud is one of the four deployment models of cloud computing which means that cloud computing infrastructure and services are designed to be provided to certain organizations whose participants have issues of common concern. It can be owned and managed by one or more organizations and only provide relevant cloud services among the organizations. The medical community cloud is an specialization of community cloud in medicine, whose structure is shown in**Figure 9**. Its purpose is to provide ubiquitous medical information systems and services. It is an upgrade on the basis of hospital private cloud that provides major medical institutions' medical information systems as services to grassroots medical institutions by using high-speed private network. The maintenance and management of cloud infrastructure and the deployment, maintenance, and design of medical information systems are all completed by major medical institutions at the background of cloud platform, making grassroots medical institutions invest in hospital informatization at zero cost.

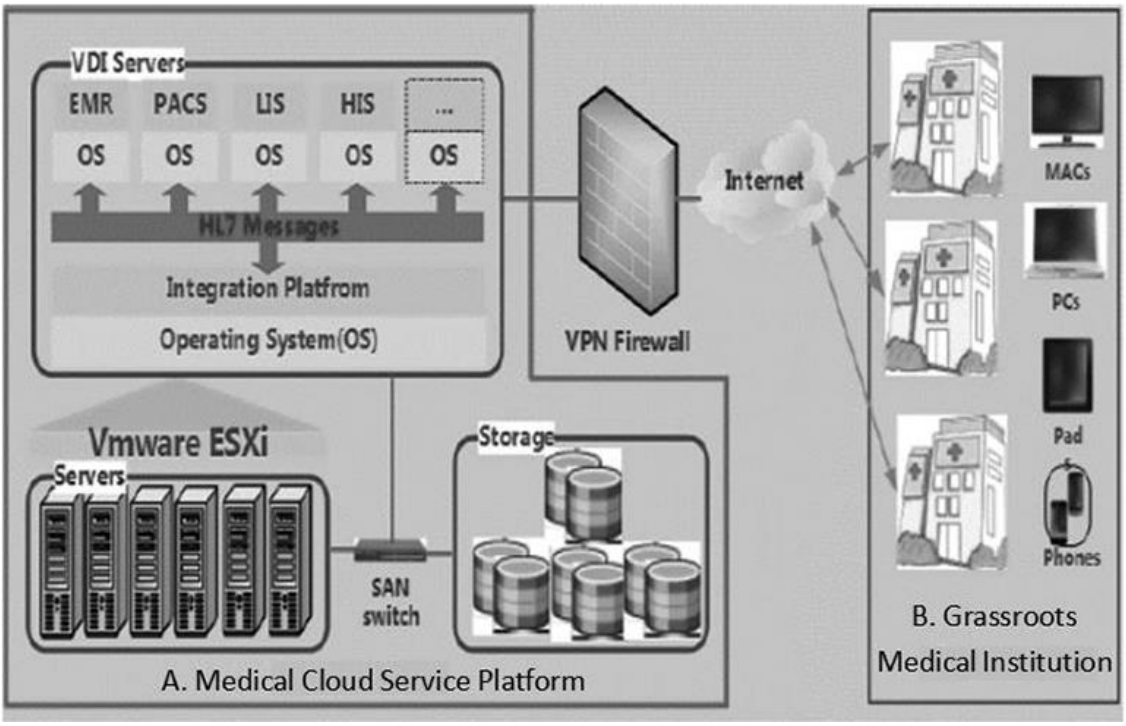

FIGURE 9.

The structure of medical community cloud.

6.1.3. MEDICAL BIG DATA SYSTEMS BASED ON DISTRIBUTED COMPUTING TECHNOLOGY

An overall architecture of a medical Big Data processing system based on Hadoop is shown in **Figure 10**. The system consists of three components: (1) Big Data collection module, (2) Big Data storage management module, and (3) Big Data analysis module. The three modules respectively correspond to three processes used in solving medical Big Data processing problems: Big Data collection, Big Data storage and management, and Big Data analysis. The Big Data collection module firstly develops Extract-Transform-Load (ETL) module based on Sqoop in order to transfer structured data from relational databases to Hadoop platform and then develops transmission function of semi-structured data and unstructured data based on Hadoop Common; the Big Data storage and management module firstly realizes the physical storage of Big Data base on HDFS and then achieves logical management and high-speed access of Big Data based on Hive; and the Big Data analysis module develops Big Data recommendation engine based on Mahout. In the application of Big Data, the system provides relatively reliable personalized recommendation to the users of medical information system by its recommendation system module and a distributed collaborative filtering algorithm which reveals the collective wisdom of the medical Big Data, in order to improve the daily work efficiency. Meanwhile, to solve the limitation that Hadoop cannot achieve ad hoc query and interactive system design, the HL7 interface between recommendation system and hospital information system is developed. The interface can transfer the results into standard HL7 message in real time, realizing real-time interaction with the hospital information system. Hospital information systems, three modules of medical Big Data processing system, recommendation systems, and its HL7 interface constitute a medical Big Data closed loop of ‘generation -collection-storage-analysis–feedback’.

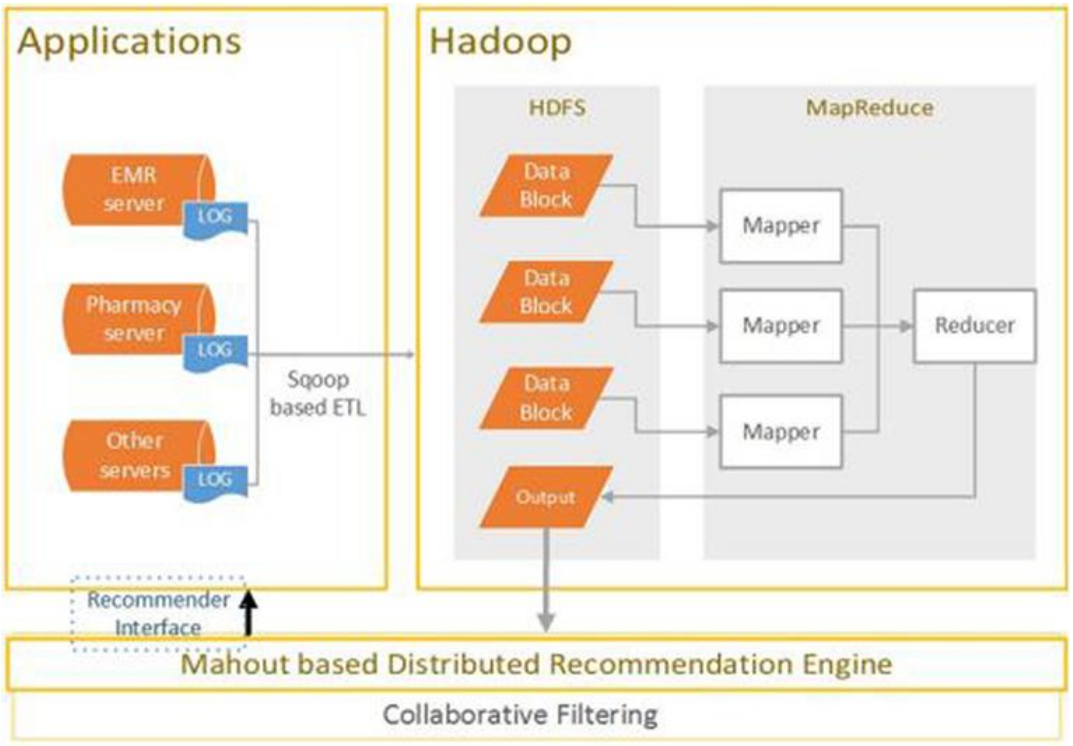

FIGURE 10.

Overall architecture of the medical Big Data processing system.

Because the three modules of the architecture are designed in the environment based on Hadoop-distributed computing, and Hadoop cluster can provide MapReduce (distributed computing) and HDFS (distributed storage), both of which are needed for the system, the system can process medical Big Data in reasonable time. Compared to the non-distributed architecture, this architecture has a huge advantage in all the three aspects: the performance in the collection, storage, and analysis of medical Big Data.

6.2. SEMANTIC FRAMEWORK DEVELOPMENT TO PROVIDE CLINICAL DECISION SUPPORT BASED ON MEDICAL BIG DATA

A clinical decision support system (CDSS) is a computer-based information system developed specifically for clinical decision-making, in which the characteristics of an individual patient are matched to a computerized clinical knowledge base, and patient-specific assessments or recommendations are then presented to the clinician or the patient for a decision [24]. A large body of evidence suggests that CDSSs can be helpful in improving clinical practice. However, to this day, CDSSs have not found wide use outside of a handful of mostly academic medical centres, and their impact on patient outcome is marginal. A major impediment to their wide adoption is the lack of standard knowledge representation formalisms and lack of efficient technologies to process medical Big Data [25]. As the knowledge used by CDSSs is typically derived from standard clinical pathways (or care plans, CPs), this section presents a CP-related case study that successfully implements the Semantic Web framework in solving the above-mentioned deficiency [26]. It proposed a data-driven clinical decision support method to improve CP practicality by applying semantic analysis and reasoning to clinical data in HIS. In addition to the standard general CP orders, detailed locally customized CP orders and mined CP orders from local treatment protocols were provided to efficiently compose hospital-specific CPs, which is beneficial for improving the practicality of CPs and contributing to improve patient-centred care quality.

6.2.1. MODEL CONSTRUCTION

The study used Protégé [27] as the ontology editor tool, OWL as the ontology representation language, and Jena Semantic Web framework as the integrated platform for semantic transforming and reasoning. Global ontology containing standard CP terms and associated relationships were constructed based on the CP specifications published by the Ministry of Health of China. Semantic mapping were created to realize the semantic mapping from standard CP terms to practical clinical data represented by local ontologies, which were built based on vocabulary databases in HIS.

Four super classes, 84 subclasses, and 98 individuals were created in the final CP ontology. As depicted in **Figure 11**, SeptumDeviationCP is an individual of the CP ontology to represent the deviated nasal septum CP. Three order events of the CP are listed with their related order terms and execution dates. Every order term is assigned a value of the property hasHZTerm. The order term 'AntisepticDrug', which is a subclass of Injection, has multiple values assigned for the property hasDrugHZTerm. Standard CP orders from the CP ontology are listed according to their execution date.

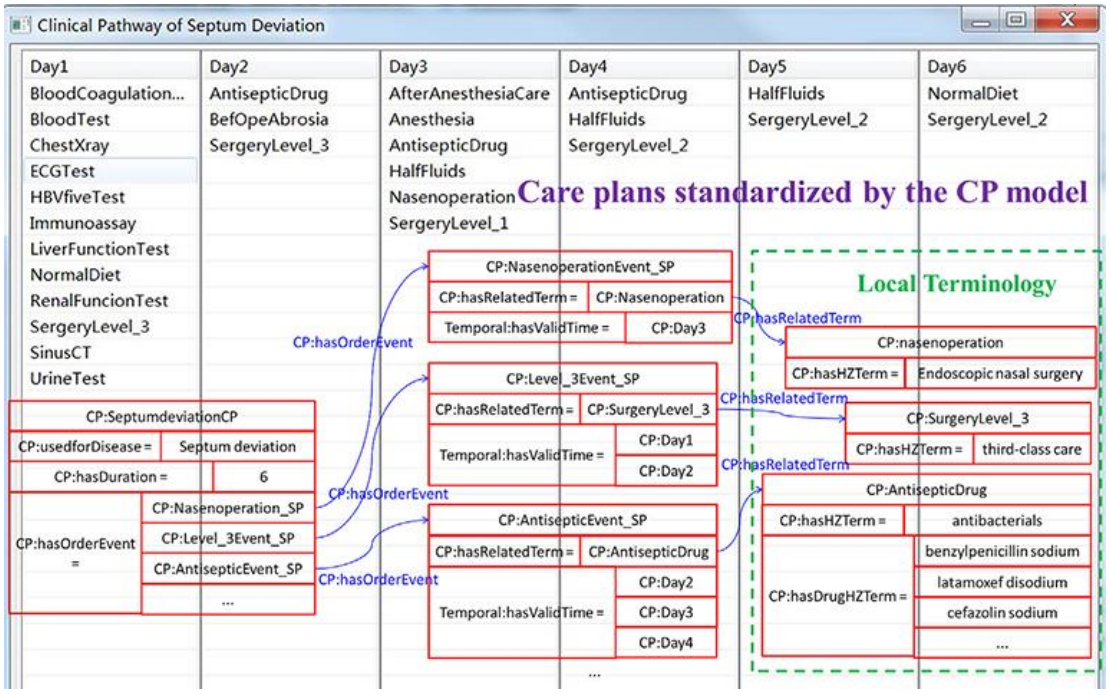

FIGURE 11.

Care plans standardized by the CP model.

6.2.2. SEMANTIC TRANSFORMATION

Semantic transformation of non-semantic relational data modelled into unified semantic data model (RDF format) solves the problem of data heterogeinity and realizes semantic level data integration; it is the foundational process of semantic reasoning and other advanced Semantic Web applications for the meaningful use of medical big data. In this study, Class OrderFact, a super class, is introduced to represent the order data. Each order record acquired by structured query language (SQL) from the relational data model is transformed to an individual of class OrderFact. Fields of order records correspond to the properties of individuals. The transformed data can be accessed and shared using the SPARQL Semantic Web query

language [28]. A statistical analysis on the repetition rate of historical clinical procedures was further conducted to derive the similarity of patient treatment.

A total of 224 individuals of class patient and 11,473 individuals of class OrderFact are imported. As shown in **Figure 12**, each individual of class OrderFact includes the following nine properties: hasPatientData, hasOrderType, hasOrderCode, hasOrderName, hasRepeatIndication, hasStartDate, hasStopDate, hasExecuteDay, and hasCPFlag. In addition, two self-defined properties, hasExecuteDay and hasCPFlag, were inserted.

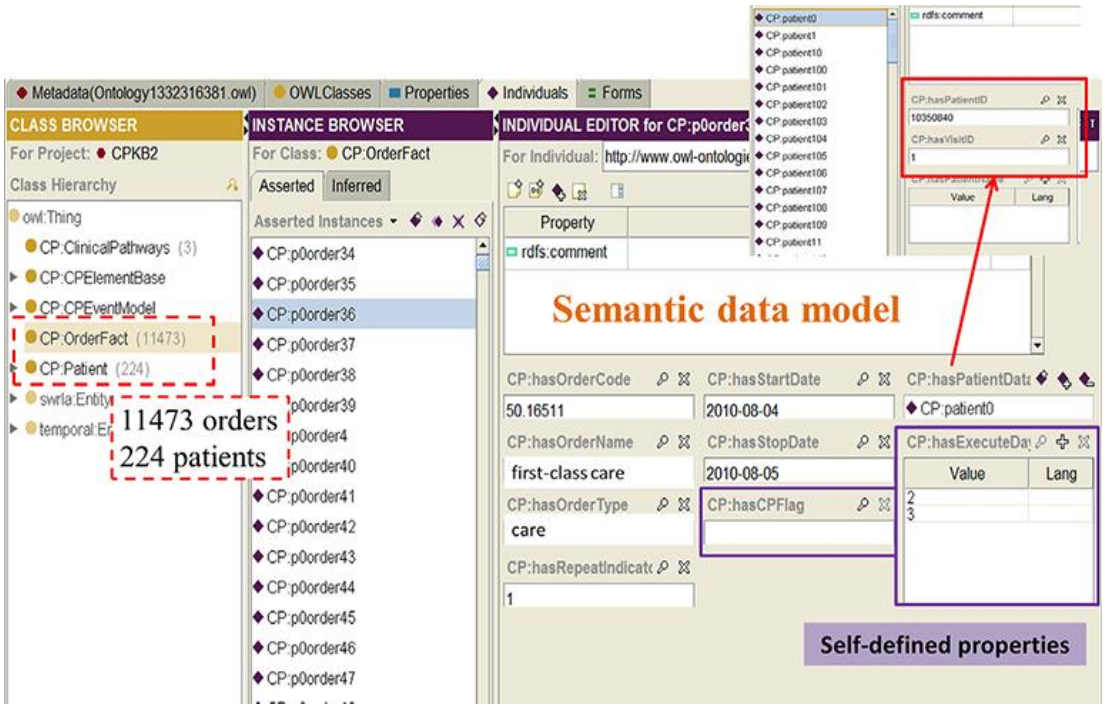

FIGURE 12.

Semantic data model after semantic transformation.

The semantic property hasExecuteDay represents the relative execution day of each order, which is used in long-term order processing aimed at distinguishing long-time orders whose validity holds for multiple hospital days, from temporary orders whose validity holds only at a specific hospital day. This process is important to avoid possible omissions in traditional statistical methods that simply count order records in the raw data tables. The results of long-term order processing are shown in**Figure 13**. The resulting differences in treatment procedures are becoming significant since the third day. For example, detailed orders such as ‘first-class care’, ‘second-class care’, and ‘third-class care’ are being added to the original general nursing orders. In practice, nursing orders, diet orders, and injection orders are typically recorded as long-term orders. Therefore, long-term order processing is necessary to keep a complete track of patient longitudinal medical records.

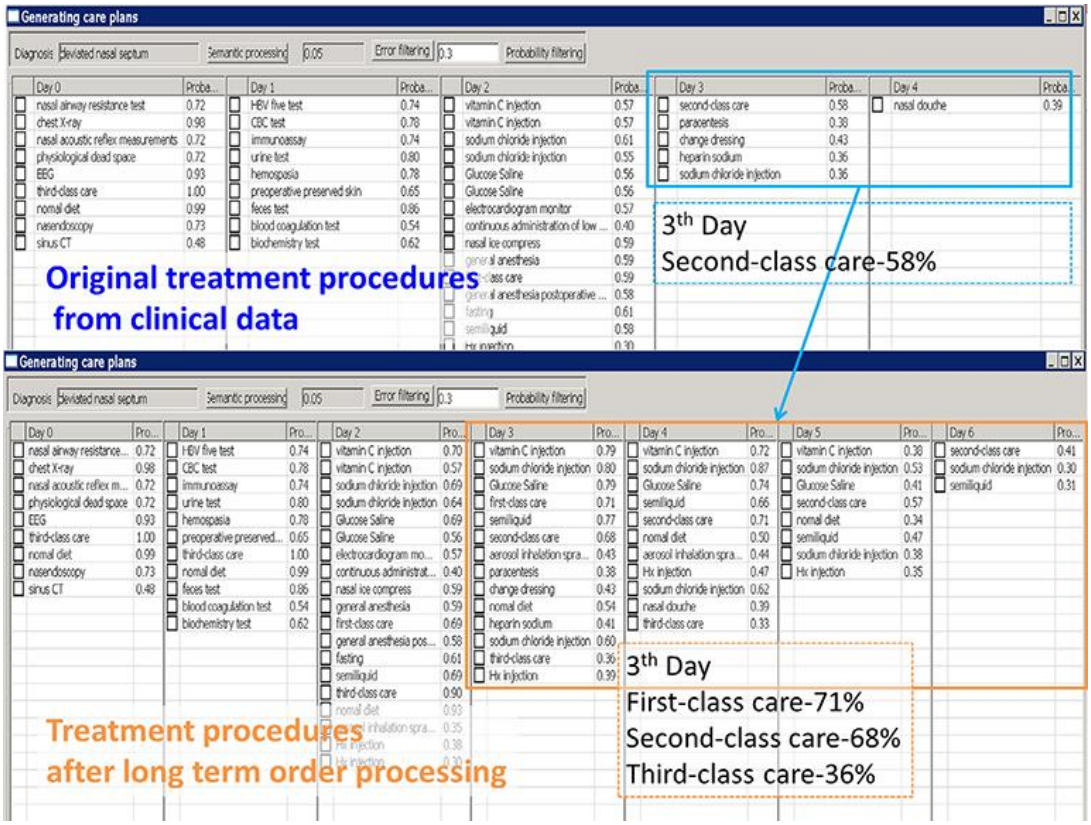

FIGURE 13.

Results of long-term order processing.

6.2.3. A SEMANTIC REASONING

The occurrence of incorrect order records is inevitable. These incorrect order records can be categorized into two types: (1) random errors resulting from recording mistakes; these errors could be eliminated by filtering out the clinical procedures with probability less than the pre-determined minimum. In addition, orders with small probability are specially given to a small number of patients, while not common to other patients. After consulting relevant domain experts, this study used 5% as the default minimum to filter out incorrect orders. A total of 2370 erroneous orders were successfully detected and removed; (2) incorrect data recorded during actual medical procedures, these errors could be eliminated by semantic reasoning. For example, in cases where equivalent long-term and temporary orders co-exist, the semantic rule Rule 1 (**Figure 14**) has been proposed to avoid repetitive ordering: ?order1 and ?order2 are instances of OrderFact, assigned with the same patient (?patient), valid execution time (?day), and title (?name). However, ?order1 is long term, while ?order2 is temporary. After semantic reasoning, the redundant execution time of ?order1 is removed.

```
Rule1:
@prefix CP: <http://www.owl-ontologies.com/Ontology1332316381.owl#>.
[ErrorData: (?order1 CP:hasPatientData ?patient)
(?order2 CP:hasPatientData ?patient)(?order1 CP:hasExecuteDay ?day)
(?order2 CP:hasExecute-Day ?day)(?order1 CP:hasOrderName ?name)
(?order2 CP:hasOrderName ?name)(?order1 CP:hasRepeatIndicator ''1'')
(?order2 CP:hasRepeatIndicator ''0'')
-> remove(2)]
```

FIGURE 14.

Rule 1.

As defined in the following OWL ontology definition, the semantic property hasCPFlag is defined to compare actual clinical workflow identified from historical data with the standardized treatment procedures defined by the CP model. A property value of '1' signifies a direct correspondence between the data order and a CP order, while '2' signifies that the data order provides more details of the CP order. Rule 2 (**Figure 15**) specifies the criteria for determining this property value by comparing the order name (?name) of a data order with the term assigned to hasHZTerm.

```
Rule1:
@prefix CP: <http://www.owl-ontologies.com/Ontology1332316381.owl#>.
[ErrorData: (?order1 CP:hasPatientData ?patient)
(?order2 CP:hasPatientData ?patient)(?order1 CP:hasExecuteDay ?day)
(?order2 CP:hasExecute-Day ?day)(?order1 CP:hasOrderName ?name)
(?order2 CP:hasOrderName ?name)(?order1 CP:hasRepeatIndicator ''1'')
(?order2 CP:hasRepeatIndicator ''0'')
-> remove(2)]
```

FIGURE 15.

Rule 2.

A common problem of implementing standard CPs in a local health care setting is the lack of details such as prescription dose and frequency, which can be mined from local data records. In Rule 3 (**Figure 16**), orders mined from data records which provide such supplemental information of standard CP orders are inferred with hasCPFlag value '2', meaning 'deduced pathway orders'.

```
Rule3:
[DrugCPOrder:(?order CP:hasOrderName ?name)
(CP:SeptumdeviationCP CP:hasOrderEvent ?order_event)
(?order_event CP:hasRelatedTerm ?order_term)
(?order_term CP:hasDrugHZTerm ?name)
-> (?order CP:hasCPFlag 2)]
```

FIGURE 16.

Rule 3.

The reasoning results of executing the Jena rule Rule 1 are shown in **Figure 17**. Take the orders in the second day as an example. There exist reduplicate injection orders for injections such as vitamin C (70 and 57%), sodium chloride (69 and 64%), and glucose saline (69 and 54%) in preoperative treatment procedures. After reasoning, recurrences are removed. Long-term order processing using Rule 1 makes the recorded treatment process more consistent with clinical practice, improving data quality for further analysis.

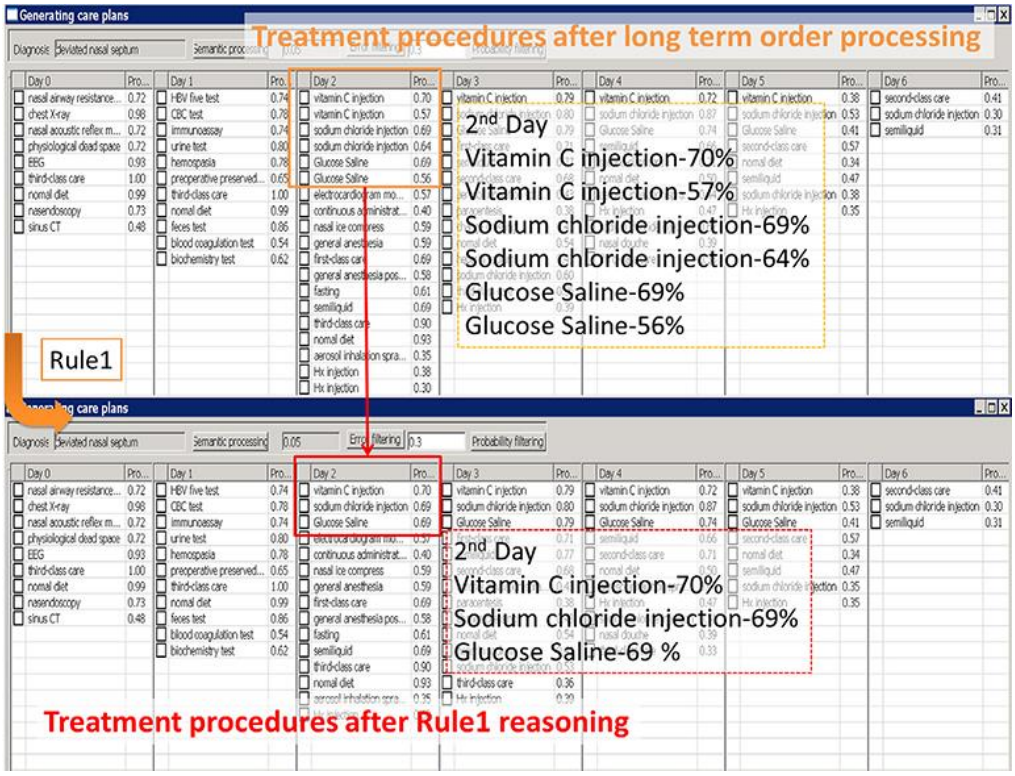

FIGURE 17.

Reasoning results of executing Rule 1.

FIGURE 18.

Reasoning results of executing Rule 2 and Rule 3.

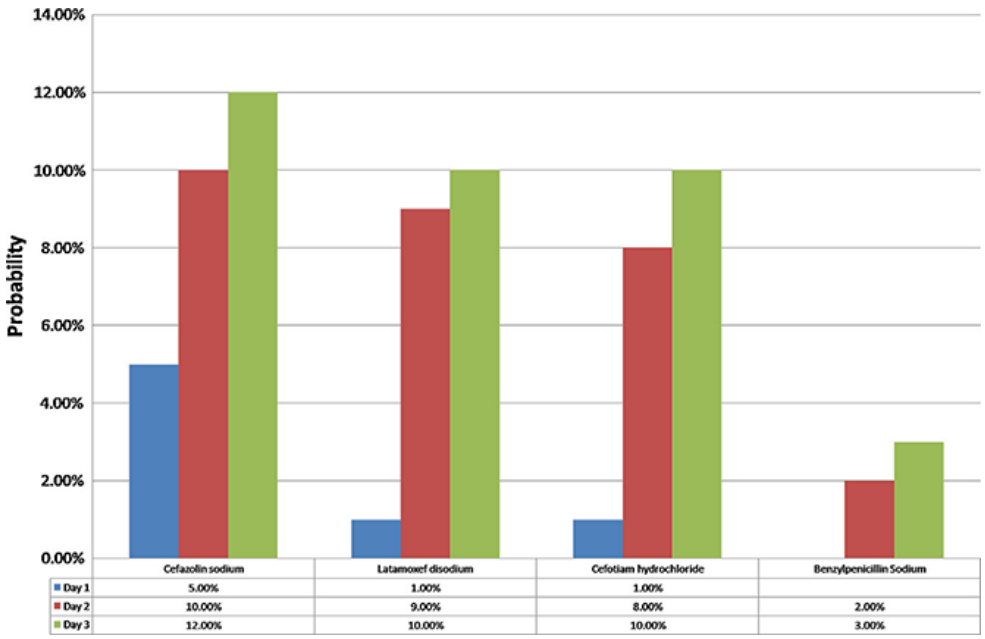

FIGURE 19.

A detailed description of the pathway order "antibacterial."

As depicted in **Figure 18**, different item backgrounds in each child table illustrate the different reasoning results after executing Rule 2 and Rule 3. Orders with a blue background are pathway orders, while orders with a red background or an asterisk are deduced pathway orders, which specify and detail the general knowledge of pathway orders in the CP model. The results show that cefazolin sodium, latamoxef disodium, cefotiam hydrochloride, and benzyl penicillin sodium are common antibacterial drugs for patients with a deviated nasal septum. **Figure 19** presents the probabilities of four detailed antibacterial drugs being prescribed from hospital day one to day three.

Probability of pathway orders refers to the probability of pathway orders that appear in historical data, while percentage of pathway orders is defined as the percentage of pathway orders with some probability in all pathway orders. After calculating the percentage of each pathway order with different probabilities, the practical statistical data are plotted. The plot results are shown in **Figure 20**.

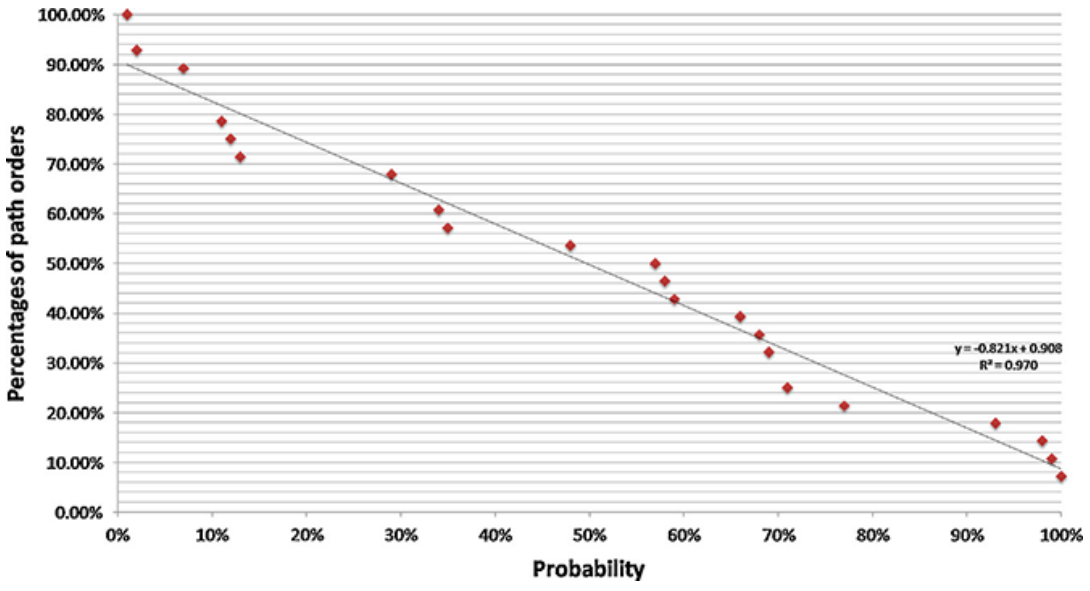

FIGURE 20.

Percentage of each pathway order with different probabilities.

By conducting curve fitting, the percentage of pathway orders and the corresponding probability demonstrates a linear relationship, as given in the following equation, where y stands for the percentage of pathway orders, x stands for the probability, respectively.

$$y=-0.821x+0.908;k=0.821;y0=0.908y=-0.821x+0.908;k=0.821;y0=0.908$$

This study combines traditional statistical methods with advanced semantic technologies to improve the practicability of CPs, which enable timely clinical decision support for healthcare practitioners in balancing evidence-based care with clinical practice, with a final goal of improving healthcare quality, efficiency, and patient satisfaction.