Approaches on improving residents electronic health record management implementation

Liufen Peng1,2,a, Bohua Feng1, Deyu Tang1,2 and Yongming Cai1,2

1School of Medical Information and Engineering, Guangdong Pharmaceutical University, 510006, P.R.China
2Guangdong province precise medicine and big data engineering technology research center for traditional Chinese medicine, 510006, P.R.China

a Corresponding author: penglf@gdpu.edu.cn

Abstract. The residents electronic health records of 8 municipalities of Guangdong province as of December 2017 were extracted. Through data analysis, the author finds out the characteristics and problems of the collected data, which are mainly manifested in the following aspects: high data missing rate, low accuracy and low authenticity of data, erroneous and illegal data caused by incorrect data input, different representation methods caused by data from different sources, inconsistencies and non-compliance with the integrity of references. And the corresponding solutions to these problems are proposed. Approaches to improve standardization, integrity, accuracy and authenticity of data, were also discussed in this paper to promote the sharing and utilization of medical information resources. It would help to build an interconnection and interoperability platform for health big data platform.

1.Introduction
In the action plan of promotion for the "Internet+ medical health" (2018-2020 years) of Guangdong province, vigorously developing "Internet+" public health services is put forwarded. We should innovate "Internet+" health management services, improve residents medical health information, strengthen residents medical and health information management, and achieve establishing residents personal medical and health information records files, continuous records, dynamic sharing updates, and develop health care information services covering the whole life cycle. Recently the informationization level of residents health records has been continuously improved, but the standardization of health records, the sharing and utilization of information resources are inadequate. It is difficult to carry out deep-seated data mining and analysis, and the advantages of information technology can not be brought into full play. In this paper the data characteristics and existing problems in residents health records will be demonstrated, corresponding solutions are put forward to explore the standardization and co-construction and sharing of residents health records. Methods for promoting standardized development of community health management is helpful to construct big data platform for interconnected health services, and to realize the information sharing among families, community health service institutions, hospitals, preventive health care institutions and health administrative departments.
2. Data sources and processing methods

2.1 Data sources
The data of urban and rural residents health records from 8 municipalities of Guangdong province as of December 2017 were collected as the research object. According to “the project implementation plan of Guangdong province health records for urban and rural residents (hereinafter referred to as the "plan"), residents health records include basic personal information, physical examination, health management records of key groups and other medical service records. Referring to the "plan", we establish two main relational tables: "personal basic information table" and "health examination table". Using the principle of database aggregation, the original database data in each region after processing are aggregated into above the two tables. After data aggregating, about 2.7 million records were recorded in the "personal basic information table" and 2.5 million records were recorded in the "physical examination table". The data of residents health records obtained have not been updated in time, the continuity of data has not been maintained well, and data faults, duplications or vacancies have occurred. On the whole, the data quality is poor, and the missing rate (the ratio of the number of vacant values in a field to the total number of records, which is calculated in this paper) and the error rate are relatively high.

2.2 Data cleaning process
After discussing of experts from data analysis, public health, clinical medicine we analyzed and decoded the health records data, which are acquired from grass-roots level. We analyzed the overall situation of the data and the relationship between the tables and the fields in the tables. The relevant tables also are merged or split, where fields and values are carefully standardized including unified format and units. The data accuracy and quantities are considered to be helpful for ensuring a more reasonable range of data, as a data cleaning standard. The data cleaning process is as below: (1) According to the requirements of the "plan" and the real situation, a standardized database is established in SQL Server. There are mainly two relational tables in the database: the personal basic information table and the health examination table, which are used to import the personal basic information and the medical examination data from all the original data. (2) Data decoding is based on data dictionary provided by the data source units. Some data dictionaries are incomplete or have ambiguous field meanings, which need to refer to relevant information and data for analysis. (3) Referring to the requirements of the database and relational table established in (1) from different sources, the initial data processing work includes adding necessary fields, splitting fields, unifying units or representations. (4) Using data transformation services (DTS). The residents electronic health record from different database types are imported into the database for further data cleansing.

3. Results and discussion

3.1 Data missing rate analysis
The data is collected complying with the "plan", but the execution of data acquisition varies from place to place. The values of fields contain illegal data or unreasonable data. These invalid values will be removed in the process of data cleaning. Therefore, the missing rate of the cleaned data will be higher. The missing rates of each field in "personal basic information table" and "health examination table" are shown in table 1 and table 2. The missing rates of several original data reach 100%, but these fields are not necessary information in this study, so it is not considered here. Table 1 and table 2 show the number and percentage of fields within a deletion rate. Among them, the "basic information table" has 86 fields, and the "health Checklist" has 248 fields.

Table 1. Missing rate of each field in basic information table.

| Missing rate range | Number of fields | Ratio |
|--------------------|-----------------|-------|
| 100%               | *               | *     |

### Table 2. Missing rate of each field in health examination table.

| Missing rate range | Number of fields | Ratio   |
|--------------------|------------------|---------|
| 100%               | *                | *       |
| 90-100%            | 127              | 55.95%  |
| 80-90%             | 17               | 7.49%   |
| 70-80%             | 8                | 3.52%   |
| 60-70%             | 13               | 5.73%   |
| 50-60%             | 5                | 2.20%   |
| 40-50%             | 7                | 3.08%   |
| 30-40%             | 5                | 2.20%   |
| 20-30%             | 10               | 4.41%   |
| 10-20%             | 22               | 9.69%   |
| 0-10%              | 13               | 5.73%   |
| Total              | 227              | 100.00% |

*: The 36 fields with missing rate of 100% were not included.

#### 3.2 Problems and suggestions for improvement

After standards are formulated to the data cleaning, the obvious error and beyond-range data are basically cleared, the irregular data are converted into a unified format. Some reasons have resulted in data redundancy and data inconsistency occurred in the data aggregation processing. For example, there is no standard data dictionary in some areas of the original data, there are both new and old versions standard of the "resident health records" in some cities, personal basic information in some cities is composed of multiple tables, and the information of the health examination table is also scattered in many tables. During cleaning process, data were found to have problems: (1) The data missing rate is high. (2) The data are incomplete. There are some missing contents in one medical record. (3) Inconsistent field names, types, formats and constraints. Data formats and units are inconsistent with basic industry specifications, and citation integrity is not complied. (4) Data entry is
arbitrary. There are spelling errors, printing errors, meaningless letters, punctuation marks or other characters (*$/,. etc.). (5) Existing illegal values. There are unreasonable values, for example, blood biochemical parameters appear negative or abnormally large values. (6) There are many forms of expression for the same entity, and the numerical expression is not completely consistent. For example, blood group is represented by A/B/C in some data, but also by 1/2/3 in the others, and the corresponding blood group of 1/2/3 (A/B/C) is not completely consistent.

It is suggested to standardize and unify the database design before data acquisition. The data collection at all stages should be regulated in order to ensure the effectiveness and usefulness. From the beginning of the establishment of the database, the unified health records standards and database standardization requirements need to be followed to establish standard relational tables. The same data should not be scattered into too many tables. Some advanced suggestions in database design and data collection are as follows: (1) To establish standardized data dictionary. The key data can not be vacant, and the form of data representation should be unified. (2) The format of variables should be standardized. The same health records number coding rules should be followed. (3) Field naming follows database variable naming rules. Data format and units should be consistent and conformed to industry standards. Data integrity constraints should be added to ensure the mutual constraints between data and the accuracy of data between tables. (4) Variables should be divided into the smallest units, such as blood pressure: left hand systolic pressure, left hand diastolic pressure, right hand systolic pressure, right hand diastolic pressure. And the past disease history according to each disease are set to a variable. (5) It is suggested that special supervisors should be appointed to improve the professional skills and professional accomplishment of medical record informants. It is helpful to avoid duplication, adulteration and random input of data. (6) To discriminate the data provided by residents initially. This may reduce erroneous or untrue data. Guiding residents to provide complete and accurate information is important.

The other main reason beyond technical factors for the high missing rate and the low accuracy rate of data is lied in the residents trust. In practical situation, residents are unwilling to provide accurate information or refuse to provide it because of their lack of understanding of information collection work, and strong alert psychology or the influence of old ideas. They are afraid of being cheated or revealing their privacy and they refuse to investigate medical staff. In this regard, we should strengthen publicity and education, strengthen grass-roots cooperation. Staff abide by professional ethics and pay attention to language skills, they should protect the residents right to know, get the greatest trust and cooperation of residents [1].

3.3 Improving health records standardization
Resident health records are an important part of hospital information construction. The problems analyzed above, such as inconsistent fields, incomplete data and multiple representations of the same entity in the data, are precisely caused by the nonstandard health records. It is necessary to solve a series of standardization problems to standardize the data collection of residents health records, improve the database of health records and promote the sharing and application of health information among the whole people. Resident health records involve five major standardization contents: (1) writing standardization; (2) using terminology, coding standardization; (3) functional standardization, including national electronic medical records rating standards and HIMSS EMRAM rating; (4) interconnection standardization including electronic medical records sharing document specifications, national standardized maturity assessment of interconnection and interoperability of medical health information; (5) standardization of management and application including data sharing and authenticity.

In order to accelerate the standardization process of health records, we need to encourage medical institutions at all levels to strengthen their application maintenance technical support in the process of promoting health care big data collection and storage, and to open up channels for data resource sharing. We will speed up the construction of basic databases for residents electronic health records,
electronic medical records and electronic prescriptions, and establish a sound standardization system for health records [2].

3.4 To perfect the health management team and the sharing of health records
We should not only enhance the level of technology and standardization of management of residents' health records, but also strengthen the personnel management and promote the co-construction and sharing of information resources. There are some problems in health records data such as artificial error data or illegal data, incomplete and inaccurate data, which can be avoided by corresponding measures. In the process of collecting residents health records information, health records administrators should be conscientious and responsible to check the data carefully to ensure its accuracy and completeness. After confirming the correctness of the information, the classification and cataloging shall be carried out according to the unified standards, formats and rules in the "plan", indicating the index number and time, etc. The relevant managers should also strengthen the data audit to ensure the authenticity and integrity of the data, and then enter the basic database.

The collection and management of health records are mainly completed by the grassroots medical record information personnel. In the process of collecting and managing health records, it is often necessary to record and distinguish various kinds of information by means of various management methods. However, at the present, there is a serious shortage of medical information personnel at the grass-roots level, the overall work ability is deficient, and the comprehensive quality needs to be improved generally[3]. In order to achieve effective management and ensure the smooth implementation of data collection proposals, we must accelerate the training of "Internet+ medical health" talents, improve the construction of health records management team, and improve the professional skills of health records managers. To establish authoritative coordination management institutions at all levels and policy guarantees, it is helpful for standardize the electronic health record collection process and management methods. By strengthening publicity, residents right to know is protected to obtain their greatest trust. Modern information technology makes it convenient to achieve information resource sharing [4].

4. Conclusion
With help of information technology, corresponding service system and mechanism, we try to minimize the avoidable data problems and data missing. While using modern information technology, applying information resource management concept to the management of residents health records can ensure the standardization of data collection and storage process, which can improve the utilization rate of data, emerging advantages of residents health records. Besides ensuring data quality and reducing data missing rate through the technical proposals put forward in this paper, there are still many problems to be solved in improving standardization and building a sharing mechanism in order to truly realize the deep-seated data mining of residents health records. Methods for guiding the practice of health records in terms of technology and theory should be furtherly discussed to realize the reuse and sharing of health records. By improving the "Internet+ medical health" service system, community residents could be full of confidence in health management.

Acknowledgement
This work was supported by Guangdong province precise medicine and big data engineering technology research center for traditional Chinese medicine. It was also supported by innovation and strengthening university project of Guangdong pharmaceutical university: exploration& practice on the training mode of applied medical X-ray engineer(2015[15]). And we thank to the funding support of the construction project of Guangdong province quality engineering for characteristic specialty in 2016 ([2016]233).

References
[1] X. Fu, H. Wang, J. Traditional Chinese Medicine Management, 20, 2 (2012)
[2] X.P. Zhou, X.M. Feng, Records Science Bulletin, 6 (2017)
[3] J. Chen, China Management Informatization, 4 (2018)
[4] M. Yu, J. Traditional Chinese Medicine Management, 12 (2018)