MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220062

Data-Sharing Gateway System Design for Large-Scale Medical Information Collection with Distributed EMR Storage

Katsuya Tanaka[a,b], Masami Mukai[a,b], Ryuichi Yamamoto[c], Naoki Mihara[a,b]

[a] Department of Medical Informatics, National Cancer Center Hospital, Tokyo, Japan
[b] IT Integrating and Support Center, National Cancer Center, Tokyo, Japan
[c] Medical Information Development Center, Tokyo, Japan

Abstract

The collection and use of large-scale medical information for developing artificial intelligence engines are actively ongoing. In Japan, collection systems have been built to collect data for medical image analysis and disease repositories. In the experimental project for the next generation medical infrastructure law, a centrally integrated basic system was developed, and standardized electronic medical record (EMR) storage data distributed to each hospital were transferred into one data center and imported to a database for secondary use. The law requires a mechanism for maintaining a list of notified or opted-out patients. To operate these systems, safe and efficient secondary use of collected information is essential not just for the law but also for large-scale data collection projects, such as multifacility clinical research. This paper considers wholesomely, requirements for providing medical care information to data collection projects and proposes additional requirements for a gateway system under development.

Keywords:
Electronic Medical Records, Standardization, Patient Data Privacy

Introduction

The collection and use of large-scale medical information for developing artificial intelligence engines are actively ongoing [1]. In Japan, collection systems have been built to collect data for medical image analysis and disease repositories. Enforcing the Next Generation Medical Infrastructure Law since May 2018 is expected to accelerate the collection of medical information. Technically, a large-scale medical information collection system is essentially used to standardize electronic medical data for immediate use upon collection. In Japan, the SS-MIX2 (Standardized Structured Medical Information eXchange ver.2) standardized storage system [2] is the established domestic standard for exchanging electronic medical information and adopted as the electronic description standard for various collection projects. The standard storage system is structured such that more than 30 types of HL7 v2 message files, containing patient numbers and date of care as sort keys, are stored in a single storage system. In the experimental project for the Next Generation Medical Infrastructure Law, a centrally integrated basic system was developed, and standardized EMR storage data distributed to each hospital were collected into one data center and imported to a database for secondary use. The law also requires a mechanism for maintaining a list of notified or opted-out patients by notifying patients to opt-in or out. Previously, we developed a gateway system with these functions in consideration and deployed it to medical institutions participating in the demonstration project.

To operate these systems in medical institutions, safe and efficient secondary use of collected information is essential not just for the law, but also for large-scale data collection projects, such as multifacility clinical research. This paper considers in detail the requirements for providing medical care information to data collection projects and proposes additional requirements for the gateway system developed so far.

Methods

Overview of the Designed System

An overview of the gateway system we designed is shown in Figure 1. The developed gateway searches and extracts patient data necessary for clinical research from the standardized storage of each medical institution. It also processes data including anonymous processing, according to the purpose of research, and extracts data to another standardized storage format.

The storage is developed based on FUSE (File System in User space), a virtual file system technology. We adopted pgfuse as the FUSE and PostgreSQL as the RDBMS. The recorded HL7 messages are stored in the DB tables as BLOB data and the transaction is tracked in real-time using RDBMS. The HL7 messages are parsed by PL/SQL and parsed medical records (HL7v2 segments, fields) are stored in user-defined tables in the RDBMS. Parsing tasks are executed periodically. Once the records are stored in the tables, we can query the minimum required items by applying view schemas individually according to the purpose of each project.

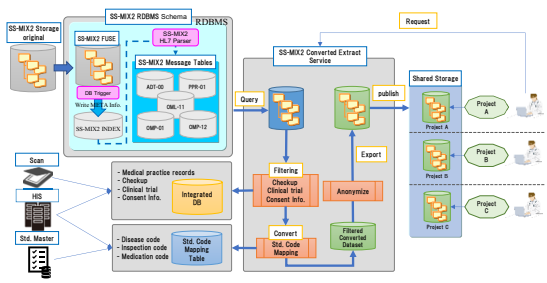

Figure 1. Overview of the developed system for data collection and analysis

Design Points

Fundamental requirements of the newly developed medical information collection gateway are listed below. Because protocols for sending and receiving data between centers differ for each project, these were excluded as basic functions.

Technical Requirements

Standardization

The SS-MIX2 standardized storage system uses HL7 v2 for standard message descriptions. The main description code in the message file includes information, such as disease names, laboratory tests, and medications; however, for the master code, we adopted the standards of the Ministry of Health, Labor, and Welfare of Japan. In practice, the order entry system is under-equipped with the standard code, so in-house codes are used to standardize storage. Therefore, creating and maintaining a separate correspondence table for code mapping between in-house and standard codes is necessary. The standard code should be added in ER7 format in HL7 v2 for investigations later.

Code Mapping

To confirm the feasibility of the above standard code, in this paper, we have conducted a mapping examination for the standard code. We mapped the disease to domestic standard code representing ICD-10 [3], the laboratory test to JLAC10[4], and the drugs to domestic standard code called HOT9. Both are coding systems defined as standard codes by the Ministry of Health, Labor, and Welfare of Japan.

Consent Information

Large-scale information collection in clinical research is based on patient consent. For this reason, opt-in and opt-out functions are required. In Japan, research participation intent is recorded on paper, and signed consent forms are digitized and linked to each extracted EMR data. Electronic representations are realized using HL7 CDA [5]. Figure. 2 shows the specific flow of handling the consent form. The consent form for clinical research is scanned for use as an electronic file stored in the electronic medical record system after signing the paper. If the scanned image is convertible to searchable XML formats using the HL7 CDA standard and stored in standardized storage, the medical and consent information of the target patient for each research can be explored simultaneously. Moreover, it is possible to determine whether the data can be used for specific clinical research secondarily.

This paper adopts a method that uses the scanned image of the consent document and exports it as an electronic file to the extended area of the SS-MIX2 standardized storage, which is an optional standard by the standardized storage, so it is searched separately. As a prerequisite, a dedicated consent form is prepared for each research project, and by identifying the research project from the printed identification number on the consent document or the QR code, we can judge the secondary availability of the relevant patient clinical data for the clinical research analysis related to the consent document.

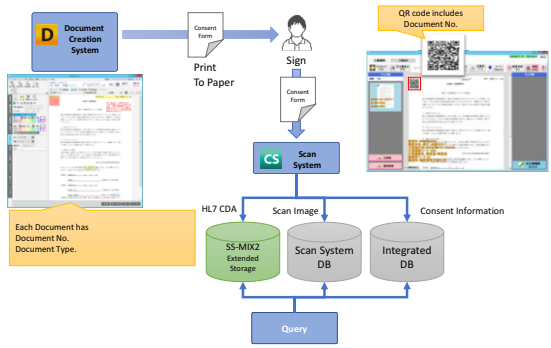

Figure 2. Overview of Information Consent Handling with Scan system and Standardized EMR storage

Multipurpose Archiving

Sometimes, multiple clinical information collection studies employ standardized storage as a data source, or data processing methods (e.g., master code master used and data processing accuracy) differ between research projects. Our study assumes only a single original standardized storage system is deployed to a medical institution and that conversion rules are defined for each data collection project. To enable different data processing rules for various projects, the retention mechanism of the processing rule and the processing area for storage was deployed to the gateway. From Figure 3, by exporting the dataset extracted for each data collection project to the FUSE-based storage in another SS-MIX2 standardized storage format in HL7 v2, the dataset can also be searched for data items. We can continue to focus only to parse the format of SS-MIX2 standardized storage represented by HL7 v2 messages.

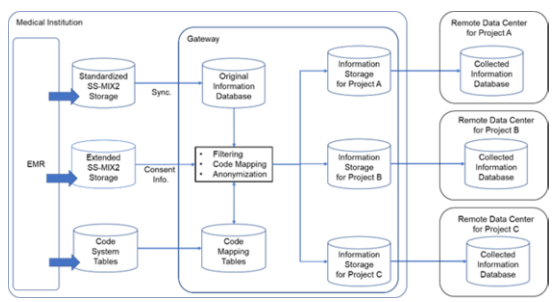

Figure 3. Overview of the gateway system for the secure medical information collection

Security Requirements

Information Protection

Various purposes are considered for data collection projects, but limiting EMR data to third parties is necessary for certain cases. This includes cases with access restrictions stipulated in a contract, such as patient's participation in clinical trials. Depending on the patient's profile and medical practice records, a mechanism for determining whether data can be shared with a third party conducting a collection project is necessary. Furthermore, since our hospital conducts cancer screening, the exported standardized storage has been containing data on patients who have undergone screening. For clinical studies, the data of screening patients correspond to that of healthy subjects, so it is necessary to treat them separately from the

patients who visit our hospital. Therefore, when using data, a search and extraction mechanism that distinguishes the screening patients is required.

Evaluation

We use the medical care data of the National Cancer Center Central Hospital to identify issues to consider when implementing the above-mentioned functions. We examined the possibilities and limitations of implementation. The number of standardized storage data used in this study is shown in Table 1. The standardized storage stores about 640,000 patients and about 4.51 million messages. With the medication implementation system not functional, there is no message output for the prescription. Additionally, due to implementation, results from physiological examinations are not stored. The number of codes in our code master is shown in Table 2. The standard code is not used but the disease name in the order entry system. Since the local code cannot be used for cross-medical collection and analysis in clinical research, the mapping mechanism to the standard code adopted in Japan was conducted.

Table 1. Message Counts by Message Types in SS-MIX2 standardized storage of our hospital

Message	Type	Counts	Message	Type	Counts
Patient's basic information	ADT-00	727,840	Foods order	OMD	17,852
Change of investigator	ADT-01	9,465	Radiological examination order	OMG-01	49,500
Reception of outpatient physical examination	ADT-12	2,208,958	Notice of radiological examination conduct	OMG-11	44,658
Hospitalization plan	ADT-21	8,453	Endoscopy order	OMG-02	5,712
Conduct of hospitalization	ADT-22	3,015	Notice of endoscopy conduct	OMG-12	4,686
Conduct of staying outside	ADT-31	1,001	Physiological examination order	OMG-03	9,541
Conduct of return from staying outside	ADT-32	999	Notice of physiological examination result	OMG-13	0
Plan of change of department/building	ADT-41	994	Specimen examination order	OML-01	59,056
Conduct of change of department/building	ADT-42	2,146	Notice of radiological examination conduct	OML-11	614,862
Plan of discharge	ADT-51	3,164	Prescription order	OMP-01	47,825
Conduct of discharge	ADT-52	2,630	Prescription conduct notice	OMP-11	0
Allergy information	ADT-61	31,637	Injection order	OMP-02	194,571
			Injection conduct notice	OMP-12	112,611
			Disease name (history) information	PPR-01	353,645

Table 2. Code Numbers used in our EMR

Code	No. of Used	Usage pattern in Order Entry System
Laboratory Test	789	Local codes are applied
Disease	10,166	820 local codes are applied Others are operated in Domestic Standard Code
Drug	2,984	Local codes are applied
Injection	513	Local codes are applied

Results

Code Mapping

For the SS-MIX2 standardized storage system, each EMR item is an output in standard code; however, not all available EMR systems are fully standardized, including subsystems. Therefore, some messages that are output to the standardized storage can include in-house codes. Not only clinical trial drugs are excluded in the master code but laboratory tests, drugs, medical materials, etc. are subject to special house codes since standard codes are not applied for these items.

Table 3. Overview of Mapping Results

Codes	Total	Not Applicable	Total Valid	Uniquely Mapped	Multiple Candidates	Needs Correction	Mapping Ratio
Disease	25987	2	25985	25967	-	18	100%
Modifiers	2259	-	2259	2257	-	2	100%
Drug	8237	4801	3436	2007	1006	16	88.1%
Laboratory Test	1964	759	1205	95	854	1	78.8%

Table 3 shows the mapping results. From the results, the disease and drug were high, followed by the laboratory test. The uniquely identifiable ratios are disease, drug, and laboratory tests. Also, some code settings were incorrect and results for each master code are described in detail below.

Disease Code

Table 4 shows the mapping results of diseases and modifiers to standard codes. Since the standard code is defined as the code used for medical fee billing, the mapping possibility rate is high.

Table 4. Mapping Results to Standard Code of Disease and Modifier Codes

Disease local code to MEDIS Standard Code.		Total	Mapping			
			Auto	Candidates	Needs correction	Failure
Disease name						
Standard code	available	25,967	25,967	0	0	0
Standard code	none, invalid	18	0	0	18	0
Modifier						
Standard code	available	2,257	2,257	0	0	0
Standard code	none, invalid	2	0	0	2	0
	Total	28,244	28,224	0	20	0

Drug Code

The mapping results of the drug codes are shown in Table 5. The accounting code, the code used by the Ministry of Health, Labor and Welfare code, and the drug price code are linked with the local code and using these codes together, it is possible to map to the standard code. However, many registrations for clinical trial drugs, including anticancer drugs cannot be mapped to standard codes.

Table 5. Mapping Results to Standard Code of Drug Codes

Medication local code to HOT code.		Total	Mapping			
			Auto	Candidates	Needs correction	Failure
MHLW code	available	2,971	1,976	987	8	0
Accounting code Claim code	available available	50	31	19	0	0
Accounting code Claim code Drug price code	available invalid available	8	0	0	0	8
Accounting code Claim code Drug price code	available invalid none, invalid	124	0	0	0	124
Accounting code Claim code Drug price code	available none, invalid none, invalid	230	0	0	8	222
MHLW code Accounting code	none, invalid none, invalid	53	0	0	0	53
Trial medication	None-target	4,801	-	-	-	-
	Total	8,237	2,007	1,006	16	407

Laboratory Test Code

Table 6 shows the mapping results of the laboratory test code. Unlike the other two codes, it was difficult to map automatically. The accounting code, medical fee-billing code, and medical treatment code are prepared and linked with the local codes in the order entry system. If we make full use of these, the standard code remains at the level difficult to match uniquely.

Table 6. Mapping Results to Standard Code of Laboratory Test Codes

Laboratory test code to JLAC10 Code.		Total	Mapping			
			Auto	Candidates	Needs correction	Failure
Accounting code Claim code Medical practice code	available available available	949	95	854	0	0
Accounting code Claim code Medical practice code	available available none, invalid	90	0	0	0	90
Accounting code Claim code Medical practice code	available none, invalid none, invalid	57	0	0	0	57
Accounting code Claim code Medical practice code	none, invalid none, invalid none, invalid	109	0	0	1	108
	Total	1,205	95	854	1	255

Data Filtering

Policies governing each research project and specifies that patients are included or excluded are based on patient consent, clinical trial patient information, and patients treatment and care. The system takes the lists of targets and excluded patients as output from the EMR system for each project and referred to from the processing of the gateway. The system aided the identification of patients participating in clinical trials and patients undergoing clinical screening tests. Patients participating in the clinical trial are evaluated from the list of patient numbers registered in the clinical trial management system. Similarly, for screening patients, a list of patient numbers with a history meeting a consultant in the screening department system was extracted. Using these lists, we determine whether the data of the patients be included at the extraction of each clinical research.

Consent Information

Table 7 is a database for managing the consent information because of considering the document items specified in the HL7 CDA and the secondary use for clinical research. By identifying each consent document, it is possible to identify the target clinical research theme, information disclosure destination, presence/absence of consent, and validity period.

Table 7. Designed schema for consent information using scanned consent form papers

Column	Unique	Type	Content
ProjectId	○	varchar	Identifier of Research Project
PatientId	○	varchar	Patient No.
Consent	○	varchar	Usage
Purpose	○	varchar	Purpose
Participant	○	varchar	Destination of Disclosure
NegationInd		boolean	Flag of Approve
StartDatetime	○	timestamp	Start of effective time
EndDatetime	○	timestamp	End of effective time
Inactive		boolean	Varidity
InsertedDatetime		timestamp	Time of registration
UpdatedDatetime		timestamp	Time of modification

Discussion

Limitations

The aggregated result in this paper are values from a hospital specializing in cancer treatment and do not represent the tendency of university or general hospitals.

Significance of the System

The gateway system we proposed and developed supports multipurpose data collection projects, such as the collection of information by legal systems and research-based data collection projects. Thus, the gateway system is developed to be versatile for several medical institutions. In particular, the mechanism for inspecting the consent of patients with standardized storage and sharing of information outside medical care is an important function of this gateway.

Code Mapping

The disease and drug codes can be automatically mapped to the standard code. However, the unique identification rate of the laboratory test code was considerably low. Also, it is presumed that the JLAC code includes inspection materials and measurement methods in the code expression. Therefore, a more detailed study is a topic for the future. By limiting the mapping target to the main test items that are frequently used in clinical research instead of all test items to be performed, the effort required for the mapping process can be reduced.

Future Issues

Maintenance Cost

A centrally integrated data collection system requires the cost of maintaining and continuously increasing storage space (e.g., cloud storage) for the research project to be met to ensure continuity of the research. However, because the gateway system proposed includes a search function for each distributed storage system, then limiting the use of the central facility as an accumulation location and collecting case information on demand suffices. Our next plan is to allow typical search queries, such as disease name, laboratory test value, and medication.

Privacy Risk Assessment

In a centrally integrated data collection system, data anonymization is considered appropriate for extracting and creating datasets according to the applicable conditions. Whether anonymization should be performed before collection depends on the target data. Because our gateway system is configured so that a repository exists for each purpose on the medical institution side, data processing before collection is possible for processing primitive fixed conditions. However, typical anonymization conditions vary widely; thus, an issue that will be solved in the future.

Conclusions

In this paper, we developed and designed an additional function of a gateway system for large-scale data collection. The purpose was to implement the function for various research projects, particularly large-scale medical information collection projects of the Next Generation Medical Infrastructure Law and other clinical research projects. We examined the basic requirements for employing the system as a data collection gateway used for both public and restricted purposes.

Acknowledgements

This study has been approved by the Research Ethics Committee of the National Cancer Center Japan (Permission number: 2019-122, 2020-008) and was partly supported by The National Cancer Center Research and Development Fund (30-A-12), Japan.

References

[1] R. Yamamoto, Large-scale health information database and privacy protection, *Jpn Med Assoc J* **59** (2016), 91-109.

[2] M. Kimura, K. Nakayasu, Y. Ohshima, N. Fujita, N. Nakashima, H. Jozaki, T. Numano, T. Shimizu, M. Shimomura, F. Sasaki, T. Fujiki, T. Nakashima, K. Toyoda, H. Hoshi, T. Sakusabe, Y. Naito, K. Kawaguchi, H. Watanabe, and S. Tani, SS-MIX: a ministry project to promote standardized healthcare information exchange, *Methods Inf Med* **50** (2011), 131-139.

[3] O. World Health, *ICD-10: international statistical classification of diseases and related health problems: tenth revision*, in, World Health Organization, Geneva, 2004.

[4] N. Nakashima, Japanese sentinel project and contribution of laboratory medicine, *Rinsho Byori* **61** (2013), 501-510.

[5] H.L.S. International, *HL7 Standards Product Brief - HL7 CDA® R2 Implementation Guide: Privacy Consent Directives*, Release 1, in, 2017.

Address for correspondence

Katsuya Tanaka, National Cancer Center
5-1-1 Tsukiji, Chuo-ku, Tokyo, 104-0045, Japan
katstana@ncc.go.jp