Chapter 6
Health Test Bed Group

Katsuya Tanaka and Ryuichi Yamamoto

Abstract Under the new law for the secondary use of medical information, which was activated in May 2018, the future expected secondary use with information anonymization may contribute to research and development in the medical field of integrated medical research and public health. On the other hand, under the revised Personal Information Protection Law and the revised ethical guidelines in medical research, privacy protection and patient consent management is a crucial issue for the management of researches. Our JST CREST project, which started in March 2014, has issued the development of technological elements and synthesized the developed methods for real-world system for the secondary use and privacy protection of big data on cloud infrastructure, including safe clinical information management, commercial cloud utilization, and privacy risk evaluation. In this paper, assuming the utilization of the Standardized Structured Medical Record Information Exchange version 2 storage, the following target issues are described: (1) effective utilization of existing standardized storage, (2) secure data collection across medical institutions, (3) privacy risk evaluation in analysis, and (4) traceability while secondary use.

K. Tanaka (✉)
National Cancer Center Japan, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan
e-mail: katstana@ncc.go.jp

R. Yamamoto
Medical Information System Development Center, Kagurazaka 1-1,
Shinjuku-ku, Tokyo 162-0825, Japan
e-mail: yamamoto@medis.or.jp

A. Miyaji and T. Mimoto (eds.), *Security Infrastructure Technology for Integrated Utilization of Big Data*,
https://doi.org/10.1007/978-981-15-3654-0_6

6.1 Overview of Legislation and Standardization for the Secondary Use of Electronic Medical Records

6.1.1 Personal Information Protection Act and Next-Generation Medical Infrastructure Act

The Personal Information Protection Act [1] was revised in September 2015 and was fully enforced in May 2017. Prior to the revision, the Personal Information Protection Act was established in 2003 and fully enforced in 2005. At the time that the previous law was established, both Houses of Councilors recognized that it was insufficient, with the establishment of separate laws being required in multiple fields, including medicine, and it was not actually reviewed. This situation remained for a decade or more. It is now high time for it to be revised. Consequently, several problems in the previous law were improved; however, a few problems still exist, and some new concerns have emerged. These include fears that secondary use, which is essential in the medical treatment field, will become problematic.

To avoid a negative impact on innovation, including drug discovery and medical equipment development, the Next-Generation Medical Infrastructure Act (official name: Act on Anonymously Processed Medical Information to Contribute to Medical Research and Development) [2], specializing in the secondary use of medical data, was enacted in April 2017 and enforced on May 11, 2018. In this paper, we investigated the issues related to the Personal Information Protection Act and the predicted effects of the revision to the law, provided an overview, and considered the impact of the Next-Generation Medical Infrastructure Act.

6.1.1.1 The Personal Information Protection System in Japan and Related Issues

The main objective of reviewing the previous law was to respond to the EU directive concerning cross-border personal information in 1995 and the concerns about privacy infringement as a result of the resident registration network brought about by revisions to the Basic Resident Registration Act. Furthermore, there were concerns regarding the prospect of eavesdropping being made possible with court approval through revisions to the Criminal Procedure Code. Also, as previously stated, this was fully enforced in April 2005. Basically, this conformed to the OECD personal information cross-border guidelines [3]; however, allusions to several issues have been identified. The problems in the medical area include the following facts: the law is a comprehensive one that does not specify the field; the characteristics of medicine frequently provided to third parties, an essential purpose, are not being considered; the definition of personal information is ambiguous, which inevitably makes anonymization difficult; and focus is placed on protection, and so, the promotion of reuse that does not violate the right of the individual, which was the original

purpose of the law, is largely ignored. There is also the fact that, in a private sense, it is aimed at the operator, and where there is an individual causing the infringement, it is an indirect regulation concerning supervision by the operator. Moreover, the penalties are light and indirect and thus lacking in effectiveness. The different systems for acquiring personal information are enforced for governments, independent administrative agencies, local governments, and private enterprises, and this is considered to obstruct the utilization of personal information across these frameworks.

In the revised Personal Information Protection Act, the concept of important information has been introduced, and as the vast majority of medical data is designated as important information, this is a step forward from a comprehensive law that does not specify fields. When acquiring important information, explicit consent is required, and third-party provision based on opting-out, which can occur when providing such information to third parties while the intentions of the person concerned are still unclear, is prohibited. This is clearly a step forward and promises to suppress provision to third parties where this is not intended by the person concerned. On the other hand, in the case of third-party provision, which is essential in collaborative medicine, while there were concerns about the explanation of symptoms to family members, consultations with specialists, etc., this is largely covered by the clear definition of opt-out consent as "implicit consent" in guidance concerning the appropriate handling of personal information by healthcare and nursing care providers (hereafter, "healthcare and nursing care guidance"), implementation guidelines issued jointly by the Personal Information Protection Commission (hereafter, "PPC") and Ministry of Health, Labor and Welfare (hereafter, "MHLW"). However, the fact that, under law, clear consent is required for provision to third parties for the purpose of drug discoveries for the development of medicine or medical equipment remains unchanged. Gaining clear consent places a considerable burden on medical sites, and even where there may be no intention to violate rights, it must be considered that this is significantly more problematic than with the previous law. In the revised law, the concept of anonymized information has been introduced, and by anonymizing data in accordance with the standards of the PPC, this may be provided without consent under certain conditions.

However, it is necessary to impose the conditions of prohibition on reidentification and safety management on the recipient of the data, and it is procedurally complex to make third-party provision with anonymized information a public duty. Additionally, to meet the anonymization standards of the PPC, a certain amount of information processing capability is required, which is not simple. While this is not a legal item, in regard to important information, another feature of the revised law is that traceability must be secured. Moreover, although a significant impact is feared in healthcare and nursing care fields where there is frequent provision to third parties, in the healthcare and nursing care guidance, this is virtually all considered as essential for healthcare and nursing care, thus avoiding a major increase in the workload of the healthcare and nursing care institutions. On the other hand, in regard to provision to third parties involved in secondary use that is not essential for healthcare and nursing care, the creation of records and their confirmation at the time of receipt are required. For genetic information, as well, having a personal identifier code specified from which

the individual can be identified, provided certain conditions are met, has immense medical significance.

The points alluded to earlier are all features of the previous and revised law as seen from the perspective of the healthcare and nursing care field. Furthermore, as the responsibility for enforcing the revised law has been centralized in the PPC and penalties have been significantly increased, it has become more effective. Major changes, such as the conditions for distributing personal information overseas being clarified, have been determined, but these will only be listed in this chapter.

The revised law promises to improve several issues in the previous law. The strengthening of punitive measures increases its effectiveness, and the introduction of the concept of important information reduces discrimination based on the illegal use of special personal information, preventing its use through provision to a third party not intended by the person concerned. However, several issues remain unresolved. The first of these is that, as operations are based on different regulations from the government, independent administrative bodies, local governments, and private enterprises, there are about 1,800 autonomous bodies and close to 2,000 statutes. Certainly, there are not any major differences in their basic thinking, but the executing body varies depending on the statute, and subtle decisions are made by each executing body. In the case of healthcare and nursing care, the body is a private company, but the local government and institutions at its rank often contribute, as do national institutions and independent administrative bodies.

For example, if one prefectural, two city, and two town hospitals and five private medical institutions collaborate to share organic patient data, it will be necessary for at least four autonomous bodies to review whether this is possible. For healthcare providers looking to move forward, this can become a significant burden. Currently, the fact that the statute may be different depending on the acquiring body has not been improved at all. For hereditary information, it is expected that genetic information will be specified with a personal identification code and handled prudently; however, under the Personal Information Protection Act, consent gives an absolute pardon. On the other hand, in the case of hereditary information, even if the person providing the information provides consent, the impact of such may extend to blood relatives such as parents and offspring. If, as a result of a parent's consent, a child became the victim of discrimination, this could not be handled under the Personal Information Protection Act. At the current time, improving this point seems to be not possible, and several people indicate that this is an issue. This should be reviewed in the near future, and it is to be hoped that it will be resolved quickly.

6.1.1.2 Review and Establishment of the Next-Generation Medical Infrastructure Act

As previously described, with the revisions to the Personal Information Protection Act, although several of the issues in the previous law were improved, secondary use, where there is no intention to violate personal rights and the aim is to use personal information for the public good, was previously possible via opt-out. However,

with the revisions, this is no longer possible. Healthcare and nursing care must be performed based on medicine, but this cannot develop without the use of patient and user data. Immediately utilizing research results obtained using the laboratory or animals in medicine and healthcare is not possible, and human knowledge is essential. In other words, if this type of usage is suppressed, the acquisition of medical knowledge itself will likely be suppressed, and this may obstruct the development of healthcare and nursing care itself. If medical institutions and nursing care providers are able to anonymize, they will be able to provide data for secondary use without consent. However, in the case of regional comprehensive care and collaborative medicine, information is distributed between multiple operators. Therefore, unless information can be concentrated in a single institution through a joint use declaration, linking and anonymizing the disparate information will not be possible. Anonymization makes reidentification impossible, and so anonymized information cannot necessarily be linked. The simple solution would entail making a joint declaration of use; however, in this case, the perimeter of information for joint use and other information must be clarified. In Japan, the healthcare and nursing care services can be freely chosen by patients and users, so setting the perimeter is essentially difficult. Additionally, it is necessary to announce the fact that anonymization is taking place, and this cannot be provided without restriction. A prohibition on reidentification is sought from the recipient, and although this is effort based at best, safety management is also required. The provider has no duty to supervise the destination, but if an incident or illegal use occurs, the complaint from the individual embodying the information will be directed at the providing medical or nursing care institution, which may result in a civil lawsuit. Supervision may be considered to be mandatory. Although it is not impossible, some preparedness and effort are required. However, it is not desirable that this situation impacts the development of medicine/medical equipment or drug discovery. Regarding academic research, in Chap. 4 of the revised law, it states that although various duties are not placed on the operators acquiring the personal information, this is limited to academic research by academic research institutions, and although there are calls to draw up and execute guidelines for those not covered by Chap. 4 of the revised law, such guidelines are difficult to implement on a statutory basis.

Faced with this situation and the awareness of the need to promote use for the public good that does not violate the rights of the individual, the Cabinet Secretariat and Office of Healthcare Policy primarily reviewed the measures, whereupon the Next-Generation Medical Infrastructure Act was submitted to the Diet as Cabinet legislation. The basis of this held that if operators with the ability to perform reliable and safe anonymization, who were able to provide safe information for the public good in a broad sense, were accredited and medical information was provided from the accredited operators to the medical institutions, consent could be provided via an opt-out system. This was established at the end of April 2017 and delivered in May.

6.1.1.3 Content of the Next-Generation Medical Infrastructure Act

This law focuses on accredited anonymizer medical data creation operators, and as previously described, operators who can perform anonymization reliably and handle and provide information safely are accredited by the government. The law intends, "through the safe and appropriate utilization of anonymized medical data, to promote cutting-edge R&D related to health and healthcare, and new industries, and contribute to the development of a society where people live healthy and long lives." The aim is not simply commercial use but use for the public good in a broad sense. Although its scope is narrow, it is positioned as an individual law from the Personal Information Protection Act, and this overwrites such Act.

Definition of Wording

This law is not aimed at general personal information but "medical data." The main target is the information related to healthcare, which is a type of important information, and the definition has been slightly expanded. The Personal Information Protection Law covers the information of living individuals, but the Next-Generation Medical Infrastructure Act includes medical information on deceased people as well. In healthcare, life and death exist consecutively, and so, this can be considered to be a reasonable extension. Additionally, in the revised Personal Information Protection Act, the guidelines for anonymization are indicated by the PPC, whereas the guidelines for the anonymizer medical data in the Next-Generation Medical Infrastructure Act are provided by the minister in charge. However, the wording in the definition is the same, and the law clarifies that this should be determined after consultation with PPC. Anonymizer information and anonymizer medical information are basically the same, but with the latter, it is possible to provide detailed guidelines depending on the case in which it is used.

Accredited Anonymizer Medical Data Creation Operators

The core of this law is the stipulation of accredited anonymizer medical data creation operators. This is limited to companies who possess appropriate anonymizer capabilities and can provide information to operators who can handle the safe anonymizer information in accordance with the law. The anonymizer work of such operators does not apply to stipulations regarding the creation of anonymizer information in Article 36 of the Personal Information Protection Act. Additionally, the safe management of information and an appropriate response to this are required. This also does contravene the concept that this is provided to contribute to R&D in the medical field, and use that exceeds the scope of achieving the objectives of the accredited operator is not recognized.

With this, no particular restriction is noted on the operator other than the accredited work. Additionally, provided that the information before the anonymization was for the operator to create the anonymizer medical data, it may be provided to other accredited anonymizer medical data creation operators within the scope of that purpose. In this way, if, for example, accredited operator A is mainly accumulating hospital information and operator B is mainly collecting clinic information, it is pos-

sible for A to provide to B and B to provide to A and create anonymizer medical data after linking the medical information of the clinics and hospitals.

Operators Handling Medical Data
This refers to medical institutions, and broadly speaking, two types of regulations when providing medical data to accredited anonymizer medical data creation operators are described. The first is notice to the patients and notification to the minister in charge. In the notice, it must be clarified that provision shall be stopped if there is a request for such from the patient or a bereaved family member. A point to note here is that this is just described as "notice" to the patient. Simply presenting it is not enough, and the content of the notice must be actually notified to the patient, etc. The second point is that if provision is stopped due to the request of the person concerned or the bereaved family, there is a duty to issue evidence in writing that there has been a request to stop provision, and a copy of this must be stored. In case there is a request to stop the provision of medical data owned by the accredited anonymizer medical data creation operator, this information may not be received.

Operators Handling Anonymizer Medical Data
Recipients provided with anonymizer medical data from the accredited anonymizer medical data creation operators are exempt from the stipulations of Articles 37 (provision of anonymizer information), 38 (prohibition on identification action), and 39 (safety management measures, etc.) of the Personal Information Protection Act. On the other hand, in the Next-Generation Medical Infrastructure Act, reidentification itself is prohibited. This should not just be a penalty stipulation for "operators handling anonymizer medical data," and the restriction of agreements with accredited anonymizer medical data creation operators is also necessary. If an actual breach occurs or the agreement conditions lack effectiveness, the application of the Unfair Competition Prevention Act should also be considered.

Accredited Medical Data Handling Contractors
Operators undertaking the work of accredited anonymizer medical data creation operators need to be accredited by the government.

6.1.1.4 Opting-Out Under the Next-Generation Medical Infrastructure Act

Provision to third parties is specified with opt-out under Article 23, paragraph 2, of the Personal Information Protection Act. When providing to a third party after notifying the party concerned or in a situation where the person concerned could easily learn of the fact, provided that there is no motion of refusal from the person concerned, it may be provided to a third party. Originally, this was prohibited in cases involving sensitive information, and so, medical data cannot be provided to a third party in this way. In contrast, in the Next-Generation Medical Infrastructure Act, as long as the third-party provision is to an accredited anonymizer healthcare information creation body, an exception shall be granted, and this may be provided

in an opt-out form. However, it is only permitted to be provided to a third party after notifying the person concerned if there is no motion of refusal. In other words, it just being a situation where they could easily learn of the fact is not enough.

6.1.1.5 Safety Management Measures

The safety management measures section stipulates the safety management measures to be taken by accredited anonymizer healthcare information creation bodies, and the contents are as follows:

1 Purpose and target of application
2 Concrete measures
2-1 Organizational safety management measures
2-2 Human safety management measures
2-3 Physical safety management measures
2-4 Technical safety management measures
2-5 Other measures

These can be considered to be typical chapter headings, and the majority of these are not particularly different from the MHLW "Security Guidelines for Medical Information Systems." However, the network is limited to dedicated lines and IP-VPNs within accredited operators. In terms of availability, while the superiority of dedicated lines is unquestionable, as they are not clearly superior in regard to completeness or anonymity, implementation may be difficult when cost is considered.

6.1.1.6 The Future of the Next-Generation Medical Infrastructure Act and Issues

If the Next-Generation Medical Infrastructure Act functions as intended, the regulations strengthened in the revised Personal Information Protection Act can be introduced in a form without the risk of violating the rights of individuals. Moreover, regarding the purpose restricted to R&D in the medical field, there are expectations that it can be promoted in a safe and significant manner. However, two main issues are identified. The first is the establishment of the system itself. Although the law has been established, we are still waiting for the establishment of a basic policy as well as the government and ministerial ordinances delegating the main part of this work. At present, only the outline has been fixed. We do not yet have a system with "meat on the bones" to be used in actual operation, and efforts by all related parties are required. Additionally, it is expected that there will be public comments once a draft of the government and ministerial ordinance or guidelines is determined, and hopefully, several people will have constructive comments from many people. The second issue is that although the accreditation of anonymizer medical data creation operators is a public work contributing, in a meaningful way, to R&D in the medical field, the law presumes the accreditation of private operators. In other words, the accredited operators must both maintain their own survival and continue work with

significant public work elements. This is certainly not simple for a private operator. Unless the accredited operator can gain trust, the medical institutions, etc., will lose enthusiasm for the provision, and the system itself may fail. If we consider the world aimed for by this law to be significant, the support of not only the administration and operators aiming for accreditation but also a wide range of people, including medical-related parties and patients, is required.

Even if the aforementioned issues can be safely overcome and operations begin, problems will remain. We have repeatedly indicated that this framework is based on accredited anonymizer medical data creation operators who are private companies. However, at present, the government, local governments, and insurers are systematically accumulating information, and much of the useful medical data is owned by the government. For example, information on life and death is the ultimate outcome of treatment, and to determine this outcome with certainty, basic resident registration information and death certificates, etc., must be accessed. Although according to this law, cooperation on consent with accredited anonymizer medical data creation operators is possible, there is no consideration at all regarding collaboration on the information owned by the government, local governments, and insurers under this system. Despite the fact that R&D in the medical field is urgent in terms of maintaining social security, it must be indicated that there is a problem in terms of efficiency, based on the Next-Generation Medical Infrastructure Act alone. It is considered necessary to establish an external system to promote a comprehensive system for using information for the public good for government and private enterprise. Additionally, the security and anonymizer standards are somewhat abstract. In the case of technology that uses individual data with Privacy Preserving Data Mining and multiparty protocols in its anonymous form for calculation purposes only, despite the fact that it has been demonstrated that a technical solution is possible, as no consideration has been given from a statutory or system viewpoint, it is difficult to judge whether this can be used under the Personal Information Protection Law. While promoting technical initiatives, it is also necessary to clarify positioning in a statutory and system sense.

6.1.2 Ethical Guidelines and Anonymization of Medical Information

6.1.2.1 Ethical Guidelines

The Ethical Guidelines for medical and health research involving human subjects [4] apply to medical research for human beings and basically requires researchers to respond to the request sought by the Act on the Protection of Personal Information. However, Chap. 4 of the Personal Information Protection Law shown in the following is exempt from clinical research:

Chapter IV Obligations, etc., of a Personal Information Handling Business Operator
Section 1 Obligations of a Personal Handling Business Operator
Section 2 Obligations of an Anonymously Processed Information Handling Business Operator, etc.
Section 3 Supervision
Section 4 Private Sector Body's Promotion for the Protection of Personal Information

It also applies to the information infrastructure for collecting and analyzing medical information that this project aims to build. In large-scale data collection research, specific responses required by the Ethics Guidelines are mainly described in the informed consent. The description in the Ethics Guidelines is as follows:

> Researchers do not necessarily need to receive informed consent however, if you do not receive informed consent, the subject of the study appropriate consent of the However, in cases where it is difficult to obtain appropriate consent, information used in other studies to conduct research. from 4(1) to (6) for the implementation of the research, if there is a particular reason for to be notified or published to the subject of the research, and to ensure that the research is carried out or continued. opportunities for research subjects, etc., to be denied. personal information may be used.

Also, the Ethical Guidelines need to notify or publish the following matters to the patient, etc:

> (1) The purpose of use and use of samples and information (including methods when provided to other organizations)
> (2) Items of samples and information used or provided
> (3) Scope of use
> (4) Name or name of the person responsible for the management of samples and information
> (5) To use samples and information to identify the subject of the research or to other research institutions at the request of the research subject or its agent
> (6) (5) How to accept the request of the subject or its agent

This is also true for information systems that deal with large-scale data. Furthermore, if the target personal information is the anonymized one, it is not necessary to notify the patients. At this time, the opportunity of the consent withdrawal is not guaranteed to the patient.

In fact, regardless of the presence or absence of anonymization processing for electronic medical record (EMR) items, in most cases, the content of research that uses medical information is made public on the homepage of each medical or research institution and patients. It is difficult for patients to understand how their own EMR items are used and provided.

6.1.2.2 Anonymizer Medical Data

Anonymizer medical data is an extension of the anonymizer information under the Personal Information Protection Act. Under this Act, the target information is limited to personal information surviving, while under the Next-Generation Medical Infrastructure Act, the information of deceased individuals may also be covered, depending on the situation. Additionally, anonymizer information is information from which the

Table 6.1 Contents of guidelines for anonymizer information

1 Positioning of these guidelines
2 Definition
2-1 Medical information
2-2 Anonymizer medical data (related to Article 2, paragraph 3)
2-3 Anonymizer medical data creation business
3 Duties of accredited anonymizer healthcare information creation bodies and bodies handling anonymizer healthcare information
3-1 Thinking behind duties regarding the handling of anonymizer medical data
4 Processing required when creating anonymizer medical data
4-1 Processing standards for anonymizer medical data
4-1-1 Deletion of descriptions, etc., from which specific individuals may be identified
4-1-2 Deletion of individual identification codes
4-1-3 Deletion of codes that interconnect information
4-1-4 Deletion of peculiar descriptions, etc.
4-1-5 Other measures based on the nature of medical information databases
4-2 Items requested for investigation when creating anonymizer medical data
4-2-1 Format for using anonymizer data
4-2-2 Possibility of identification by referring to other information
4-3 Anonymizer medical data creation process
4-4 Method of anonymization based on medical data categories
4-5 Medical data-specific anonymization
4-5-1 Medical images
4-5-2 Genome data
5 Safety management measures, such as anonymizer medical data
6 Prohibition on identification actions
7 Provision registration

individual cannot be identified by ordinary people, whereas anonymizer medical data is information from which the individual cannot be recognized by general healthcare-related people. As this fulfills the stipulations of the guidelines on anonymizer information determined under the Personal Information Protection Act, processing based on additional risk analysis is required. Furthermore, another characteristic can be considered to be the fact that even after the provision of the information, there is a duty to follow up, including confirming how it is used. The content of these guidelines is shown in Table 6.1.

Another feature is that medical information is categorized from a risk perspective, which is shown in Tables 6.2 and 6.3.

Table 6.2 Categorization of the risk of individual identification in medical data

Category	Overview
Identifier	Information directly linked to an individual (name, number of insured, etc.)
Quasi-identifier	Information that, when multiple types are combined, can lead to the identification of the individual (date of birth, organization, etc.)
	*The medical institution code is considered a quasi-identifier
Static attributes	Highly invariant information (height, blood type, allergies, dates [such as consultation dates], etc.)
	Information related to external characteristics such as disabilities
	*Handling of information on chronic illnesses with a high level of invariance needs to be reviewed
Semi-static attributes	Data with universality for a fixed period (weight, etc.)
	It is assumed that this relates to information on diseases, procedures, administered medicines, etc.
Dynamic attributes	Information that is constantly changing (data on inspection values, food, other treatment, etc.)

Table 6.3 Anonymizer examples through the categorization of medical data

Category	Example of anonymization method
Identifier	Deleted or irreversible pseudonymization
Quasi-identifier	Generalization (date of birth -> year born, address -> prefecture) or micro application that satisfies k-anonymity
	Delete data items
	Add attributes (geographical, scale, etc.), such as medical institution codes, and convert codes into an unidentifiable form
Static attributes	Numerals are top-to-bottom coding
	Generalization or micro application
	Generalization or offset based on treatment date, etc.
Semi-static attributes	Numerals are top-to-bottom coding
	Delete sensitive diseases when not necessary
Dynamic attributes	Anonymization not required, but where necessary, numbers are top-to-bottom coding
	In consideration of the significance of abnormal values, look at the distribution of values and carry out processing, such as rounding of upper and lower % values

6.1.3 Standardization of EMRs

The Standardized Structured Medical Record Information Exchange (SS-MIX) [5, 6] aims to promote/develop the results of the standardized electronic medical chart information exchange system development commission project conducted by the Health Policy Bureau of the MHLW in FY 2006 in Japan.

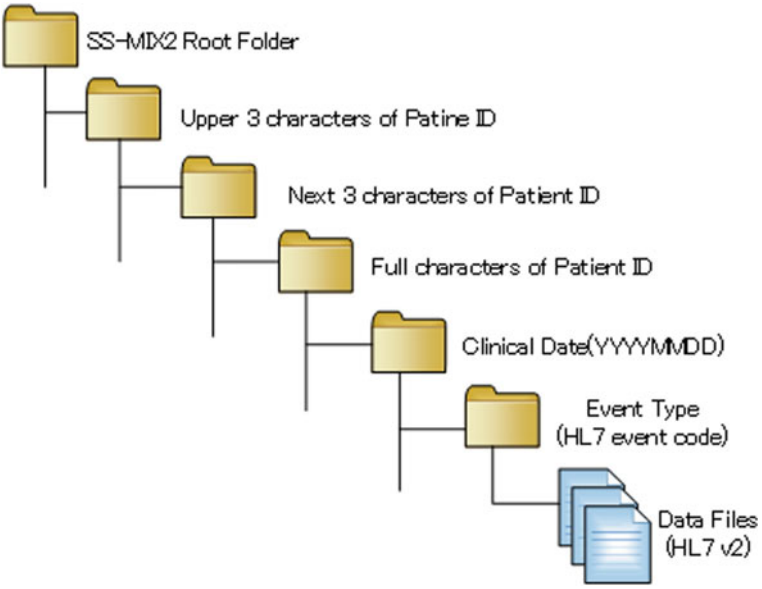


Fig. 6.1 Overview of SS-MIX2 directory layout

SS-MIX includes the following:

(1) Hospital information system information gateway telegraphic message specification
(2) "Standardized Storage Specification" directory structure
(3) Electronic medical information CD and patient referral document CD specification

Furthermore, the scenes where the utilization of this standardized storage is expected are as follows:

Ensuring continuation of medical information Repository in community healthcare coordination Information sharing among multiple vendors Utilization as backup information

Figure 6.1 shows the file system layout of SS-MIX2 storage. Directories are sorted by patient ID, clinical date, and event type.

Table 6.4 shows the clinical event types covered by SS-MIX2 storage represented by HL7 v2. We can represent 30+ clinical events using this storage [6].

6.2 Medical Test Bed Concepts and Requirements

Considering the current situation surrounding medical information as described earlier and the development of future utilization, the public cloud is used for the secondary use of medical data scattered through medical institutions across the organization. We are developing a secure information utilization base test bed in the medical information field, assuming the utilization promotion by adopting.

Table 6.4 SS-MIX2 data types

No	Data type	Name	HL7 message type
1	ADT-00	Update of patient's basic information	ADT^08
2	ADT-00	Deletion of patient's basic information	ADT^23
3	ADT-01	Change of investigator	ADT^54
4	ADT-01	Cancellation of investigator	ADT^55
5	ADT-12	Reception of outpatient physical examination	ADT^04
6	ADT-21	Hospitalization plan	ADT^14
7	ADT-21	Cancellation of hospitalization plan	ADT^27
8	ADT-22	Conduct of hospitalization	ADT^01
9	ADT-22	Cancellation of conduct of hospitalization	ADT^11
10	ADT-31	Conduct of staying outside	ADT^21
11	ADT-31	Cancellation of conduct of staying outside	ADT^52
12	ADT-32	Conduct of return from staying outside	ADT^22
13	ADT-32	Cancellation of conduct of return from staying outside	ADT^53
14	ADT-41	Plan of change of department/building (change of room/bed)	ADT^15
15	ADT-41	Cancellation of plan of change of department/building (change of room/bed)	ADT^26
16	ADT-42	Conduct of change of department/building (change of room/bed)	ADT^02
17	ADT-42	Cancellation of conduct of change of department/building (change of room/bed)	ADT^12
18	ADT-51	Plan of discharge	ADT^16
19	ADT-51	Cancellation of plan of discharge	ADT^25
20	ADT-52	Conduct of discharge	ADT^03
21	ADT-52	Cancellation of conduct of discharge	ADT^13
22	ADT-61	Registration/update of allergy information	ADT^60
23	PPR-01	Registration/update of disease name (history) information	PPR^D1
24	OMD	Food order	OMD^03
25	OMP-01	Prescription order	RDE^11
26	OMP-11	Prescription conduct notice	RAS^17
27	OMP-02	Injection order	RDE^11
28	OMP-12	Injection conduct notice	RAS^17
29	OML-01	Specimen examination order	OML^33
30	OML-11	Specimen examination result notice	OUL^22
31	OMG-01	Radiological examination order	OMG^19
32	OMG-11	Notice of radiological examination conduct	OMI^23
33	OMG-02	Endoscopy order	OMG^19
34	OMG-12	Notice of endoscopy conduct	OMI^23
35	OMG-03	Physiological examination order	OMG^19
36	OMG-13	Notice of physiological examination result	ORU^01

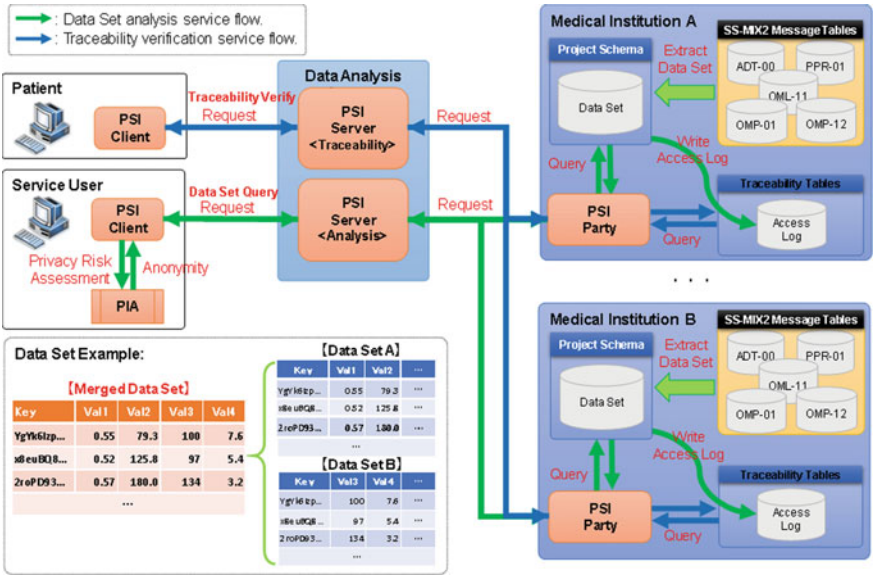


Fig. 6.2 Overview of the system developed for secure data collection and analysis

The main points in the development of a medical test bed are as follows:

- Unnecessary sensitive information not used for research should not be leaked outside the medical institution.
- Securely extract and combine medical information across organizations.
- Information extraction control linked with patient consent is possible.
- Privacy risk can be evaluated for the extracted dataset.
- Patients can verify the history of information utilization on the platform.

Figure 6.2 presents an overview of the developed system [7]. The key concepts are the following:

1. Each medical institution has EMR data in SS-MIX2 storage, including billions of HL7 v2 messages.
2. HL7 v2 messages are periodically parsed and stored to relational database management system (RDBMS) tables, maintaining synchronization with the billions of message files in SS-MIX2.
3. Analysis requests from researchers and data collection are managed by the private set intersection (PSI) service on the cloud, which communicates with a client agent located at a client terminal and PSI agents located at each medical institution.
4. Target data criteria, such as diseases, age, and gender, must be defined before the PSI executes data collection. The PSI party agent deploys the target dataset in advance from the local RDBMS to memory.

5. Data collection is achieved using PSI software, which is based on Bloom filter technology for record verification across institutions. The application of bloom filter technology is aimed at realizing data matching in which personal information does not leak outside each hospital during the data collection process.
6. The collected dataset can be verified considering the possibility of patient identification using the extracted attributes.
7. Patients can trace the use of their medical records during data collection.
8. If they choose, patients can withdraw consent for the secondary use of their data. Consent withdrawal information is assumed to be an input to existing the EMR system and exported SS-MIX2 storage in each hospital.

6.3 Features and Implementations of Secondary Use Infrastructure Development

This section describes key features and implementation details of our developed test bed for medical field.

6.3.1 SS-MIX2 Standardized Storage

6.3.1.1 Objective

In Japan, SS-MIX2, which is the domestic standard of exporting whole EHR data as HL7 v2 message files to the external storage for the purposes of backup, regional collaboration, disease repository, and others, is common. In this standard, EHR data is exported to a storage in a directory structure using patient id and clinical date and event type. Therefore, the use of the exported storage for cross-patient analysis such as epidemiological studies is challenging. We are applying an RDB-based virtual file system technology to the storage to achieve cross-patient/cross-institution analysis without collecting data files.

6.3.1.2 Methods

The overview of the system is shown in Fig. 6.3. The storage is developed based on Filesystem in Userspace (FUSE), a virtual file system technology. We adopted pgfuse and PostgreSQL as the FUSE and RDBMS, respectively. The recorded HL7 messages are stored to the DB tables as BLOB data, and the RDBMS traces the transaction in real time. The HL7 messages are parsed by PL/SQL, and parsed medical records (HL7 segments, fields) are recorded to user-defined tables in the RDBMS. Parsing tasks are intended to be executed periodically. Once the records have been stored to

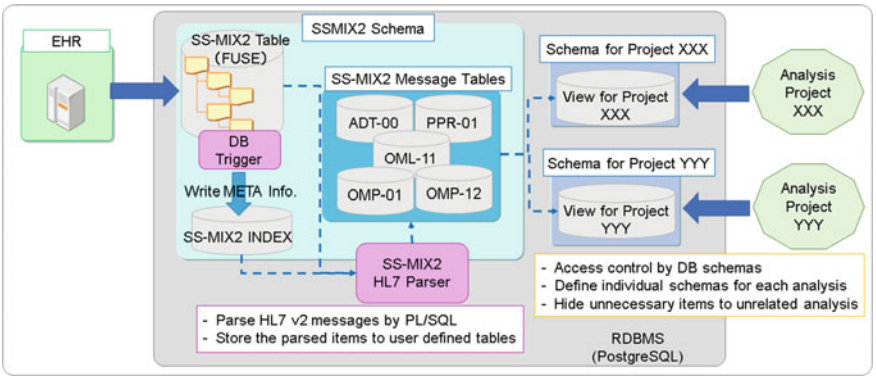


Fig. 6.3 Overview of the developed storage with a virtual file system

Table 6.5 Evaluation results (s) for various numbers of medical records

COPY (VFS to HDD)	COPY (HDD to VFS)	Delete (VFS)	Parse files (VFS)
3,261 s	1,466 s	775 s	1,989 s
5.03 Mbps	2.26 Mbps	9.52 Mbps	54.9 files/s

the tables, the minimum required items can be queried through individually applied view schemas according to the purpose of each analysis project. Performance tests are executed with dummy messages of 109,174 files (922 MB in total, 1,689 patients) including 27 clinical events defined by SS-MIX2 standard, such as ADT-00, OMP-01, and OML-11.

6.3.1.3 Results

Performance test results are shown in Table 6.5. All types of messages could be parsed by PL/SQL. Based on the performance, this storage can process the daily generated medical records of our hospital in less than 2 h.

6.3.1.4 Discussion

The developed storage enables the rapid cycle for the secondary use of medical records analysis among institutions and also prevents the disclosure of unnecessary patient information to each analysis by the regulations of applying view schemas for queries. Moreover, using the developed storage, exported medical records and parsed result tables can be more easily backed up to a remote place in real time using DB replication technology, compared with synchronizing the enormous number of files.

6.3.1.5 Summary

This section describes the development of a standardized storage for the purpose of cross-patient/cross-institution analysis based on the domestic EHR data exporting standard. We will try to develop a secure data collection infrastructure assuming the distributed environment of the developed storages.

6.3.2 Secure Collection of Distributed Medical Information

In this section,[1] we propose an alternative method of collecting and storing EMR data, wherein only necessary items are included in collected data, eliminating the need for individual identifiable information to spread outside the medical institution. The system facilitates EMR data distribution within each medical institution, enabling cross-patient or cross-facility data collection and analysis. The PSI library developed by Miyaji [8] is used for the data integration and encryption of the extracted EMR data. This paper aims to provide an overview of the system and its major technical elements and evaluate the transaction performance of data extraction and collection from the distributed SS-MIX2 storage.

6.3.2.1 Methods

Experimental Environment
The transaction performance of data extraction and collection from the distributed SS-MIX2 storage was evaluated using an experimental environment comprising a server (PSI Server), three data stores (PSI Party), and a client (PSI Client). The Server and Party machines were deployed as VMware ESXi virtual machines. The PSI Client can be deployed on any machine that can run Java.

Experimental data were virtually produced by anonymizing laboratory test result data in the SS-MIX2 storage exported from the EMR system of The University of Tokyo Hospital (Tokyo, Japan). Storage assumed to have 10% overlap between each node was arranged and used for the evaluation tests. The hash value of the character string combining the patient's name, date of birth, and sex was used as the key attribute of each record for the bloom filter.

[1] This section is reprinted from "Studies in Health Technology and Informatics, Vol 255, Katsuya Tanaka, Ryuichi Yamamoto, Kazuhisa Nakasho, Atsuko Miyaji, Development of a Secure Cross-Institutional Data Collection System Based on Distributed Standardized EMR Storage, pp. 35–39," Copyright (2018), with permission from IOS Press. The publication is available at IOS Press through http://dx.doi.org/10.3233/978-1-61499-921-8-35.

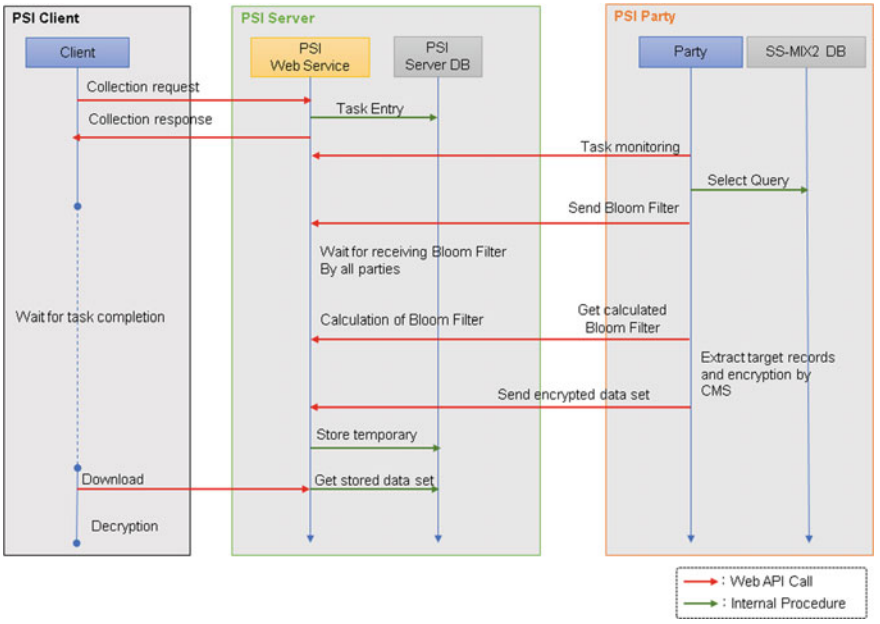


Fig. 6.4 Overview of the transaction flow during data collection

Data Collection with PSI

Figure 6.4 presents an overview of the transaction flow during a secure data collection using the system. The entire system was designed as a Web service so that in the future the service could be available via a commercial cloud. The PSI application programming interface was developed in Java using SOAP Web services and deployed on an Apache Tomcat. All Web communications were implemented with client authentication under TLS 1.2. Extracted EMR data are encrypted by Cryptographic Message Syntax and can be decrypted only by the user requesting the collection.

6.3.2.2 Results

Table 6.6 summarizes the evaluation test results for data queries for calculations, bloom filter calculations, and result data extraction for increasing numbers of EMRs. The processing time linearly increased with the number of records.

Table 6.6 Evaluation results (s) for various numbers of medical records

Records	20,000	40,000	80,000	160,000	320,000	640,000	960,000	1,280,000
Query data	0.1	0.2	0.2	0.3	0.6	1.0	1.5	2.0
Bloom filter processing	0.5	0.5	1.0	2.0	2.9	7.4	13.5	20.7
Data extraction	3.3	3.0	3.1	3.4	4.0	10.1	15.4	23.5
Total	3.8	3.7	4.3	5.7	7.4	18.5	30.4	46.2

6.3.2.3 Discussion

Significance of the System

The system was completely achieved using Web service architecture with the encryption of the extracted EMR data, indicating that medical institutions participating in research would not need to maintain a secure connection to the specific service provider if the developed PSI services are operated on the commercial cloud. The encryption of EMR data avoids any disclosure of the extracted information to the cloud service providers. Furthermore, because the infrastructure makes it unnecessary to connect an EMR storage to the Internet, this eliminates the possibility of experiencing network attacks to the data storage. To meet the requirements of a given analysis, the PSI can execute not only intersection operations but also union operations on distributed datasets.

Performance

The experimental results showed that an intersection operation involving approximately 1 million records was completed within a minute. With this level of processing performance, there should not be any problems with actual operations. We now intend to verify this with larger datasets.

Future Work

The remaining issues for development include (1) the management of consent information, (2) risk assessment for the extracted dataset, and (3) traceability management against data collection. The first issue can be addressed by scanning paper-based consent information related to patients opting-out of the secondary use of their data and storing the scanned data files to the SS-MIX2 storage. We intend to represent consent information as XML files, such as HL7 CDA Privacy Consent Directives, Release 1 [9]. The other two issues are under discussion.

6.3.2.4 Summary

This section describes the underlying concepts and implementation of a secure data collection infrastructure with distributed standardized EMR storage. Using the PSI data collection technology, the experimental results demonstrated high performance. A few issues remain for future implementation.

6.3.3 *Privacy Risk Assessment of Extracted Datasets*

6.3.3.1 Overview

This section describes a prototype of a Web service that enables a series of operations to perform privacy risk evaluation against a dataset extracted from multiple storages by the PSI service developed.

6.3.3.2 Method

The PSI and privacy impact assessment (PIA) libraries are applied using SS-MIX2 standardized storage that adopts FUSE, one of the virtual file systems developed so far. As a FUSE, pgfuse corresponding to PostgreSQL was adopted. Assuming that a service for finally collecting data safely will be operated in the public cloud, the server and client, which will be the nodes of the data collection infrastructure based on the PSI and PIA libraries, are configured as Web services using SOAP. The configuration of the experimental system is shown in Fig. 6.5.

The data for verification was constructed by virtually distributing the HL7 v2 format data obtained by anonymously processing the SS-MIX2 standardized storage data held by The University of Tokyo Hospital to three storages and constructing a virtual multi-facility environment. Storing 1 million specimen test result messages for each storage, creating a dataset using the PSI library, and developing a user interface that can apply the extracted dataset to the risk assessment function in a one-stop manner. In addition, patients between SS-MIX2 standardized storages were artificially adjusted with 10% duplication as a count of the patients.

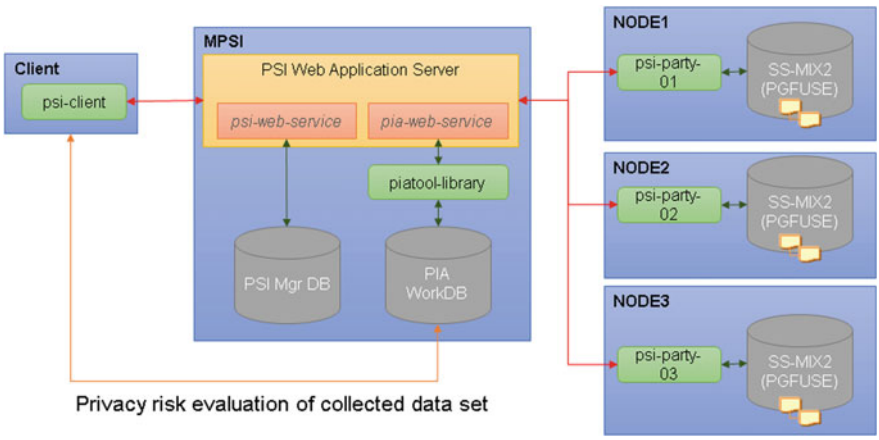


Fig. 6.5 Overview of experimental settings for privacy risk assessment

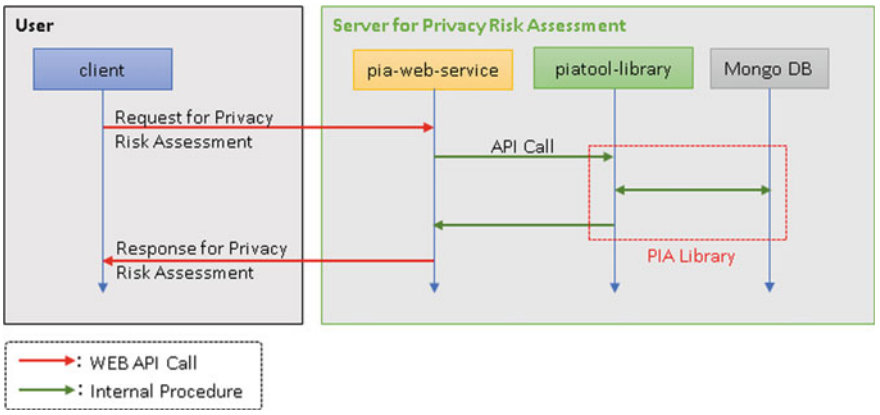


Fig. 6.6 Operation flow of the developed privacy risk assessment service

Fig. 6.7 Overview of the developed GUI for privacy risk assessment

The privacy risk evaluation function is configured as a separate Web service and positioned so that it can be operated as data extraction processing by PSI and data processing after acquisition. The system operation flow of the risk evaluation function is shown in Fig. 6.6. While checking the maximum and minimum values and the number of data in each data item of the extracted dataset on the screen, top and bottom coding and generalization processing (processing of numerical data with the specified division accuracy) were constructed.

Figure 6.7 shows the user interface for evaluating privacy risk developed in our project. The dataset extracted by the PSI service can be read, and the maximum and minimum values of each data item can be confirmed. For numerical data, processing

can be performed by specifying the upper and lower limits and division unit. In addition, it is possible to calculate the degree of overlap after processing the target dataset using the attribute value group specified on the screen and have an interface for confirming an index for privacy risk evaluation.

6.3.3.3 Results and Discussion

The privacy risk evaluation service operates with a response of up to several tens of seconds when numerous attribute values are specified for a post-extraction dataset with a scale of 100,000. Although there is no problem in performance, it is a configuration in which functions are centrally arranged on the service side regarding the processing of numerical data and evaluation of redundancy, and it does not function unless the original data is exposed to the service side. From the viewpoint of data concealment, it remains a problem, and the functional layout needs to be reconsidered.

6.3.4 Secondary Use and Traceability

This section describes how to implement the capability of traceability in the developed system for secure data collection and analysis.[2]

6.3.4.1 Objective

This section describes how to implement the capability of traceability in the developed system for secure data collection and analysis. Blockchain technology has been recently applied in healthcare fields, including primary patient care, data aggregation for research purposes, and connecting healthcare providers [10–12]. The system that we are developing has a second purpose: to secure the traceability of EMR data, methods to disclose the logs of secondary use are needed. In the present situation, where patients do not have any common ID, it is difficult for a patient to audit all the secondary use logs across the distributed hospital storages that he/she visited. By blockchain technology, we expect to provide patients a common search infrastructure with immutable secondary use logs. Thus, we plan to apply blockchain technology to the aggregation of data extraction log records. This method has several possible implementations, and they must be evaluated assuming operations in real use.

[2]This section is reprinted from "Studies in Health Technology and Informatics, Vol 264, Katsuya Tanaka, Ryuichi Yamamoto, Assessment of Traceability Implementation of a Cross-Institutional Secure Data Collection System Based on Distributed Standardized EMR Storage, pp. 1373–1377," Copyright (2019), with permission from IOS Press. The publication is available at IOS Press through http://dx.doi.org/10.3233/shti190452.

The following experimental results mainly concern data structure and transaction performance compared with traditional implementation for achieving the aggregation of distributed log records of EMR data extraction.

6.3.4.2 Methods

Traceability for Patients

EMR storage for the developed secure data collection system is supposed to process queries from clinical researchers using the standard interface implemented by the PostgreSQL database. EMR data are extracted by data extraction requests handled by the PSI service. Thus, the selected records are identifiable based on each query result, and the records represent the disclosure history of EMR data during data collection through the use of the developed PSI service. By making the log record of extraction searchable by patients, we suppose that traceability in the secure data collection system will be achieved. However, because storage is supposed to be distributed at each hospital, log records must be aggregated by some secure method to be made auditable.

Log data is assumed to be represented by a combination of the following attributes:

1. Identifier of target patient (patient identifier)
2. Storage source (medical institution identifier)
3. Disclosed destination (extracting user identifier)
4. Purpose of use
5. Type of extracted EMR data
6. Extraction timestamp

Attribute 1 (patient identifier) is mandatory for patient identification. In Japan, at present, universal patient identifiers are not available. We assume that insurance numbers may be desirable for searching log records across medical institutions because the patient ID at one medical institution is only applicable for searching log records at that medical institution.

Attribute 2 (medical institution identifier) is used to distinguish the institution storing the extracted EMR data.

Attribute 5 (type of extracted EMR data) is represented by HL7 v2 message types such as “ADT-00,” “OMP-01,” and “OML-11.”

Attributes 3, 4, 5, and 6 are used to distinguish the secondary use of target EMR data by patients. By verifying these attributes, patients can determine whether actual secondary uses meet their consent.

Data Structure for Query

A query for EMR storage may extract the records of several patients at one time. For disclosing extracted history to patients, the extracted history should be sorted by patient, and each history should include the aforementioned attributes.

Table 6.7 Sample data representing extraction history

```
{
 "patientID":"781e5e245d69b566979b86e28d23f2c7",
 "insuti-tionID":"aabd258c8894b996e8d8561fa868364d",
 "disclosedDestination":"AnalysisUser001",
 "purposeofUse":"DrugDevelopment",
 "typeofRecords":"OMP-01",
 "extractionTime":"2018/11/12 01:23:45"
}
```

By focusing on one patient, the extracted history grows as queries hit the target patient EMR record. Moreover, this extracted history is distributed at each EMR storage site across the participating medical institutions.

For achieving desirable response, the aggregation of extracted history should be obtained in a realistic time. This is closely related to the data structure and size of each log record. Future studies should focus on the data size of stored log records.

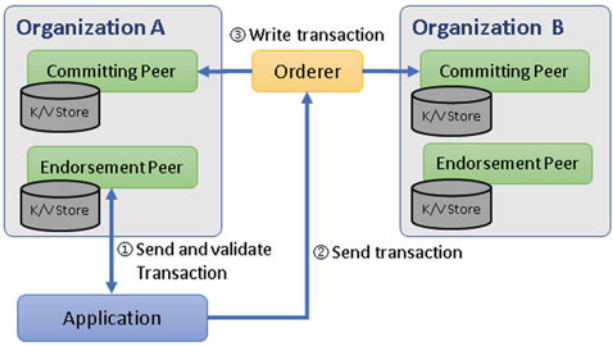
In the performance test, a simple message structure is defined as a JSON (shown in Table 6.7). The identifiers of patient and institution are represented as hash values. Each log record can be stored separately in the blockchain (separate style) or aggregated in a block by a patient appending records to the corresponding block (appending style). In the former method, the pieces of the records related to the patient of interest must be gathered. In the latter method, the block size grows as the system is used. We examined performance differences when the data size of a record to be written is changed.

Experimental Setups

We evaluated the following three approaches to implement traceability function. Of these, two are based on blockchain technology. The last approach uses the same method of secure data collection as PSI against log records stored in distributed PostgreSQL databases.

- Hyperledger Fabric [13]
- BigchainDB [14]
- PSI (Bloom filter)

The experimental settings for each approach are described as follows. Between Hyperledger Fabric and BigchainDB, key/value store implementation for search use differs from each other.



VM Host Machin	
CPU	Intel Xeon Bronze 3106 (16 CPUs @1.70GHz)
Mem	64GByte
OS	VMWare ESXi 6.7.0

VM * 2	
CPU	4 CPUs
Memory	8GByte
OS	Ubuntu 18.04.1 LTS

Middle / Soft Ware version	
HyperLedger Fabric	1.3.0
Docker	18.06.1
CouchDB	2.1.1

Fig. 6.8 Experimental settings (hyperledger fabric)

1. Hyperledger Fabric
   Figure 6.8 shows the experimental setup using Hyperledger Fabric to store query log records during data collection. Assuming two participating institutions, two nodes were set for the performance test. Native implementation only offers key-value storage and is applicable to a separate style. Furthermore, we evaluated Hyperledger implementation with CouchDB [15], which enables query against the value of the JSON message described earlier. Thus, both separate and appending styles can be implemented.
2. BigchainDB
   Figure 6.9 shows the experimental settings for using BigchainDB to store log data. As mentioned earlier, two nodes were prepared for evaluation. MongoDB [16] was selected as the backend database. In this case, both separate and aggregated structures are possible on the same implementation.
   Query key candidate is the transaction ID of the stored block or stored JSON value.
3. PSI (Bloom filter)
   Figure 6.10 shows the experimental settings in the case of PSI implementation. The log records of data extraction are recorded at the time of extraction. Using the same method of EMR data collection, we can gather the log records against distributed storages under encryption. Particularly, although the search is performed by specifying the insurance number, date of birth, and gender by patients, since the matching is performed using the bloom filter, these values are not directly disclosed on the infrastructure.
   In this test, three nodes were prepared for evaluation, but the performance test measurement was executed on only one node.

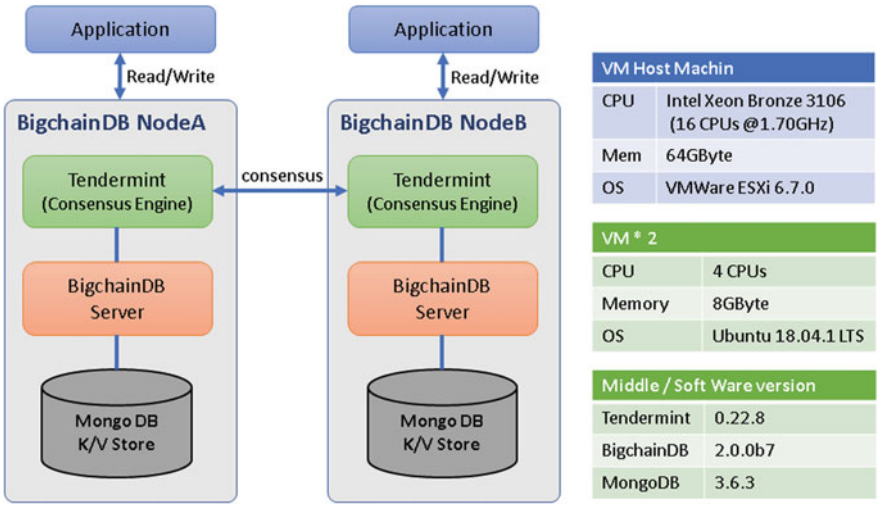


Fig. 6.9 Experimental settings (BigchainDB)

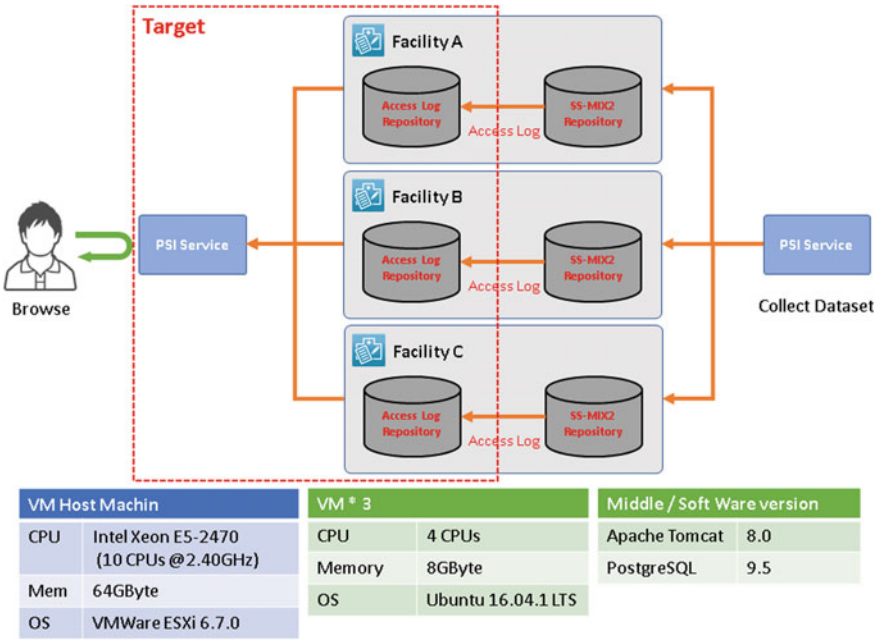


Fig. 6.10 Experimental settings (PSI)

6.3.4.3 Results

Performance by Data Size

Figure 6.11 shows the performance of writing records to the blockchain storage by record size for Hyperledger. In the experimental environment, it worked normally for records with a size of 7 MB or smaller. As the record size grew, the response became unstable.

Figure 6.12 shows the same test for BigchainDB. The maximum record size was 0.6 MB, which was much lower than that for Hyperledger. However, the transaction time to commit was larger than that for Hyperledger.

By contrast, the data size for PSI can be as large as allowed by the database system.

Transaction Performance

Figure 6.13 shows the performance results of writing records to the blockchain storage for Hyperledger with/without CouchDB and BigchainDB under one or five thread processings. In all cases, processing by threads contributed to storage performance, but the throughput did not increase linearly with the number of threads.

Comparing the three implementations, BigchainDB was slightly faster than Hyperledger. Hyperledger with CouchDB had the worst performance; this is likely caused by the cost of indexing within CouchDB. In the best case, 1 million records were written to the blockchain storage in 3–4 h. This performance is equivalent to writing 10 million records or less in one day.

Comparing these implementations using blockchain technology, the performance of PSI was equivalent to the "insert" performance of the PostgreSQL database used. The necessary time for inserting 1 million records to the database was below 10 min. This performance is about 1,000 times faster than the blockchain implementations.

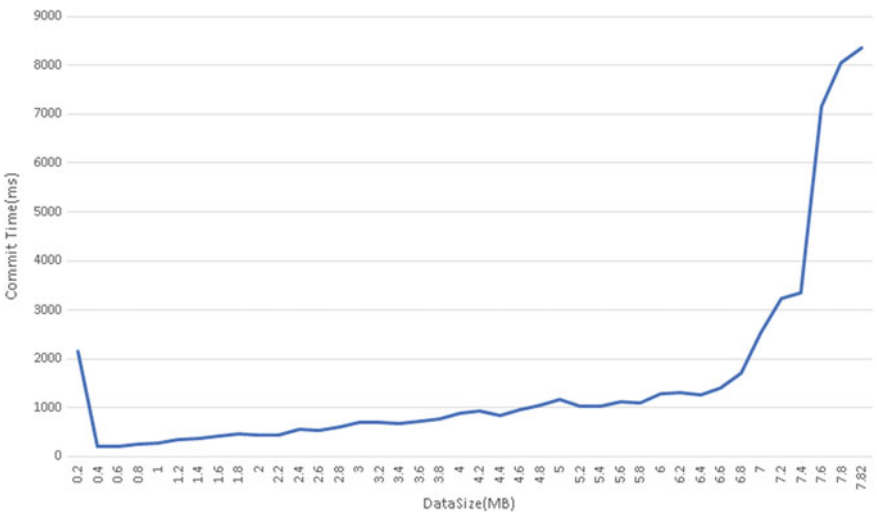


Fig. 6.11 Performance results by record size (Hyperledger)

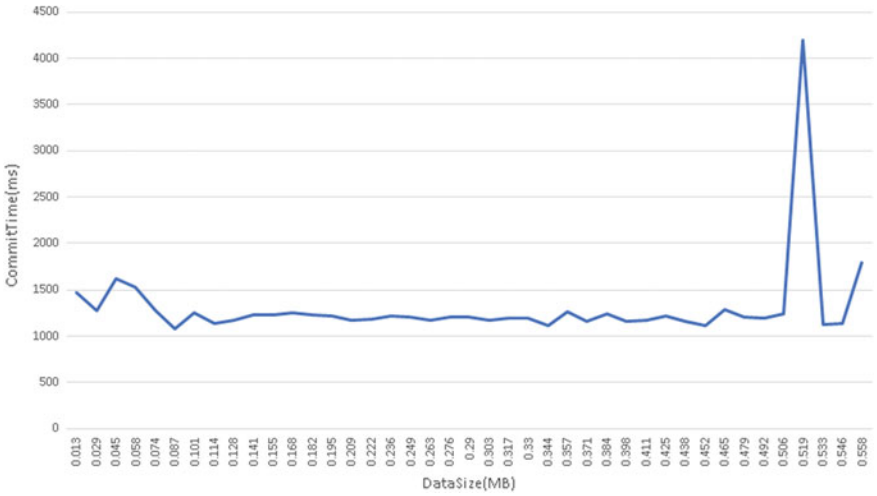
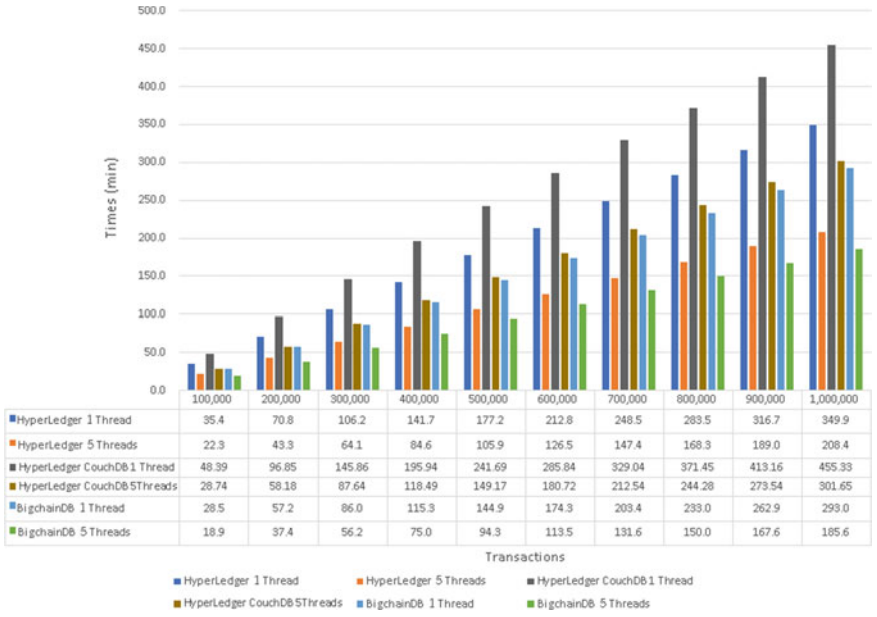


Fig. 6.12 Performance results by record size (BigchainDB)



	100,000	200,000	300,000	400,000	500,000	600,000	700,000	800,000	900,000	1,000,000
HyperLedger 1 Thread	35.4	70.8	106.2	141.7	177.2	212.8	248.5	283.5	316.7	349.9
HyperLedger 5 Threads	22.3	43.3	64.1	84.6	105.9	126.5	147.4	168.3	189.0	208.4
HyperLedger CouchDB 1 Thread	48.39	96.85	145.86	195.94	241.69	285.84	329.04	371.45	413.16	455.33
HyperLedger CouchDB 5 Threads	28.74	58.18	87.64	118.49	149.17	180.72	212.54	244.28	273.54	301.65
BigchainDB 1 Thread	28.5	57.2	86.0	115.3	144.9	174.3	203.4	233.0	262.9	293.0
BigchainDB 5 Threads	18.9	37.4	56.2	75.0	94.3	113.5	131.6	150.0	167.6	185.6

Fig. 6.13 Performance result of writing records

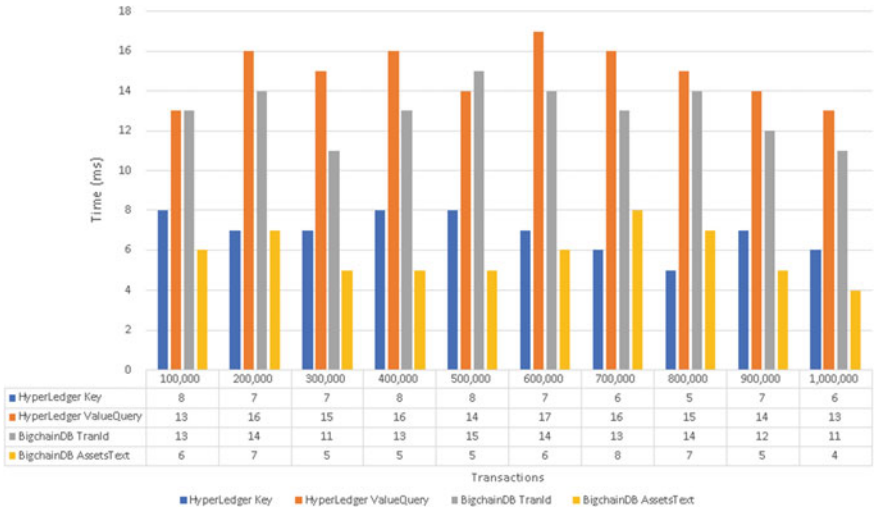


Fig. 6.14 Performance result of querying records

Query Performance

Figure 6.14 shows the performance test results of retrieving one record from the blockchain storage using four types of implementation. No significant differences were noted in the query response times between Hyperledger and BigchainDB. "Hyperledger Key" and "BigchanDB transid" represent the separate style of storage, whereas "Hyperledger Value" and "BigchainDB AssetsText" represent the aggregated style.

Query response is fast enough for actual use in the case of 1 million records in the storage. This result shows hitting 1 record, and the response time linearly increases as hit records increase.

On the other hand, PSI implementation needs 1 min or less to aggregate the extracted results across the distributed databases.

6.3.4.4 Discussion

Based on the initial evaluations, the following recommendations are made.

Transaction Performance

The transaction performance of a blockchain network was quite low for storing massive numbers of log records generated by queries in the developed system. In the case of blockchain, at most 100 transactions per second is best for a node to register to storage. Compared with implementation with PostgreSQL, the total transactions per day will be 1,000 times smaller. If we do not implement any aggregation of log records, it will be impossible to process the enormous numbers of log records generated for each EMR item. Some patient-based aggregation of log records should be considered to overcome performance limitation.

Data Size
The results by data size show the upper limit for storing log records to the blockchain storage. As writing large records to storage makes the system unstable, writing in the appending style is not suitable because of the long operation time of the system. Considering the transaction performance test results mentioned earlier, the total number of transactions to the blockchain network per day should be limited.

Query Response
As the amount of storage increases, the search function must query all storage in the network. The whole log records thus require some possible indexes for searching by patient. The query performance test results show a good response for searching for a log record in the blockchain network despite the increase in the number of log records.

Proposed System for Future Implementation
Based on the performance evaluation results, we decided to implement the following policy as the basis for making the search log history visible to patients when using the developed secure data collection system:

- Aggregate log data by patient in each facility.
- All log records are stored at each facility.
- Record the minimum amount of data, such as the log record identifier key and facility identifiable key, for retrieving index data in the blockchain.
- For query log data, use personal identification information, such as insurance number, date of birth, and gender.

By following these policies, a patient can search the blockchain and find the storage facility. Moreover, the number of records that must be recorded per period can be reduced to the number of related patients. Figure 6.15 shows an overview of the proposed log search system. The log records should include the following:

- Facility identifiable key
- Log record identifiable key
- Digest to audit each log record
- Key to identify each patient (this could be generated by encrypting a patient identifier such as insurance number, date of birth, and gender)

We plan to develop a log search system with the described structure.

6.3.4.5 Limitations

Because we did not have sufficient time to set up larger records, performance tests were executed for 1 million records or less. As the number of records increases, the test results and system stability may change. Performance tests with more records are required in the future work. Similarly, performance should be estimated for larger numbers of nodes.

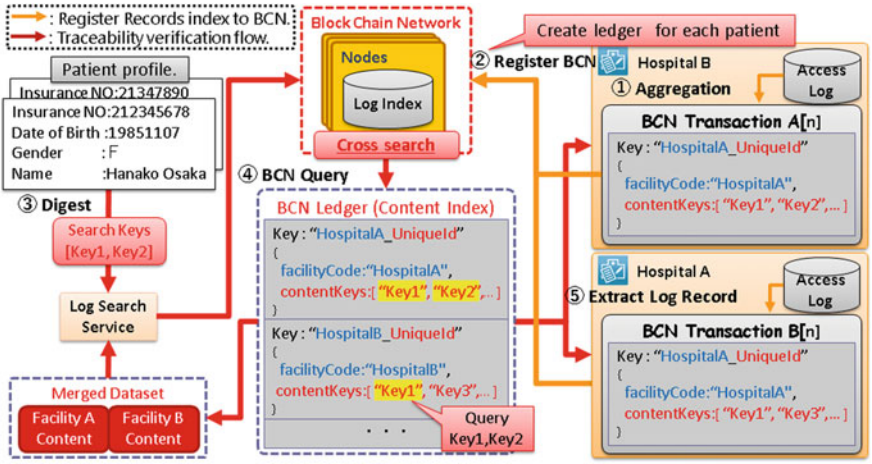


Fig. 6.15 Overview of the proposed log search service using a blockchain network

6.3.4.6 Summary

This section reports the initial performance results related to traceability for a secure data collection system under development. The desired data structure and system infrastructure were examined. Although blockchain implementation is a strong candidate for establishing an audit infrastructure to verify the use of EMR data for clinical research, there are some challenges for maintaining long-term operation as the amount of data increases. Thus, we proposed a data structure and querying implementation to overcome the implementation performance.

6.4 Integration and Prospects

As described earlier, the implementation and verification of the following element function have been carried out for a secure secondary use of medical data with the capability of access control by consent information and secondary use status confirmation by traceability function. The key features of our medical test bed are the following:

1. Improvement of the searchability of medical data in SS-MIX2 standardized storage
2. Safe medical data extraction function from SS-MIX2 standardized storage using PSI
3. Electronic description of consent information and mechanism for checking consent information when extracting data
4. Privacy risk assessment function for the extracted dataset
5. Traceability function that can be verified by patients.

Currently, the development of the aforementioned functions is being integrated and developed with the in mind that it can be used as a Web service applicable to public cloud.

If our developed system is ready on public cloud, it would help clinical researchers to conduct cross-institutional data collection and analysis with a certain level of security guaranteed.

References

1. Japan Personal Information Protection Commission. Amended act on the protection of personal information (2017), https://www.ppc.go.jp/files/pdf/Act_on_the_Protection_of_Personal_Information.pdf
2. Japan Ministry of Justice. Act on anonymously processed medical information to contribute to medical research and development (2017), http://www.japaneselawtranslation.go.jp/law/detail/?re=01&dn=1&x=33&y=11&co=1&ia=03&yo=&gn=&sy=&ht=&no=&bu=&ta=&ky=%E5%8C%BB%E7%99%82%E5%88%86%E9%87%8E&page=1
3. Organisation for Economic. OECD privacy guidelines (2013), https://www.oecd.org/internet/ieconomy/privacy-guidelines.htm
4. Ethical guidelines for medical and health research involving human subjects (2015), https://www.mhlw.go.jp/file/06-Seisakujouhou-10600000-Daijinkanboukouseikagakuka/0000080278.pdf
5. 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, S. Tani, SS-mix: a ministry project to promote standardized healthcare information exchange. Methods Inf. Med. **50**(2), 131–139 (2011)
6. Japan Association for Medical Informatics. "SS-mix2 standardized storage" explanation of the structure and guidelines for implemenation ver. 1.2 (2014), https://www.jami.jp/jamistd/docs/SS-MIX2/descript-implemglonSS-MIX2_V1.2.pdf
7. K. Tanaka, R. Yamamoto, K. Nakasho, A. Miyaji, Development of a secure cross-institutional data collection system based on distributed standardized EMR storage. Stud. Health Technol. Inform. **255**, 35–39 (2018)
8. A. Miyaji, K. Nakasho, S. Nishida, Privacy-preserving integration of medical data. J. Med. Syst. **41**(3), 37 (2017)
9. Health Level Seven International. HL7 standards product brief - HL7 CDAR R2 implementation guide: privacy consent directives, release 1 (2017), http://www.hl7.org/implement/standards/product_brief.cfm?product_id=280
10. A. Dubovitskaya, Z. Xu, S. Ryu, M. Schumacher, F. Wang, Secure and trustable electronic medical records sharing using blockchain. AMIA Annu. Symp. Proc. **2017**, 650–659 (2017)
11. M.N. Kamel Boulos, J.T. Wilson, K.A. Clauson, Geospatial blockchain: promises, challenges, and scenarios in health and healthcare. Int. J. Health Geogr. **17**(1), 25 (2018)
12. J.M. Roman-Belmonte, H. De la Corte-Rodriguez, E.C. Rodriguez-Merchan, How blockchain technology can change medicine. Postgrad. Med. **130**(4), 420–427 (2018)
13. Hyperledger fabric - hyperledger (2018), https://www.hyperledger.org/projects/fabric
14. @bigchaindb. Bigchaindb - the blockchain database (2018), https://www.bigchaindb.com/
15. Apache couchdb (2018), http://couchdb.apache.org/
16. Open source document database (2018), https://www.mongodb.com/index