Large-scale Health Information Database and Privacy Protection*1

JMAJ 59(2&3):91-109, 2016

Ryuichi YAMAMOTO[1]

Abstract

Japan was once progressive in the digitalization of healthcare fields but unfortunately has fallen behind in terms of the secondary use of data for public interest. There has recently been a trend to establish large-scale health databases in the nation, and a conflict between data use for public interest and privacy protection has surfaced as this trend has progressed. Databases for health insurance claims or for specific health checkups and guidance services were created according to the law that aims to ensure healthcare for the elderly; however, there is no mention in the act about using these databases for public interest in general. Thus, an initiative for such use must proceed carefully and attentively.

The PMDA*2 projects that collect a large amount of medical record information from large hospitals and the health database development project that the Ministry of Health, Labour and Welfare (MHLW) is working on will soon begin to operate according to a general consensus; however, the validity of this consensus can be questioned if issues of anonymity arise. The likelihood that researchers conducting a study for public interest would intentionally invade the privacy of their subjects is slim. However, patients could develop a sense of distrust about their data being used since legal requirements are ambiguous. Nevertheless, without using patients' medical records for public interest, progress in medicine will grind to a halt. Proper legislation that is clear for both researchers and patients will therefore be highly desirable.

A revision of the Act on the Protection of Personal Information is currently in progress. In reality, however, privacy is not something that laws alone can protect; it will also require guidelines and self-discipline. We now live in an information capitalization age. I will introduce the trends in legal reform regarding healthcare information and discuss some basics to help people properly face the issue of health big data and privacy protection with a sense of ownership.

Key words Big data in health field, Large-scale health information databases, Privacy, The Act on the Protection of Personal Information, Secure computing

Overview and Problems of Health Big Data

When we talk about health big data, it can be rather difficult to understand what the term actually means. So, I would like to start by giving a clear overview of health big data and its problems.

People often say that we now live in a data-oriented age. I will give you a brief overview of what this actually means and introduce some of the preparations currently being made toward

*1 This article is based on the lecture presented in Japanese at the Japan Medical Association Research Institute (JMARI) Symposium on "Current Status and Future of Health Big Data in Japan" held on February 12, 2015. Unless otherwise stated, the figures in the article are originally made by the author in Japanese and translated for JMAJ.

*2 PMDA: Pharmaceuticals and Medical Devices Agency.

1 Project Associate Professor, Department of Health Management and Policy, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan (yamamoto@hcc.h.u-tokyo.ac.jp).

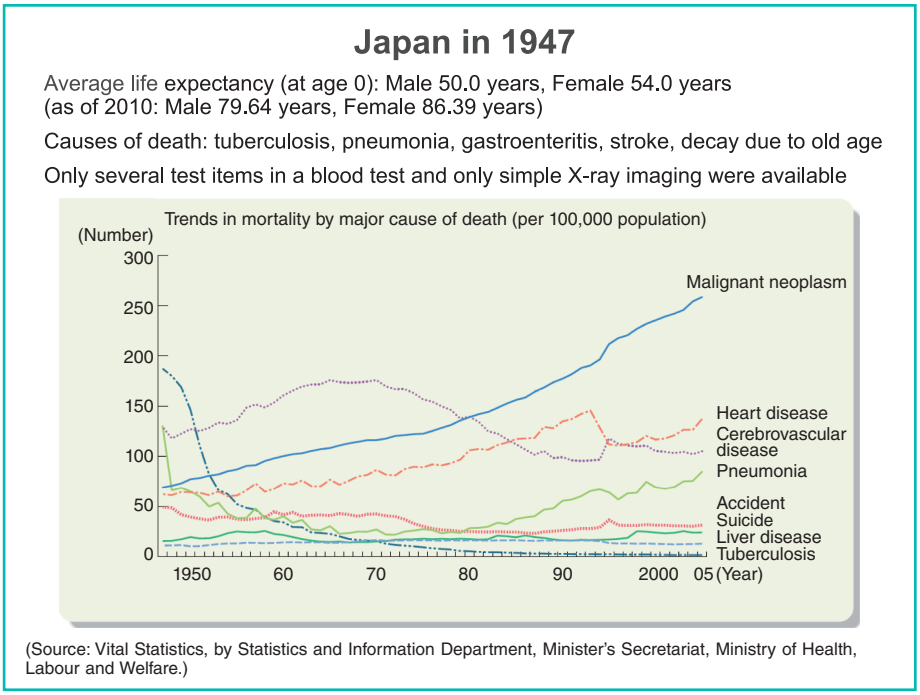


Fig. 1 Japan in 1947 and trends in mortality by major cause of death

establishing health information databases in Japan.

I will also discuss what legislation will be required and then go into detail regarding various informatics measures that should be promoted, including possible data management options and what more can be done in terms of informatics measures.

The Disappearance of Large Amounts of Medical Information

As everyone probably knows, 1947 is the year that the Constitution of Japan came into effect (**Fig. 1**). Article 25 of the Constitution states "All people shall have the right to maintain the minimum standards of wholesome and cultured living."

So, what was 1947 like? The number one cause of death was tuberculosis, followed by pneumonia and gastroenteritis. So, the top 3 causes of death were all infectious diseases. Stroke was ranked the 4th and the 5th was decay due to old age. Most of these conditions progress quickly and usually resulted in death. The medical technology available back then was limited; only simple X-ray imaging was available, and blood tests were simple ones such as counting blood cells under a microscope or reading the blood sedimentation rate. The average life expectancy at age 0 in 1947 was 50 years for men and 54 years for women. These figures, however, already exceeded 80 in 2010.

The top 3 causes of death have also shifted to malignant neoplasm, heart diseases, and cerebrovascular diseases. Heart diseases and cerebrovascular diseases are essentially the end results of lifestyle-related diseases. Malignant neoplasm, or simply cancer, is often said to have progressed for more than 10 years before it is found, and many patients must endure many years of cancer treatment. In other words, people do not die quickly from cancer. The 4th greatest cause of death, pneumonia, is quite different from what it was back in 1947, as you already know. The next one is accidents, followed by suicide. Thus, it is evident that most of the current leading causes of death are diseases with very long courses.

I made a request to an organization related to the Ministry of Internal Affairs and Communications so that I could investigate what kinds of information concerning medical care and

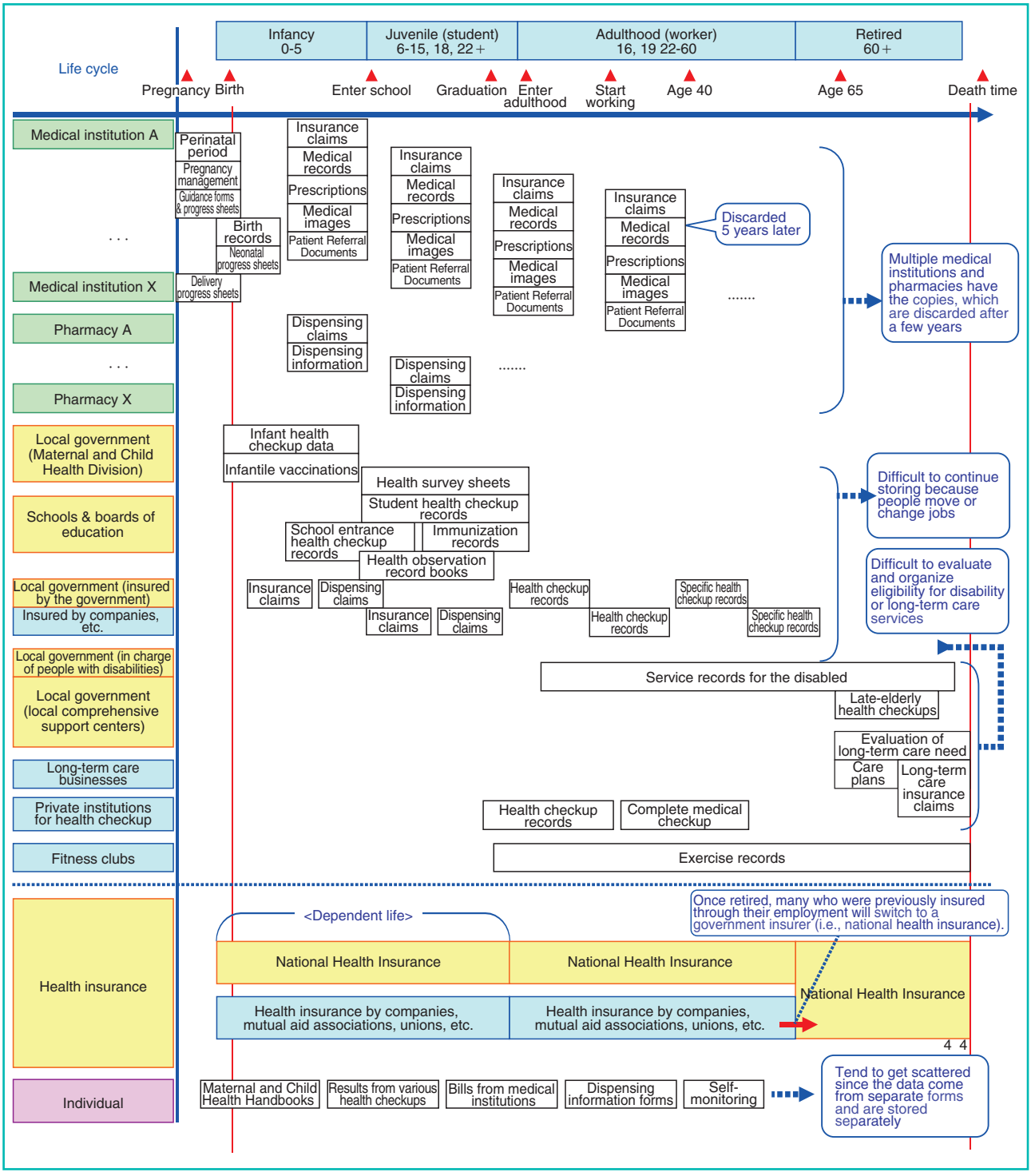


Fig. 2 Medical and health care information produced in life

health are produced in one's life. There are indeed various sorts of information (**Fig. 2**).

The problem, however, is that all this information often disappears altogether, or it is somewhere that the individual does not know and thus has no way of accessing. In essence, a large

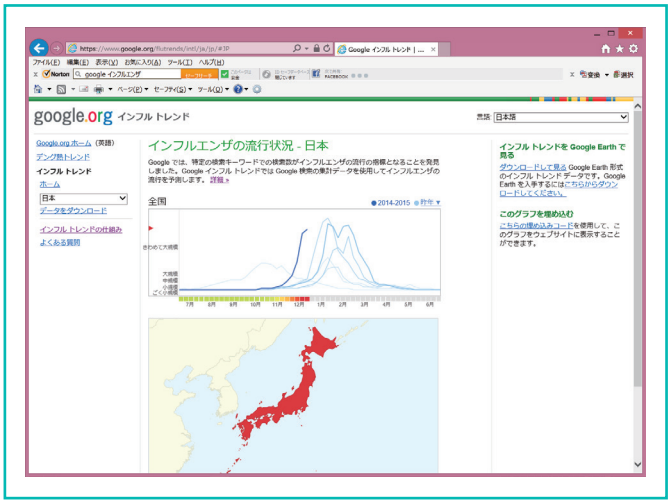


Fig. 3 Influenza outbreak trends in Japan by Google

amount of information is produced but completely disappears. Thus, deciding policies relating to research, medical care, and health in this situation is quite a challenge.

Finding New Value by Capitalizing a Large Amounts of Data

Of course, dealing with large amounts of data using analog methods such as an abacus or lined sheets of papers manually is impossible to begin with; the handling of a large amount of data was made possible by computer technology and IT advancements.

In January 2011, during a popular television quiz show "Jeopardy!" in the US, a computer system called WATSON, created by an American IT company, won against human opponents. This was the first time that a computer system had won against a human in a knowledge-based competition. It was quite the news.

Actually, WATSON was not really developed by the IT company; it was the result of a collaboration between informatics experts all over the world. What WATSON does is read vast amounts of available documented information scattered across the Internet in various natural languages, process this information—not online but in an isolated environment—and analyze it earnestly to derive knowledge. This is what WATSON did to win the quiz show.

In this quiz show, panels are shown to all participants and a presenter asks participants to choose a question depicted in the panel—for example, "The number 30 in 'politics,' please." There was a similar quiz show in Japan, wherein the text of the questions was often not stated in an easily comprehensible manner and could be quite confusing even for humans. Still, WATSON won.

At present, the IT company in question is focusing on WATSON's application in medical areas, and is independently conducting empirical projects at institutions such as the University of Pittsburgh.

Figure 3 shows Google Flu Trends, which first started when researchers began estimating the number of people with influenza using Google search data. Now, it incorporates other various factors in the prediction. This work has become somewhat world-famous since it allows prediction of specific local flu trends sooner than does fixed-point monitoring.

So, we are now capable of processing more and more data. So, what can we do with such a large amount of data? This is the essence of the concept of big data. In reality, big data analysis is still in its infancy, and not many attempts are currently in progress. However, there is no doubt that new projects are being developed one after another.

Twelve years ago in the US, this kind of poster was quite popular (**Fig. 4**). "Data is the New Oil." For someone like me, who learned classical English grammar, this phrase "Data *is*" draws the attention a bit. Anyways, the point of

1) High-resolution (Capable of extracting specific data in accordance with the subject's personality);
2) Highly frequent generation (Capable of almost real-time analyses); and
3) Diverse and non-structured (Capable of associating various types of data)

The data that were discarded because they were generated or stored but remained unused for a while actually hold value.

Fig. 4 This kind of poster was popular in the US twelve years ago

Data-oriented Age

- National Insurance Claim and Health Check-up Database (DB) (*aka* NDB)
- Mid-Net Project (PMDA & MHLW)
- KDB
- Long-term Care Need Evaluation DB
- National Cancer Registry
- Cardiac Catheterization DB
- Heart Failure Case DB
- National Clinical DB
- .....

Fig. 5 Examples of healthcare databases in Japan

such a poster is that data has been thrown away because we had no way to keep it. By capitalizing on it, we can give it new value. Data capitalization is the keyword of the present era, and that is why the current age is called the data-oriented age.

Development Status of Databases in Japan

So, are there any databases on healthcare in Japan? Actually, yes.

In terms of IT development in healthcare fields, Japan was in the top position until about 2000 or so, thanks to all the forerunners in the IT area. Every medical institution had a computer, for financial processing for national health insurance claims to the government. No other country had almost all of its medical bill information in an electronic form in one way or another.

The issue of social security is, of course, an urgent matter in many countries, and there has been a trend to introduce IT and streamline social security systems. However, few of those involved in introducing IT to social security had thought of accumulating and capitalizing on data, save for some notable trailblazers.

Then, things began to change around 2005 or so. We had planted the seeds of IT development in healthcare, so why not harvest the fruits as well? The tendency to accumulate and analyze data was suddenly growing, along with the urge to accomplish something. The databases shown in **Fig. 5** have already been developed or are being developed.

The first one is what we commonly call the "National Database (NDB)," or the national health insurance claim and health check-up database, of the Ministry of Health, Labour and Welfare (MHLW). I will discuss this in more detail later since I serve as the chairman of the Assessment Committee on the evaluation of database use.

The second, MID-NET*3 Project lead by Pharmaceuticals and Medical Devices Agency (PMDA) and MHLW's Safety Division is a database of drug safety events from 10 million people or so, which will contain data not from the claim data but from electronic medical records and service/test ordering systems from medical institutions; furthermore, it will include information such as all test results. This project aims to discover less common side effects to drugs more quickly.

KDB is also a database of national health insurance (*Kokuho*, for short). Unlike the MHLW's database (NDB), KDB puts some burden on local governments. Furthermore, this database is linked to databases for nursing care services, which contain information almost entirely extracted from the elderly. The linked database is already operational as a tool to analyze geriatric medical and long-term care, and local governments have started to use it.

*3 MID-NET: Medical Information DB NETwork.

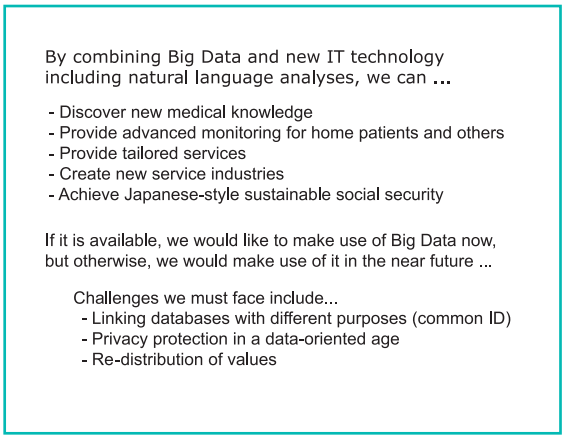


Fig. 6 Possibilities and challenges concerning health big data

The Nursing Care Need Evaluation DB is also done electronically from the beginning, and the information is all in the database. The National Cancer Registry, which will begin in 2015, will also be in a database, and the Japanese Circulation Society is in the process of developing databases for cardiac catheterization and heart failure cases. I hear that they already have data on quite a number of cases. Listed next is the National Clinical DB, which is the database managed and used by various surgical societies. This one was created many years ago, and they have since accumulated a considerable number of cases.

Issues to Be Solved Concerning Health Big Data

Although these databases and the information contained within are not fully standardized, they show some potential through IT techniques including natural language analysis. We all want to make use of what is available. Evidence-based policymaking and healthcare will be more important in the future, so databases will definitely have a role to play. However, the use of databases is not to be taken lightly, and there are still some challenges to be resolved (**Fig. 6**).

The National Cancer Registry, for example, will register any cancer case once a diagnosis is made, but will still only include one main treatment and the rest is untouched. This cancer registry, which begins later this year, is based on current law and will cross-reference with the Basic Resident Register (*Jumin Kihon Daicho*). So, we would be able to tell from the database if a person dies, but we will not be able to find the details of the treatment he/she received before passing away.

If this database were to be connected with the NDB, we would be able to tell what kind of care a patient received after their diagnosis until they passed away. The data will not include test results, but the medical services he/she received will be available from the linked database. At present, however, there is no means of connecting these databases.

If possible, this information should be integrated and analyzed to produce better results. A common ID system would be necessary for that purpose. On the other hand, it would mean that someone could find almost all of an individual's medical information using these databases. We cannot deny the possibility that such information would be used to violate individuals' rights or for discrimination.

Therefore, there are challenges in considering what privacy protection is appropriate for such a data-oriented age (**Fig. 6**). The "re-distribution of values" is listed at the bottom, but I would like to skip this part today.

Utility Value and the Characteristics of a Large-scale Health Database

I would like to quickly introduce the NDB, which was designed in 2006 and has been accumulating data since 2009 (**Fig. 7**). This database, which can be used both online and offline, contains all data from health insurance claims and specific health check-up and guidance information in an electronic form. I was part of the discussion briefly when the design of this database was still being debated, and I can tell you that the data are processed before storage so that patients' identities cannot be traced.

The NDB was created according to the Act on Assurance of Medical Care for Elderly People and the law ordains its use for optimization of healthcare expenses. Thus, application of its pre-determined use generally proceeds smoothly, but this database can be utilized for other various purposes, too.

Those who wish to use the database for purposes other than those for which it was intended must first apply to the database's authority, after

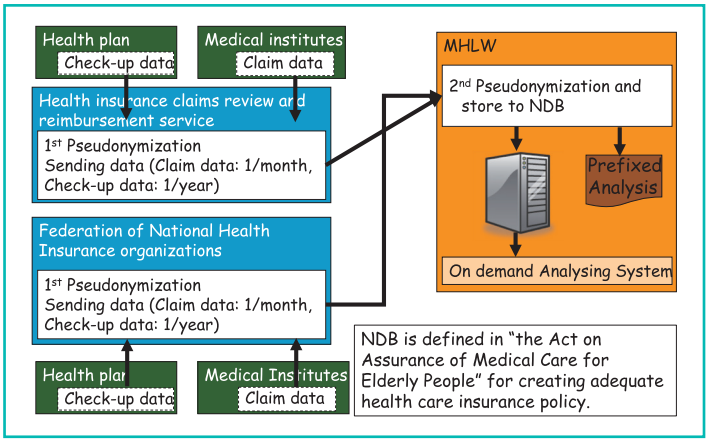


Fig. 7 National Health Insurance Claims and Health Check-up Database (NDB)

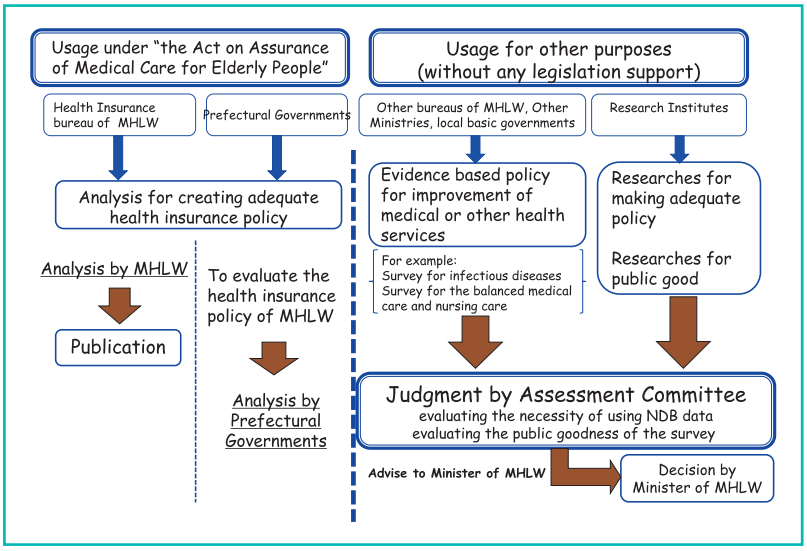


Fig. 8 Usage of NDB

which the public interest and data safety of the stated use will be fully reviewed before any data are disclosed. Meetings by experts, called the Assessment Committee, are held to evaluate these uses and offer advice to the Minister of Health, Labour and Welfare (**Fig. 8**).

So, what do the data include? All information from health insurance claims are included with the identities removed. The insurance policyholder's number, his/her date of birth, and gender are all erased in one step by creating 1 hash value from these 3 numbers. This is done using an encryption program to create a short string of numbers from a longer string—in our case, the number created is still quite long.

Another value is then produced by transforming the policyholder's name, date of birth, and gender into a single value. This transformed value cannot be transformed back into the original numbers (policyholder number, date of birth, and gender); however, the original and the transformed values match nearly one-to-one.

These 2 hash values, which I will call the ID number for now, will be used in the database. The patient's gender can be determined from the ID, but not his/her policy number or date of

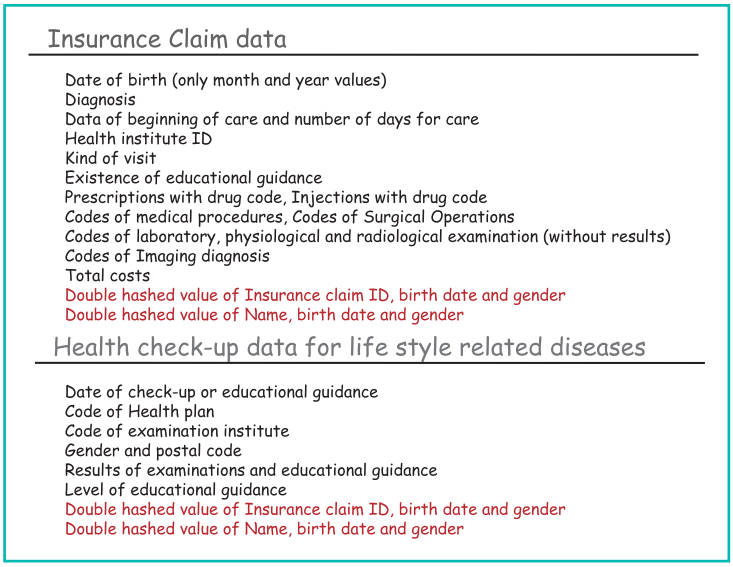


Fig. 9 Data included in NDB

birth. The ID number of the medical institution is included in health insurance claims and specific health check-up data (**Fig. 9**).

To put it shortly, such a database is very comprehensive. In Japan, over 95% of health insurance claims and almost 100% of prescription data exist in an electronic format. Dental claims, which were once a little behind in digitalization, have recently been rapidly catching up, and the database containing such claims now covers almost all people in Japan. Such nationwide databases are relatively common in Asian countries. South Korea and Taiwan started a little earlier than Japan did. Each nation's database is unique, but I will not go into details here since there is little point in it.

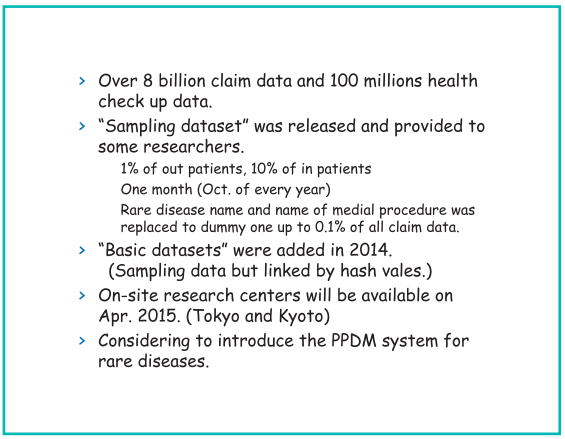


Fig. 10 Present Status of NDB

Types of Databases and System Management

So, how big is the NDB now? It now contains data from 8 billion insurance claims and close to 100 million specific health check-up and guidance cases (**Fig. 10**).

Because this database is quite heavy and troublesome to handle, 2 subsets were developed for convenience. One is commonly referred to as the "sampling dataset," and consists of a random sample of 1% of outpatients and 10% of inpatients, for which health insurance claims over 1 month and prescription claims for the same month and the following month are consolidated.

Among these, uncommon data such as rare diseases, rare medical practices, or rarely prescribed medicines and medical supplies—all occupying less than 0.1% of the data—are replaced with dummy data. As such, about 90% of disease names, 80% of medical practices, and 90% of prescriptions and medical supplies are omitted from the data subset. The remaining data in the subset, which consequently only comprise common diseases and procedures, are then replaced with dummies and made available for approved users.

The second subset is called the "basic dataset." This also comprises sampling data, but the data from insurance claims for the same persons are all linked. This subset is made from a sample of about 5% of the claims, and it is currently available for use. Starting from April 2015, on-site research centers become operational.

In the health insurance claim data, the names of policyholders are erased but can be easily guessed for certain people. The database allows physicians to follow up on someone with a very rare disease for an extended period of time. So, for example, one would be able to tell which month a patient went to see a physician over a 5-year period. Someone close to this patient, say a daughter, would be able to infer that the data is maybe that of her father. Therefore, the database is made available to individuals only on the premise that it will be securely managed because the data are not completely anonymous.

However, this "fully secure data management" cannot be easily achieved by most researchers or the general public. We frequently decline applicants for not filling out the items on secure management properly in the application forms.

The on-site research centers will provide secure management for data, so that people could conduct their research by visiting a center and looking up the data. There are 3 locations—2 in Tokyo and 1 in Kyoto—and they have been operational since April 2015. People can look up data quite freely at these centers, but no data can be brought outside.

If someone wants to bring the data out of the center, he/she must fill out an application, claiming that "I processed the data these ways and created such and such datasets, and I would like the copies." The dataset requested will not be readily available at the center; the requested dataset will be re-extracted by the center staff to verify its safety before it is given to the applicant.

To further protect the privacy of the data in the on-site research centers, a PPDM*4 system, which is actually an umbrella term, is expected to be introduced at some point in the future so that center users can analyze data while ensuring data privacy.

What Can We Tell from Health Insurance Claim Data?

So, what can we tell from health insurance claim data? **Figures 11 and 12** show the data based on a test study that a member of the Assessment Committee conducted by filing an application for the current NDB use to make sure that the application procedures we developed would actually function properly.

These charts illustrate the relationship between the Secondary Healthcare Zones and inpatients' diseases. In **Fig. 11**, for example, the darkest blue is the north Kyushu Healthcare Zone. This shows the proportions of inpatients in the north Kyushu Healthcare Zone who actually live in that area.

So, we can tell that most brain infarction inpatients come from the same healthcare zone. For breast cancer, however, the figure is very different (**Fig. 12**). Not many breast cancer patients are going to medical institutions in the same healthcare zone in which they live. This means that many are going outside of their own Secondary Healthcare Zone.

This study on health insurance claim data suggests that the concept of the Secondary Healthcare Zone should be seriously re-evaluated for some diseases. This is just an example of what health insurance claim data can tell us.

Protection of ID and Privacy by Legislation

We need proper legislation for the protection of ID information and privacy in order to use these data without unjustly discriminating against patients or the medical institutions that provided data.

For ID data, MHLW has already started reviewing several possibilities for capitalizing on the numbering system in healthcare. This numbering system*5, which begins in October 2015, is very similar to Taxpayer Identification number

*4 PPDM: Privacy-Preserving Data Mining.

*5 Since October, 2015, Individual Numbers (nicknamed "My Number") have been issued to residents in Japan under the Social Security and Tax Number System. "My Number" is required for administrative procedures related to social security, taxation, and disaster response, beginning in January, 2016. http://www.cas.go.jp/jp/seisaku/bangoseido/english.html.

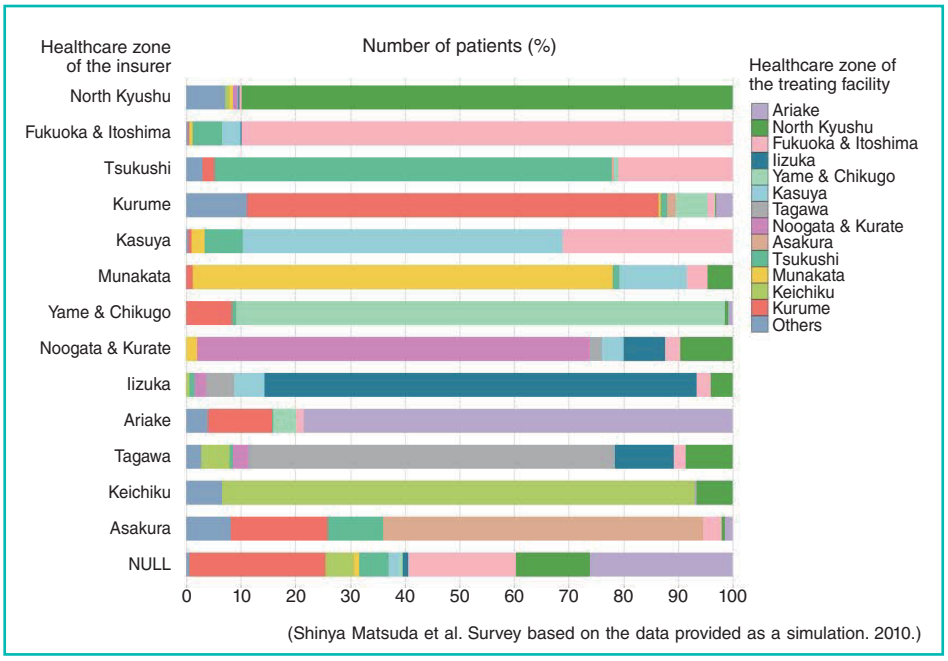


Fig. 11 Healthcare zones in which patients received care by the zones in which they reside
(Cerebral infarction; all ages; inpatients; national health insurance; Longevity Health Plan; and social welfare programs in total; shown in %)

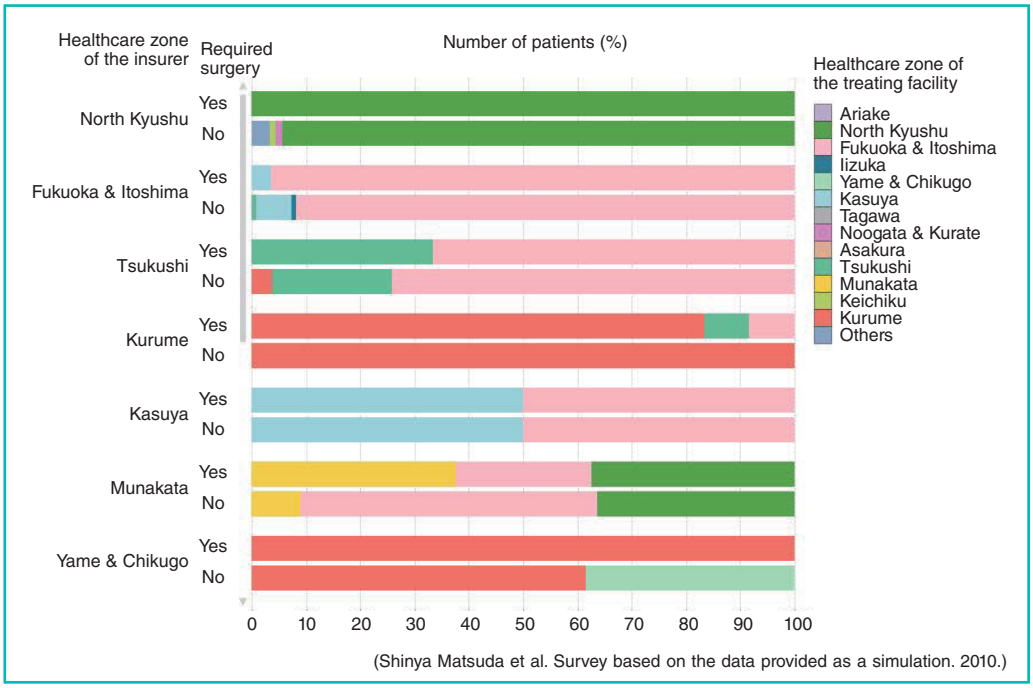


Fig. 12 Healthcare zones in which patients received care by the zones in which they reside
(Breast cancer; all ages; inpatients; national health insurance, Longevity Health Plan, and social welfare programs in total; shown in %)

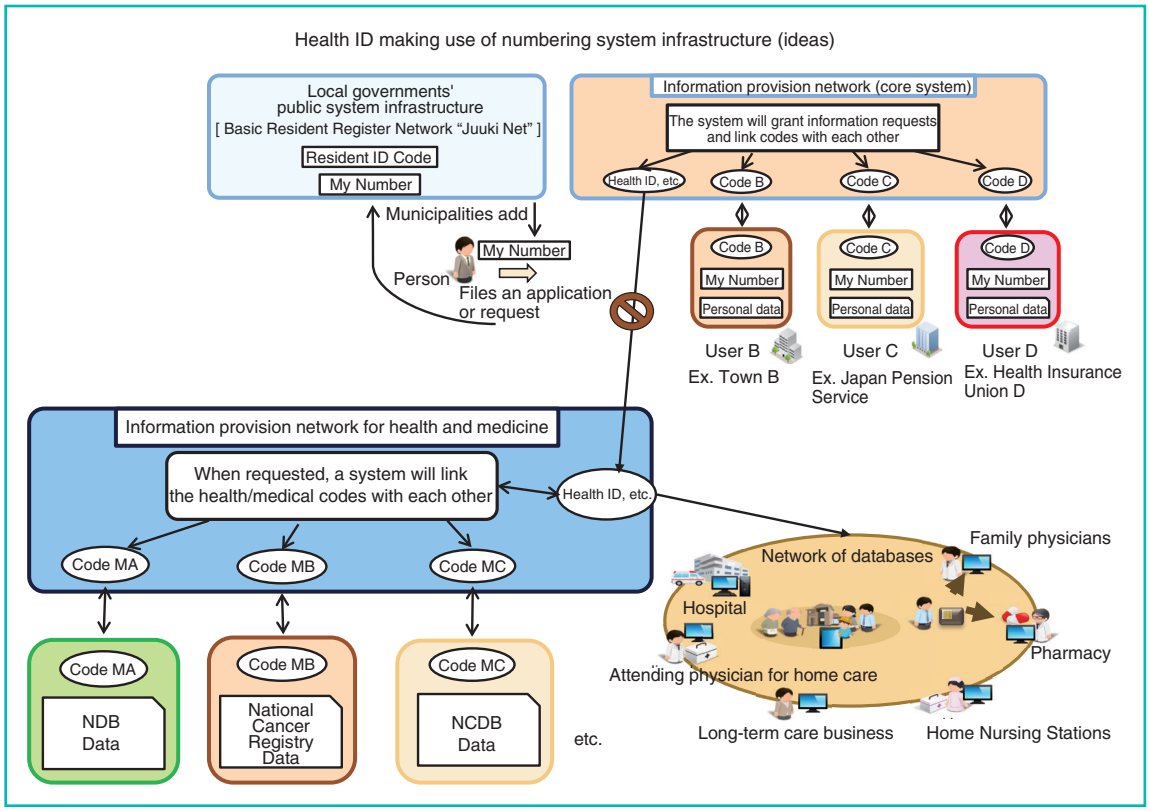


Fig. 13 Health ID making use of numbering system infrastructure (personal ideas)

System. The Taxpayer Identification Number System aims to understand individuals' income so that the burden of taxes and social security costs is shared equally among people. Legally speaking, it is not allowed to be used in healthcare practice or for in-kind benefits in nursing care.

On the other hand, those involved in healthcare liaison systems would have considerable trouble keeping track of each patient separately without the use of a numbering system. The Big Data that I have been talking about cannot be connected without a numbering system since each database is completely independent.

For example, there is a database on claims for nursing care insurance. However, claims for nursing care insurance and those for health insurance are completely unrelated currently, and there is no way to link them. They can be linked partly at the level of local governments, but not at the national level. Anyone would question whether healthcare issues could be properly addressed if things remain as such.

The Study Group on the Use of Numbering Systems in Healthcare and Other Fields prepared their interim report in December 2014, and suggested that the use of numbering systems should be promoted within the scope of current laws. In healthcare fields, for example, insurers should be allowed to use the numbering system. Moreover, the report suggests predetermining the specific occasions for linking information and reviewing the existing numbering systems and the nature of their numbers for further reevaluation.

Personal Opinion on the Use of Numbering Systems in Healthcare Fields

The upper half of **Fig. 13** shows the current numbering systems in the healthcare fields, but the rest is completely my personal ideas as to how these numbering systems could be better applied. Devising a system for ID numbering is surprisingly difficult, and there is no way we should not make use of the existing numbering systems. So,

I am starting with these numbering systems for healthcare ID numbering.

A single healthcare ID number will be created in a way that it cannot be easily traced back to the original data, so that the information linked with the other ID numbers of the same person cannot be easily connected. This healthcare ID will be used as needed in information networks such as collaborations between nursing care and medical care.

Patients have the freedom to choose, of course, so how his/her ID is used will be up to each patient. Meanwhile, codes reflecting various databases are prepared to be added to the healthcare ID as needed, such a code for the NDB or a code for the National Cancer Registry. This is the information provision network for health and medicine, and these codes are linked to each other only within this network.

It is therefore impossible to link these databases to each other in general. If someone filed a formal application of his/her plan, stating a reasonable cause to link the databases—a very important project for cancer management in Japan, for example—the matched data would be made available. That is what I picture the ideal system to be. Of course, we still have more to debate on this issue.

Act on the Protection of Personal Information and Big Data

Next, I would like to talk about the Act on the Protection of Personal Information. It is commonly understood that the current act is insufficient for addressing Big Data. The progress schedule of the government says that proper rules will be established sometime in 2015 (**Fig. 14**). Actually, there is no reason to prohibit the use of healthcare information in the first place.

Figure 15 shows the image of famous textbooks in medicine, which I also used when I was a student. Almost all of the knowledge in these textbooks comes from patient information; it does not come from test tubes or laboratory mice. So, the knowledge acquired from very sensitive medical records is written in these kind of textbooks. There is no way for medicine to advance if all this information is not available for use. Yes, it must be used.

That, however, does not mean that some patients or healthcare workers should be discriminated or suffer harm. For this reason, personal information protection and legislation for privacy protection began to develop all over the world from the late 1990s.

According to a report issued by the Academy of Medical Sciences in the United Kingdom in 2006, the data protection law was amended in 1998 and there are now rather strict requirements in healthcare and other fields.[1] The report, which is a book of over 200 pages, continues on to state that the amendment of the data protection law made medical research in the public interest very difficult to conduct, that financial demand for such research has increased, and that the privacy of patients is still not protected. The report also offers quite specific suggestions as to how these problems should be addressed.

Unlike the UK, Europe, or Japan, the United States has no comprehensive personal information law; however, the U.S. Department of Health & Human Services did enact the HIPAA*6 Privacy Rule for patients' medical records. This privacy rule came into effect in 2003 for large medical institutions and then for all medical institutions in 2004. According to the 2009 Report by the Institute of Medicine, since the implementation of this privacy rule, epidemiological research has become very difficult or nearly impossible to conduct, or, when it is conducted, it is exceedingly expensive, and still patient privacy is not protected.[2]

So, it is a process of trial and error for any country to make a progress. Reports suggest that although rules are made, they have proven to be rather ineffective. Moreover, the use of data has become extremely difficult even when it is for fair use and there is no intention of infringing on anyone's privacy.

This is true for all countries, but privacy protection laws prioritize the protection of data and tend to be insensitive about the consequences of the decision to not use data. If no data were allowed to be used for writing the textbooks I mentioned, for example, progress of medicine would grind to a halt (**Fig. 16**).

*6 HIPAA: Health Insurance Portability and Accountability Act.

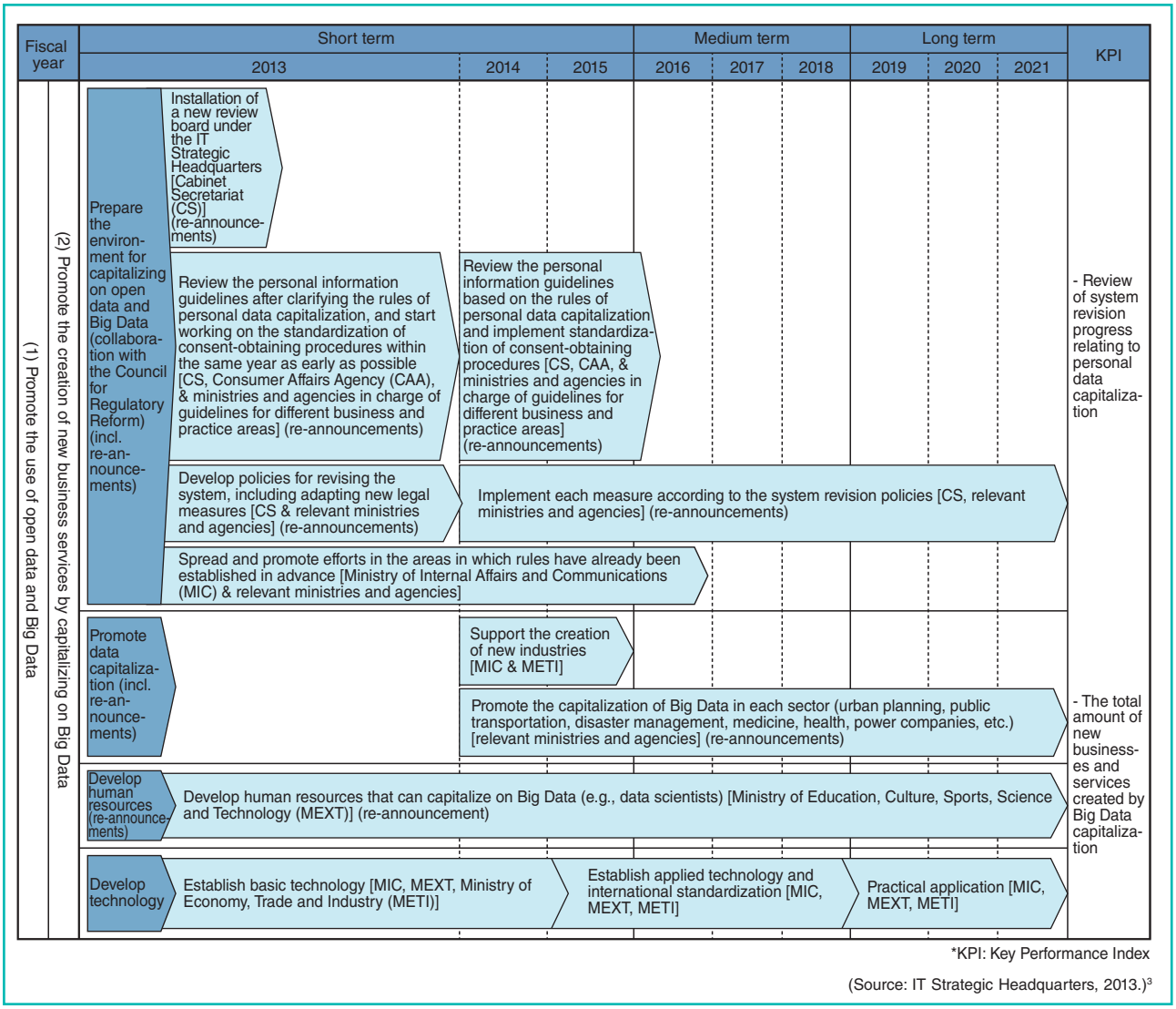


Fig. 14 Implementation Schedule (1. Creating new innovative industries and services and realizing a society that promotes the growth of all industries)

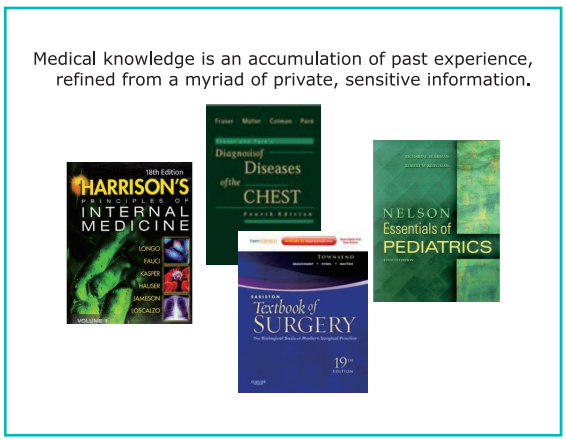


Fig. 15 Examples of medical textbooks

Current Situation of and Challenges Regarding the Act on the Protection of Personal Information in Japan

Japan is unique in that the rules of the Act on the Protection of Personal Information that apply to different sectors—the private sector, independent administrative corporations, government administrations, or local municipal governments—actually vary. This is the so-called "2,000 problems"; there are 2,000 rules for personal information protection in Japan, and each is slightly different from the rest. You might think that a slight difference will not pose a

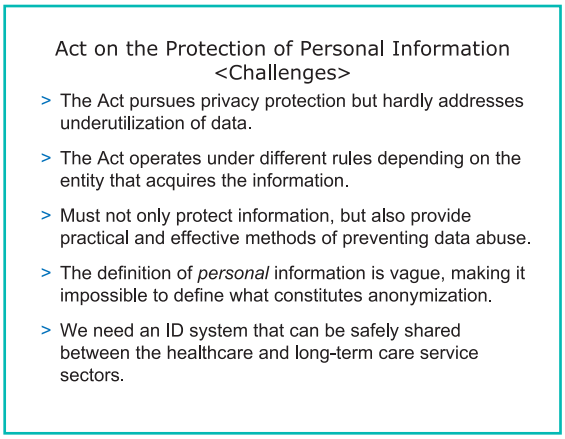


Fig. 16 Challenges regarding the Act on the Protection of Personal Information in Japan

problem, but in reality, someone has to be responsible for following the rules within each institution (**Fig. 16**).

Healthcare liaison systems in particular require the exchange of highly private and sensitive information concerning patients' health and medical care across institutional boundaries. When exchanging information between institutions, for example, from a prefectural hospital to a private hospital or from a private hospital to a university hospital that is an independent administrative corporation, each institution has its own staff members in charge. So, each staff member in charge must file an application with the review committee for evaluation. If an independent administrative corporation, a national medical institution, a private hospital, a prefectural hospital, and a private clinic are launching a healthcare liaison project together, each of the 5 institutions has to apply and be approved; this is actually happening in reality. This is quite an impossible task for physicians in clinical practice, and it is becoming a major problem.

I should also point out that there is no effective measure against unjust use of information. There is a law against it, but it is quite difficult for this law alone to stop wrongful use of data with malicious intent in a practical sense.

Furthermore, the official definition of personal information, which commonly refers to any information with a distinct identity, is rather vague. To compare, there is no definition of what exactly data safety is. The fact that there is no common ID is also a major problem because a person has no means to search for available information concerning his/her health or medical care. Therefore, the person cannot investigate whether that information has been used appropriately or whether it has been misused in any way.

When I said that there is no effective means of stopping unjust use, I am talking about the Unfair Competition Prevention Act as shown in **Fig. 17**. This is a law aimed at controlling industrial espionage.

There was a man in 2014 who was indicted for extracting a list of customer names from Benesse[*7]'s database and selling it to name list traders, and his charge was the violation of the Unfair Competition Prevention Act. It was not violation of the Act on the Protection of Personal Information. Apparently, the penalty for violating the Act on the Protection of Personal Information is not as significant as that of the Unfair Competition Prevention Act. That was the logic of the prosecution, but I feel somewhat questionable about his charge since I am not sure if what he stole constituted a business secret such as a blueprint for a new car.

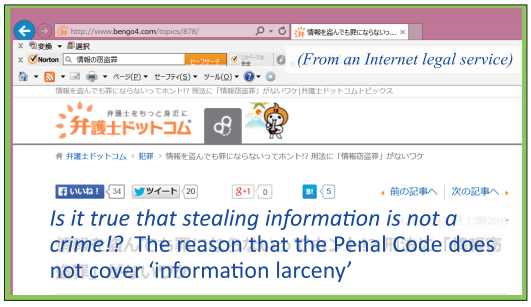
Anyhow, my point is that imposing a penalty is quite difficult under the current legislation for personal information protection. I would seriously question whether the current system is capable of handling medical or health information appropriately.

Future Trends for Amending the Act on the Protection of Personal Information

Discussion on amending the Act on the Protection of Personal Information began in 2013, and soon the bill will be finalized and is expected to be submitted to the Diet during a regular session in 2015. The Japan Medical Association is holding a National Healthcare Information System Liaison Council meeting this coming Sunday, and I believe Mr. Uryu will be there as a presenter; nevertheless, the bill will soon be ready (**Fig. 18**).

The motivation for amending this law has actually come from the unsatisfied voices of the people involved on how it is difficult to capital-

*7 Benesse is a major correspondence education provider for children in Japan.



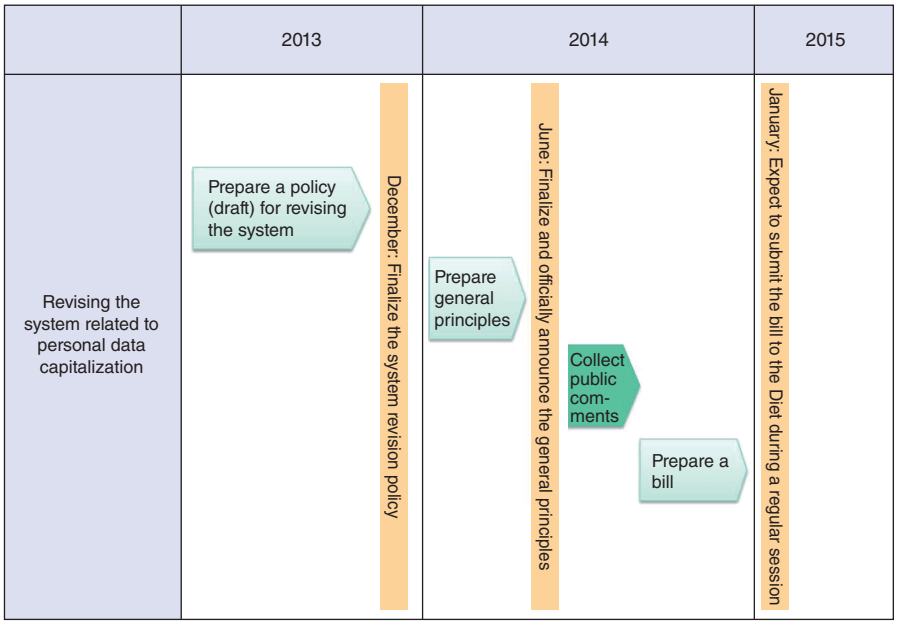
Legally stipulating information larceny in the Penal Code is difficult because it is difficult to limit the methods of categorizing different types of information.

There is a regulation that punishes illegally obtaining information in the Unfair Competition Prevention Act. So, an industrial espionage can be punished. But what about others?

Is it really possible to determine the value of information only in terms of privacy?
Is it possible to prevent discrimination from health and medical information (including genetic information)?

Perhaps we need an anti-discrimination law for physical/health information?

Fig. 17 Description of information larceny on a portal site for legal services



* Other nations including European countries and the USA are in the process of system revision, so it will likely take some time to ensure conformity with their efforts.
(Example: Proposed EU data protection rules, which the European Parliament is expected to adopt in the April 2014 Plenary Session)

Fig. 18 Road map to revise the system related to personal data capitalization

ize on personal data in business or the current law is too strict in practice (**Fig. 19**).

The amendment thus aims to make it easier for people to use data. Presumably, the amendment is also meant to protect what must be protected. However, if we are not careful, there is a chance that we will end up simply relaxing the regulations and increasing the potential risk. I

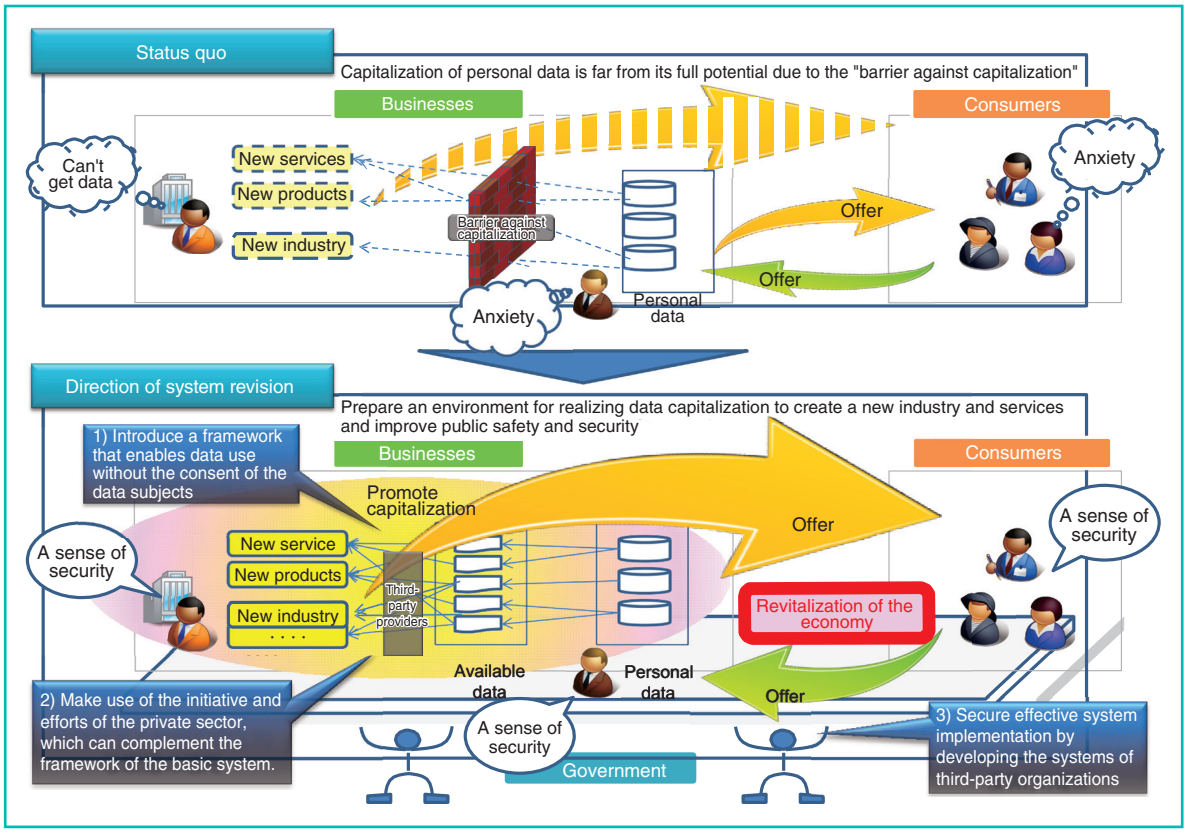


Fig. 19 Revising the system to capitalize on personal data

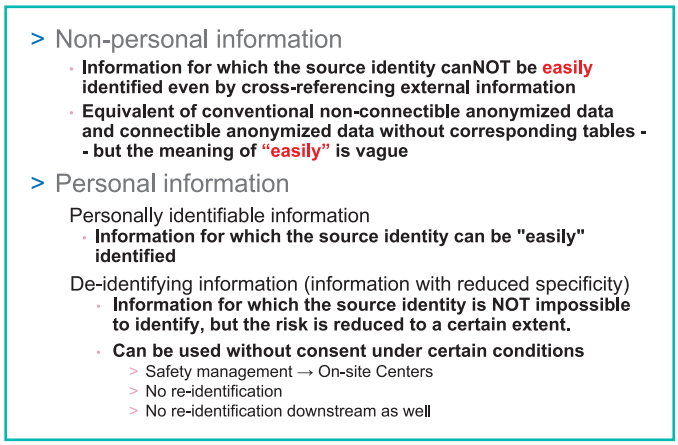


Fig. 20 Information derived from an individual

am afraid that I cannot go into specific details since the text of the bill is not finalized, which makes me a little irritated.

I can give you one example, nevertheless. The information derived from a person was considered either personal information or non-personal information before. In the bill, the personal information is further divided into 2 categories (**Fig. 20**).

One category refers to information from which an individual's identity can be easily identified, and the other category refers to informa-

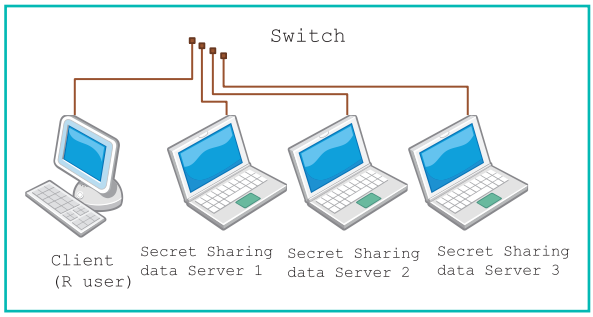


Fig. 21 Configuration diagram of a secure computation system

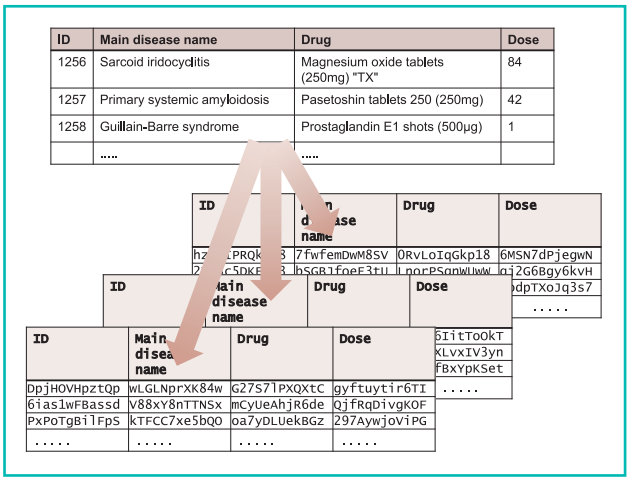


Fig. 22 Values are secretly shared

tion from which an identity cannot be easily identified but is not impossible. The latter is called de-identifying information.

The former is really personal information, so using it requires consent from the individual concerned—that has not changed. In the upcoming bill, the de-identifying information, on the other hand, can be used without consent if certain conditions are met. These certain conditions include that the data analysis has a proper purpose, that this purpose is clear, that the use of the data will not violate the rights of the individuals concerned, and that the safety and security of the data management is fully considered because the data can be analyzed or used by anyone if they are stolen. Moreover, the user must never attempt to re-identify the data. The draft version of the bill also stated that data processing will likely be entrusted to a third party and that the safety and security of the data must be maintained by that third party as well.

There are other issues to consider besides these. What about using the data for non-intended purposes, for example? These questions need to be addressed carefully in the future.

Other Informatics Measures That Should Be Further Promoted

Finally, I would like to discuss what more can be done in terms of informatics measures. This work, called a secure computation, is something that my laboratory carried out (**Fig. 21**). A secure computation involves dividing the original data into 3 separate sets (**Fig. 22**). Once divided, each dataset turns into 3 columns of data. They carry no information value at all, and they are not encrypted but rather are truly divided, so there

Function	Description
sec.mean	Average
sec.var	Unbiased variance
sec.median	Median
sec.max	Maximum value
sec.min	Minimum value
sec.subset.eq	Filtering by a conditional expression (= equal to)
sec.subset.gt	Filtering by a conditional expression (> larger than)
sec.subset.ge	Filtering by a conditional expression (>= equal to or larger than)
sec.subset.lt	Filtering by a conditional expression (< less than)
sec.subset.le	Filtering by a conditional expression (<= equal to or less than)
sec.subset.ne	Filtering by a conditional expression (!= does not equal)
sec.shuffle	Randomly replace the order of records
sec.xtabs	Cross tabulate

Fig. 23 Various statistical functions available via secure computation

		Computation time	(Computation time: Breakdown)	
		[sec]	Secure computation	Others
Average		6.14	5.71	0.43
Median		82.30	81.88	0.43
Conditioning filter	(Dose 1) = 10	7.76	5.56	2.20
	(Dose 1) > 10	38.41	36.20	2.20
Cross tabulate	Main disease name	86.74	86.15	0.59
	Main disease name, Gender	172.66	171.81	0.85

Number of records in the target data: 50,001

Fig. 24 Processing times for various statistical calculations via secure computation

is no way to transform them back into the original data. It is still mathematically possible to carry out statistical calculations using such non-transferable data. This method allows us to compute the total healthcare expenditure or total dose used for a given drug in a situation where no patient case can be seen.

For example, there are diseases for which very few cases exist and yet they have a significant social impact. Let us assume for now that the Ebola hemorrhagic fever arrived in Japan, and 10 patients presented with it. The media would desperately cover the news, and the identities of those 10 people would be revealed eventually. If someone tried to use or analyze the Ebola patients' data, they could easily identify who those patients were from just the data. If a researcher wanted to know the total dose of a drug and the patients' prognosis, this method of secure computation would enable such an analysis without actually seeing the original data.

We implemented secure computing method as the functions of common statistical language "R," and functions implemented are shown in **Fig. 23**. We can perform these kinds of searches and compute various statistical values, including the mean, unbiased variance, median, and maximum and minimum values. I believe most statistical procedures are feasible.

As for the computation speed, it is true that it takes longer when compared to the original, non-processed data. It may take 10 times longer or more, but one can muddle through in less than 100 times as long (**Fig. 24**).

There are other procedures similar to secure computation, too. It is actually a very hot topic of study right now. For example, it should be possible to trace back the original identity of certain data without actually looking at the data. It is theoretically possible, and is called secure traceability. In fact, my colleagues and I are in the process of bringing 3 or 4 organizations together to work on a project to study secure traceability on a continuous basis now.

I am afraid that my talk may have become slightly incoherent. I would like to thank the audience for their attention.

References

1. Personal Data for Public Good: Using Health Information in Medical research. London: The Academy of Medical Sciences; 2006. http://www.acmedsci.ac.uk/policy/policy-projects/personal-data/.
2. Institute of Medicine. Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research. Washington, DC: The National Academies Press; 2009. http://www.ncbi.nlm.nih.gov/pubmed/20662116
3. Strategic Headquarters for the Promotion of an Advanced Information and Telecommunications Network Society (IT Strategic Headquarters). Declaration to be the World's Most Advanced IT Nation. June, 2013 (in Japanese). http://www.kantei.go.jp/jp/singi/it2/dai62/siryou04.pdf