Medical Big Data for Research Use: Current Status and Related Issues*1

JMAJ 59(2&3):110-124, 2016

Koichi Benjamin ISHIKAWA[1]

Abstract

Advances in the computerization of information and development of technology have mitigated restrictions on handling of a large amount of information. This has resulted in growth of expectations for the use of large-scale databases, or so-called "big data." This is also the case in the field of healthcare. Projects that involve building of the national receipt database (NDB) of medical fee bill (receipt) information and special health check-up information based on the Act on Assurance of Medical Care for Elderly People and the development of medical information databases have been pursued by the national government, and considerable attention has also been focused on researches conducted through the secondary uses of publicly collected data.

Aside from these trends, there are numerous projects which collect diagnosis procedure combination (DPC) data to build large-scale databases for research purposes. Following to the ethics guidelines for epidemiologic studies, they collect and analyze anonymized DPC data from cooperating institutions.

This communication concentrates on the use of DPC data, and outlines the scale of data currently available for research use. Examples on the use of DPC data will be shown for analysis on the current status of clinical practice from the microscopic perspective and macroscopic analysis of community medical care provision. Additionally, potential for extending studies to long-term outcomes research, limitations and issues related to the use of medical big data will also be discussed.

Key words Large-scale medical databases, DPC data, Open data, Clinical processes, Community care provision

Usage of Big Data in the Medical Field

This presentation will describe the use of big data for research, including current applications and future expectations and needs. The focuses will be placed on the collection and processing of big data related to provision of clinical services, conducted by the national government, or within epidemiologic studies. The use of data in the private sector and other types of data such as genomic and lifestyle data should be left to a different occasion because uses of such data are still limited.

Building of Databases Promoted by the National Government

As mentioned by Dr. Ryuichi Yamamoto, the government's efforts to build large-scale databases in the field of healthcare have progressed tremendously. The government is collecting data under relevant laws such as the acts on assurance of medical care for the elderly and nursing care for the elderly, as shown in the left portion of **Fig. 1**. In addition, the national cancer registry is scheduled to launch in January of 2016, according to the Act on Promoting Cancer Registries.

*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.

1 Head, Economics Section, Division of Surveillance, Center for Cancer Control and Information Services, National Cancer Center, Tokyo, Japan (kishikaw@ncc.go.jp).

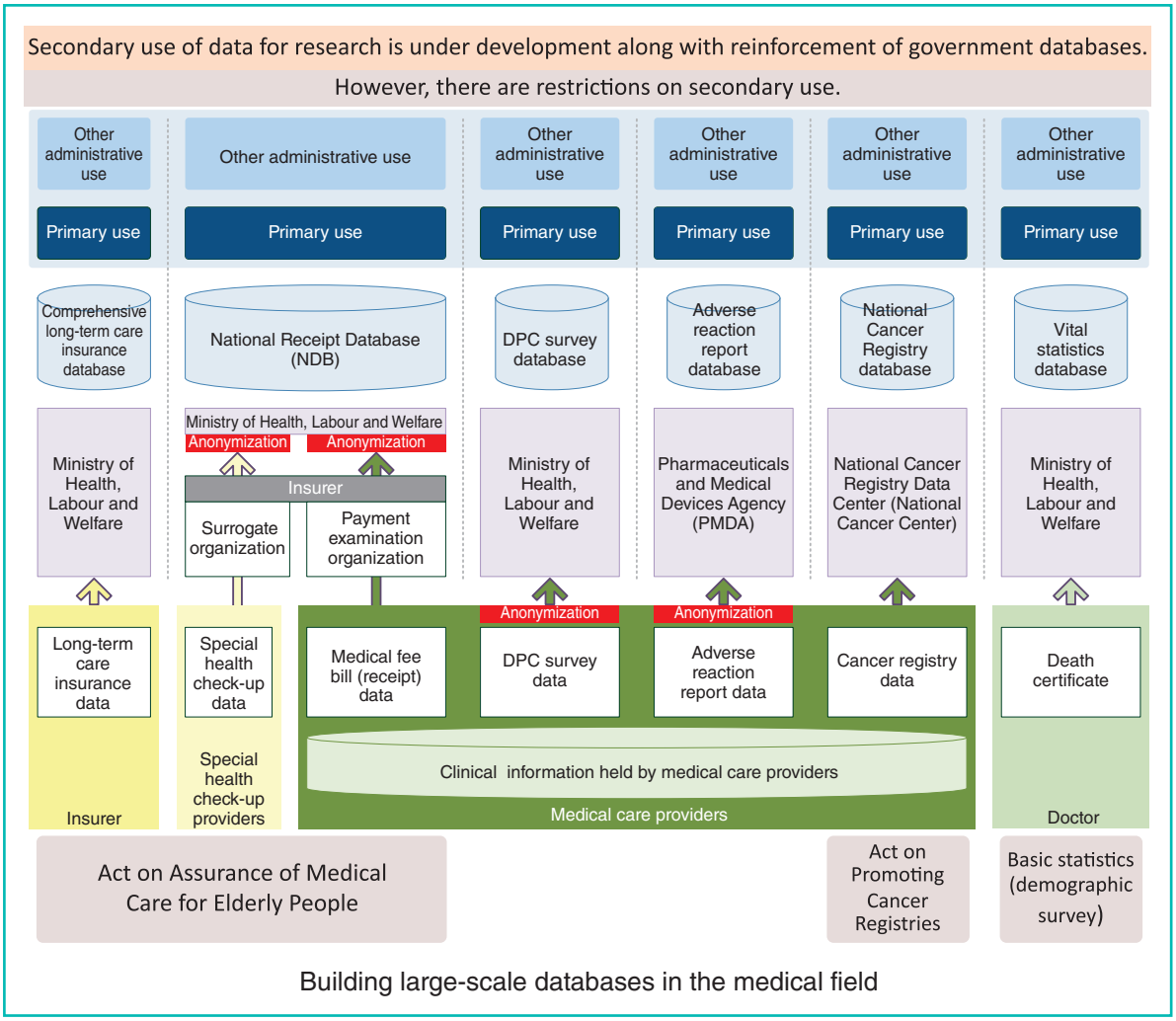


Fig. 1 Medical databases promoted by the Japanese government

Demographic data is collected to serve as the nation's core statistics.

To illustrate some examples, under the Act on Assurance of Medical Care for Elderly People, data on long-term care insurance are submitted by each local government to the Ministry of Health, Labour and Welfare (MHLW), forming a database for Health and Welfare Services for the Elderly. As for the national receipt database (NDB), medical fee bill (medical fee receipt) data are first submitted by each medical care provider to a payment examination organization. The MHLW then collects such data, anonymizes them, and forms the database. In the same manner, data on special health check-ups are obtained from providers and are sent to insurers via surrogate organizations. They are then submitted to the MHLW, where they are anonymized to form databases. The data, once compiled into databases, can be used in various ways.

In this array of databases, there is a new project of the National Cancer Registry. The National Cancer Center is in charge of the data management under commission from the government. Data on cancer cases, at the time of occurrence, newly found in hospitals and clinics with beds are compiled into the National Cancer Registry database, and used for reporting on cancer incidence and outcomes.

Current Status of the Use of Medical Databases

Within this framework, there is another trend,

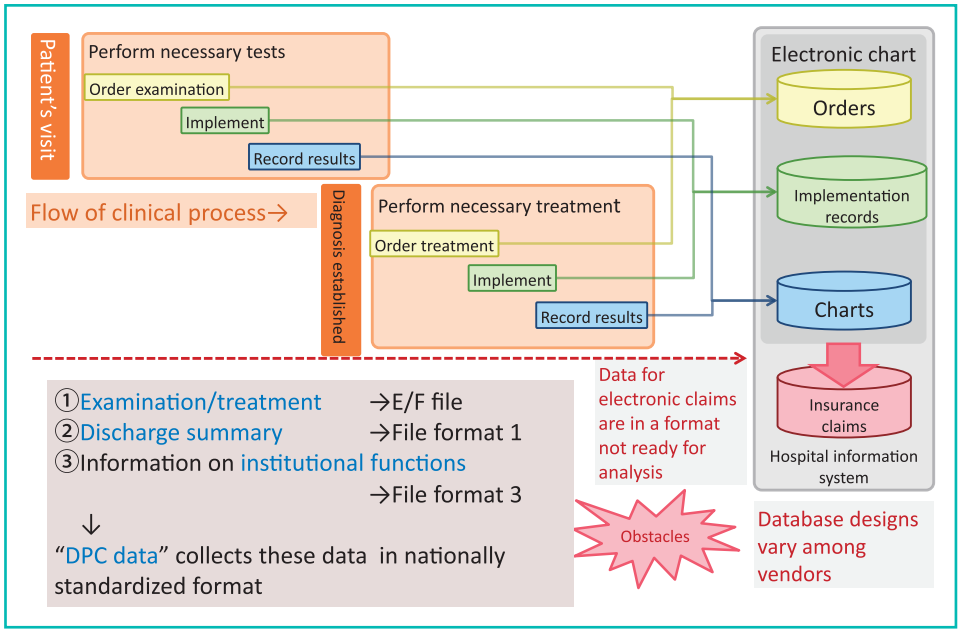


Fig. 2 Flow of clinical process and data

i.e., collection of data not directly prescribed by law. One example is the diagnosis procedure combination (DPC) survey. Anonymized data are collected from medical institutions by the MHLW, as part of an information gathering process conducted under the Health Insurance Law. In a sense, this is similar to conducting studies. The MHLW uses these data as source materials for revision of DPC classifications and price setting. For the purpose of collecting information on adverse drug reactions, the adverse drug reactions observed at the scene of clinical practice are submitted to form a database, and the data is used for various usages including revision of the package inserts.

Recently, the private sector, healthcare related industry, and researchers are requesting to use such databases, so the government is experimenting in secondary data use of public data. However, as Dr. Ryuichi Yamamoto mentioned in his lecture, in reality, there are many restrictions on the secondary use of data.

Relationships between Medical Services and Data

Patients who visit medical institutions undergo necessary examinations according to physicians' orders; healthcare professionals working in various specialties perform examinations under these orders. The results of the examinations, such as direct outputs from medical apparatuses, diagnostic imaging and reports or diagnosis based on such data are then recorded in patient charts. Thus, medical information is obtained through 3 steps comprising orders, implementation records and results (**Fig. 2**).

After this examination phase, a diagnosis is established, and a treatment plan is then devised according to the algorithm or guidelines for the diagnosis. Next, the physician in charge gives instructions pertaining to the treatment, and the results are recorded after implementation of the treatment.

This flow of information is captured by hospital information systems (HIS). Physicians' orders are recorded and collected in an order-entry database, data from each process performed under physicians' orders are recorded as implementation records, and the results and summaries are recorded in medical charts.

The HIS have been very actively developed in Japan since the 1980s. Electronic recording of physician order entry became available in almost every hospital during the last century. In large-scale hospitals, supporting systems were developed to manage the ongoing process of examinations and treatments because these tasks were monumental. Later, development of electronic medical charts was started, and a consid-

erable number of medical institutions had obtained such databases by the beginning of this century.

Medical institutions are required to charge and collect medical treatment fees. So they sent charge data to payers, based on the data stored in the three databases.

Obstacles to Secondary Use of HIS Data

Although in-hospital systems were functioning efficiently, there were barriers to the secondary use of these data.

For instance, the structures of the databases varied according to the vendor (company) that had developed the hospital information system. Even when the databases of hospital A and hospital B were developed by the same vendor, the databases were often incompatible due to the differences in versions of the product.

Sometimes, the code for recording a service in hospital A may differ from that in hospital B, bearing obstacles to compare the data from two hospitals. To solve such issues, standardized codes for electronic processing of medical fee bills were developed and efforts were made to promote the use of such codes.

Even when above issues were resolved, the data format for electronic claims was not suitable for analysis. They were recorded in a way mimicking paper claims and careful processing was needed to transform them to tabular data. This problem persisted for a very long time, until DPC data format was made available.

DPC Data Collected in Nationally Standardized Format

The core concept of DPC data is the submission data in an itemized list format. Required items in the medical fee bill, such as tests and procedures performed, drugs administered are exported as a list from in-hospital database. This format is referred to as E/F files, recording information on when, where, and to whom a certain examination (e.g., x-ray examination) or a certain drug at a certain dose is provided.

In DPC-based reimbursement system, information about the diagnosis and treatment of patients is needed to classify patients into case mix. To collect these data, discharge summaries carrying patients' basic demographics, diagnosis and other items in a standardized manner are recorded in file format 1.

It is also important to evaluate the functions of each medical institution in the DPC system. Relevant data on institutional functions and organizational structure is collected in file format 3. Using these standardized formats, data have been collected since 2006.

Reasons for Producing DPC Data

The DPC data is primarily produced to gather information for the prospective payment system for acute-phase inpatient hospital care [DPC/per-diem payment system (DPC/PDPS)] (**Fig. 3**).

The DPC/PDPS system is basically comprised of two technologies: grouping of patients by DPC, and bundled payment per day according to DPC classification.

However, another important technology in DPC/PDPS is the DPC data. They were produced in more than 1,800 hospitals in 2014. Although they are only prepared by about a quarter of general hospitals nationwide, their coverage of general hospital beds now exceeds 60% and they cover 70% of discharges, i.e., 10 million of 15 million discharges per year.

Currently, DPC/PDPS prospective payment system is only applied to acute inpatient care, but inpatient care is not independent from outpatient care. The MHLW research group on DPC/PDPS began to collect outpatient care data in E/F files since 2006. Afterwards, the MHLW started to do so since 2012. This makes the DPC data format a de facto standard for collection of clinical information in both inpatient and outpatient settings.

Information Collection through DPC Data

The basic reason for the DPC data format to spread across the nation to present level is that submitting data was required for claiming medical fees by DPC/PDPS. However, starting this year, not only DPC-paid hospitals but all facilities that claim basic inpatient care charges with higher nursing standards (7-nurses-per-patient nursing) are required submit inpatient and outpatient DPC data. Also, hospitals with regional comprehensive care wards (i.e., hospital wards

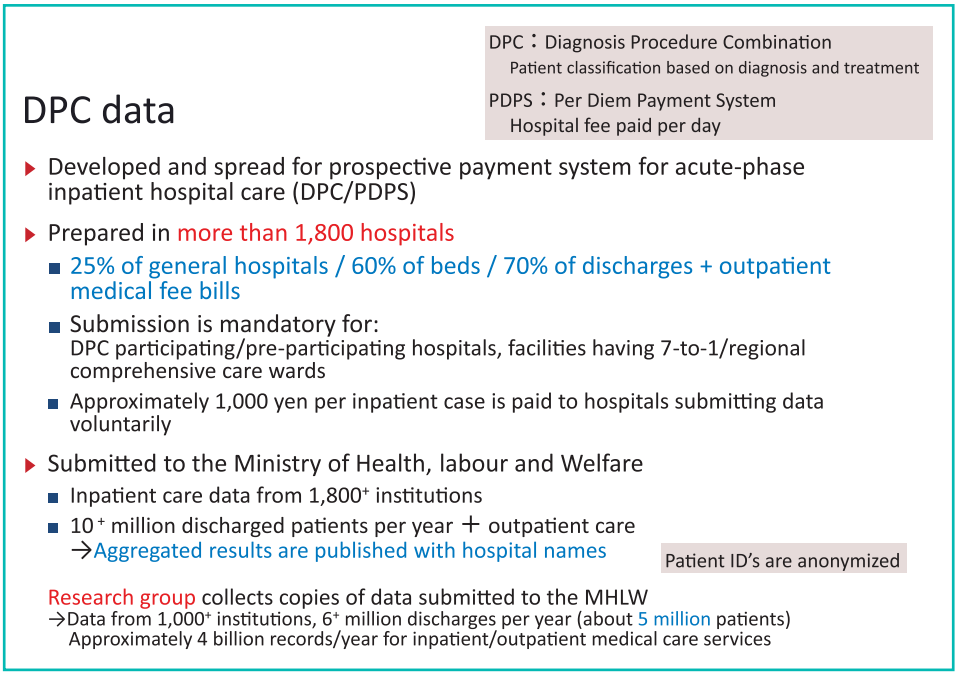


Fig. 3 DPC data

that accept patients in need for medical care after the acute phase and provide support to return to their homes or to nursing facilities), are required to submit DPC data.

In addition to these hospitals that are obliged due to claiming special inpatient services, other hospitals can receive 1,000 yen per admission if they submit DPC data voluntarily. For such reasons submission of DPC data has grown to a level previously mentioned.

It should be noted again that DPC data are anonymized at the time of submission from medical institutions. The MHLW collects these data and analyzes them according to the medical institution code. The results of analysis are published together with the names of hospitals to provide reference information for evaluation of hospitals within the DPC system. They are made available to the public as an open data and consumers can see which hospitals are participating in DPC/PDPS and how they provide care.

Big Data Usable for Multiple Purposes

DPC data are not solely used by the MHLW, but are also used by hospitals for administrative and other purposes. They can analyze DPC data either internally or with support from third-party organizations such as consulting companies. At the same time, there is a research project funded by the MHLW Grants-in-Aid for Scientific Research that collects DPC data from hospitals. The research group has a long history starting from introductory phase of DPC/PDPS.

Currently the research is led by Professor Kiyohide Fushimi of Tokyo Medical and Dental University. Participating researchers like myself work under his supervision. Out of 1,800 hospitals that submit DPC data to the MHLW, two-thirds, about 1,000 hospitals provide copies of the data to our research group. The data currently holds more than 6 million discharged cases for 5 million anonymized patients per year.

Detailed data for inpatient and outpatient clinical services (what treatment was given to which patient, on which date, and in which hospital) sums up to 4 billion records per year. In recent years, these data have become available all year round. In our research group, the data is available since 2010.

These data are not collected pursuant to the law or as part of the government project. Data are collected for research after obtaining approval from the ethics board of research organization. The study complies with the personal information protection policy for research and

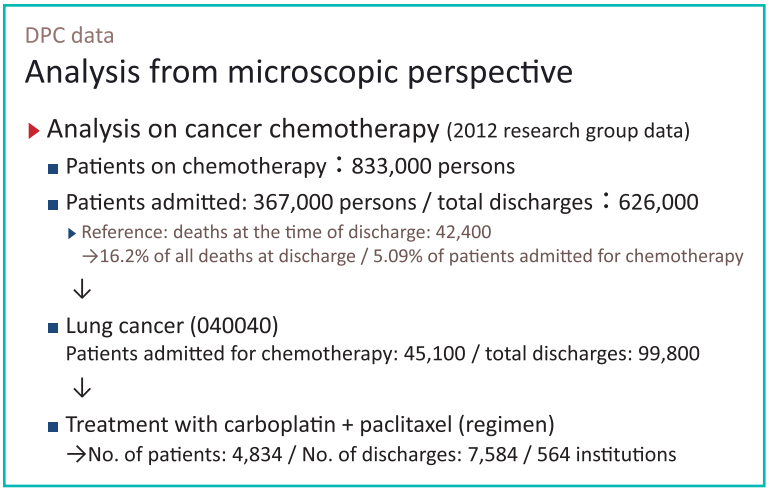


Fig. 4 Analysis from microscopic perspective using DPC data

the protection of trial subjects stated in the Declaration of Helsinki. As the research does not involve medical interventions, and uses only anonymized retrospective data, the study protocol is devised according to the ethics policy for epidemiologic research. For researchers, a large standardized data at reasonably low cost is already available for analysis.

So let me show you some findings from the research group.

Analysis from the Microscopic Perspective

Here, I will show you an example from microscopic perspective (**Fig. 4**).

Because I work at the National Cancer Center, I have been in charge of analyzing data on chemotherapy for cancer in our research group.

In 2012, the research group had data on approximately 6 million inpatient care cases or 5 million inpatients. Including outpatient data, the number of unique identifications (IDs) reached tens of millions. Using that data, we identified 830,000 patients who received chemotherapy related drugs at either inpatient or outpatient settings. The diagnosis or the site of cancer was not available for outpatients, but for 360,000 inpatients diagnosis with ICD-10 codes were available. We linked 630,000 admission and outpatient records to the 360,000 inpatients for analysis.

We found about 42,000 deaths for cancer chemotherapy patients over 12-month period. These deaths account for about 16% of all deaths observed in DPC data used by our research group. About 5% of patients who have undergone chemotherapy had died.

Analysis by Disease Name Is Feasible

With DPC data we can not only tabulate patients by chemotherapeutic agents but also by diagnosis. Among the 360,000 patients, 45,000 were lung cancer patients. The total number of admission was nearly 100,000. There, we can see how the hospitals provided chemotherapy.

The results of this type of analysis have been published or reported since 2007. Two-drug chemotherapy using carboplatin and paclitaxel is common in the treatment of lung cancer, and 4,800 patients, corresponding to a little more than 10% of the 45,000 patients with lung cancer, receives this regimen. They had 7,500 admissions in a year and 564 hospitals provided the therapy.

Although knowing the chemotherapy regimen by diagnosis is in itself very informative and important, specific analysis of the E/F files of these patients provides the following useful information.

Figure 5 shows the data from patients with lung cancer (040040) who underwent two-drug chemotherapy with carboplatin and paclitaxel. It is a complex graph, and it may be difficult to find the point most worth paying attention to, but attention should first be focused on the light blue, orange, and gray bars. This bar graph has the length of stay on the horizontal axis and the

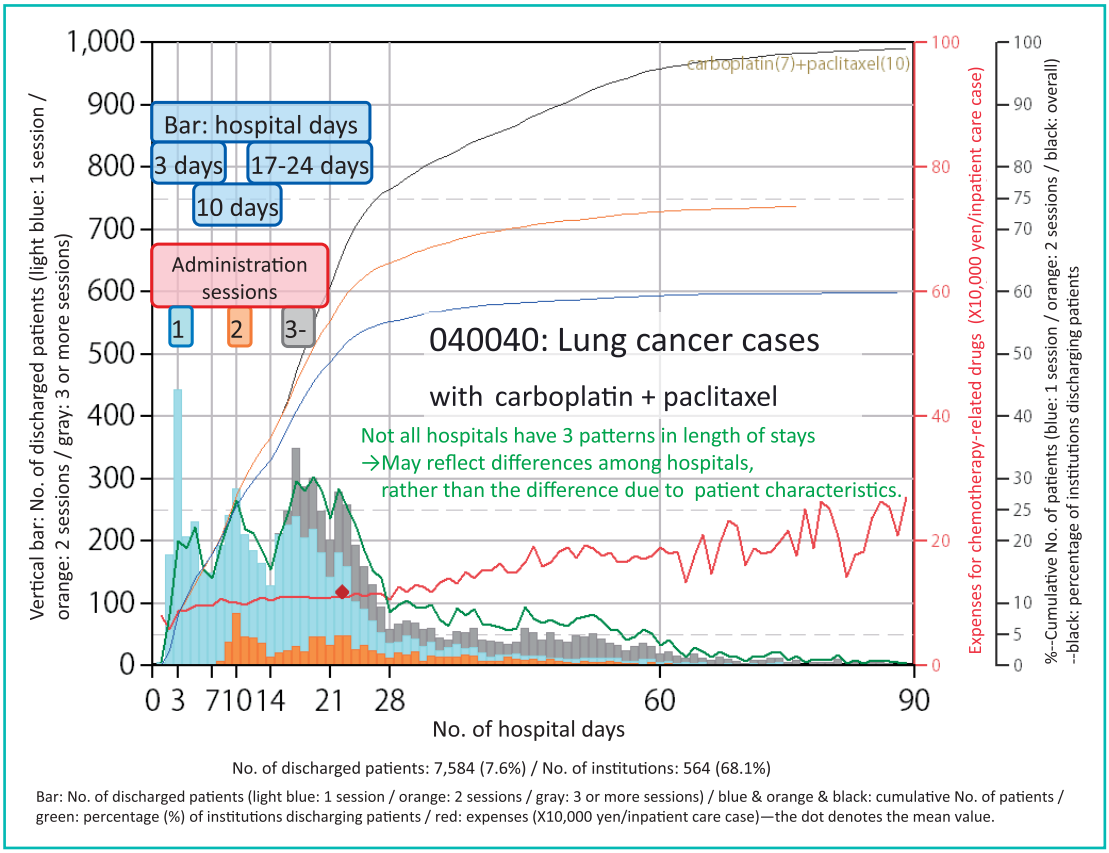


Fig. 5 Lung cancer chemotherapy admission (treated with carboplatin and paclitaxel)

number of patients on the vertical axis.

As is clear from the graph, a hospital stay of 3 days and 2 nights was most common among approximately 7,000 inpatient cases. There were nearly 400 patients. But we can see a peak in length of stay at 10 days and another at 17-24 days, showing a total of 3 peaks.

Therapeutic Processes in Medical Care as a Whole Can Be Understood

Many anticancer drugs are not continuously administered for many days. The standard method of administering this chemotherapy relies on giving the drugs in a cycle of 1 or 3 weeks.

Let us pay attention to the number of administrations of anticancer drugs during a hospitalization. The light blue part of the bar graph denotes patients who received only one session per hospitalization, whereas the orange and gray parts denote those who underwent 2 sessions and 3 sessions, respectively. For example, the orange part at 10 days shows patients who had the first session and spent a week in the hospital and were discharged after the second session. For patients in gray part, they received 3 sessions of one-week-cycle chemotherapy. But for light blue bars representing a considerable number of patients, they received only 1 session of chemotherapy in an admission.

The necessary length of stay usually varies according to the patient's disease condition. However, these 3 patterns in length of stay were not necessarily found in all hospitals. Let us pay attention to the green line in **Fig. 5** that denotes the percentage of hospitals with patient discharged at specific length of stay. There is a scale of 0-100 on the right side. On this scale, the peak percentage of hospitals reaches only up to about 30% regardless of length of stay.

If majority of hospitals had 3 peaks, it may be reasonable to assume that hospitals choose the length of stay according to the patients' conditions. However, under the situation that each peak reaches only 30%, it seems to indicate that three patterns are inherent to the practice pat-

terns of treatment at each hospital, rather than being attributable to the characteristics of the patients.

This is an analysis that yielded outstanding results. But please be relieved to know that therapeutic strategies do not differ to this extent for other types of cancer or regimens among hospitals. Even in cases with lung cancer, regimens using other drugs are more standardized. For women with uterine cancer or ovarian cancer, the same combination of drugs, carboplatin and paclitaxel are more standardized; a hospital stay of 3 days and 2 nights is adopted in more than 50% of hospitals.

Analysis of DPC data shows which practice patterns are currently common and how much variability exists between hospitals. It gives guidance on what kind of care we can guarantee for patients in Japan, and that had not been cleared up to now. These features will be visible through the use of medical big data.

Analysis from the Macroscopic Perspective

Then what can we see from the macroscopic perspective, if we use DPC data (**Fig. 6**).

We have seen difference in patterns of care, and let us turn to see how care is provided from a wider perspective.

In DPC survey data, 7-digit postal code for the patient address was included as an item in file format 1, which holds discharge summaries. If we can map the 7-digit postal codes on a map with latitude and longitude, we will be able to know from where patients are coming. In addition, if we can see which hospital patients are admitted to, we will be able to know which hospital is responsible for the community and what role they play.

As a pioneering example of such analysis, Dr. Yamamoto presented the findings by Professor Shinya Matsuda, from an analysis of NDB data describing a cross-section of location of patients (or payers) and location of hospitals. With DPC data processed properly, we can do more detailed analysis like I show you next.

Discharge summaries recorded in file format 1 include information on whether or not a patient was transferred by ambulance on admission. Extracting acute myocardial infarction (AMI) patients transferred by ambulance and plotting them would produce a map shown in

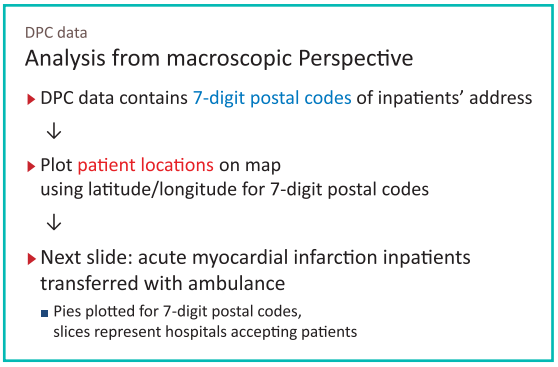


Fig. 6 Analysis from macroscopic perspective using DPC data

Fig. 7.

In order to protect the privacy of patients following ethics guidelines, and to adhere to the data use agreements of participating hospitals, background map is not shown on **Fig. 7**. It shows relative positions of the 7-digit postal codes, and color represents hospitals accepting patients.

On the upper right side of the figure, there are 2 medical institutions shown in dark and pale pink that cover a region. For other regions, dense plots of postal codes reflect area with high population density. In such areas, there are multiple hospitals to which patients are admitted to. It shows that geography divides hospitals market. In urban areas hospitals as a group is responsible for care for AMI in a region.

These are the results from analysis of DPC data, from microscopic and macroscopic perspective.

Strengths of DPC Data

Using DPC data, which constitute an example of medical big data, analysis on clinical care processes will provide statistics on hospital care by DPC (patient category) or by provider (hospitals). It is possible to see the length of stay for chemotherapy admissions using particular drugs, allowing comparisons among different medical institutions (**Fig. 8**).

There is an analysis by members of our research group, comparing conventional and laparoscopic surgery that found peculiar difference in the percentage of patients who required blood transfusion and the actual transfusion volume required. Because DPC data records clinical tests or interventions with the computerized

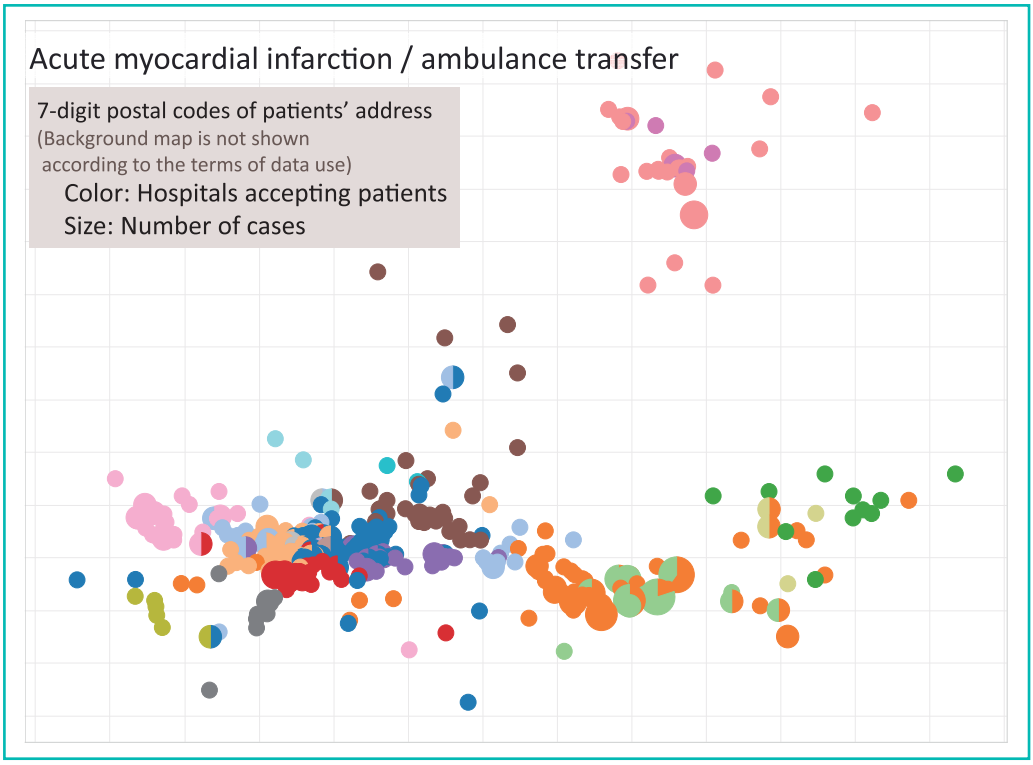


Fig. 7 Map of acute myocardial infarction patient address by accepting hospital

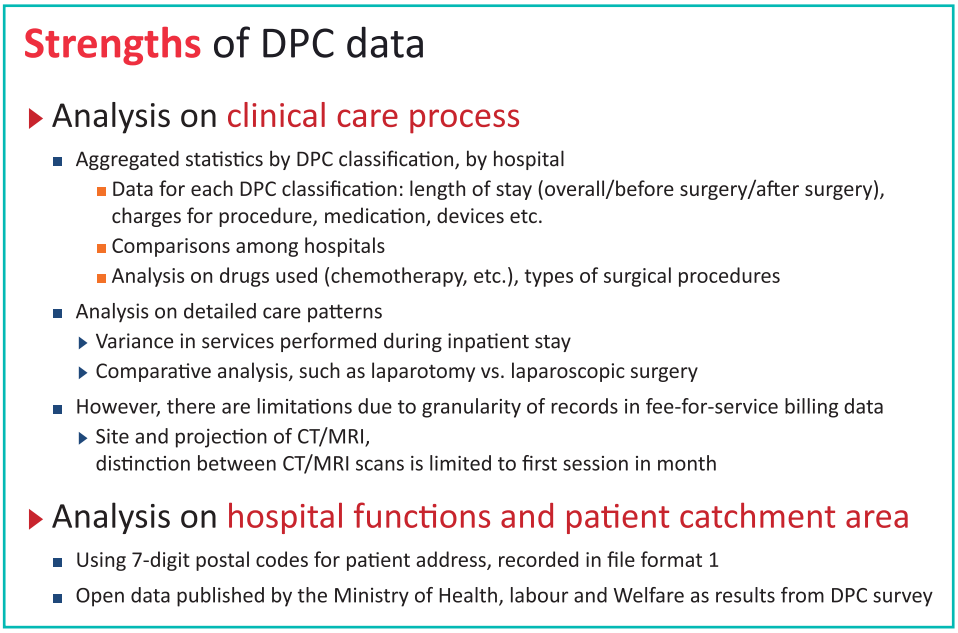


Fig. 8 Strength of DPC data

code for reimbursement, there are limitations for analysis at a detailed level.

With regard to diagnostic imaging tests, patients with lung cancer initially undergo CT examination to determine the tumor size, and then brain and/or liver may be examined for

Hospitals for lung cancer in Fukuoka Prefecture (10 or more patients per year)
https://public.tableau.com/views/DPCH25-24/sheet2

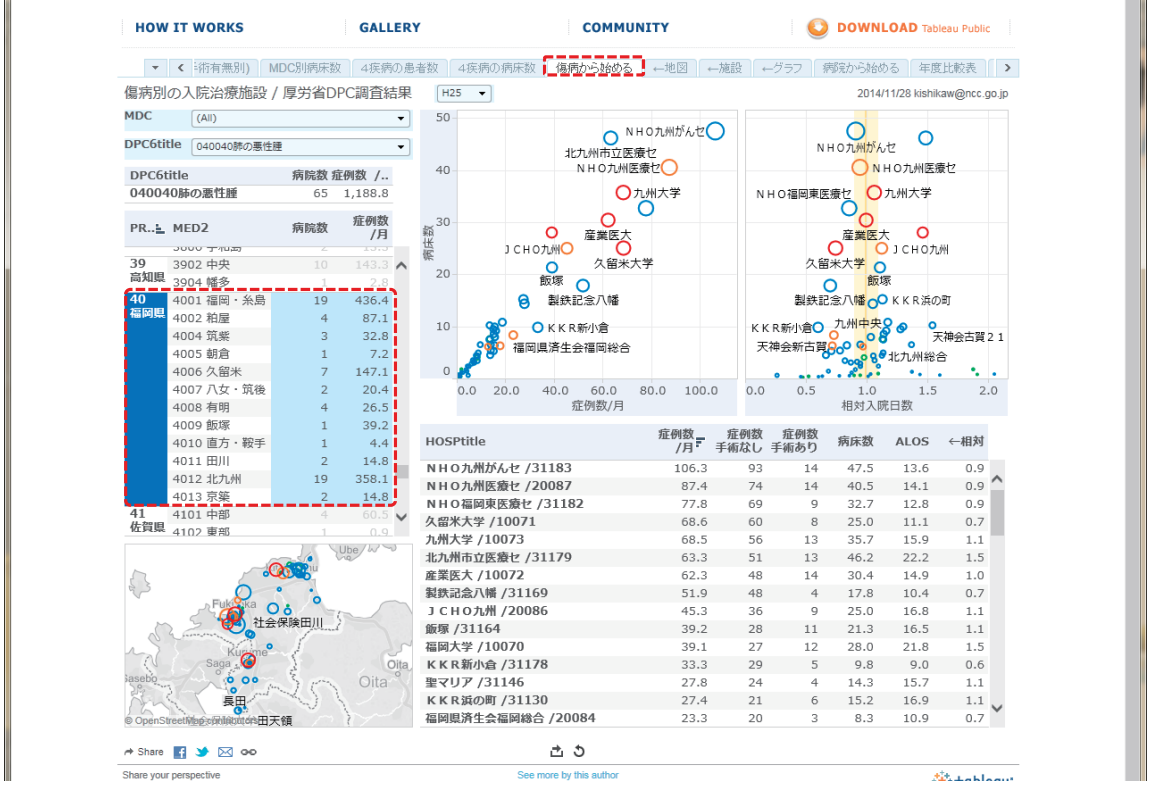


Fig. 9 DPC survey results by the MHLW (2013): Hospitals for lung cancer inpatient care in Fukuoka Prefecture

metastasis if they are suspected. For such occasions, DPC data allows us to analyze the number of examination. However, the site of interest for a CT examination is not recorded in the current format of the medical fee bill or items covered by DPC data. So there is a limitation in studies only using DPC data.

Analysis of DPC Data at a Higher Level

DPC data also allows us to analyze the functions of hospitals within a geographical area. If raw data were available, we can obtain maps like **Fig. 7**. More advanced analysis will be possible, if we use the results of DPC survey published by the MHLW.

The DPC survey results are available from the MHLW website. Published data include list of participating hospitals for the survey, distributions of patients by gender and age group by DPC patient classification, chemotherapeutic regimens and drug combinations used, and so on. In addition, detailed data for participating hospitals are also available concerning the type of treatment, number of treated cases, and type of disease by diagnosis or by major diagnosis category.

For instance, **Fig. 9** shows the results for lung cancer admissions. The table in the lower right panel of the figure shows the list of hospitals in Fukuoka prefecture, ranked by the number of treated cases per month. If we focus on the number of hospitals, we see that there are many hospitals for lung cancer in this prefecture.

In Fukuoka Prefecture, there are 62 institutions with 10 or more lung cancer discharges per year. **Figure 10** is a map produced by plotting each hospital on it, and coloring according to the time required for residents to reach the nearest hospital.

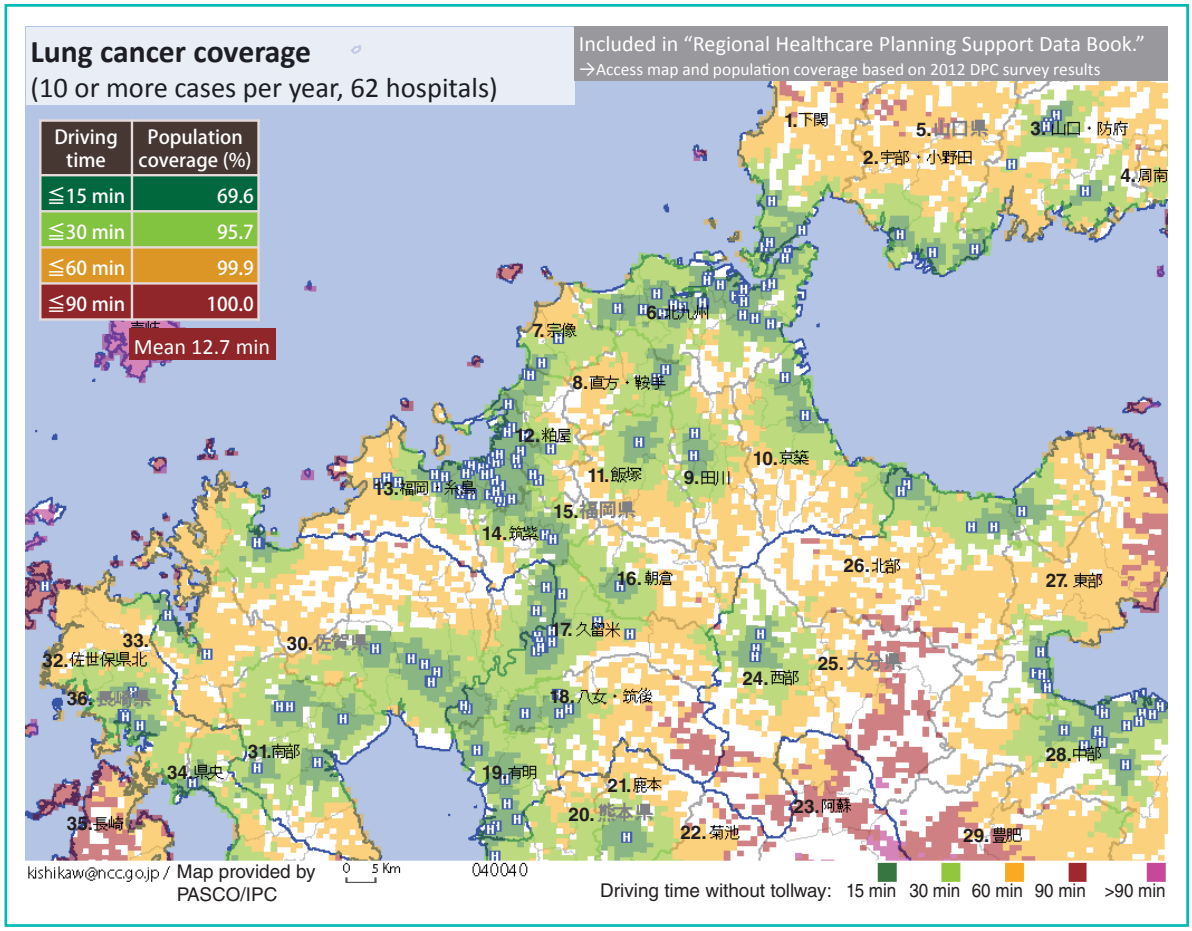


Fig. 10 Lung cancer inpatient care coverage

1. Shimonoseki; 2. Ube/Onoda; 3. Yamaguchi/Hofu; 4. Shunan; 5.Yamaguchi Prefecture; 6. Kitakyushu; 7. Munakata; 8. Nogata/Kurate; 9. Tagawa; 10. Keichiku; 11. Iizuka; 12. Kasuya; 13. Fukuoka/Itojima; 14. Chikushi; 15. Fukuoka Prefecture; 16. Asakura; 17. Kurume; 18. Yame/Chikugo; 19. Ariake; 20. Kumamoto Prefecture; 21. Kamoto; 22. Kikuchi; 23. Aso; 24. Western area; 25. Oita Prefecture; 26. Northern area; 27. Eastern area; 28. Central area; 29. Hohi; 30. Saga Prefecture; 31. Southern area; 32. Sasebo; 33. Northern area; 34. Central area; 35. Nagasaki; 36. Nagasaki Prefecture

If we sum up the population by driving time, 70% in the prefecture can be admitted to a hospital within 15-minute range. If patients may travel up to 30 minutes, the coverage of the population rises to 96%. We can use microscopic DPC data in a macroscopic analysis to see what can be done in local areas and how it impacts the population.

For lung cancer, the population coverage is high, and we may satisfy with the level of care provision in Fukuoka. On the other hand, for uterine cancer, there are only 26 institutions with 10 or more discharges per year (**Fig. 11**). In this case, people who can receive inpatient care for uterine cancer in a local hospital within 15 minute drive account for little less than 50% of the population. A quarter of patients may drive to an adjacent town for 30 minutes to be treated, but the remaining quarter of prefecture has to drive over 30minutes. That is a remarkable difference from lung cancer, and it is evident that accessibility to inpatient care differs by types of cancer.

Weakness of DPC Data

Today, we saw the strengths of DPC data with examples from microscopic and macroscopic analyses. However, there are research areas not suitable for DPC data (**Fig. 12**). For instance,

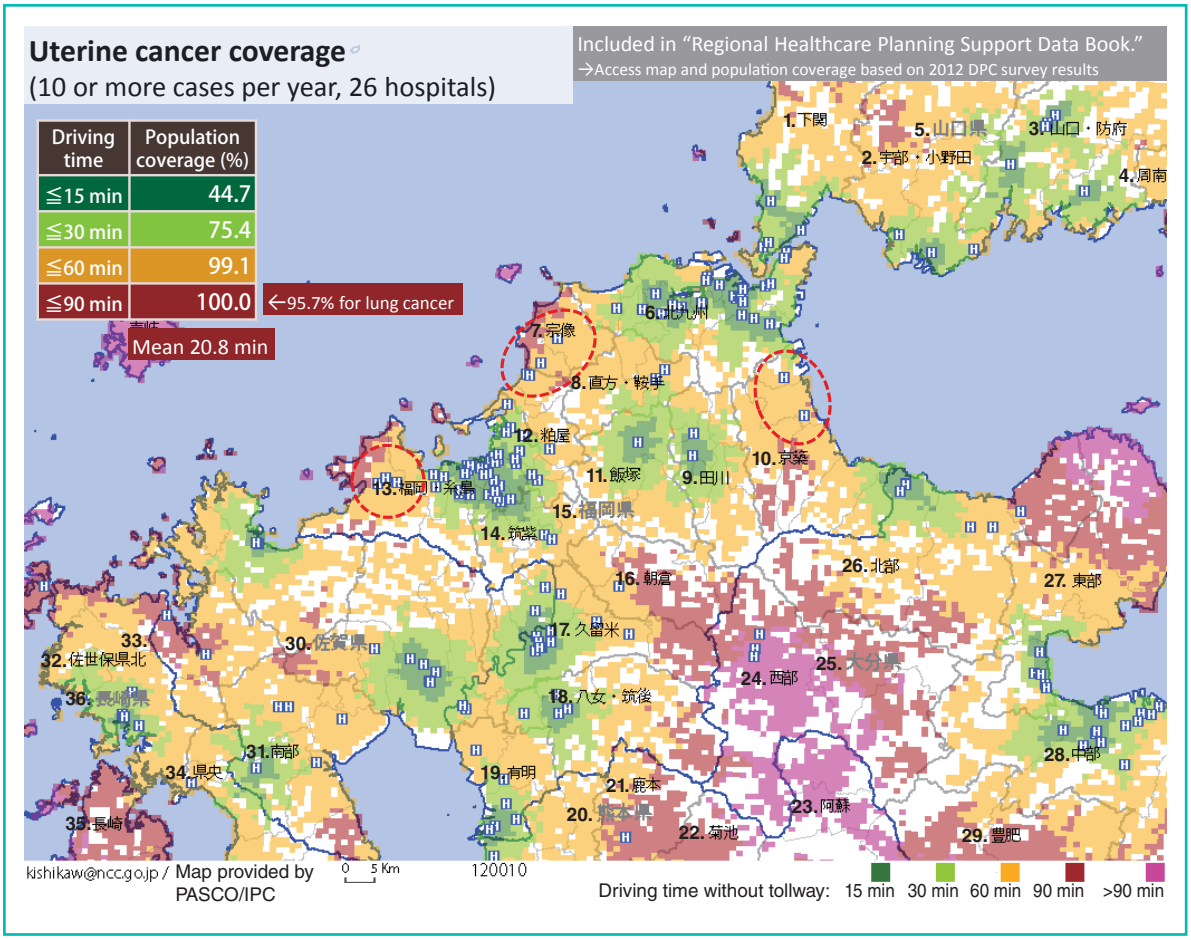


Fig. 11 Uterine cancer inpatient care coverage

1. Shimonoseki; 2. Ube/Onoda; 3. Yamaguchi/Hofu; 4. Shunan; 5.Yamaguchi Prefecture; 6. Kitakyushu; 7. Munakata; 8. Nogata/Kurate; 9. Tagawa; 10. Keichiku; 11. Iizuka; 12. Kasuya; 13. Fukuoka/Itojima; 14. Chikushi; 15. Fukuoka Prefecture; 16. Asakura; 17. Kurume; 18. Yame/Chikugo; 19. Ariake; 20. Kumamoto Prefecture; 21. Kamoto; 22. Kikuchi; 23. Aso; 24. Western area; 25. Oita Prefecture; 26. Northern area; 27. Eastern area; 28. Central area; 29. Hohi; 30. Saga Prefecture; 31. Southern area; 32. Sasebo; 33. Northern area; 34. Central area; 35. Nagasaki; 36. Nagasaki Prefecture

DPC data include deaths at the time of discharge, but there is no information about long-term prognosis after discharge. To obtain such data, additional effort is needed to collect them. In another case, if we want to investigate on adverse reactions of a certain drug, it may be possible to ascertain the presence of neutropenia, from a diagnosis recorded with the ICD10 code, but specifics on the severity of neutropenia remains unclear.

To overcome such weaknesses, we believe that the key to the better use of DPC data is to use them in rapid identification of patients and gathering of basic data, and extend them by adding data needed to answer specific research questions.

Collaborative use of data is an important concept for practical use of big data, as mentioned in Dr. Yamamoto's lecture. In collaborative research project with Chugai Pharmaceutical Co., Ltd., the National Cancer Center collected anonymized DPC data over a span of 5 years. There, we identified 13,000 patients on chemotherapy and proceeded to chart review for 884 colorectal cancer patients to clarify the type of treatment and their survival. Because we were successful with that attempt, we are now extending to 20 cancer centers around the nation to enlist colorectal cancer patients with DPC data and to enforce the data with additional clinical

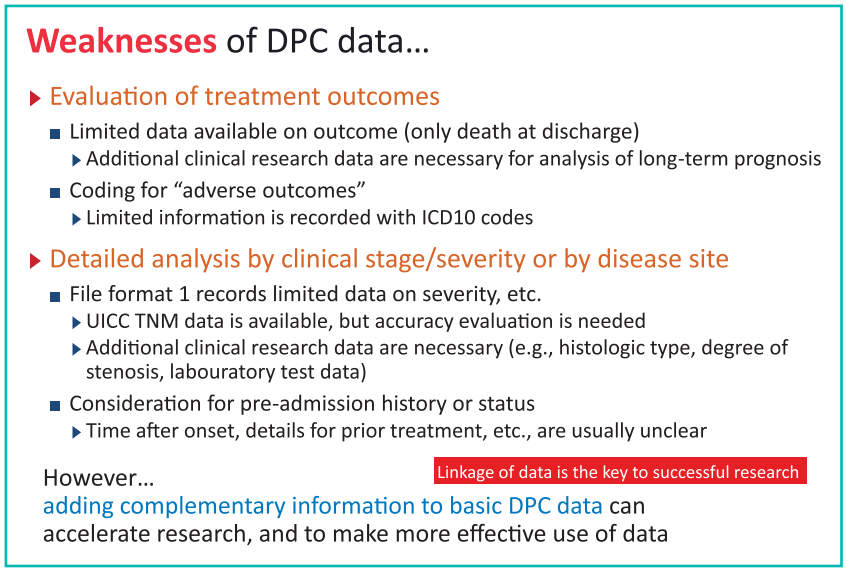
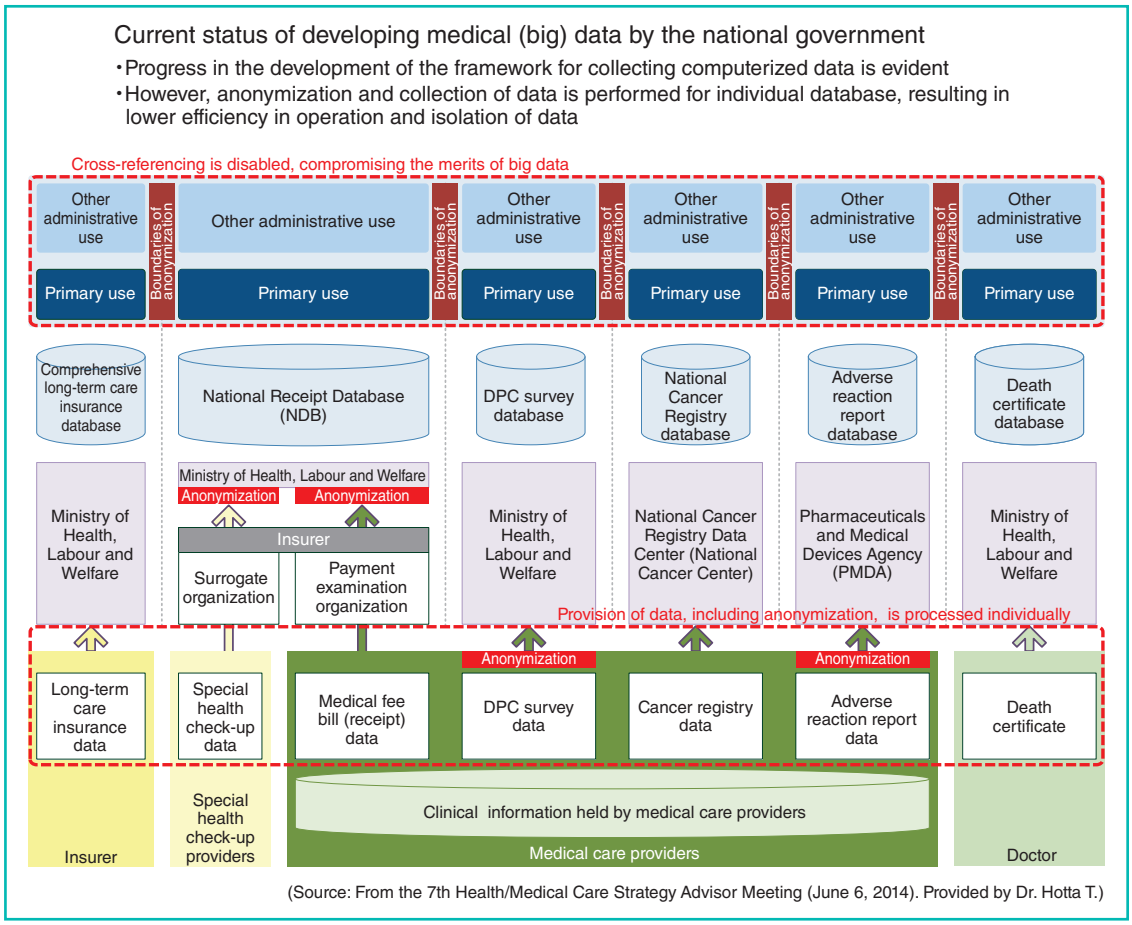


Fig. 12 Weaknesses of DPC data



(Source: From the 7th Health/Medical Care Strategy Advisor Meeting (June 6, 2014). Provided by Dr. Hotta T.)

Fig. 13 Current status of developing medical (big) data by the national government

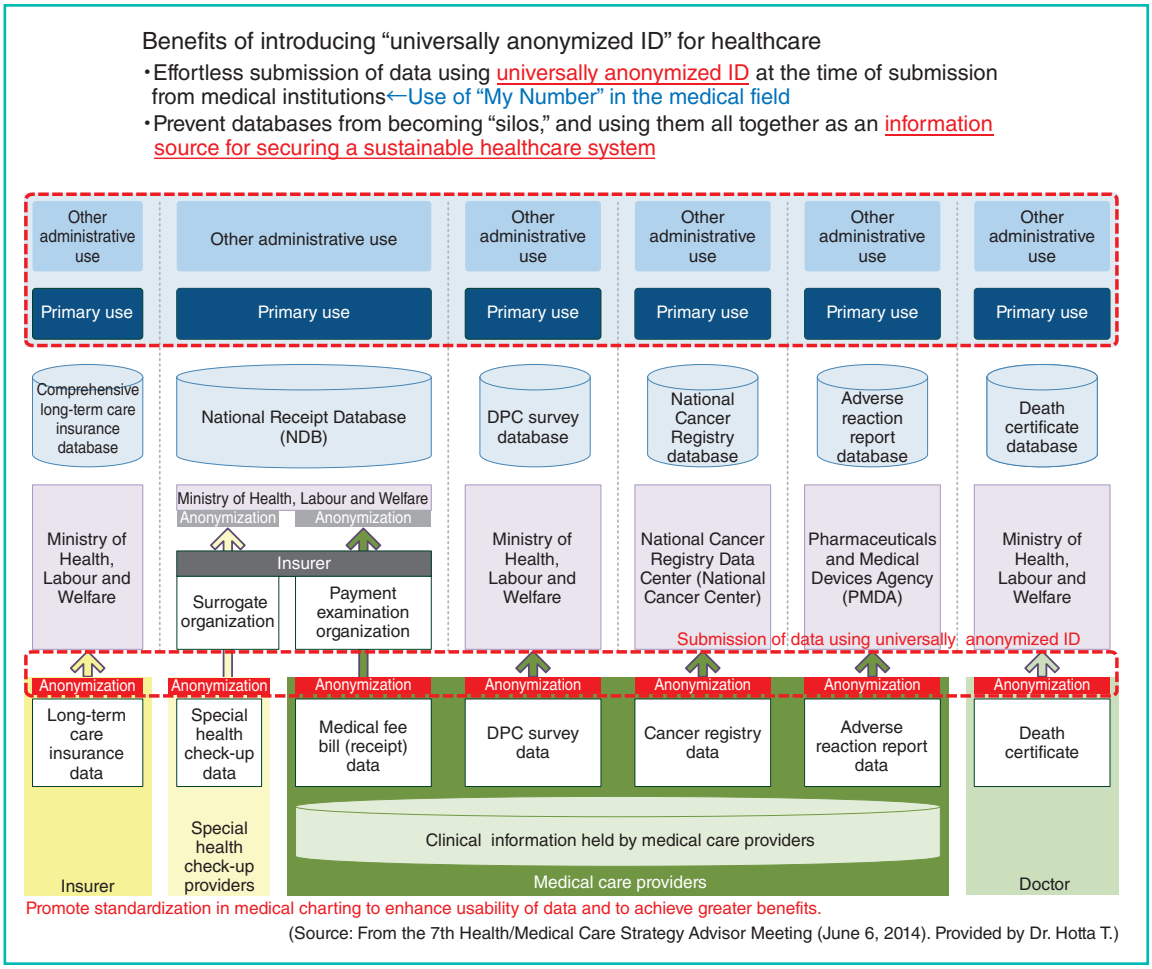


Fig. 14 Benefits of introducing "universally anonymized ID" for healthcare

and prognostic information from hospital cancer registry.

State of Development of Medical Big Data

What can the government do to promote the advancement of this type of research in the future?

Figures 13 and 14 are the materials presented by Dr. Hotta, President of the National Cancer Center, at the Health/Medical Care Strategy Advisor Meeting in June of last year.

As shown in the slide, various data are accumulated in this country through the development of medical big data promoted by the government (**Fig. 13**).

This raises the issue of the burden on the provider of data. They must either choose from non-anonymized or anonymized data. If the data needs to be anonymized, the methods differ from a project to another.

At the same time, analysis of data is focused to project-dependent purposes and each process is segregated from a project to another, lacking in overall harmonization. Therefore, if we want to integrate the findings from multiple projects, we can only combine their results, but are unable to combine individual data for more streamlined analysis. That compromises the merit of using big data.

Recognizing such issues, it is desirable to allocate a universally anonymized ID at the time of data submission from medical institutions. If non-anonymized data are necessary for a project, it is a good practice to add minimal non-anonymized data to anonymized data. When assigning universally anonymized ID, we can

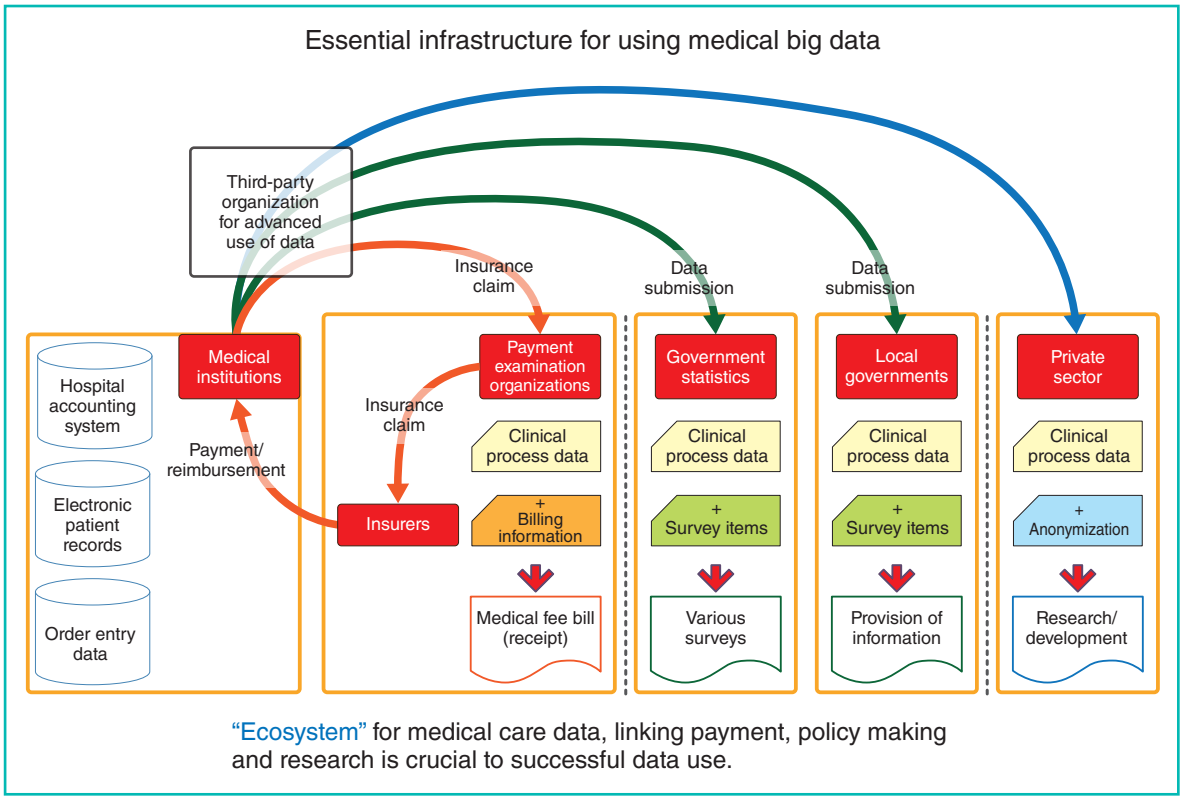


Fig. 15 Essential infrastructure for using medical big data

generate them for use in the healthcare sector, derived from the social security and tax number system (My Number System), as proposed by Dr. Yamamoto.

If the basic data are to be anonymized in this manner, and they are to be handled securely, they will become an essential information infrastructure for the nation. They will eliminate currently existing boundaries in anonymization and data processing, and will be an invaluable source of information to realizing a sustainable health care system (**Fig. 14**). But we should not forget the need for standardization in patient charting.

Major Issues in Using Medical Big Data

Another aspect that needs to be mentioned is on how data is submitted from providers. If possible, health care data should be once collected by a single organization and then routed to where they are needed. Currently, providers submit the data individually for each purpose, i.e. reimbursement of medical fees, national government statistics, and local government administrative procedures or for other private purposes (**Fig. 15**).

In order to make maximal use of digitized health care data, we must build an ecosystem that is effective and efficient. It should cover activities in the government and research community as well as insurance claims processing. To make such system come true, special consideration should be paid in the process of future revision of the Private Information Protection Law or the Social Security and Tax Number Law. Then the fruit of the ecosystem can be shared among health care providers, patients and the public.

This is the end of my presentation. Thank you for your kind attention.