Contribution of Medical Big Data Analysis Technology to Medical Innovations

● Tadaaki Katsuda ● Hayato Isshiki

Recently, utilization of big data by ICT has been attracting a great deal of attention in various fields in Japan and overseas. In the medical field, for the purpose of realizing personalized medicine, innovations ranging from diffusion of electronic medical records to accumulation of medical care information and further to integration of genome information have been promoted up to now. However, there is a limit to manually processing the large volume of information accumulated by these, and technology utilizing medical big data that can be effectively and efficiently processed is considered necessary. Accordingly, Fujitsu has developed medical big data analysis technology. This technology meets the requirements specific to the medical field such as execution of analysis with promptness, high reliability, and accuracy ensured and an appropriate interface independent of the analysis skills of users. This technology has been confirmed to be capable of contributing to medical innovations through actual medical care information analysis conducted jointly with the National Hospital Organization's Nagasaki Kawatana Medical Center. This paper presents the medical big data analysis technology developed by Fujitsu.

1. Introduction

A paradigm-shift is taking place globally in recent years concerning the use of big data through ICT. Ever since 2012, the USA and major countries in Europe and Asia have mentioned the big data utilization, and certain government policies and government-led projects are underway.[1] In the USA, for example, the Office of Science and Technology Policy under the Executive Office of the President (EOP) has been leading a big data R&D initiative since 2012, providing a total budget of more than 200 million dollars for the development, which is shared among six governmental institutes over five years. South Korea has also prepared a big data master plan, which secured 500 billion won over four years starting in 2012 to be allocated to the promotion of big data use. These are only a couple of the many cases in which national governments are pursuing projects leveraging big data.

In Japan, meanwhile, the government is developing national strategies and legislations, aiming to establish Japan as a so-called "Data Nation." In the 2016 Japan Revitalization Strategy, 10 Strategic Public-private Joint Projects have been advanced revolving around the key factors of AI, IoT, and big data, aiming to bring the GDP to the 600-trillion-yen level.[2] The White Paper of the Ministry of Internal Affairs and Communications also nominates the manufacturing, transport, and service industries as the fields of application, and various industries are joining big data utilization.[3]

The medical field is also anticipated to benefit from the use of big data. The social security system in Japan is in a critical situation today, as the national population continues to age quickly while the birth rate declines and medical costs grow. Given these issues, the national government has been considering the utilization of medical big data in a system that supports technological R&D and enables the efficient provision of effective medical services made widely available to the population. The government submitted a bill in 2017, the so-called Medical Big Data Act, concerning the handling of anonymously processed medical data used for R&D in the fields of medical and health care. This bill, when passed, will enable institutions approved by the government to collect a large volume of anonymously processed medical data without having to obtain patients' consent. It will support not only basic research and pharmaceutical developments but also the realization of personalized medicine. It is therefore expected to improve the situation facing the country's social security system.

This paper describes the recent trends in big data utilization, with a focus on personalized medicine in Japan. Then, stating the technology required for the use of big data in medical care, we will describe the characteristics of the medical big data analysis technology that Fujitsu has developed, and the results of a trial study.

2. Trends in medical big data utilization

As electronic medical records systems have become prevalent, it is expected that the accumulated medical data may be leveraged to realize personalized medicine. An electronic medical records system replaces conventional paper-based medical records, which physicians used to manually prepare, with an electronic system in which data can be recorded, edited, and stored electronically. So far, 72.9% of hospitals in Japan with 400 beds or more have adopted electronic medical records.[4] A large volume of patient-centered medical data has been compiled through these systems. It is thus expected that, leveraging these data, personalized medicine may be advanced, making it possible to select medical care individually tailored to specific patients.

In personalized medicine, the utilization of genomic information is as important as that of medical data. For example, sometimes the same drug does not lead to the same outcome when it is prescribed to two patients who suffer from a symptomatically identical cold. This is often due to differences in individual genetic and constitutional characteristics of the patients, but it is difficult to understand this based on medical information alone. This is where genomic information is important in addition to the medical data. Genomic information contains vital information from which a human being is composed, such as DNA sequences, which form a recipe for life. Therefore, genomic information provides crucial information that reveals differences between different patients.

Thanks to the diminishing costs of genomic sequencing in recent years, the integration of medical and genomic information is advancing. A new technology has been developed, called the next-generation sequencer, that realizes a super-fast analysis of genomic information, leading to a reduction in costs for conducting the sequencing.[5] As a result, more and more medical practices introduce genomic information and integrate it into their medical records. For example, Tohoku Medical Megabank Organization of Tohoku University is pursuing the integration of medical records and genomic information for 150,000 local residents, as they aim to rebuild community medical care following the damage from the Great East Japan Earthquake and establish personalized medicine.[6]

While the growing diffusion of electronic medical records and integration of genomic information has helped to significantly increase the volume of electronic medical data, this massive body of information is too much to process manually. Medical big data analysis is one of the methods to facilitate the processing of the massive data and propel the realization of personalized medicine forward.

3. Purposes of medical big data analysis

The number of medical records accumulated in recent years through the use of electronic medical records systems has reached several hundred thousand. Analyzing this massive data puts a significant load on the system. With genomic information integrated, the number of variables used in the analysis can possibly be in the tens of thousands or even tens of millions. The more variables involved, the more contaminated the data can become. For example, there are risks of false positives and/or false negatives being included in the analyzed data. A false positive is when a person's test returns a positive result despite the absence of the illness, and a false negative is when a person's test result is negative, depending on which test methods were used, despite being afflicted with the illness. Furthermore, each patient undergoes different sets of examinations. This leaves many examination items with no values. Data sets with void values impede machine learning and analysis engines from working properly. However, as it is developed for medical care, which deals with human lives, there is no excuse for delayed responses before an analysis yields results, analysis results that include false positives/negatives, or an analysis engine that malfunctions on certain data sets. Therefore, medical big data analysis must be able to return highly accurate analysis outcomes to the field quickly.

In the medical field, it is also important to ensure high reliability with the knowledge derived from big data analysis. The analytical algorithm for realizing personalized medicine may have an impact on patients' lives, and the knowledge in particular needs to be supported by clear reasoning. With machine learning, however, the knowledge production process is in most part a black box without much clear logical reasoning. Given this, guaranteeing the credibility of the knowledge obtained through the medical big data analysis is a must.

Furthermore, an interface must be able to facilitate analyses based on medical knowledge regardless of the analytical skills possessed by the system users. Big data analyses involve methods based on complex formula manipulations such as machine learning. For this reason, specific analytical skills, with prerequisites of complex formula manipulations, are required of users to be able to interpret the obtained outcomes. However, while most medical institutes are well-staffed with medical experts, they are less familiar with analytical skills. Therefore, there needs to be an interface that allows users to conduct big data analysis and interpret the outcomes based on medical knowledge.

4. Features of developed medical big data analysis technology

For the realization of personalized medicine, Fujitsu has developed medical big data analysis technology that utilizes medical data compiled in a large volume to extract a group of cases similar to the condition of a particular patient and predict the progress of the condition and effectiveness of treatments.

4.1 Search engine for similar cases and variable selection feature

The similar case search engine quickly processes large volumes of data that include missing values by conducting an analysis on the computer memory. As shown in **Figure 1**, one CPU can process 1 million records to ascertain the degree of their relevance in about 2 seconds. The search engine is also equipped with a feature to select the variables that are likely to be medically significant when processing the genomic information from tens of millions of variables available for the analysis, as stated earlier. By filtering out irrelevant variables, this feature helps to enhance the accuracy of the analysis. These features enable the search engine to handle a large volume of medical big data with missing values and return analysis results quickly and accurately back to the medical care site.

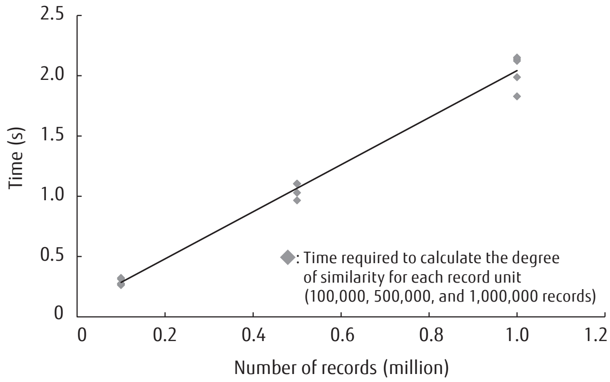

Figure 1
Processing time for analyses of number of records and similar case search.

4.2 Knowledge extraction feature

When obtaining the outcome from the analysis results based on certain premises, the knowledge extraction feature, which indicates a probability value (p-value) for speculating the premises based on observational results, increases the credibility of the analysis results. The knowledge referred to here is an abridged body of information about patients in similar conditions. Thus, this feature employs the primary data of these patients as the basis of the knowledge in order to ensure credibility. The similar case group is treated as a statistical cohort, which enables statistical verification. The knowledge is extracted from this, and the p-value is obtained by the likelihood, as stated above. The smaller the value is, the more credible the knowledge, and vice-versa.

4.3 Example of knowledge extraction feature application

The knowledge extraction feature has two functions: prediction of the progression of illness conditions, and prediction of the effectiveness of treatments. An example of these functions in practice is described in a case where a diabetic patient had a complication with diabetic nephropathy (**Figure 2**).

Based on the results of the condition progression prediction in patients with similar conditions, the median affliction period was 2,200 days, which turned out to be significantly shorter than other patient groups by 500 days. This means that the result for this particular patient is certainly bad. Meanwhile, the treatment effectiveness prediction makes it clear that patients in the similar case group with poor recuperation are found to be more likely not to suffer from diabetic nephropathy if they are treated with insulin injections coupled with GLP1 injections than if they are given other treatments. It is thus suggested in this prediction that this patient is highly likely to have complications with diabetic nephropathy, which may be prevented with a high probability by undergoing a treatment with insulin and GLP1 injections.

We also add that, in this particular case, the interface employed was designed so that users could operate the system easily with only medical knowledge and understand the analysis outcomes equally easily. **Figures 3** and **4** illustrate the characteristic features of these functions in the similar case search engine screen and the condition progression prediction screen. These screens do not require users to have analytical skills such as an understanding of complex formulas and data processing. They can interpret the outcomes in similar cases or analytical results displayed on the screen using their medical knowledge.

Take Figure 3 for example; this similar case search engine screen is designed with an interface that allows users to judge similarities based on the differences in the values of tested items. Figure 4 depicts the condition progression prediction screen, which displays the analysis results, incorporating the statistical methods such as the log-rank test and the Kaplan-Meier curve, which is often used in medical guidelines and case reports. These usage specifications and screen displays have received favorable feedback from physicians in

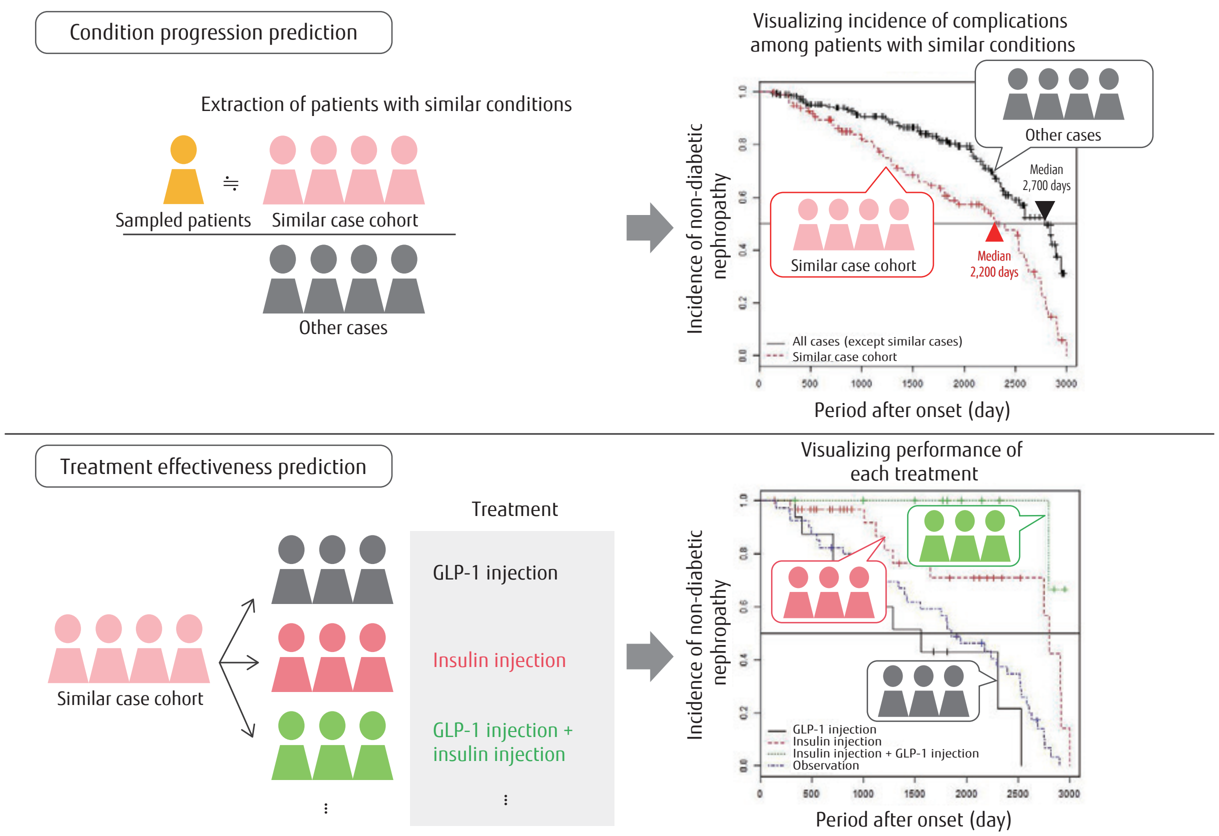

Figure 2
Examples of condition progression prediction and prediction of treatment effectiveness.

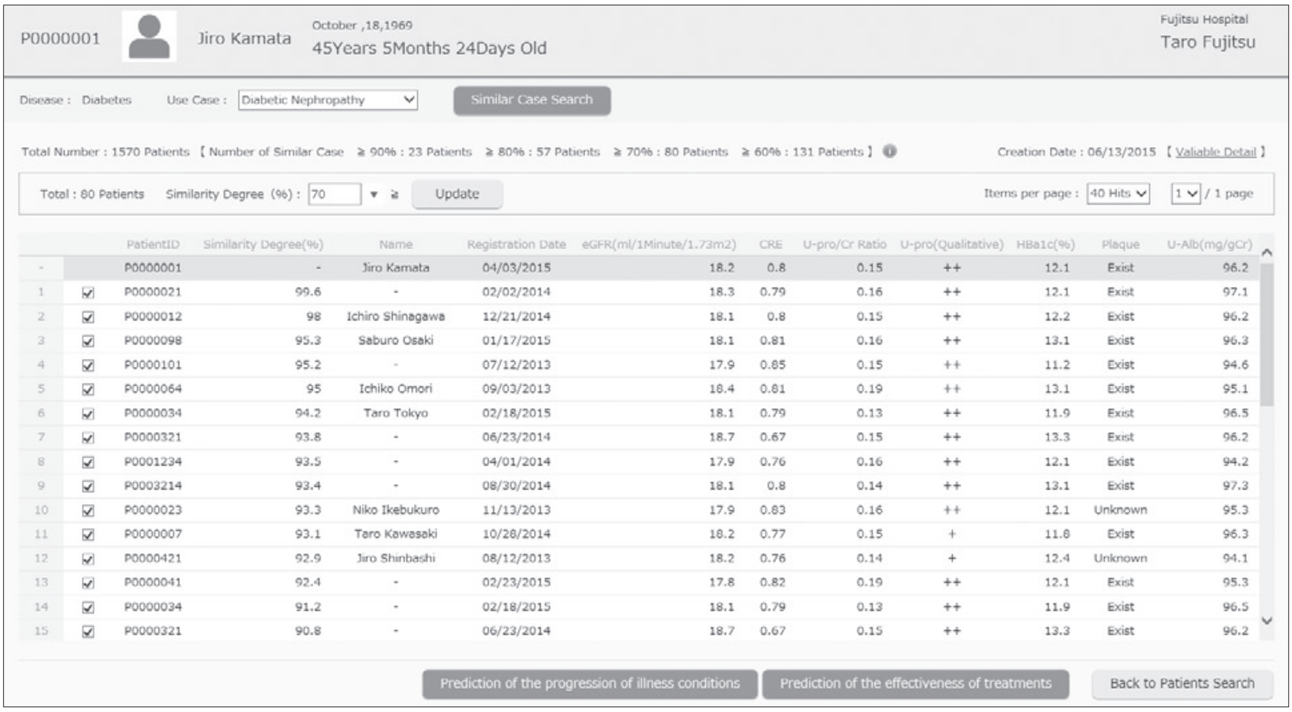

Figure 3
Similar case search engine screen.

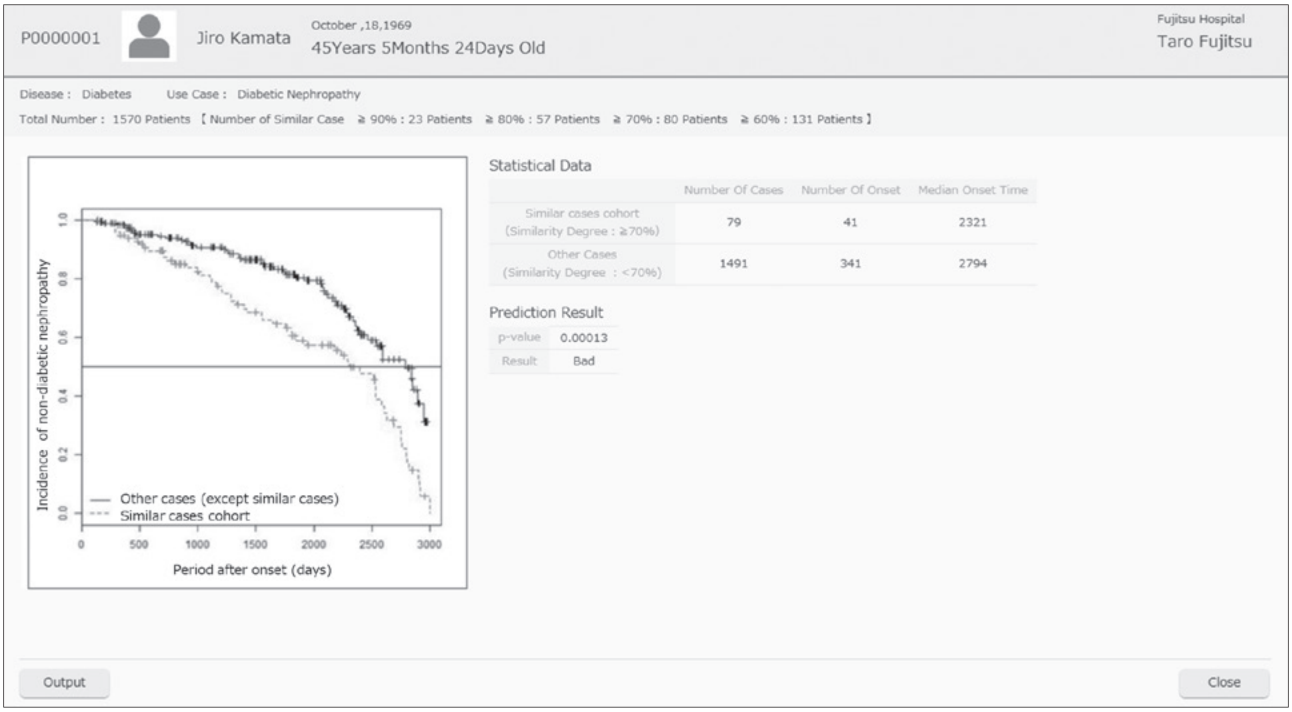

Figure 4
Condition progression prediction screen.

the field regarding their ease of understanding and operation.

5. Evaluations through trial studies

The number of diabetic patients is on the rise in Japan, presenting challenges in terms of the growing medical costs and the shortage of specialists. In FY2014, Fujitsu developed an illness management system to serve as a database platform to compile and manage information about diabetic patients, such as details of the given treatments, in the local medical care network, which facilitated the secure data exchange of medical information between communities. In October 2015, Fujitsu and the National Hospital Organization's Nagasaki Kawatana Medical Center (hereafter, the Nagasaki Kawatana Medical Center) started a research project, aiming to leverage the clinical data compiled in the illness management system in the prevention of diabetes progression and facilitation of effective treatments. As a result, we developed a similar case search engine system by applying the medical big data analysis technology. Subsequently, we have been conducting trial studies since FY2016 with patients who have given their consent to the study, working with the Nagasaki Kawatana Medical Center and other four medical institutes.[7)]

The medical big data analysis technology has proved to be effective in predicting the incidence of diabetic complications from actual clinical data with high accuracy. We took 30 cases of diabetic nephropathy patients from 2014 up to the present, organized and managed the clinical records by each patient in the following periods, conducted condition progression prediction, and evaluated the outcomes (**Figure 5**).

i) Three months prior to the onset of the condition
ii) Three to Six months after onset
iii) Six months after onset and thereafter

Given that these 30 cases are of patients who had diabetic nephropathy, a "bad" result is correct while a "good" result is incorrect. Where the analysis result is "not available," this means that the "analysis is invalid" due, for example, to an insufficient number of similar cases, and the outcome is not statistically credible. As shown in Figure 5, the clinical information involved missing values in about 50% of the data, but almost all analyzable data returned "bad" results—except those "not available" results—yielding the correct answer. We thus verified from these outcomes that the medical big data analysis technology facilitated predictions of the progression of illnesses with high accuracy even with trial records that had many missing values.

6. Future prospects

The medical big data analysis technology currently only handles structured and quantified information, such as blood test results. However, clinical records include much non-structured information, such as texts and CT scan images. This type of information bears higher importance over structured information for patients with certain ailments. For example, CT images are more important than blood test results in the case of cancer patients. In the future, non-structured information will need to be included in the analysis to widen the scope of illnesses the system is able to handle.

Also, AI must be leveraged in the medical big data analysis technology. AI is very suitable for handling non-structured information such as texts and images. In particular, by utilizing deep learning, machines are now capable of picking up very subtle characteristics in images, voices, and texts. Fujitsu offers Zinrai

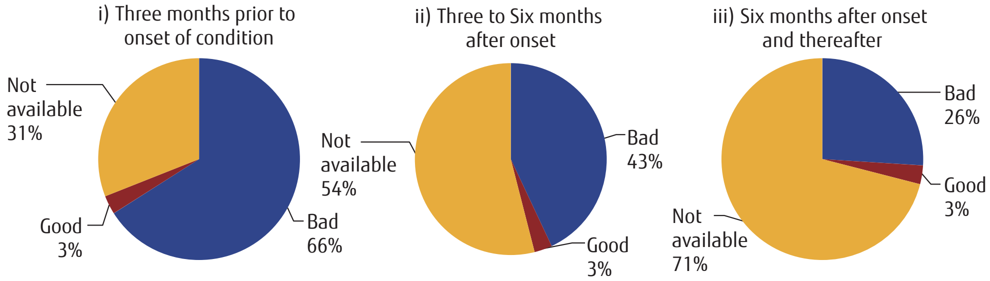

Figure 5
Results of condition progression prediction through trial studies.

Deep Learning, a service that leverages deep learning technology.[8] Medical big data analysis technology will incorporate these technologies and services in the future, enabling the handling of non-structured information in order to be able to respond to a diverse range of illnesses and usage situations.

7. Conclusion

This paper presented the contributions to medical innovations by medical big data analysis technology, describing the recent trends in the use of big data with a focus on the realization of personalized medicine, the necessary technologies and concepts as well as the efforts made at Fujitsu, and the evaluation of trial studies. The volume of information will further grow in the medical field, and we expect that our medical big data analysis technology will offer an effective approach to realizing human-centric medical/healthcare solutions and services that support enriching and convenient lifestyles.

We close this report by expressing our deep gratitude to Nagasaki Kawatana Medical Center, the National Hospital Organization Nagasaki Medical Center, Nakamura Clinic, Meiwakai Izaki Clinic, and Chikuba Clinic for their cooperation in the evaluation of the trial studies.

References

1) Ministry of Internal Affairs and Communications: Information and Communications in Japan 2014, Chapter 3, Section 1 (in Japanese). *http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h26/pdf/*
2) Cabinet Office of the Government of Japan: Japan Revitalization Strategy - Japan is Back 2016 (in Japanese). *http://www.kantei.go.jp/jp/singi/keizaisaisei/kettei.html*
3) Ministry of Internal Affairs and Communications: Information and Communications in Japan 2015, Chapter 5, Section 4 (in Japanese). *http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h27/pdf/*
4) Japanese Association of Healthcare Information Systems Industry: JAHIS Order Entry and Electronic Medical Record System Introduction Survey Report (2016 survey) (in Japanese). *https://www.jahis.jp/action/id=57?contents_type=23*
5) K. Wetterstrand: DNA Sequencing Costs. *https://www.genome.gov/sequencingcostsdata/*
6) F. Nagami, et al.: Success in an integrated high-accuracy genomic analysis of 1,070 Japanese samples. Tohoku Medical Megabank Organization (in Japanese). *http://www.megabank.tohoku.ac.jp/news/11873*
7) Fujitsu: Fujitsu embarking on trial studies of a system to support the prevention of diabetes progression (in Japanese). *http://pr.fujitsu.com/jp/news/2016/07/8.html*
8) Fujitsu: Fujitsu Launches Deep Learning Platform System Offering World-Class Speed. *http://www.fujitsu.com/global/about/resources/news/press-releases/2017/0516-04.html*

Tadaaki Katsuda
Fujitsu Ltd.
Mr. Katsuda is currently engaged in the development of medical big data analysis solutions.

Hayato Isshiki
Fujitsu Ltd.
Mr. Isshiki is currently engaged in the development of medical big data analysis solutions.

