# Strategies for Using Open Innovation to Establish a Next-generation NDB Data Research Platform

## What Sort of System is Right for the Era of Big Data?

While making the vast amounts of data contained in the National Database of Health Insurance Claims and Specific Health Checkups of Japan(1) available in a convenient and appropriate form might be considered a duty to society, enabling the simple, high-speed, and accurate analysis of this dataset (which is among the world's largest) is no easy task. Through open innovation with the Institute for Health Economics and Policy of the Association for Health Economics Research and Social Insurance and Welfare and the Institute of Industrial Science at the University of Tokyo, Hitachi has developed the super-fast super-interdisciplinary Japanese medical insurance claim big data analytics platform system, a next-generation platform for studying this data that helps extract new evidence from the information it contains. This article describes three different activities by Hitachi aimed at ensuring that the system will continue to find use in the future together with details of how the data is being used by a research project of the Institute for Health Economics and Policy.

Yoshinori Sato
Hideyuki Nomura
Shunsuke Ito
Koichiro Kimotsuki
Yasuhiro Tahara
Shuji Watanabe, Ph.D.

## 1. Introduction

On the basis of the Japanese government's vision of a future Society 5.0, it has been suggested that health-care should move away from generalized treatments toward more personalized medicine, also shifting its focus from the treatment of disease to pre-emptive care and prevention(2). Given these proposals, the key to things like improving the quality of medical services and controlling ever-rising social security costs lies in the huge amount of data contained in the National Database of Health Insurance Claims and Specific Health Checkups of Japan (NDB).

The NDB data[*1] includes information on health insurance claims collected from claimants and other

[*1] The database contains information on health insurance claims and on specific health checkups and specific health advice. The information on health insurance claims is made up of detailed billing data submitted by medical institutions (hospitals and pharmacies) to insurers (municipalities and other health insurers) for treatment covered by insurance. As Japan has introduced universal health insurance cover, meaning that most medical care supplied by healthcare providers in Japan is covered by insurance, the analysis of health insurance claims should provide a clear picture of the current state of healthcare for the people of Japan.

sources based on article 16 of the Act on Assurance of Medical Care for Elderly People amended in 2008, the quantity of which totaled about 15 billion records (covering approximately nine years) as of the end of March 2018. Access to NDB data has been available for public-good research since FY2013, with provision by third parties to research groups that have been approved by a committee of experts having also commenced[3]. The NDB is one of the largest datasets of this type in the world and establishing the research infrastructure for its convenient and appropriate use might be considered a duty to society.

Accordingly, in a project titled The Development of a Next-Generation Super-fast Super-interdisciplinary NDB Data Research Infrastructure to Enable the Rapid Accumulation of Evidence for Healthcare Policy commissioned by the Japan Agency for Medical Research and Development (AMED) in FY2016, the Institute for Health Economics and Policy (IHEP) has embarked on research into development of a next-generation NDB data research infrastructure through open innovation with the Institute of Industrial Science (IIS) at the University of Tokyo and Hitachi[4].

Through collaborative creation by IHEP, IIS, and Hitachi, the excellent know-how and technology of each partner[5], [6] combined with a strong shared desire to contribute to society have delivered significant results, having succeeded at the research project's objective of developing next-generation NDB data research infrastructure while also putting in place the research and development structures needed to ensure that the infrastructure is maintained and further developed[7]. This article describes the system that provides this next-generation NDB data research infrastructure, called the super-fast super-interdisciplinary Japanese medical insurance claim big data analytics platform system (SFINCS).

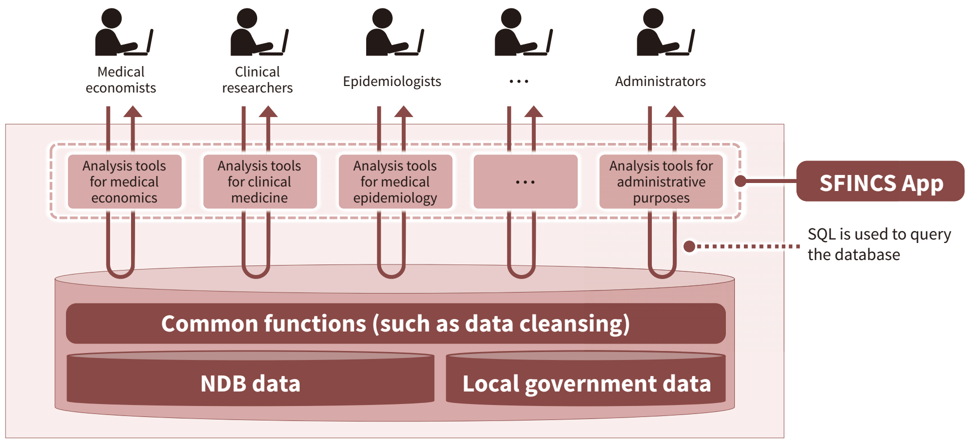
## 2. How to Ensure a System Able to Stand the Test of Time

SFINCS is based on a high-speed claim analysis system developed and operated by IIS since 2012[8]. Users access the database for analysis using the SFINCS Apps (see **Figure 1**). SFINCS commenced operation in 2017. It currently provides access to six years of NDB data (2009 to 2014) and four years of local government data (health checks, treatment, and nursing care).

A key requirement during the development of SFINCS was to ensure that the system would stand the test of time and remain in use for a long time to come. The system development was a major challenge for Hitachi, paying close attention not only to the issues of the day but also to issues that could potentially arise in the future. This article describes how Hitachi went about the development of SFINCS.

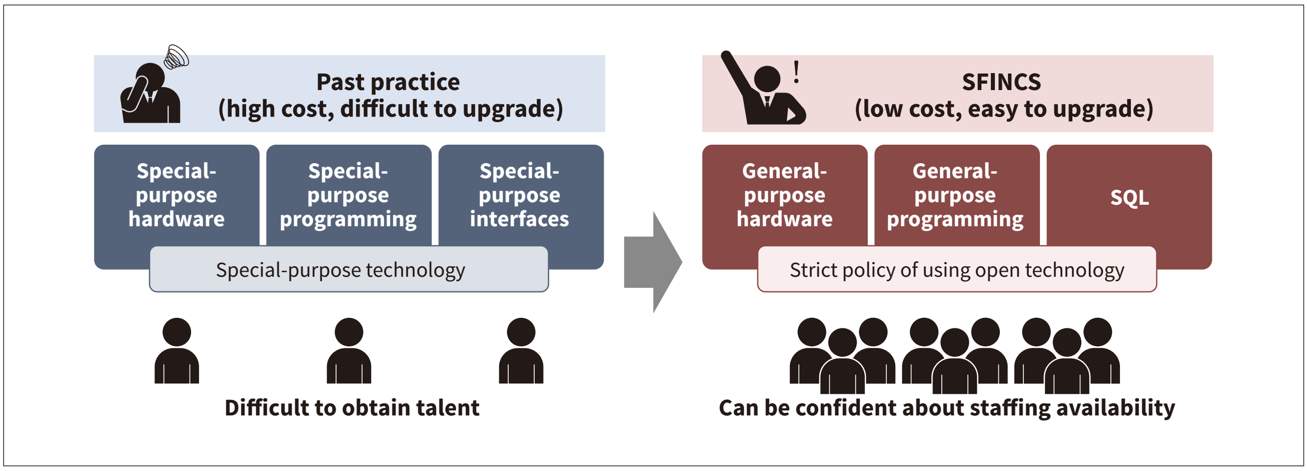
**Figure 1 — SFINCS: A Database of More than 200 Billion Records and a Suite of More than 20 Analysis Tools**

SFINCS provides a variety of users with simple, high-speed, and accurate analysis of the vast amounts of data in the NDB.


Medical economists
Clinical researchers
Epidemiologists
…
Administrators
Analysis tools for medical economics
Analysis tools for clinical medicine
Analysis tools for medical epidemiology
…
Analysis tools for administrative purposes
SFINCS App
SQL is used to query the database
Common functions (such as data cleansing)
NDB data
Local government data

SFINCS: super-fast super-interdisciplinary Japanese medical insurance claim big data analytics platform system SQL: structured query language
NDB: National Database of Health Insurance Claims and Specific Health Checkups of Japan

Figure 2 — Uncompromisingly Open Technology: Provides Confidence about Staffing and Ease of System Upgrading
By utilizing open technologies such as SQL in development, Hitachi created a system that will continue to be easy to upgrade in the future.


Past practice
(high cost, difficult to upgrade)
Special-purpose hardware
Special-purpose programming
Special-purpose interfaces
Special-purpose technology
Difficult to obtain talent
SFINCS
(low cost, easy to upgrade)
General-purpose hardware
General-purpose programming
SQL
Strict policy of using open technology
Can be confident about staffing availability

## 2. 1 Development Based on Standardized and Open Technology

The first step in Hitachi's development of SFINCS was to design the system so as to allow for easy upgrading.

The six years of NDB data currently held in SFINCS represents around 200 billion records. Past practice for enabling the ultra-high-speed analysis of such large datasets was to use special-purpose hardware and programming. For example, expensive hardware costing as much as a supercomputer and highly specialized custom-written programs would be used in an effort to shorten execution times. The problem with this approach is that a large number of people with a high level of specialized development skills need to be kept on hand to deal with functional upgrades or other system enhancements. Use of specialized technology and people not only makes the system more difficult to upgrade, it also increases costs. This was one of the problems with past system development (see the left side of **Figure 2**).

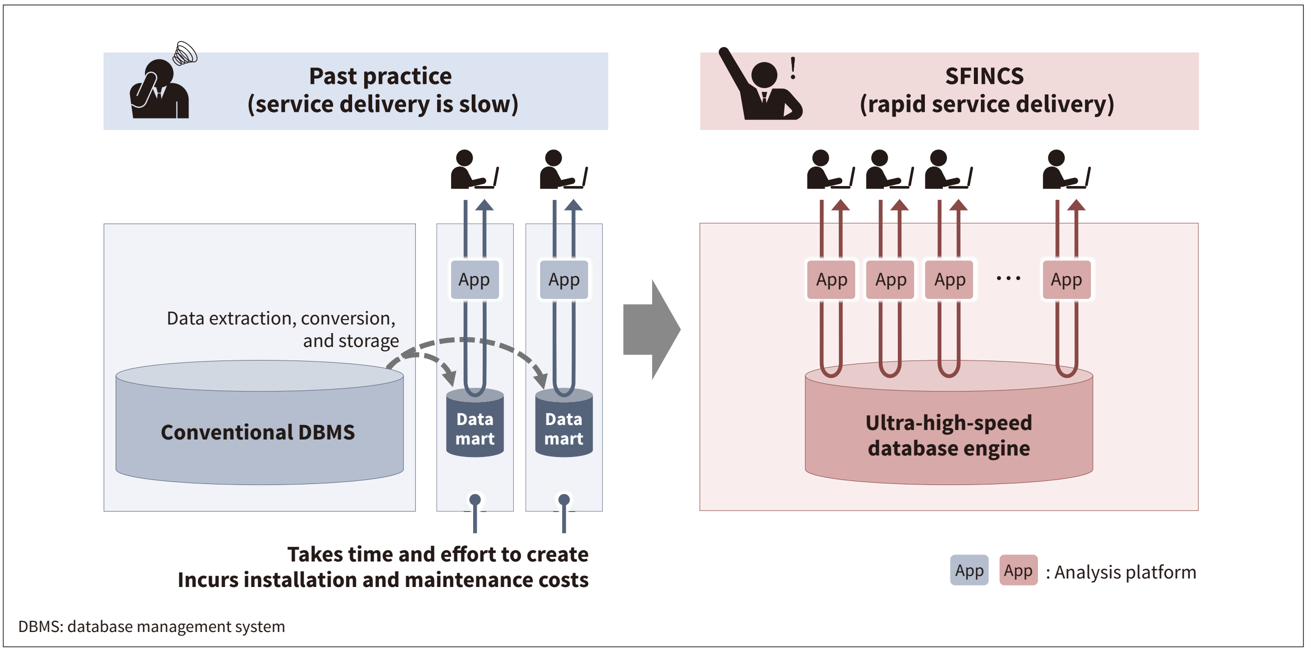
To overcome this, Hitachi adopted a strict policy of basing system development on open technology, without needing to resort to special-purpose hardware or programming (see the right side of **Figure 2**). For example, structured query language (SQL), an open industry standard, is used by the SFINCS App analysis tools to access the database. This facilitates functional upgrades and modifications to the SFINCS Apps. Moreover, a survey found that around 60% of IT developers have experience with SQL[9], indicating that a high degree of confidence can be had regarding future staff recruitment. Keeping staffing risk to a minimum is an essential requirement if the system is to continue to grow in the current environment where talent shortages in the IT sector have become a problem for society. In this way, Hitachi took account of staffing as well as technology considerations, successfully creating a system that will continue to be easy to upgrade in the future.

By making the system easy to upgrade, a variety of different SFINCS Apps have been released to suit user requirements, the number totaling more than 20 as of June 2019. The SFINCS Apps provide an easy way to analyze data using the app screens, without the need for knowledge of the particular structure of the NDB or the ability to develop data extraction programs. The intention is to continue expanding the range of SFINCS Apps to suit different user requirements.

## 2. 2 Establishment of Platform for Dynamic Service Delivery

Hitachi followed this by finding ways to combine high-speed analysis with dynamic service delivery. Microsoft Corporation founder Bill Gates wrote in his book that, "the company's survival depends on everyone moving as fast as possible"[10]. In other words, the value of speed manifests in a wide range of situations. Being able to put the analysis techniques they devise into practice quickly is also of importance to

**Figure 3—A Platform that Does not Rely on Data Marts: High-speed Data Analysis and Dynamic Service Delivery**
Hitachi succeeded in creating a system that combines high-speed analysis with dynamic service delivery without relying on data marts.



Past practice (service delivery is slow)
SFINCS (rapid service delivery)
App
App
Data extraction, conversion, and storage
Conventional DBMS
Data mart
Data mart
App
App
App
App
Ultra-high-speed database engine
Takes time and effort to create
Incurs installation and maintenance costs
App App : Analysis platform

DBMS: database management system

NDB users. Given that, from a user's point of view, an analysis is for all practical purposes impossible if it takes too long to perform, it is essential to make the time taken from devising an analysis technique to obtaining results as short as possible.

The speed of analysis execution depends on how much data is being analyzed, and the NDB will continue to grow in size at a rate of about 35 billion records annually. Past practice for enabling an ever-growing set of data to be analyzed at high speed was to extract the required data from the dataset to generate a purpose-specific database (a data mart). The problem with this approach is that a new data mart needs to be generated whenever a user decides they want to try something new, requiring that time and effort be spent on both program development and database design. Furthermore, additional storage hardware is needed to hold the data marts, adding further costs for installation and operation. In other words, use of data marts is problematic for the dynamic delivery of services (see the left side of **Figure 3**).

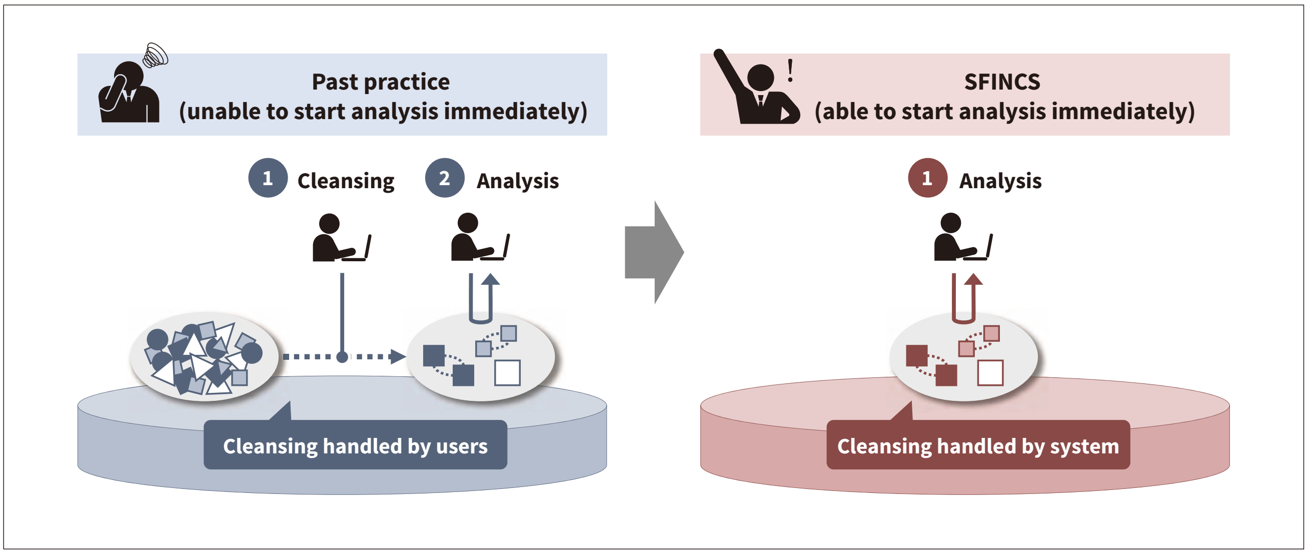
Hitachi addressed this problem by developing a platform capable of high-speed analysis without resorting to data marts (see the right side of **Figure 3**). First of all, execution times were significantly shortened by selecting Hitachi Advanced Database[*2], an ultra-high-speed database engine developed through open innovation with IIS, as the data processing platform for SFINCS(11). Further work was then conducted looking at how the data was structured in the database to ensure that it would allow for both the ever-increasing size of the dataset and upgrades to the SFINCS Apps. Database experts followed a process of trial and error that drew on a variety of information about analysis processing to devise a data structure that could cope with both the growing volume of data and further development of the SFINCS Apps. In this way, by looking for solutions without being held back by accepted practice, they succeeded in creating a system that combines high-speed analysis with dynamic service delivery.

This high-speed data analysis capability means that analyses that in the past would have taken hours or days can now be completed in minutes or seconds. Similarly, the ability to analyze the entire dataset at high speed without resort to data marts means that it is easier to conduct comprehensive and in-depth studies. Meanwhile, dynamic service delivery meant that, for example, user requests could be implemented quickly during SFINCS development.

[*2] This work utilizes results from the Development of the Fastest Database Engine for the Era of Very Large Database and Experiment and Evaluation of Strategic Social Services Enabled by the Database Engine (Principal Investigator: Prof. Masaru Kitsuregawa, The University of Tokyo/Director General, National Institute of Informatics), which was supported by the Japanese Cabinet Office's Funding Program for World-Leading Innovative R&D on Science and Technology.

Figure 4 — Data Cleansing Function: Users Can Get Started on Analysis Immediately

The inclusion of a data cleansing function provides a system that allows users to focus on the job of analysis.


Past practice
(unable to start analysis immediately)
1 Cleansing
2 Analysis
Cleansing handled by users
SFINCS
(able to start analysis immediately)
1 Analysis
Cleansing handled by system

## 2. 3 Provision of Functions that Help Boost Data Analysis Productivity

Hitachi also took steps to ensure that users would be able to focus on the job of analysis. Rather than the time taken by the analysis itself, a high proportion of the time spent on the analysis of big data tends to be accounted for by preparatory work, such as the collection and collation of data or establishing the system configuration needed for the analysis. Indeed, surveys have reported that data cleansing (getting the data into a condition where it is amenable to analysis) makes up 50 to 80% of the total time spent[12].

This preprocessing is also needed when analyzing NDB data to clarify interrelationships in the data and get it in a suitable state for analysis, such as collating on a per-patient basis claims data that has been recorded on a monthly basis by individual medical institutions, or cross-matching different types of claims, such as those for medical treatment and those for drug dispensing. This situation, where each user needed to do their own preprocessing and it took them a long time to get to the point where their actual analysis work could begin, was a problem for NDB data analysis in the past (see the left side of **Figure 4**).

Hitachi addressed this problem by implementing this preprocessing as shared functions in SFINCS rather than in the SFINCS Apps, thereby eliminating the need for users to do this for themselves and allowing them to proceed directly with data analysis without having to worry about preprocessing (see the right side of **Figure 4**).

Hitachi also took steps to provide flexibility in how the preprocessing could be done. In the example given above, collating data by individuals is called patient-matching and a variety of different techniques have been investigated [13], [14]. A survey conducted as part of work on incorporating this function into SFINCS asked about the different available patient-matching techniques and found that the best technique to use depended on factors such as how much rigor is required in the matching results. Accordingly, various different patient-matching techniques were added to the system to suit different analysis objectives, thereby giving users a range of techniques to choose from. By doing so, and by equipping the system with more than just a limited range of patient-matching options, Hitachi succeeded in providing a system that allows users to focus on the job of analysis.

## 3. Applications Made Possible by Use of NDB Data

As of June 2019, 15 universities, six academic societies, and more than 60 experts were involved in the IHEP research project centered around SFINCS. That the scope of people analyzing NDB data is so wide and that the data is being put to use in so much research can be seen as a major success for the SFINCS development. This section describes what some research projects have achieved using NDB data.

## 3. 1

## Announcement at Japan Diabetes Society of Results of Collaborative Work by Four Medical Societies

Research over recent years has found that lifestyle diseases like hypertension (high blood pressure) and diabetes have a major impact on medical costs as well as on lifespan, with a variety of preventive measures having been adopted by the Ministry of Health, Labour and Welfare (MHLW), health insurers, local governments, and other such organizations.

An analysis of the number of patients with lifestyle diseases (hypertension, diabetes, hyperlipidemia, and kidney failure) and the state of medical care for these conditions (both drug and other treatments) conducted as part of a survey of NDB data by Naohiro Mitsutake, Associate Director at IHEP, with aims that included assessing the quality of medical care, found similar trends to those seen in patient number estimates based on the National Health and Nutrition Survey and Patient Survey conducted by the MHLW[7] (see **Figure 5**). In other words, analysis of the NDB data produced similar results to those obtained by sending out researchers to conduct surveys directly, indicating the potential for an alternative survey conducted using high-speed analysis of the full dataset.

Furthermore, having found that only 3.5 million of the 10 million diabetes patients were receiving drug or other treatments in accordance with the guidelines, the analysis-based survey could also be said to have provided knowledge able to be incorporated back into the treatment guidelines of the Japan Diabetes Society.

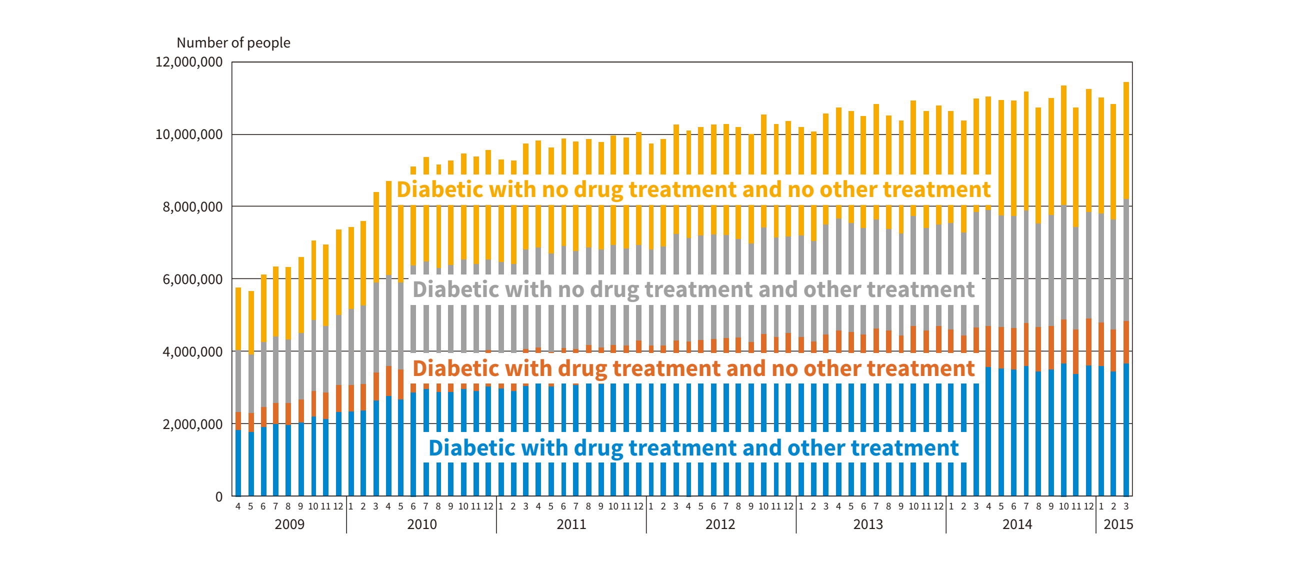
## 3. 2

## Identification of Evidence Needed for Designation as Intractable Disease

Diabetes exists in a number of different forms, but Type 1 diabetes in particular imposes a large social and economic burden on its sufferers. Type 1 diabetes is a condition in which damage to the $\beta$ cells of the pancreas results in a lack of insulin, meaning that sufferers, to stay alive, need to inject daily doses of insulin for the rest of their lives. The high cost associated with this has led to calls for the condition to be designated an intractable disease, which would qualify sufferers for financial assistance. However, to be considered for designation, the disease needs to satisfy two requirements: that the number of people in Japan with the

**Figure 5 — Use of NDB Data to Analyze Number of Diabetes Patients in Japan and State of Medical Care (Permutations of Disease, Drug Treatment, and Other Treatment)**

Analysis of NDB data provides a quantitative overview of drug and other treatments for diabetes patients.



condition is small (less than 0.1% of the population) and that objective diagnostic criteria exist.

While no clear figure existed for the number of Type 1 diabetes sufferers in the past, a survey of NDB data by the research group of Professor Naoki Nakashima at Kyushu University estimated the number in 2014 at approximately 117,000. The survey of NDB data produced a number that supports the case for Type 1 diabetes being a rare disease, thereby providing one piece of the evidence needed for its designation as an intractable disease.

Through work like this, the IHEP research project is delivering a steady stream of new results from use of the NDB data.

## 4. Conclusions

There is currently a growing push, especially from the government, behind analyses that combine the NDB with various other forms of big data. Data on nursing care insurance is one such example. Combining NDB data on things like the types of medical interventions performed and to whom together with nursing care insurance data recording the outcomes of medical interventions and what subsequently happened to the patient would allow for the analysis of things like the relationship between medical interventions and the progress of dementia or the trends and totals for medical and care expenses. This has attracted attention not only from researchers but also from the national and local governments that have the role of medical and nursing care insurance providers.

SFINCS is taking the lead in this initiative to combine NDB and nursing care insurance data. In Mie Prefecture, for example, with help from IHEP and IIS, activities include establishing methods for predicting medical and nursing expenses, analysis of how to boost efficiency and increase the proportion of people receiving specific health checkups, and provision of information on the balance of supply and demand for medical and nursing care based on a broad survey encompassing health check, medical, and nursing data. In a future where the populations continues to dwindle, it seems likely that organizations will need to adopt a policy of making rigorous use of the data they hold when considering medical and nursing care policies and effective funding practices that are tailored to their regional circumstances.

Another initiative involves making SFINCS an intrinsically cloud-based service, including the accompanying high-speed anonymized data handling functions. In both cases, technology successfully developed by IIS and Hitachi through open innovation as part of a project involving a data processing engine called the Ultra Big Data Platform for Reducing Social Risks, a research topic of the Impulsing Paradigm Change through Disruptive Technologies Program (ImPACT)(15), will play a vital role in the future provision of data analysis via the public cloud(16), (17). Accordingly, Hitachi intends to contribute to the health and wellbeing of the Japanese people through the continued development of SFINCS as a powerful platform for supporting the extraction of new evidence from a variety of big data sources.

Hitachi is drawing on its extensive experience with system development for the healthcare industry to support progress on measures for using the Internet of Things (IoT) in the health, medical, and nursing care sectors. It is also confident that the SFINCS it developed through open innovation with IHEP and IIS will serve as a useful example of how to go about configuring systems in the era of big data in applications beyond the healthcare sector.

Hitachi intends to push ahead with open innovation in a variety of different fields as it works toward creating a sustainable society in which everyone can live vibrant and comfortable lives.

### Acknowledgements

The SFINCS system described in this article was developed through collaborative creation with the Institute for Health Economics and Policy (IHEP) and Institute of Industrial Science (IIS) at the University of Tokyo on the basis of assistance from Development of a Next-Generation Superfast Super-interdisciplinary NDB Data Research Infrastructure to Enable the Rapid Accumulation of Evidence for Healthcare Policy, a research project of the Japan Agency for Medical Research and Development (AMED) for establishing the information and communications technology infrastructure for clinical research and other fields. Considerable advice was received from all concerned, including in particular Naohiro Mitsutake, Associate Director of IHEP, and Professor Masaru Kitsuregawa and Associate Professor Kazuo Goda of IIS. The authors would like to express their deep appreciation.

### References

1) Ministry of Health, Labour and Welfare (MHLW), "First NDB Open Data Japan," https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000139390.html in Japanese.

2) Japan Business Federation (KEIDANREN), "Healthcare in Society 5.0," https://www.keidanren.or.jp/en/policy/2018/021_overview.pdf

3) MHLW, "Homepage for The Provision of Information on Medical Fee Receipts/Health Checkups," https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/iryouhoken/reseputo/index.html in Japanese.

4) Institute for Health Economics and Policy, "Annual Report 2017," http://www.ihep.jp/business/annual/2017.html in Japanese.

5) K. Umemoto et al., "A Prescription Trend Analysis using Medical Insurance Claim Big Data," Proceedings of 35th IEEE International Conference on Data Engineering (ICDE2019), pp. 1928–1939 (Apr. 2019).

6) J. Sato et al., "Novel Analytics Framework for Universal Healthcare Insurance Claims Database," Proceedings of the 17th World Congress of Medical and Health Informatics (MedInfo 2019), (Aug. 2019) to appear.

7) Institute for Health Economics and Policy, "Annual Report 2018," https://www.ihep.jp/business/annual/ in Japanese.

8) H. Yamada et al., "Performance Evaluation of Out-of-Order Parallel Data Processing System in 128-Node Storage Intensive Cluster," The IEICE Transactions on Information and Systems, 98-D, pp. 728–741 (May 2015) in Japanese.

9) stack overflow, "Developer Survey Results 2018," https://insights.stackoverflow.com/survey/2018#technology

10) Bill Gates et al., "Business @ the Speed of Thought," Warner Books (Mar. 1999).

11) "Big Data Analysis Platform for the Medical Field," Hitachi Review, 67, p. 328 (Mar. 2018).

12) "For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights," The New York Times: Digital subscription (Aug. 2014), https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html

13) S. Kubo et al., "National Database of Health Insurance Claims and Specific Health Checkups of Japan (NDB): Outline and Patient-Matching Technique," (Apr. 2018), https://doi.org/10.1101/280008

14) J. Sato et al., "Enabling Patient Traceability Using Anonymized Personal Identifiers in Japanese Universal Health Insurance Claims Database," Proceedings of the AMIA 2019 Informatics Summit, American Medical Informatics Association, pp. 345–352 (Mar. 2019).

15) "The Ultra Big Data Platform for Reducing Social Risks," A research topic of the Impulsing Paradigm Change through Disruptive Technologies Program (ImPACT), https://www.jst.go.jp/impact/en/program/16.html

16) A. Okuno et al., "Dynamic Computing Resource Adjustment in Shared-storage Database Engine," Information Processing Society of Japan (IPSJ) Transactions on Database (TOD), 11, pp. 30–43 (Jul. 2018) in Japanese.

17) N. Nishikawa et al., "Design and Preliminary Evaluation of Interactive Anonymization of Large Scale Data using Out-of-order Database Engine," Institute of Electronics, Information and Communication Engineers (IEICE), The 11th Forum on Data Engineering and Information Management/17th Annual Conference of the Database Society of Japan (DEIM2019), J3-2, (Mar. 2019) in Japanese.

### Authors

**Yoshinori Sato**
First Department, Healthcare Solutions Division, Healthcare Business Unit, Hitachi, Ltd. *Current work and research:* Digital health business including healthcare big data platforms, etc.

**Hideyuki Nomura**
First Department, Healthcare Solutions Division, Healthcare Business Unit, Hitachi, Ltd. *Current work and research:* Digital health business including healthcare big data platforms, etc.

**Shunsuke Ito**
First Department, Healthcare Solutions Division, Healthcare Business Unit, Hitachi, Ltd. *Current work and research:* Digital health business including healthcare big data platforms, etc.

**Koichiro Kimotsuki**
New Business Development Center, Healthcare Solution Division, Healthcare Business Unit, Hitachi, Ltd. *Current work and research:* Business development in healthcare informatics section.

**Yasuhiro Tahara**
Database Department, Data Management, IoT & Cloud Services Business Division, Service Platform Business Division Group, Hitachi, Ltd. *Current work and research:* Technical support of Hitachi Advanced Database.

**Shuji Watanabe, Ph.D.**
Database Department, Data Management, IoT & Cloud Services Business Division, Service Platform Business Division Group, Hitachi, Ltd. *Current work and research:* Proposal, design, and development of data management systems including medical information systems.