OHE

Data Governance Arrangements for Real-World Evidence in Japan

CONSULTING | REPORT
SEPTEMBER 2019

Hideo Yasunaga
Hayato Yamana
Marina Rodes Sanchez
Adrian Towse

ohe.org

OHE

SEPTEMBER 2019

Data Governance Arrangements for Real-World Evidence in Japan

Hideo Yasunaga
University of Tokyo

Hayato Yamana
University of Tokyo

Marina Rodes Sanchez
Office of Health Economics

Adrian Towse
Office of Health Economics

OFFICE OF HEALTH ECONOMICS
CONSULTING

Please cite this report as:
Yasunaga, H., Yamana, H., Rodes Sanchez, M. and Towse, A., 2019. Data Governance Arrangements for Real-World Evidence in Japan. OHE Consulting Report, London: Office of Health Economics. Available at: https://www.ohe.org/publications/data-governance-arrangements-real-world-evidence-japan

Corresponding Author:
Marina Rodes Sanchez
mrodessanchez@ohe.org

For further information please contact:

Professor Graham Cookson
Chief Executive, OHE
Honorary Visiting Professor in Economics at City, University of London

Tel +44 (0)207 747 1408
Email gcookson@ohe.org

About OHE Consulting Reports

Many of the studies OHE Consulting performs are proprietary and the results are not released publicly. Studies of interest to a wide audience, however, may be made available, in whole or in part, with the client's permission. They may be published by OHE alone, jointly with the client, or externally in scholarly publications. Publication is at the client's discretion.

Studies published by OHE as OHE Consulting Reports are subject to internal quality assurance and undergo a rapid external review, usually by a member of OHE's Editorial Panel. Any views expressed are those of the authors and do not necessarily reflect the views or approval of the OHE's Editorial Panel or Research and Policy Committee, or its sponsors.

Funding and Acknowledgements

This consultation briefing study was commissioned and funded by Eli Lilly and Company.

Table of Contents

Executive summary

Lilly's objective in commissioning the research presented in this OHE Consulting Report was to understand how Real-World Data (RWD) is accessed or generated in Japan, and used to produce or generate Real-World Evidence (RWE), to see whether use is aligned with international standards.

This paper follows two earlier reports covering the specifics under which RWE has been used in a total of nine different countries. Evidence used to support decision-making in health care is becoming increasingly diverse, while the importance of considering the impact of health care interventions in real-world settings is being more and more recognised. As such, appropriate governance arrangements for RWE are imperative to facilitate evidence collection to meet the demands of regulators and HTA bodies, and to enable health care information to fulfil the role it can play in improving patient care.

This report briefly introduces the Japanese health system before presenting in more detail the arrangements behind the collection and management of patient data in Japan. We outline models of data governance in the country. There are various types of RWD regularly collected by insurers and/or hospitals. Different entities – public or private, non-commercial or commercial –gather the data to create their own bigger, anonymised databases, which are eventually made available to various recipients (e.g. limited to selected academic researchers, or open to private companies). Sources of RWD include, but are not limited to, administrative claims data for healthcare services, national health check-up data, the Diagnosis Procedure Combination (DPC) data, and Electronic Medical Records (EMR). Datasets are available from different sources depending on the provider and recipient type.

The report outlines the core legislation and governance arrangements for the collection and use of patient data. RWD must be handled under the legal framework for using personal data, which includes the Act on the Protection of Personal Information (APPI) and related government guidelines. The legislation and related governance arrangements cover two important aspects. Firstly, patient consent for collecting and using routinely collected data: under current conditions, most Japanese RWD are open only to academic researchers and societies, who don't need to obtain prior individual consent, and are closed to the private sector (e.g., pharmaceutical companies and consulting companies), except for several commercially available databases which use de-identified patient data. Secondly, de-identification of routinely collected data. An option that allows more flexibility within the Act on the Protection of Personal Information (APPI) is to obtain data through a de-identification process undertaken within the medical institution itself, so that patient data are no longer deemed to be personal information and are therefore no longer subject to the Act.

Japan is still facing some challenges in developing a clear set of nationally agreed and implemented standard rules to optimise interoperability of health system records. This limits the ability to benefit from achievements in the areas of de-identification, privacy rules, data control, and access. However, new legislation, the Next-generation Healthcare Infrastructure Act (NHIA), was put into force in 2018 with the potential to fundamentally change the governance arrangements for RWD in Japan. According to this law, "certified operators for de-identifying medical data" are allowed to collect patients' personal information directly from healthcare providers; link all the data at individual level; and provide de-identified data to academic researchers. The implication of this being that NHIA can improve individual-level data linkage which has the potential to enhance RWD studies. However, the Ministry of Health, Labour, and Welfare (MHLW) is still searching for candidates to act as the certified operators, and thus the issue of data accessibility has not been fully addressed. As such, it is still unclear whether this new development will help deal with the asymmetry of the position of commercial and non-commercial researchers.

1 Arrangements for data governance in Japan

1.1 Introduction, objectives and context

In 2015, OHE Consulting published a report on data governance arrangements for real-world evidence (RWE) covering the specifics under which RWE was used in eight different countries: Australia, France, Germany, Italy, Sweden, the Netherlands, the UK and the US (Cole et al., 2015). Two years later, Lilly commissioned a second report based on the South Korean setting, following the same method and structure as the original (Lee et al., 2017).

This third report focuses on data governance arrangements for RWE in Japan. On this occasion, relevant information has been collected by the Japanese authors Hideo Yasunaga and Hayato Yamana through desk research, using the same pre-specified pro-forma as for the previous reports (see Appendix 1).

Real world-data (RWD) and real-world evidence (RWE) are playing an increasing role in clinical studies and healthcare decisions. In Japan, as well as other countries, studies using RWD have been increasing, and medical societies are using RWE to revise guidelines and to create medical decision tools useful in clinical practice. Lilly's objective in commissioning the research presented in this OHE Consulting report is therefore to understand the core principles that govern how RWD is accessed or generated, and used credibly to produce or generate RWE in Japan while comparing it with a set of "international standards".

The authors summarise the information on Japan in Table 8, using the same structure as in Table 6 of the original 2015 report. The table is structured along four main headings: "Data Protection – Health"; "Data Linkage"; "Access" and "Governance ideals and changes in the environment". The team has also outlined a "heat map" for Japan (Table 9), comparing Japan's governance arrangements with the "ideal governance framework" suggested by Cole et al. (2015).

RWD, RWE and data governance

RWD can be in various forms – but two of its key characteristics are that it is collected outside a clinical trial and is used for health care decision making (Garrison et al., 2007). Broadly, it could consist of either data that are already routinely collected in a health care system (electronic medical/health care records, administrative reimbursement databases, pharmacy data used to fill prescriptions, etc.) or data that is collected specifically for the purposes of a project (e.g. new patient registries for a disease or clinical procedure or pragmatic clinical trials)[1]. Other uses of RWD can include achieving appropriate levels of access and reimbursement, improving safety surveillance and risk management, supporting better outcome measurement, and informing drug development decisions throughout the product lifecycle. These represent important and increasing applications of RWD.

In order for this data to be used credibly to inform clinical practice, an appropriate governance framework is needed for its access and use. Governance has been described as covering: *"…the*

[1] *While important to consider other sources of healthcare data, such as digital devices and social media, these are outside the scope of our remit.*

processes, roles, standards and metrics that ensure the effective and efficient use of data and information in enabling an organisation to achieve its goals" (Gartner, 2014).

The ultimate goal of the information that is collected around health care is to provide evidence for health care interventions and thereby to influence clinical practice and the treatment of patients. In the present report, we describe the current status of RWD in Japan, including core legislation and governance arrangement, data linking, data access, and data use, as compared with the "ideal framework" proposed in Cole et al. (2015).

1.2 Brief overview of the health system & collection/management of patient data

1.2.1 Health system: Organisation of health institutions involved in the financing and provision of care

In 1961, Japan established a universal healthcare insurance system and by law all residents must be enrolled in a health insurance programme. Enrolees have no choice of programme as plans are designated according to their employment status, age, and residence.

There are two main types of health insurance: Employees' Health Insurance System and National Health Insurance (NHI). Employees' Health Insurance covers public servants and those who work in companies, while the NHI covers the self-employed and unemployed. Employees' Health Insurance is further divided into major categories: Japan Health Insurance Association (JHIA), Society Managed Health Insurance (SMHI), Mutual Aid Association, and Seamen's Insurance, the last one covering only a very small percentage of the population. In addition, in anticipation of increasing medical expenditure with an aging population, the Advanced Elderly Medical Service System for people aged 75 or older was enacted in April 2008. At the same time, in order to adjust the imbalance among the insurers due to the uneven distribution of the elderly aged between 65 and 74, a system to adjust the finances of insurers was introduced.

Table 1. Insurers

Principal schemes	Number of insurers	Insured	Number of Enrolees (M)
National Health Insurance (NHI)	1,888 municipalities	Individual proprietors, Pensioners	39
Japan Health Insurance Association administered Health Insurance (JHLA)	1	Salaried employees in small companies	35
Society-Managed Health Insurance (SMHI)	1,458	Salaried employee in large companies	30
Mutual Aid Association	85	Civil servants	9
Advanced Elderly Medical Service System	47 prefectures	People aged 75 or older	14

The current system is financed by a combination of social insurance premiums, tax revenues, and co-payments. The national uniform fee schedule (i.e., amount of reimbursement, including the patients' co-payment) covers most healthcare procedures and products, including drugs. Infants, elderly people and people on low incomes are partially or completely exempted from out-of-pocket payments, while the rate for other people is 30% with a maximum amount of out-of-pocket payments determined according to income level. Overall, the health insurer pays 70–90% of the cost at the point of use, with the remainder paid by the insured patient as co-payment. While all health insurers are not-for-profit organisations, health services are provided both by public providers, and non-profit private ones. Patients enjoy 'free access' to healthcare institutions; that is, freedom to select any healthcare facility without a gatekeeper system operated by family physicians. They face no restriction arising from either their choice of insurer or severity of illness.

Providers submit health insurance claims to the 'Examination and Payment Agency' to claim treatment expenses covered by insurance. One health insurance claim is created for each patient every month, and the Agency checks the bills and judges the appropriateness of payment. The Agency then sends the checked bills to the insurers (Figure 1). Although there are multiple insurance schemes, the tariff per service is national, set by the Ministry of Health, Labour, and Welfare (MHLW). The system is linked to a lump-sum payment system for inpatients in acute care hospitals, called the Diagnosis Procedure Combination/Per Diem Payment System (DPC/PDPS); providers are paid a flat-rate prospective fee per day of inpatient hospital stay for certain DPC services and paid fee-for-service for non-DPC services.

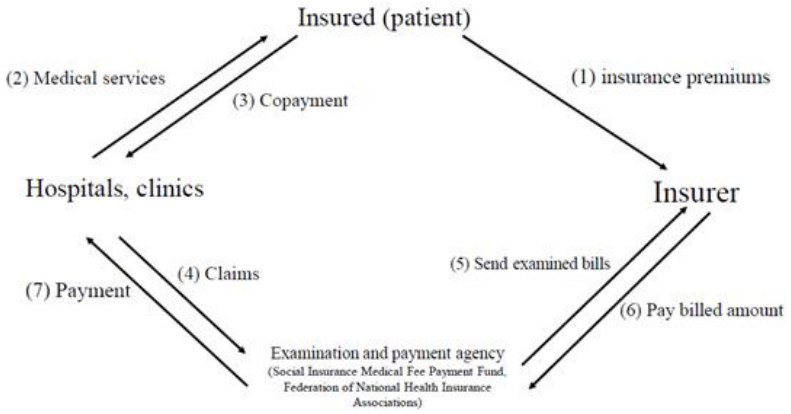

Figure 1. Public insurance system in Japan. Source: MHLWa, n.d.

1.2.2 Collection/management of patient data

In Japan, there are various types of RWD regularly collected by insurers and/or hospitals. Different entities – public or private, non-commercial or commercial – then gather the data to create their own bigger, anonymised databases, which are eventually made available to various recipients (e.g. limited to selected academic researchers, or open to private companies).

Sources of RWD include (i) administrative claims data for healthcare services, (ii) national health check-ups data, (iii) the Diagnosis Procedure Combination (DPC) data, and (iv) Electronic Medical Records (EMR). In this section, we start by explaining these four basic components of RWD before introducing in more detail the public and private entities gathering and providing the data. There are other types of RWD that are not necessarily collected by insurers and/or hospitals. These are not covered in this report, and include drug and disease registries, surveillance systems, genomics databases, death registration databases or natural records, and PRO data.

Administrative claims data

The Japanese administrative claims data is an essential source of RWE and presents crucial information on people's health and disease status. Claims include information on outpatients, inpatients, prescriptions and dental treatments, following the structure shown in Table 2. Information on the use of several types of devices is also available, including coronary stents and orthopaedic prosthesis. For example, it is possible to get information on types of coronary stents (that is, bare metal stents or drug eluting stents) and how many stents are inserted per patient. However, information on the individual types of stent used is not available (Cypher, Taxus, Nobori, etc.).

Table 2. Structure of administrative claims data

Includes	Insured Number of Enrolees (M)
Patient characteristics	Anonymised patient identifiers (Hash ID1 and ID2), sex, age group, prefecture codes, types of insurance
Outpatient claims	Diagnoses; date of visits, types of visits (day-time, out-off-hours, holiday, or night-time visit), home visit; medications (oral medication, topical medication, injection); examinations (laboratory, pathological, radiological, etc.); procedures; surgery and anaesthesia; dates of medications, examinations, procedures, surgery and anaesthesia; the number of days treated or prescribed; death; costs for medications, examinations, procedures, surgery and anaesthesia
Inpatient claims	Diagnoses; dates of admission and discharge; types of visits (day-time, off-hour, holiday, or night-time visit); medications (oral medication, topical medication, injection); examinations (laboratory, pathological, radiological, etc.); procedures; surgery and anaesthesia; dates of medications, examinations, procedures, surgery and anaesthesia; the number of days treated or prescribed; death; costs for medications, examinations, procedures, surgery, anaesthesia, basic hospitalisation fee, and specific hospitalisation fee
Prescription claims	Names of drugs, doses, dosage form, administration, dates of prescription, costs incurred
Dental Care	Diagnoses; date of visits, types of visits (day-time, off-hour, holiday, or night-time visit); medications (oral medication, injection); X-ray tests; procedures; surgery and anaesthesia; crown prosthesis and prosthodontic treatment; costs incurred

National Health Check-ups for lifestyle diseases in workers

The Japanese nationwide health screening and intervention programme was introduced in 2008 by the MHLW. It targets those aged 40-74 years. Every year, targeted individuals have the opportunity to receive health check-ups and those identified with metabolic and pre-metabolic syndromes are encouraged to receive advice and guidance on how to change their lifestyles to prevent the occurrence of lifestyle related diseases. The participation expectation of MHLW was 70%, which has not yet been attained, but the proportion of participants is gradually increasing (currently at around 40%, with around 25 million participants per year) (Tamura and Kimura, 2015). The screening includes blood pressure measurement, anthropometric measurements, chest X-rays, electrocardiographs, blood tests, and a self-reported health questionnaire to assess physical activity and eating behaviours.

Diagnosis Procedure Combination (DPC) data

Table 3 shows the structure of the DPC data, which includes discharge abstract data (so called Format 1) and administrative claims data for every inpatient.

Table 3. Structure of the DPC data

Format 1	Data
Diagnoses	– main diagnosis – admission-precipitating diagnosis – most resource-consuming diagnosis – second most resource-consuming diagnosis – comorbidities present on admission – complications arising after admission
Hospital data	– unique identifiers of the hospital – location of the hospital
Patient characteristics	– zip code – type of admission (urgent or elective) – type of psychiatric admission (voluntary or involuntary) – ambulance service use – dates of admission and discharge – age and sex – body weight and height – smoking index (pack years) – pregnancy – discharge status
Clinical data	– Japanese Coma Scale – TNM Classification and Stage for cancer – modified Rankin scale – Hugh–Jones classifications – New York Heart Association classification – Canadian Cardiovascular Society classification for angina pectoris – Killip classification for myocardia infarction – A-DROP scoring system for pneumonia – Child–Pugh classification for liver cirrhosis; – Japanese severity classification for acute pancreatitis – the date of stroke onset – Burn Index – Activity of Daily Living scores – Global Assessment of Functioning Scale – Mechanical restraint for psychiatric patients.
Administrative claims data	– Anaesthesia, surgery, rehabilitation and other procedures – Duration of anaesthesia (min) – Volume of blood transfusion (ml) – Pharmaceuticals and devices used – Dates of procedures – Dates of using drugs and devices – Estimated costs

The Diagnosis Procedure Combination (DPC) is a case-mix patient classification system, which was originally developed in 2002. The key objectives of introducing DPC data collection were to implement a standardised electronic claims system and to achieve transparency of hospital performance, so that the MHLW could track national trends in healthcare utilisation. All 82 academic hospitals (80 university hospitals, the National Cancer Centre and the National Cerebral and Cardiovascular Centre) were obliged to adopt the DPC system, but adoption by community hospitals is voluntary. The more than 1600 acute-care hospitals that participate in the DPC system are called DPC hospitals, and all of them must create and submit "DPC data" to the MHLW.

As presented in Format 1, diagnoses are recorded with text data in the Japanese language and ICD-10 codes. Using ICD-10 codes on comorbidities, researchers can calculate the Charlson comorbidity score for each patient. The dates of procedures and the dates of using drugs and devices are all recorded, and thus an interval between the start and the end of any process can be calculated (e.g. duration of mechanical ventilation, duration of chest tube drainage). The database also includes estimated total costs based on reference prices from the Japanese national fee schedule which determines item-by-item prices for surgical, pharmaceutical, laboratory, and other inpatient services.

Electronic Medical Records

EMRs have the potential to provide detailed patient information that can be useful for clinical studies and health services research. Compared with administrative claims data, EMRs offer richer clinical information (e.g. blood test results), which enables researchers to better control for bias in clinical studies due to "confounding by indication".

1.2.3 Public and private provision of patient data

Datasets are available from different sources. Table 4 summarises provision by RWD source.

Table 4. Provision of patient data

RWD source	RWD provided by the government, quasi-governmental organizations, or academic societies	RWE provided by private companies
Administrative Claims data	NDB database	JMDC Claims database and Medi-Scope
National Health Check-ups	NDB database	JMDC Claims database and Medi-Scope
Diagnosis procedure combination (DPC) data	MHLW-DPC, MID-NET, NHO database, The DPC Study Group database and JROAD-DPC	Medical Data Vision (MDV) database
Electronic medical records	MID-NET and NHO database	
Registries*		
Surveillance data (safety)*		
Vital data (i.e. death records)*		

**Elements of RWE beyond the scope of this report.*

RWD provided by the government, quasi-governmental organisations, or academic societies

Public provision of administrative claims and health check-ups data

In 2011, the MHLW created a huge national database, the National database of Health Insurance Claims and Specific Health Check-ups of Japan (NDB) (Matsuda et al., 2012). The primary purpose in establishing the NDB was to plan health policies to regulate national health expenditure by using national data, but the MHLW also started to facilitate the secondary use of the database for research purposes.

The NDB covers approximately 98% of data on healthcare services provided by healthcare institutions, with the exception of accidents covered by automobile liability insurance and workers accident compensation. Claims data include the following items: (i) prefecture codes, (ii) sex, (iii) age group, (iv) dates of admission and discharge, (v) procedural codes and dates of procedures, (vi) diagnostic codes, and (v) codes for pharmaceuticals and high-cost medical devices. An enormous number of health insurance claims and specific health check-ups of all residents have accumulated in this national database - approximately 1.6 billion claims have been added annually to the NDB since its launch in April 2011.

Public provision of DPC data

Several entities independently collect anonymised DPC data from the DPC hospitals to create secondary databases, extract datasets for specific purposes and provide them for researchers or the private sector free of charge. Data recipients can use the DPC data to track and analyse healthcare utilisation, access, quality, outcomes, and costs in acute care hospitals. Furthermore, DPC data can be utilised for clinical epidemiology and health services research because it includes a range of diagnostic and other clinical data. Table 5 shows the public entities that collect DPC data.

Table 5. Public entities that collect DPC data from multiple DPC hospitals

Entities	Collected Data
Government or quasi-governmental organisations	
Ministry of Health, Labour, and Welfare (MHLW)	MHLW-DPC database: DPC data provided compulsorily from all DPC hospitals (more than 1600)
Pharmaceuticals and Medical Devices Agency (PMDA)	Medical Information database Network (MID-NET): DPC data along with EMR from 10 institutions
National Hospital Organisation (NHO)	NHO database: DPC data from 143 national hospitals along with EMR from 60 national hospitals
Academic groups	
The DPC Study Group (a government-funded academic group)	The DPC Study Group database: DPC data provided voluntarily from more than 1000 DPC hospitals
The Japanese Circulation Society	JROAD-DPC: DPC data on patients with cardiovascular disease, provided voluntarily from more than 900 DPC hospitals

DPC data provision from government or quasi-governmental organisations:

- **MHLW-DPC database:** The MHLW electronically collects DPC data for health service planning, including the refinement of case-mix classification and the revision of DPC-based fee schedules. In 2019, the MHLW started to provide aggregated data from the MHLW-DPC database to researchers based on "the guideline for providing DPC data" to which only this database is subject to (MHLWb, 2019).

- **MID-NET database:** Medical Information database Network (MID-NET) is a national project initiated by the MHLW and PMDA (Pharmaceuticals and Medical Devices Agency) to establish a database network using electronic healthcare data for drug safety. The key features of MID-NET are adopting 1) distributed dedicated database systems among 23 hospitals related to 10 hub medical institutions, and 2) SS-MIX2-based standardized database systems retrieved from electronic health records in the hospitals for analysis and evaluation of ADR (Adverse Drug Reactions) (Kazuhiko, 2017).

- **NHO database:** The National Hospital Organisation (NHO) established the Medical Information Analysis database (MIA) in 2011 to collect administrative claims data and the DPC data from all the 143 hospitals affiliated with NHO. The database includes about 17% of all the acute care inpatients in Japan. In 2016, NHO started to collect part of EMR (including laboratory test results and vital signs data) from 61 hospitals, which is called NHO Clinical Data Archives (NCDA). Data are used for clinical research by physicians affiliated with NHO.

DPC Data provision from academic groups:

- **The DPC Study Group database:** The DPC Study Group is a government-funded academic group that collects copies of the DPC electronic data for research purposes, independently of the MHLW (Yasunaga et al., 2013). The duration of data collection was 4 months from fiscal year (FY) 2002 to FY2005, 6 months from FY2006 to FY2009, 9 months in FY2010, and 12 months from FY2011 on. The number of participating hospitals and the number of inpatients whose data is included has steadily increased; since FY2010, the number of participating hospitals has been more than 1,000 and the coverage rate of inpatients in the DPC database of all acute-care inpatients in Japan has reached more than 50%. The DPC Study Group database partially corresponds to the Nationwide Inpatient Sample (NIS) in the US but has several unique features. Table 6 shows the comparison between the DPC data and the NIS data. There are several advantages of the DPC database, for example regarding coded diagnoses, as complications that occurred after admission are clearly differentiated from comorbidities that were already present at admission. Furthermore, the DPC database includes a variety of measures for severity, which can be utilised in clinical studies.

- **JROAD-DPC:** The Japanese Registry of All Cardiac and Vascular Diseases (JROAD) was launched in 2004 by the Japanese Circulation Society to assess the treatment of cardiovascular diseases at each hospital. All participating hospitals provide data on resources (hospitals, beds, and cardiologists), burden (number of inpatients), and outcomes (cardiovascular mortality and autopsy). Since 2014, the DPC data on patients with cardiovascular disease have been collected voluntarily from more than 900 DPC hospitals (Yasuda S, et al. 2018).

Table 6. Comparison between the DPC data and the US NIS

	DPC Database	US NIS
Scale	Approx. 7 million inpatients per year	Approx. 8 million inpatient hospital stays per year*
Sampling methods	Approx. 50% non-stratified sample of the Japanese academic and community hospitals	20% stratified sample of the US community, academic and some specialist hospitals
Population-representativeness	Not representative	Representative
Data elements		
Diagnosis codes	ICD-10	ICD-10 CM/PCS
Procedure codes	Japanese original codes	ICD-10 CM/PCS
Cost or charge	Estimated cost	Charge
Age, Sex	Available	Available
Race	N.A.	Available
Length of stay	Available	Available
Admission and discharge status	Available	Available
Hospital characteristics (ownership, size, teaching status)	Available	Available
Hospital identifier	Available	N.A.
Physician identifier	N.A.	Available
Median household income for patient's zip code	N.A.	Available
Patient's zip code	Available	N.A.
Body weight, height	Available	N.A.
Smoking index	Available	N.A.
Severity measures**	Available	N.A.
Laboratory data	N.A.	N.A.

Valid as at July 2019. DPC, Diagnosis Procedure Combination; NIS, Nationwide Inpatient Sample; ICD, International Classification of the Diseases; CM, clinical modifications.

*The data can be weighted to provide national estimates of more than 35 million hospitalizations nationally.

**Severity measures in the DPC data are seen in Table 3.

RWD provided by private companies

Since the mid-2000s, private companies have created databases of administrative claims data, DPC data, and EMR data. These commercially available datasets are available not only to academic researchers but private sector organisations, such as pharmaceutical or consulting companies.

Private provision of administrative claims and health check-ups data

Each insurer has its own administrative claims data and health check-up data, which can be internally anonymised and made available for a small fee. Private companies as well as researchers

can use these data to create independent secondary databases and use for their own purposes, in a process independent from the MHLW. The most popular commercially available databases are "JMDC Claims database" and "Medi-Scope", yet their size is considerably smaller than that of the NDB. They contain very limited data on elderly people aged 65 or more, given that most individuals included in the two databases are employees of Japanese companies and their families.

- **JMDC Claims database:** JMDC Inc. is contracted with more than 60 insurers of "Society-Managed Health Insurance (SMHI)" (see Table 1), and collects administrative claim data linked with health check-ups data from approximately 5,600,000 insured individuals to create JMDC claims data (JMDC, n.d; Kimura et al., 2010).

- **Medi-Scope:** KYOWA KIKAKU Ltd. collects administrative claim data from approximately 5,000,000 insured individuals to create Medi-Scope (Kyowa Kikaku, n.d.).

Private provision of DPC data

DPC data are made available for commercial uses by Medical Data Vision (MDV) Co., Ltd. (MDV, n.d.). It includes approximately 4,400,000 patients, around 3% of the population (Cheung et al., 2018). Data are extracted from hospital electronic information systems, derived from more than 300 acute hospitals throughout the country. Accessing the data involves a fee, and is available to pharmaceutical companies, medical device manufacturers and researchers for marketing, pharmacovigilance and epidemiology studies.

1.3 Core legislation & governance arrangements for the collection and use of patient data

1.3.1 Routinely collected patient data

RWD must be handled under the legal framework for using personal data, including the Act on the Protection of Personal Information (APPI) and related guidelines developed by the government. The legislation and governance arrangements have two important aspects: (i) patient consent for collecting and using routinely collected data, and (ii) de-identification of routinely collected data.

Consent for collecting and using routinely collected data

The APPI aims to protect an individual's rights and interests by protecting personal privacy from collection, leakage, misuse and abuse of an individual's information, while considering the utility of personal information. According to the Act, business operators who handle personal information shall not provide personal data to a third party without obtaining prior consent from the individual and should not acquire "sensitive personal information" without obtaining in advance an individual's consent (Article 76). However, after the 2017 NHIA legislation (discussed later), the Ministry of Education, Culture, Sports, Science and Technology (MEXT) and the MHLW discussed Article 76 and the ethical guideline for use of medical information was revised. It was agreed that Article 76 would be interpreted as not applying to academic research conducted by researchers in colleges, universities, or other organisations engaged in academic studies. This means that academic researchers do not have to obtain written informed consent from each individual when they use routinely collected data secondarily for academic studies. Under these conditions, most Japanese RWD are open only to academic researchers and societies, and are closed to the private sector (e.g., pharmaceutical companies and consulting companies), except for several commercially available databases which use deidentified patient data.

Common process for de-identification

An option that allows more flexibility within the Act on the Protection of Personal Information (APPI) is to obtain data through a de-identification process undertaken within the medical institution itself, so that patient data are no longer deemed to be personal information and are therefore no longer subject to the Act.

Identifiers must be removed from the database or replaced with alternative values before extracting datasets for secondary use. Labels that need to be anonymised include: (i) names, (ii) address, telephone number, and e-mail address, (iii) identification numbers (resident registration number, passport number, driver's license number, health insurance number, etc.). Quasi-identifiers include certain variables that are necessary for analysis, such as age, sex, race, body weight, body height, socioeconomic status, diagnoses, etc. The retention of this information might allow identification of a person when combined with other information. Hence, statistical anonymisation methods are recommended[2].

The process of de-identification is addressed in similar ways for various databases, as RWD contains personal information. In the NDB, all the identifiers are removed, and Hash identifiers are created to anonymise data and to be able to combine data by insured individuals. Two types of Hash identifiers are created; the "ID1" is generated from the insurance identification number, birth date, and sex, and the "ID2" is generated from name, birth date, and sex. Using Hash identifiers, researchers can trace individual patients longitudinally without using personal information. Until recently, the information entered against these ID variables was prone to change for a variety of reasons which made individual tracing more challenging (NTT DATA Corporation, 2019).

Other databases also follow the general principle that all identifiers are removed inside each cooperating medical institution or insurer before sending them to the data centre. Administrative claims data and health check-ups data are collected from insurers, and insurance identification numbers are replaced by dummy identifiers. Patients can be traced longitudinally without any identifying personal information, even if they change hospitals or clinics. This is not the case for DPC and EMR data, which are collected by hospitals, and hospital identification numbers are replaced by dummy identifiers. In this case, when patients change their hospitals, they can no longer be traced.

The Next-generation Healthcare Infrastructure Act

There have been many issues related to accessing and using NDB data. It has a rigid structure, which is a characteristic of relational databases, and reworking has been required at each step of data use. Cumbersome data conversion work was needed to reorganize data into structured data suitable for research. An especially significant issue was solving the "ID problem", which prevented continuous analysis when a patient's ID changed due to marriage, retirement, and other life events, as well as typographical errors. NTT DATA implemented the ID matching algorithm provided by Nara Medical University on the distributed platform. As a result, obstacles that blocked continuous data tracing were cleared, increasing the accuracy of patient ID-based data aggregation (NTT DATA Corporation, 2019).

In April 2017, the National Diet passed a new piece of legislation called "Act regarding anonymised medical data to contribute to research and development in the medical field". Because this name is too long, this new legislation is officially called as "the Next-generation Healthcare Infrastructure Act"

[2] These include grouping of categories, local suppression, adding noise, micro-aggregation, and top-and bottom coding. These should be adopted, taking into consideration appropriateness and the study purpose. Furthermore, the propriety of de-identification can be assessed with the k-anonymity model, a privacy model commonly applied to protect the data subjects' privacy in data sharing scenarios.

and is commonly referred to as NHIA or "Healthcare Big Data Act". This law was put into force in May 2018 with the following key points:

1. The government is responsible for implementing necessary policies to enhance medical research using anonymised medical data.
2. Only "certified operators for de-identifying medical data" who have high security measures and high technical abilities to anonymise the data for optimal use would be entrusted with managing patients' personal information.
3. Medical institutions are required to post a notice announcing that anonymised patient data will be secondarily used for research purposes. Unless patients opt out, the institutions are permitted to provide their personal information to certified operators for de-identifying medical data.

According to NHIA, operators for de-identifying medical data are certified by the MHLW and are the only organisations allowed to collect patients' personal information directly from healthcare providers; collate these data; identify the same patients across different databases; give unique identifiers to the same patients; and link all the data at the individual level. Then, the certified operators exclude all the personal information from the database so that they can provide de-identified data to academic researchers. They have a legal responsibility to restrict their collection and management of the data solely to providing de-identified data for academic researchers and to prevent personally identifiable data leakage. After this law was enacted, no third party other than certified operators can receive personal information from medical institutions without patient consent.

Dispersed data in multiple institutions can be synthesised for research and development of medical care. Providing data from institutions to certified operators is voluntary.

1.3.2 Collecting de novo patient data

In Cole et al. (2015), the process and rules around collecting new information from patients are distinguished from those on the use of routinely collected patient data. In Japan, the APPI and the ethical guideline also apply to studies collecting de novo patient data. The ethical guideline requires that researchers obtain oral consent when using specimens collected from humans (for example, blood samples). When researchers are not using specimens but are collecting "sensitive personal information" including medical information, agreement by participants is required, unless situations do not allow for appropriate agreement. In such cases, researchers may conduct studies by posting a notice that research is being conducted and patients who receive medical care can refuse the use of their information.

1.3.3 Data linkage

Patient data can be linked across datasets only within the same institution. The multiple-payer health insurance system in Japan has made it difficult to gather data from the entire population, as each payer collects data for their insured members only. Only the DPC data and EMR in a hospital can be linked together, with the patient identifier used in the hospital as a linkage key.

It is not possible to link data from different databases at patient level. For instance, datasets extracted from the NDB cannot be linked with any other patient-level dataset. Although, many hospitals and clinics collect patient information electronically to create EMR; formats of EMR differ greatly between medical institutions (hospitals and clinics), so even if de-identification is done following the same structure, integrating EMRs from multiple institutions will remain challenging.

To improve the situation, in 2006, the MHLW started a programme that enhanced "Standardised Structured Medical Information eXchange (SS-MIX). SS-MIX is a standardised rule to exchange and

share EMR among medical institutions (SS-MIX, n.d.). Based on this rule, several databases have been established containing EMR as well as administrative claims data from multiple hospitals; e.g., the Medical Information Database Network (MID-NET) and the National Hospital Organisation (NHO) Database.

1.4 Data access

1.4.1 Access to NDB data

The MHLW provides individual level data extracted from the NDB and aggregated data created with the NDB to researchers with academic or public policy affiliations free of charge. The NDB expert council reviews the applications for providing the NDB data based on "the guideline for providing the NDB data" (MHLWc, 2016); the documents to be attached to complete the application are listed in Table 7 below.

There are three types of datasets available: (i) a specifically extracted dataset, (ii) a sample dataset and (iii) aggregated data. Despite the process for applying for data being clear and precise, the structure of the general dataset and access requirements in place have made it difficult for researchers to use the data. This has made it more challenging to analyse health insurance claims data and health check-ups data and some modifications have been recently implemented, which are described below.

Tables 7. Documents that are attached to the application for providing NDB data

	Specifically extracted dataset	Sample dataset	Aggregated data
Application for providing the NDB data	Required	Required	Required
Certificate of approval from the president of the institution	Required	Required	Required
Copy of identification card	Required	Required	Required
Copy of employee identification card	Required	Required	Required
Flow chart for managing the provided data	Required	Required	Required
Document on how to deal with risks	Required	Required	Required
Operating management regulations	Required	Required	Required
Self-monitoring rules	Required	Required	Required
Privacy policy or information security policy in the affiliated institution	Optional	Optional	Optional
Notification of receiving a grant from the MHLW	Optional	Optional	Optional
The applicant's research achievement (copies of published papers)	Optional	Optional	Optional
Certificate of the IRB approval	Required	Unnecessary	Unnecessary
Detailed explanation for data extraction criteria	Required	Unnecessary	Required
Detailed explanation for formats of tables and figures that will be presented on reports and manuscripts	Required	Unnecessary	Required

Specifically extracted dataset

Applicants can submit an application to the MHLW to extract a dataset from the NDB database for a specific research purpose. An application form can be downloaded from the MHLW website, and researchers need to consult the MHLW before submitting the application. They must ensure they use the NBD data in accordance with the guidelines. Figure 2 below outlines the review process to apply for a specifically extracted dataset, which is described below.

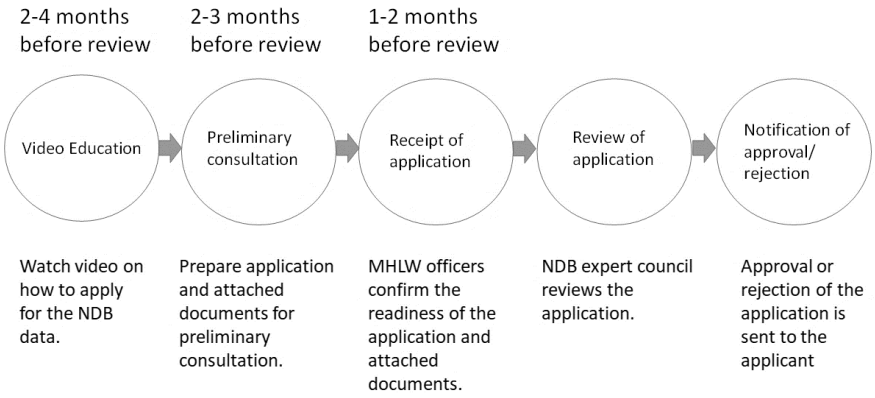

Figure 2. Review process to apply for a specifically extracted dataset.
Source: MHLWc, 2016.

Before the research teams prepare the application form and the required documents needed for a preliminary consultation by the administrative officers in the MHLW, all applicants must watch an educational video[3].

The MHLW officers will then confirm the readiness of the application and attached documents. The application is then reviewed by the NDB expert council, which deliberates on the propriety of data provision. Conferences for review of application by the NDB expert council are held four times a year (in March, June, September, and December), and are not open to the public. Criteria for reviewing the study application are as follows:

- The study is beneficial to society.
- The NDB data are suitable for achieving the study aim.
- The data extraction criteria are appropriate for implementing the study.
- The applicants have a study environment that assures security for data use to prevent data loss or leakage.

After this review, a document of approval or rejection of the application is sent to the applicant. The names of the applicants who received acceptance notices, their affiliations and study titles are disclosed. In some cases, the decision may be suspended and revision by the applicants required.

When the application is accepted, researchers enter into a contract with the MHLW to use a dataset extracted from the NDB for a specific study. MHLW officers extract data. After the security measures for data extraction are handled and completed, the MHLW will send a hard disc in which datasets are stored. Users of the NDB data or MHLW-DPC enter into a contract with the MHLW. The users must agree to the Terms of Use and sign the written oath before receiving data free of charge.

[3] *Video in Japanese: http://www.youtube.com/playlist?list=PLMG33RKISnWjiXXj6lpX7t5FbsPcjeD1b*

Until recently, to use specifically extracted datasets, researchers had to prepare a secure room in which the dataset was stored, and data analyses performed. This posed a large burden on the researcher so On-site Research Centres of the NDB were established in Kyoto University and The University of Tokyo for those researchers/analysts or institutions that could not afford to have such facilities. Data users are able to visit these centres and conduct their analyses on-site. However, researchers and analysts still have to go all the way to the on-site research centre even to understand how to analyse the data. Users unfamiliar with the NDB system that has more than 10 billion records may not be able to obtain a result, even after several days of work depending on their data search and analysis method. Further change is therefore needed, and two open-source software packages, Apache Hadoop and Apache Spark have been selected as the core software of the next-generation NDB. The new software is expected to make it easier to flexibly cope with the increasing amount of data and data types, while satisfying extremely high security requirements, Furthermore, e-learning contents to learn the basics of health insurance claims have been prepared, together with an NDB hands-on environment where users can search and test their analysis logic using NDB dummy data. With these enhancements, users can now learn and understand the logics of the system (NTT DATA Corporation, 2019).

Sample dataset

A sample dataset is created with one-month of data per year (in June), instead of providing numbers based on health insurance claims per year. The NDB provides more accurate information than the sample dataset, but the sample datasets has statistics from 2005 instead of 2014, so it can provide data to analyse longer trends (Katano et al., 2018).

After receiving the user's application for a sample dataset, the NDB expert council assesses the propriety of data provision and notifies the researchers of the result. If the application is successful, the MHLW will send a digital disc (DVD) in which the dataset is stored to the researcher by mail. In this case, preparing a secure room is not required.

Aggregated data

The MHLW publishes aggregated summaries of NDB data which are intended to be comprehensible to the general public, called NDB Open Data Japan (NDB-ODJ). Following the NDB data structure, statistics are based on claims for outpatients, inpatients, prescriptions, dental treatments and specific health check-ups (MHLW, 2016).The first version was made available in October 2016, consisting of spreadsheets that sum up the claims data for the fiscal year 2014 and specific health check-up data for the fiscal year 2013. The NDB-ODJ provides not only the total number of prescriptions and procedures in Japan but a separate number for each of the 47 prefectures. An updated version of the NDB-ODJ is to be published every year.

Following a different process, and based upon researchers' requests, the MHLW also creates aggregated tables based on NDB data. In this case, after receiving the user's application, the NDB expert council assesses the propriety of data provision and notifies the researchers of the result. If the application is accepted, the MHLW creates aggregated tables of the NDB data and provides them to the applicants. A secure room for analysing the data is also not required in this case.

1.4.2 Access to other databases

- **MHLW-DPC database:** Governmental officers and academic researchers can apply for provision of aggregated data from the MHLW-DPC database. As of 2019, the MHLW does not provide individual-level DPC data; only aggregated data are provided, and the application form and documents to be attached are similar to those needed for the NDB aggregated data request.

- **MID-NET:** data are open to pharmaceutical companies and academic researchers who conduct investigations and research in the public interest, such as those for drug safety. Those who want to use the data must submit an application to the PMDA. The application will be reviewed by the PMDA expert council. When the application is accepted, the researchers enter into a contract with the PMDA to use a dataset extracted from the PMDA for a specific study. Researchers can access and analyse the extracted data only in the on-site centre in the PMDA.

- **NHO dataset:** accessible only to researchers affiliated with the NHO.

- **DPC Study Group:** The DPC data collected by the DPC Study Group are accessible only to the members of the group.

- **JROAD-DPC** The JROAD-DPC is accessible only to the members of the Japanese Circulation Society.

- **JMDC, Medi-Scope, and MDV databases**: While access to the NDB is strictly limited to selected academic researchers or governmental officers, the JMDC database, Medi-Scope, and MDV databases are commercially available, and open to the private sector.

1.5 Data use

1.5.1 The NDB data and MHLW-DPC

The MHLW provides the NDB data and MHLW-DPC data. Aggregated summaries of the data are disclosed on the website of the MHLW so that both patients and providers can freely access the information and see the numbers of patients and the average length of stay for each DPC category in each hospital. This information can be utilised by patients to select hospitals based on clinical information and by hospitals to improve their clinical practice.

Those who can apply for provision of aggregated data from the NDB or the MHLW-DPC database include: (i) officers in the national government, local governments and quasi-governmental agencies, (ii) academic researchers affiliated with colleges, universities or research institutes, (iii) federation of health insurance organisation (including All-Japan Federation of National Health Insurance Organisations, Japan Health Insurance Association, and National Federation of Health Insurance Societies), (iv) Public Interest Corporations that aim to improve quality of health services and (v) holders of national academic grants for studies using the NDB data or MHLW-DPC aggregated data. Pharmaceutical or consulting companies cannot request data access, as requests for administrative claims data and health check-ups data to undertake studies with a commercial interest are not allowed.

The applicants must have approval from an institutional review board before requesting NDB data from the MHLW. The institutional review board discusses conflict of interest issues, while the NDB expert council discusses privacy protection issues relating to the study. More specifically, the NDB expert council, composed of academic authorities and representatives of insurers and healthcare providers, deliberates on matters related to the provision of the NDB data for the study.

All the reports (including original manuscripts and abstracts for academic meetings) must be submitted to the MHLW, before publicising them. This is to ensure that, when publicising the results of analyses using the NDB, the users take into consideration that individuals must not be identified. Although the identifiers in the NDB are removed, combining multiple quasi-identifiers might result in identification of an individual. To avoid this, the users must present the results so that the number of any unit must not be less than 10. Age should be categorised to every 5 years and patients aged 85

years or more should be categorised into a single group. The minimum unit of regions should be municipalities. Furthermore, users must submit tables and figures before publicising them to the MHLW, and the NDB expert council reviews the results and makes a decision on the propriety of publicizing them. Original papers using the NDB database are gradually increasing; the cumulative number of original articles published in peer-review journals is around 30, as of 2019. After the study period is over, all the data must be destroyed. Inappropriate use includes losing data, leaking data, or using data for purposes other than approved. For users who violate the terms of use, further provision of data is prohibited, and their names and affiliations are disclosed.

1.5.2 MID-NET

PMDA provides MID-NET data. The PMDA expert council deliberates on matters related to the provision of the MID-NET data. This council of experts is composed of third party experts independent of PMDA, and its purpose is to protect the privacy and other rights and interests of the patients while promoting the appropriate use of MID-NET. The council confirms that the purpose of data use, method for using data, and method for making public the results are all appropriate. Data users must pay a fee to use MID-NET data. For pharmaceutical companies, the fee for use of data for one product is set as 42,123,000 yen (384,105 USD).

1.5.3 Other databases

The DPC data collected by academic groups are accessible only by the members of the groups. As such, users of these data are limited to selected academic researchers. However, a lot has been done and the cumulative number of original articles published in peer-review journals is more than 300. Commercially available databases, including JMDC database, Medi-Scope, and MDV database, are open to the private sector. Information on costs for purchasing these data is not disclosed.

2 Data governance arrangements in Japan and comparisons with other countries

In Cole et al. (2015), the governance arrangements for RWD in eight countries (Australia, France, Germany, Italy, Sweden, the Netherlands, the UK and the US) were presented in a framework. For each area of the framework, which is presented below, the authors outlined key elements and provided their proposed for the ideal scenario. In addition, they developed a heat map of how the individual countries perform against the key criteria that they set out in the governance framework. A description of all the ideal framework elements can be found in Appendix 2.

In this section of the report, we provide a table that provides summary of data protection, data linkage, access and governance objectives in Japan (Table 8). This is followed by a heat map of how Japan performs against the key criteria established in 2015 by Cole et al. This is based on the authors' assessment of the information obtained as part of this project.

In Japan, various entities, governmental and non-governmental, academic and non-academic, collect and use routinely collected data (administrative claims data, health check-up data, DPC data, and EMR data). Accessibility varies widely between the databases. Overall, access to governmental databases (NDB and MHLW-DPC) is restricted to academic researchers and governmental officers. Quasi-governmental organisations or academic groups collect de-identified DPC data and EMR data from multiple hospitals, independently of the MHLW. Access to these datasets is limited to researchers affiliated with these organisations or academic groups, except for MID-NET data, provided by the PMDA, which is provided upon request with a charge for academic researchers or pharmaceutical companies. Several private companies collect de-identified administrative claims data or the DPC data, independently of the MHLW. Aggregated or individual-level datasets are commercially available not only to researchers but also to pharmaceutical companies and consulting companies, although the sizes of the databases are much smaller than those of the governmental databases.

Cole et al. (2015) showed that a framework often used in the eight countries for their RWD governance arrangements was 'consent or anonymisation'. That is, informed consent should be generally obtained when using patient data to for research; however, if consent cannot be obtained, then data must be completely anonymised. In reality, most countries make data accessible when research would be beneficial, even if consent is impossible, although in many countries access is highly restricted. Administrative claims data are already collected by health insurers for reimbursement purposes, and it is unrealistic to obtain each individual's written informed consent for each new use. This appears to be the case also in Japan. According to the APPI (amended in 2017), medical information corresponds to "sensitive personal information", and business operators must not acquire medical information without obtaining in advance an individual's consent. However, the law does not apply to academic research conducted by academic researchers. The framework in Japan seems to emphasise anonymisation of personal data to a greater extent than consent. This can protect privacy but, depending on how it is implemented, can limit the opportunities to link databases. This issue may be improved by potential change in the governance environment with the Next-generation Healthcare Infrastructure Act.

In general, Japan has clear data protection requirements that recognise the legitimacy of health care data utilisation beyond the direct care of patients, and patient identifiers which conform to national standards are used. However, even if it is clear that responsibility for the data after collection passes to the data controller, the role of patients in the permissions / management of their data is still unclear. Arguably the biggest challenge of the Japanese system is to develop a clear set of nationally agreed and implemented standard rules to optimize interoperability of health record systems, which is key for datasets to be compatible with one another.

Table 8. Data governance in Japan

Data Protection – Health	Data Linkage	Access	Governance ideals and changes in the environment
The Act on the Protection of Personal Information aims to protect an individual's rights and interests by protecting personal privacy from collection, leakage, misuse and abuse of individual information, while considering the utility of personal information. According to the Act (amended in 2017), Article 76, business operators who handle personal information shall not acquire medical information without obtaining in advance an individual's consent. However, the Article will not be applied for academic research conducted by researchers in colleges, universities, or other organisations engaged in academic studies. That is, academic researchers do not have to obtain written informed consent from each individual when they use routinely collected data secondarily for academic studies.	Patient data can be linked across datasets only within the same institution. For example, the DPC data and EMR in a hospital can be linked together, using the patient identifier used in the hospital as a linkage key. All the patient data must be de-identified in each hospital or by each insurer before taking out the data from the hospital or the insurer. All the databases include de-identified data from multiple hospitals or insurers. We cannot link data from different databases at individual patient level. For instance, datasets extracted from the NDB cannot be linked with any other individual patient level dataset.	Access to government databases (NDB and MHLW-DPC) is restricted to academic researchers and governmental officers. Quasi-governmental organisations or academic groups collect de-identified DPC data and EMR data from multiple hospitals, independently of the MHLW. Access to datasets is limited to researchers affiliated with these organisations or academic groups, except for MID-NET provided by the PMDA. MID-NET data are provided upon request with a charge for academic researchers or pharmaceutical companies. Several private companies collect de-identified administrative claims data or the DPC data, which are commercially available for not only researchers but pharmaceutical companies and consulting companies	The Next-generation Healthcare Infrastructure Act was put into force in May, 2018. The government is responsible for implementing necessary policies to enhance medical research using anonymised medical data. Only "certified operators for de-identifying medical data" who have high security measures and high technical abilities to anonymise the data for optimal use would be entrusted with managing patients' personal information. Medical institutions are required to post a notice announcing that anonymised patient data will be secondarily used for research purposes. Unless patients opt out, the institutions are permitted to provide their personal information to certified operators for de-identifying medical data.

Table 9. Heat map of data governance arrangements in Japan

Routinely collected /De novo Raw data		
Government as Regulator	Data protection legislation (health 'special case')	The general law governing data protection is the Act on the Protection of Personal Information. According to the Act (amended in 2017), Article 76, business operators who handle personal information shall not acquire "sensitive personal information" without obtaining in advance an individual's consent. Medical information (including patients' past history and present status regarding their health) corresponds to "sensitive personal information". However, the Article will not be applied to academic research conducted by researchers.
	Equitable patient selection and the protection of vulnerable subjects	N/A
Data subjects: Patients	Patient consent	Researchers do not have to obtain written informed consent from each individual when they use routinely collected data secondarily for academic studies.
	Facilitative opt-in / opt-out consent models for research	Cooperating medical institutions must post a notice of the fact that routinely collected medical information is utilised for research purposes. Patients who receive medical care at cooperating medical institutions can refuse to allow such use of their medical information by informing the medical institution of such.
Data Collectors	Unique patient identifiers (UPIs)	To ensure the anonymity of a patient, UPIs are de-identified, and replaced with an alternative serial-number ID.
	Patient information	N/A
	Data quality assurance	MHLW is under an obligation to maintain up to date and accurate records of the NDB and DPC.
	Data ownership: responsibility for data?	Unclear role for patients in the permissions / management of their data.
Data Users	Approval of data collection activities to be based on intended use	Researchers must have approval from the Institutional Review Board (IRB), which takes into account intended use.
	Clear and transparent criteria for de novo data projects	No clear framework for the collection of de novo observational data.
Cleaning and managing data		
Government as Regulator	Data management: Recognised data stewardship entities	The collectors of the data have responsibility for its data security, management and access.

Data Controllers	Process for de-identification	Statistical anonymisation methods are recommended, such as grouping of categories, local suppression, adding noise, micro-aggregation, and top- and bottom coding.
	Security arrangements: 'Privacy Enhancing Techniques and Procedures' (PETs)	N/A
	Training of staff	N/A
	Specified arrangements for how long data are kept	N/A
Linkage and aggregation		
Government as Regulator	Privacy rules	Clear process of assessment of privacy risk.
	Develop a clear set of nationally agreed and implemented standard rules to optimise interoperability of health record systems	Datasets cannot be linked between databases (e.g. cannot link the NDB and DPC data).
Data Controllers	Unique patient identifiers	Alternative serial-number ID, which is a de-identified version of the unique patient identifiers.
	Pseudonymisation	Process for pseudonymisation exists where it is necessary to keep identifying information.
	Preparation for sharing	Processes outlined by the data controllers.
Access / use of data		
Government as Regulator	Managing re-identification risk	All reports using NDB data provided by the MHLW must be submitted for review before publishing, and users must ensure that individuals will not possibly be re-identified. The expert council will review the submissions and make a decision on its appropriateness.
	Criteria for different uses (& different users)	More abundant data are provided to users with academic/public policy research purposes, whereas limited data are provided to users in private sectors, such as pharmaceutical companies or individual entities.
Data Controllers	Approval panels	The NDB expert council.
	Confidentiality and data use agreements	N/A
	Balancing benefits of linkage for research with risk for re-identification	N/A

Data users	Audit / Service evaluation and quality monitoring	N/A
	Degree of access, level of data, and mode of access	Processes clearly set out, but access to data is limited to academic purposes.
	Cost of access	Cost of governmental data are free; cost of other data are clearly outlined.
	Appropriate experience/qualifications, and funding to conduct research	N/A

Colour Key: green = aligned with recommended; amber = ok but with room for improvement; red = very problematic/ barrier. Squares are blank where it was felt that there was insufficient information to make a judgement. Source: Based on authors' interpretation.

3 Concluding remarks

Real world-data (RWD) and real-world evidence (RWE) are playing an increasing role in clinical studies and healthcare decisions. In Japan, as well as other countries, academic studies using RWD have been increasing, and academic societies are using RWE to revise guidelines and to create medical decision tools useful in clinical practice. The MHLW also uses RWD to support decisions on inclusion in universal healthcare coverage. The PMDA has just started using RWD and RWE to monitor adverse drug events and to make regulatory decisions. Pharmaceutical companies are using commercially available RWD for post-marketing studies and for supporting drug development, including clinical trial designs (e.g., pragmatic clinical trials) and observational studies to better understand patients and generate innovative drugs.

In the present report, we described the current status of RWD in Japan, including core legislation and governance arrangement, data linking, data access, and data use, as compared with the "ideal framework" proposed in Cole et al. (2015).

The framework includes two aspects: patient consent to use data and de-identification of the data. According to Japanese legislation, "sensitive personal information" including medical information must not be acquired without written informed consent for business purposes; however, academic studies are not required to comply with this regulation. Therefore, most Japanese RWD are open only to academic researchers and societies, and are closed to the private sector (e.g., pharmaceutical companies and consulting companies), except for several commercially available databases. This means that under the current setup less research than may be optimal is carried out by private sector organisations where this is not subcontracted to academic institutions. However, Japan is not an isolated case in respect of this. In other countries the costs of accessing and ability to use data generally differ for research conducted by non-profit versus for-profit organizations.

Governance arrangements in Japan emphasise de-identification. Hitherto, all the RWD have been de-identified inside the hospitals or by insurers; personal information is strictly prohibited from being taken outside the hospitals or insurers. In particular, the government and quasi-governmental organisations have very strict regulations for protecting personal information, sacrificing usability of the RWD (e.g., traceability of patients over time and across different data sources). Other entities that collect RWD face similar regulations, following the government's practice.

Japan is still facing some challenges in developing a clear set of nationally agreed and implemented standard rules to optimise the interoperability of health system records, which limits the ability to benefit from achievements in the areas of de-identification, privacy rules, data control, and access. Also, there is currently lack of clarity for the collection of de novo observational data, given that there are no specific guidelines to be followed. According to the 2015 ideal framework, for national projects, there should ideally be a central ethical review board whose decision is accepted by the relevant national and local parties; this would reduce duplication of effort and promote consistent coverage.

The ability for central linkage of datasets may be impeded in countries where there are multiple data custodians each managing distinct datasets, and this is the case in Japan. The ability to link data across datasets is incredibly important for research, and thus the country should prioritise this in the RWD agenda. This may be facilitated by NHIA but further change will be needed if not. For example, countries are using a unique patient identifier, which may either have been created specifically for health care, or be an identifier used more broadly for other services such as social security numbers or national person numbers. Most likely progress in this area will come with a greater re-identification risk, which will have to be considered at the later data sharing stage.

New legislation, the Next-generation Healthcare Infrastructure Act (NHIA), was put into force in 2018 and has the potential to fundamentally change the governance arrangements for RWD in Japan. According to this law, "certified operators for de-identifying medical data" are allowed to collect patients' personal information directly from healthcare providers; link all the data at individual level; and provide de-identified data to academic researchers. The implication of this being that NHIA can improve individual-level data linkage and has the potential to enhance RWD studies. However, given that the new legislation has just been launched, the MHLW is still searching for candidates to act as the certified operators, and thus the issue of data accessibility has not been fully addressed yet. As such, it is still unclear whether this new development will help deal with the asymmetry of the position of commercial and non-commercial researchers.

References

Cheung S., Hamuro Y., Mahlich J., Nakayama M., Tsubota A., 2018. Treatment pathways of Japanese prostate cancer patients - A retrospective transition analysis with administrative data. PloS ONE, 13(4).

Cole A., Garrison L., Mestre-Ferrandiz J., Towse A., 2015. Data Governance Arrangements for Real-World Evidence. OHE Consulting Report, London: Office of Health Economics.

Garrison L.P., Neumann P.J., Erickson P., Marshall D., and Mullins C.D., 2007. Using Real–World Data for Coverage and Payment Decisions: The ISPOR Real–World Data Task Force Report. Value in Health, 10(5): 326-335.

Gartner, 2014. IT Glossary. Information Governance. Gartner, Inc. Available: http://www.gartner.com/it-glossary/information-governance

JMDC, n.d. Website. Available: https://www.jmdc.co.jp/en/

Katano H., Koga H., Ozeki N., Otabe K., Mizuno M., Tomita M., Muneta T., Sekiya I., 2018. Trends in isolated meniscus repair in meniscectomy in Japan, 2011-2016. Journal of Orthopaedic Science, 23: 676-681.

Kazuhiko O., 2017. MID-NET: Medical Information Database Network Project to Utilize Electronic Healthcare Data for Drug Safety. J-STAGE, Transactions of Japanese Society for Medical and Biological Engineering, 55(4): 159-164.

Lee E.K., Park J.A., Cole A., Mestre-Ferrandiz J., 2017. Data Governance Arrangements for Real-World Evidence in South Korea. OHE Consulting Report, London: Office of Health Economics.

Matsuda S., Fujimori K., 2012. The Claim Database in Japan. Asian Pacific Journal of Disease Management, 6: 55-59.

MDV, n.d. Introducing MDV Database. Available: https://www.mdv.co.jp/solution/pharmaceutical/english/

MHLWa, n.d. Guidance 1. Available: https://www.mhlw.go.jp/bunya/iryouhoken/iryouhoken01/dl/01_eng.pdf

MHLWb, 2019. Guidance 2 (in Japanese). Available: https://www.mhlw.go.jp/content/000498905.pdf

MHLWc 2016. Guidance 3 (in Japanese). Available: https://www.mhlw.go.jp/file/05-Shingikai-12401000-Hokenkyoku-Soumuka/0000135460.pdf

MHLW, 2016. NDB Open Data Japan. Available: http://www.mhlw.go.jp/file/06-Seisakujouhou-12400000-Hokenkyoku/0000193322.pdf.

NTT DATA Corporation, 2019. National Database of Health Insurance Claims and Specific Health Checkups of Japan (NDB). Available: https://www.nttdata.com/global/en/-/media/nttdataglobal/1_files/success-stories/2019/amed.pdf

SS-MIX, n.d Website (in Japanese). Available: http://www.ss-mix.org/cons/ssmix2_about.html

Tamura T., Kimura Y., 2015. Specific health checkups in Japan: The present situation analysed using 5-year statistics and the future. Biomedical Engineering Letters, 5(1): 22-28.

Yasunaga H, Matsui H, Horiguchi H, Fushimi K, Matsuda S., 2012. Clinical Epidemiology and Health Services Research using the Diagnosis Procedure Combination Database in Japan. Asian Pacific Journal of Disease Management 7: 19-24.

Yasuda S., Miyamoto Y., Ogawa H., 2018. Current Status of Cardiovascular Medicine in the Aging Society of Japan. Circulation, 138(10): 965-967.

Kimura S., Sato T., Ikeda S., Noda M., Nakayama T., 2010. Development of a database of health insurance claims: standardization of disease classifications and anonymous record linkage. J Epidemiol, 20: 413–9.

Kyowa Kikaku, n.d. Medi-Scope (in Japanese). Available: https://www.kk-kyowa.co.jp/medi_scope.html

Appendix 1. PRO-FORMA

Data Governance Arrangements for Real-World Evidence in Japan

1. **Brief overview** of the health system and collection / management of patient data. Specify the key data sources available for RWD.
    - Health system: Organisation of health institutions involved in the financing and provision of care.
    - Collection/management of patient data:
        - Source: claims databases? Track of utilisation of health services? National health insurance companies collecting data? Electronic medical record data?
        - Area: mortality, medical examination, treatment, prescription, disease burden, etc.
2. **Core legislation and governance arrangements for the collection and/or use of patient data**

    a) Routinely collected patient data:

    Core legislation governing the collection / use of routinely collected patient data. Review and summarise key documentation outlining principles of governance and data protection.
    - Collection and protection of personal/patient information
    - Legal frameworks

    b) Collecting de novo patient data.

    Governance arrangements for research to collect new data. Review and summarise key documentation outlining research ethics and governance for the collection of new patient data and governing principles of the committees that grant approval.
    - Process and rules around collecting new information from patients
    - Patient consent

3. **Data linking.** To what extent can patient data be linked across datasets? What are the organisations involved, and what are the core governing principles under which they operate?
    - Linking data from different datasets: is it possible? If so, how? Using which variables? If not, is there an alternative way to combine different datasets?
4. **Data access.** To what extent is data shared, with whom, and what are the principle governance issues in the preparation / sharing of this data?
    - Can anyone access the datasets? Only academic/research purposes? How is access structured?
5. **Data use.** What, if any, are the rules governing the use of RWD? [To cover contract arrangements between data suppliers and recipients, rules around use for HTA, etc.]
    - Type of data, definition, data recipient, cost, operation guide
6. **Governance ideals and changes to the environment.** Summarise any key national documentation that contains advice or commentary on ideal governance frameworks, as well as information on any imminent changes to the governance environment.
    - National documentation in place?

Appendix 2. Ideal framework content

As proposed in Cole et al. (2015).

IDEAL FRAMEWORK for raw data:

- **Data protection legislation.** Clear data protection requirements that recognise the legitimacy of health care data utilisation beyond the direct care of patients.
- **Data quality assurance.** Requirements that records are accurate, and up-to-date. Patient identifiers which conform to national standards should be used and stored with the record.
- **Patient consent.** Where patient consent is not feasible, the collection of data for purposes beyond direct care can be supported with relevant legislation. Requirements that new legislation be passed for each new dataset poses prohibitive restraints on legitimate and worthwhile data collection activities. Greater flexibility can be administered through a legislative framework that grants statutory exemption for the requirement of consent where this would be too burdensome and where the purpose of the exemption is in the interest of the public. This should be decided after careful assessment by an ethical review board. This kind of regulation can be government-sanctioned but privately administered by a government entity. Where data collection is to be collected on a routine basis across a large patient cohort, an opt-out, rather than opt-in, system of patient consent may serve to maximise coverage and allow patients to contribute data more easily.
- **Patient information:** There must be clear communication to data subjects of potential future uses of their data. Not explaining simply and clearly the rights of patients to opt-out or 'object' to their data being collected and later used for purposes not aligned with their own care can damage public trust (HSCIC, 2015).
- **Approval of data collection activities to be based on intended use.** This relates to de-novo data collection. The requirements for new data collection activities should be cognizant of the future intended use of the data. For example data collection activities that often form part of MEAs or risk-sharing arrangements between payers and manufacturers should be recognised as essential to the appropriate and optimal treatment of patients. Clear and transparent roles for the various actors in the collecting and eventual sharing of data should be well set out, which will enable access to data without harm or impact on privacy and public interest positions.
- **Clear and transparent criteria.** The criteria of Ethics Committees for data collection projects ('de novo' data) should be clear, transparent, and replicable. For national projects, there should ideally be a central ethical review board whose decision is accepted by the relevant national and local parties; this would reduce duplication of effort and promote consistent coverage.
- **Data ownership.** Responsibility (to be distinguished from 'ownership') for the data after collection passes to the data controller, who must act in the interest of patients and the public as specified by law.

IDEAL FRAMEWORK for cleaning and managing data:

- **Recognised data stewardship entities.** Data stewardship entities that manage the acquisition, storage, aggregation, and de-identification of data. The interests of those entities must be aligned with those individuals whose data is being collected. These come under various names, for example 'Trusted Third Parties'. These organisations must comply with the relevant legislation for the countries in which they operate.

- **De-identification of data.** Where appropriate, data can be de-identified by removing any personally identifiable information and replacing the unique patient identifier (which in some countries is used across different sectors of the economy and therefore highly sensitive) with a pseudonym. Where data is not managed by one single entity, care should be taken that the algorithm for the pseudonymisation process is replicable for other datasets so that they may be linked, or else that the pseudonymisation process be reversible when desirable.

- **Data quality.** In the same way that individuals and organisations collecting data from patients have a responsibility to ensure that the data are relevant, up-todate, and accurate, so should those organisations processing patient data ensure that the quality and integrity of the data is maintained.

- **Security arrangements.** Security arrangements for the protection of confidential patient data should be assured through sound security processes, ranging from physical and technical computing protections and to the legal, security, and confidentiality training of staff involved in processing the data. Such processes and techniques are often called 'Privacy Enhancing Techniques and Procedures' (PETs), which should be implemented for the anonymisation of data as well as in preventing loss of anonymity at a later date.

- **How long data are kept.** In many countries, it is specified through data protection legislation that data should be kept 'no longer than necessary'. This is difficult to define, but the importance of rich longitudinal data that follows a patient over time through the care pathway and its benefits for research should be considered.

IDEAL FRAMEWORK for linkage and aggregation:

- **Develop a clear set of nationally agreed and implemented standard rules to optimize interoperability of health record systems.** This is key for datasets to be compatible with one another.

- **Data linkage by trusted third party.** Common organisational and technical barriers to data linkage arise when there is no single group or organisation that has the responsibility or technical expertise required to manage the linking process. This could be minimised if linkage is undertaken by a single trusted third party. Where pseudonym IDs are created to facilitate the sharing of data with reduced risks whilst still allowing for linkage of datasets, the pseudonym IDs are common to the linked datasets and indicate that the records belong to the same person while protecting anonymity. This is more feasible in systems where management of this process is centralised. Whilst pseudonymisation helps to reduce the potential identifiability of data, there will always remain some residual risk of jig-saw re-identification. Therefore, it is still appropriate for requests for non-aggregated data to be examined by information governance panels (often through ethics committees) which consider the balance between the risk to patient confidentiality and the public interest in the research. This process is considered below.

IDEAL FRAMEWORK for access / use:

- **Forms of data access.** Different access arrangements may be employed to achieve the needed balance between protection of private information and informing real-world research:
  - An often used model involves the potential data user applying for access and following privacy review and contracting, from the data provider. In this scheme, the data provider may offer information at varying levels of detail and scrutiny:
    - Data may be provided at the aggregate level in which there is no information about individual patients. Data at this level may be provided freely since the risks are low.
    - Where the data provider has capacity for such services, analyses may be conducted in-house by the data provider, the results of which are then shared with the applicant. Similarly, this would involve minimal risk to privacy.
    - Data may be provided at the level of individual patients but with most or all individually identifying elements removed (e.g. social security numbers). This level should require a routine data use agreement form in which the data user agrees to protect the privacy of individuals in the dataset and not attempt to discover their identities.
    - Data may be provided at the level of individual patients with most or all individually identifying elements intact. Clearly, this level of information carries greater risk to privacy. However, this may be justifiable in some cases when investigators specifically need the patient identifiers to link the dataset to other data sources for research. This level of data should require the highest level of scrutiny, including a data use agreement, justification that the benefits of research outweigh the risks, review by a privacy board, and perhaps ongoing scrutiny for the duration that the data user possesses the data.
    - Data at the individual patient level could alternatively be provided to researchers in a physical space, which allows for direct control and monitoring of data use in cases where those data are highly sensitive.
  - Another model which is able to allow access to individual patient data, data linkage across data providers, while protecting individual privacy, is the distributed network model. This could help to overcome the difficulties that can arise when there are multiple data custodians.
    - o In this model, a consortium of data providers mutually agrees to share data and work to develop a common data framework. Data is coded uniformly across the consortium (e.g. date of birth would be coded: "MMDD-YYYY"). Each data provider stores their own data behind a firewall protected server. Data users may write standardized code which is sent to each data provider, analysed on site, within each data provider's server (protecting patient privacy) and the aggregate results are sent back to the data user.
- **Approval panels / ethical review.** Ethical review boards (also called institutional review boards) which grant access to health care data must be assured that the interest to society of the research project significantly outweighs the risk of violation of personal integrity of the individual that the processing may involve. A 'consent or anonymise' approach is too polarised and not a proportionate system. This risk of re-identification can be minimised with Ideal Framework 111

requirements for security procedures, training of staff that will process the data, and carefully written confidentiality agreements which assure correct use and reporting of data and which carry with it sanctions for inappropriate use. Approval panels should be composed of representatives with a broad range of relevant expertise and standpoints. The criteria used by committees to grant access to data should be clear, consistent, and transparent.

- The onus should be on data custodians to communicate how information is being shared and with whom in order to ensure public trust and transparency.

- **Data use agreements and confidentiality requirements.** Permission for data access should be granted with contractual requirements around the protection of confidentiality. The agreement should clearly define the scope and define duration of use.

- **Affiliation of the data user.** The type of organisation requesting access to data may influence the potential risk associated with its distribution (both realised and perceived). However, whilst the organisation's remit may influence their motivation for requesting access, this should not be the only consideration by data providers. Where the appropriate safeguards are in place, authorisation should be based on careful consideration of the motivation for and outputs of the research facilitated, rather than on the basis of the organisation's status. This is particularly important where manufacturers are tasked by HTA agencies or regulators with assessing the evidence for their products in routine practice.

- **Access costs.** Arrangements for the cost of data access will vary according to the nature of the data controller. For many datasets collected and held on a national basis, data charges are based only to recover the costs of data extraction and cleaning. Cost of access should be fair and not excessive, but in recognition of the need for the sustainability of the system.

About us
Founded in 1962 by the Association of the British Pharmaceutical Society, the Office of Health Economics (OHE) is not only the world's oldest health economics research group, but also one of the most prestigious and influential.

OHE provides market-leading insights and in-depth analyses into health economics & health policy. Our pioneering work informs health care and pharmaceutical decision-making across the globe, enabling clients to think differently and to find alternative solutions to the industry's most complex problems.

Our mission is to guide and inform the healthcare industry through today's era of unprecedented change and evolution. We are dedicated to helping policy makers and the pharmaceutical industry make better decisions that ultimately benefit patients, the industry and society as a whole.

OHE. For better healthcare decisions.

Areas of expertise

- Evaluation of health care policy
- The economics of health care systems
- Health technology assessment (HTA) methodology and approaches
- HTA's impact on decision making, health care spending and the delivery of care
- Pricing and reimbursement for biologics and pharmaceuticals, including value-based pricing, risk sharing and biosimilars market competition
- The costs of treating, or failing to treat, specific diseases and conditions
- Drivers of, and incentives for, the uptake of pharmaceuticals and prescription medicines
- Competition and incentives for improving the quality and efficiency of health care
- Incentives, disincentives, regulation and the costs of R&D for pharmaceuticals and innovation in medicine
- Capturing preferences using patient-reported outcomes measures (PROMs) and time trade-off (TTO) methodology
- Roles of the private and charity sectors in health care and research
- Health and health care statistics

CONSULTING | REPORT
SEPTEMBER 2019

ohe.org

The Office of Health Economics A Company Limited by Guarantee of Registered No. 09848965
OHE Consulting Ltd Registered Company No. 09853113
OHE is a Charity Registration No. 1170829
Registration Office Southside, 105 Victoria Street, London SW1E 6QT