Annals of Clinical Epidemiology

STUDY PROTOCOL

A new electronic medical record database linked to claims data and discharge abstract data (the RWD database) in Japan: Study design and profile

Yasuyuki Okumura[1,2], Takashi Fujiwara[3], Hironobu Tokumasu[4], Takeshi Kimura[1], Shiro Hinotsu[5,6]

ABSTRACT

BACKGROUND

This article aims to introduce the Real World Database—a new clinical database in Japan.

METHODS

The Health, Clinic, and Education Information Evaluation Institute and Real World Data Co., Ltd. began developing the Real World Database in 2015. This is an electronic medical record database linked to claims data and discharge abstract data from medical institutions in Japan. The institutions agreed to collect data from 218 medical institutions as of June 2021.

RESULTS

In 2019, 82 medical institutions provided data, which showed that 2,184,666 patients received treatment at medical institutions. There were also 334,437 inpatients with at least one hospital stay and 2,011,628 outpatients with at least one visit. More than 200 laboratory test results were available.

DISCUSSION

This database is a potential data source for producing descriptive studies, comparative effectiveness studies, studies of adverse effects, and prediction studies.

CONCLUSIONS

The Real World Database provides an opportunity and strategy to produce real-world evidence for Japan.

KEY WORDS

electronic medical records, Japan, database management systems, data collection

[1] Real World Data Co., Ltd.
[2] Initiative for Clinical Epidemiological Research
[3] Department of Otolaryngology, Kurashiki Central Hospital
[4] Department of Management, Clinical Research Center, Kurashiki Central Hospital
[5] Health, Clinic, and Education Information Evaluation Institute (HCEI)
[6] Biostatistics and Data Management, Sapporo Medical University

Corresponding author: Takeshi Kimura
Real World Data Co., Ltd., Shiseido Kyoto Bld.4F, 480, Aburanokojidori, Kizuyabashi-sagaru, Kitafudondocho, Shimogyo-ku, Kyoto, Kyoto, Japan
E-mail: kimura@rwdata.co.jp

Received: March 30, 2024
Accepted: April 3, 2024
J-STAGE Advance published date: April 11, 2024

No. 24009

INTRODUCTION

Numerous healthcare stakeholders are becoming increasingly interested in using real-world data (RWD) to produce real-world evidence [1–3]. In 2019, the US Food and Drug Administration defined RWD as "data relating to patient health status and/or the delivery of healthcare that are routinely collected from a variety of sources" [1]. Major sources of RWD include administrative data, electronic medical records (EMRs), and clinical registries. The US Food and Drug Administration also defined real-world evidence as "the clinical evidence regarding the usage and potential benefits or risks of a medical product derived from analysis of real-world data" [1]. Real-word evidence is perceived to be a potential cost-saver and to have greater generalizability than evidence generated from traditional clinical research [4].

In Japan, the most widely used source of RWD is administrative data, including claims data and discharge abstract data (called "the Diagnosis Procedure Combination [DPC] data") [5–7]. Administrative data include clinical and procedural information, such as diagnosis, drug, and medical practice codes. Nevertheless, there are significant limitations to administrative data, including inaccuracies in diagnostic information and the absence of records for laboratory test results. Furthermore, drug codes are not recorded in the claims reimbursed under certain bundled payment plans.

Data derived from EMRs can overcome these limitations, and several attempts have been made to create EMR databases. For example, the National Hospital Organization created an EMR database derived from data from 66 national hospitals in 2016 [8]. The Pharmaceuticals and Medical Devices Agency created an EMR database linked to administrative data derived from 23 hospitals in 2009 [9, 10]. This study introduces a new EMR database linked to administrative data derived from 90 medical institutions in Japan (the RWD database). The most prominent feature of the RWD database is that any medical institution with an EMR system can participate in the RWD database. This feature is based on the novel technology called "multi-language system" that can retrieve clinical and procedural information from several types of EMR systems.

METHODS

OVERVIEW

In 2015, the Health, Clinic, and Education Information Evaluation Institute and Real World Data Co., Ltd. (HCEI-RWD) developed the RWD database. This EMR database is linked to claims data and DPC data from medical institutions in Japan. The primary purpose of establishing the RWD database was to assess the healthcare quality of each participating medical institution. The secondary purpose was to provide a subset of the RWD database to stakeholders for research purposes. This study protocol (No: RI2020026) has been reviewed and approved by an independent ethics committee at the Research Institute of Healthcare Data Science.

PATIENT AND PUBLIC INVOLVEMENT

Patients and the public were not involved in the development of the RWD database.

ELIGIBILITY CRITERIA AND SETTING

Any medical institution with an EMR system in Japan could participate in our project to establish an RWD database. In Japan, there are approximately 8,000 and 100,000 hospitals and clinics, respectively, and EMR systems are available in 44% and 42% of the hospitals and clinics, respectively [11, 12]. The data recording format of EMR systems varies by vendor, whereas that of claims information is nationally standardized. The data recording format of discharge abstract data is also nationally standardized in DPC hospitals, which comprise 36% of all hospitals with acute care wards [13].

DATA COLLECTION PROCESS

Anonymized clinical and procedural information was obtained from three data sources: EMRs, claims data, and discharge abstract data, without restricting the follow-up period. The HCEI-RWD agreed to collect data from each medical institution and obtained approval from the Personal Information Protection Commission of each medical institution. Furthermore, each medical institution was asked to use the multi-language system to extract and anonymize data from the three types of data sources and upload these data through a secure web system. All patients who declined to participate in the study were excluded from the data collection process. Anonymous identification numbers were generated using a cryptographic hash function.

DATA CONVERSION PROCESS

Data derived from the EMRs were reviewed and converted to standardized forms using in-house standardized operating procedures.

Diagnosis information

Where a disease name was recorded but the diagnosis code was unknown (called "uncoded disease name") in the EMRs, an appropriate diagnosis code was assigned where possible. For example, when "renal failure, hypertension" was recorded as an uncoded disease name, a diagnosis code for hypertensive renal failure (diagnosis code for claims data: 88334270) was added. Another example is when "angina (after intracoronary stenting)" was recorded as an uncoded disease, the diagnosis code for the presence of coronary angioplasty implants and grafts (diagnosis code for claims data: 8844391) was added.

Drug information

Where a drug code for claims data was not recorded in the EMRs, other drug information was used to standardize the data in the following order: YJ code, code for drug price list, and drug name. In addition, records of the quantity and unit of a drug varied by medical institution, such as a record of quantity and unit for aspirin 100 mg tablet could be "one tablet" and "100 mg." In such cases, the quantities and units were converted into "one tablet."

Laboratory testing

The record of the test name, unit, sample, and results for a laboratory test varied by medical institution (**eTable 1**). For example, a record of the microliter unit abbreviation for white blood cells could be "μL," "μl," "uL," and "ul." In such cases, the unit abbreviations were converted to "μL." Each laboratory test was mapped to the original master based on the Japan Laboratory Analysis Code version 10 Master [14].

RELATIONAL DATA TABLES

Eight relational data tables were created based on information from the EMRs, claims data, and discharge abstract data (**Table 1**). The patient profile file included information about the patients and the medical institutions where they received care. The admission file contained admission and discharge dates, whereas the drug file included information on prescribed drugs. When available, drug file also includes information about bring-in drugs (i.e., drugs that patients bring). The laboratory testing file includes information about all types of specimen tests (e.g., blood, urine) although physiological function tests (e.g., electroencephalogram, electrocardiogram) and imaging tests (e.g., computerized tomography, magnetic resonance imaging) are not recorded in current stage. There are two files for diagnosis information: one was derived from EMRs and the other from claims data. The procedure file contained all medical practices under a uniform national fee schedule. The discharge abstract file included information about diagnosis, surgery, and clinical conditions (e.g., activity of daily living and coma at admission) in some acute care hospitals (i.e., "DPC" hospitals). All tables contained a unique patient identifier and, therefore, they could be linked together.

Table 1 Relational data tables based on information from electronic medical records, claims data, and discharge abstract data

File name	Data source	Item
1. Patient profile	EMRs	patient identifier, birth date (year-month format), sex (male/female), death status (no/yes), death date (if applicable), medical institution identifier, number of beds (0–19/20–99/100–299/300–499/≥500), region of medical institution (Hokkaido/Tohoku/Kanto/Chubu/Kinki/Chugoku/Shikoku/Kyushu)
2. Admission	EMRs	patient identifier, admission date, discharge date
3. Drug	EMRs	patient identifier, drug code for claims data, drug code for drug price list, drug name, daily dose, unit (e.g., tablet/mg), department name, administration date, end date (i.e., administration date plus days of drug supply)
4. Laboratory testing	EMRs	patient identifier, test name, test date, sample (e.g. blood, urine), result, unit
5. Diagnosis-EMRs	EMRs	patient identifier, disease name, ICD-10 code, diagnosis code for claims data, start date (i.e., the date when a diagnosis is recorded), end date (i.e., the date when a diagnosis is removed), primary diagnosis (no/yes), suspected diagnosis (no/yes), department name (i.e., the department where a diagnosis is recorded)
6. Diagnosis-claims	Claims data	patient identifier, disease name, ICD-10 code, diagnosis code for claims data, year-month for treatment, start date (i.e., the date when a diagnosis is recorded), primary diagnosis (no/yes), suspected diagnosis (no/yes), outcome (none/recovered/remitted/transferred to other institution/transferred to other department/death)
7. Procedure	Claims data	patient identifier, administration date, medical practice code for claims data, medical practice name
8. Discharge abstract	DPC data	patient identifier, diagnosis, surgery, clinical conditions

DPC, Diagnosis Procedure Combination; EMRs, electronic medical records; ICD, International Classification of Diseases.

STATISTICAL ANALYSES

The data used in this study were obtained on July 15, 2021. The number of patients per quarter and data sources between 2000 and 2020 were calculated. Patients who received treatment in 2019 were selected, and their data was used to describe the characteristics of the medical institutions and patients. The following numbers were calculated from data of the identified patients: newly hospitalized cases (including multiple admission episodes per patient), newly discharged cases, (unique) inpatients with at least one hospital stay, (unique) outpatients with at least one visit, and number of in-hospital mortalities.

The top 30 major laboratory tests were identified for both inpatients and outpatients. The number of inpatients and outpatients in the major diagnosis category was identified based on the International Classification of Diseases (ICD)-10. Diagnostic information was selected using the following criteria: (1) definitive (non-suspected) diagnoses derived from EMRs and (2) diagnoses recorded in 2019. All diagnoses were selected when multiple diagnoses were recorded.

The frequency, proportion, and mean of the data were calculated. No statistical significance testing has been carried out. The R version 4.0.2 (R Foundation for Statistical Computing) software program was used for all the analyses.

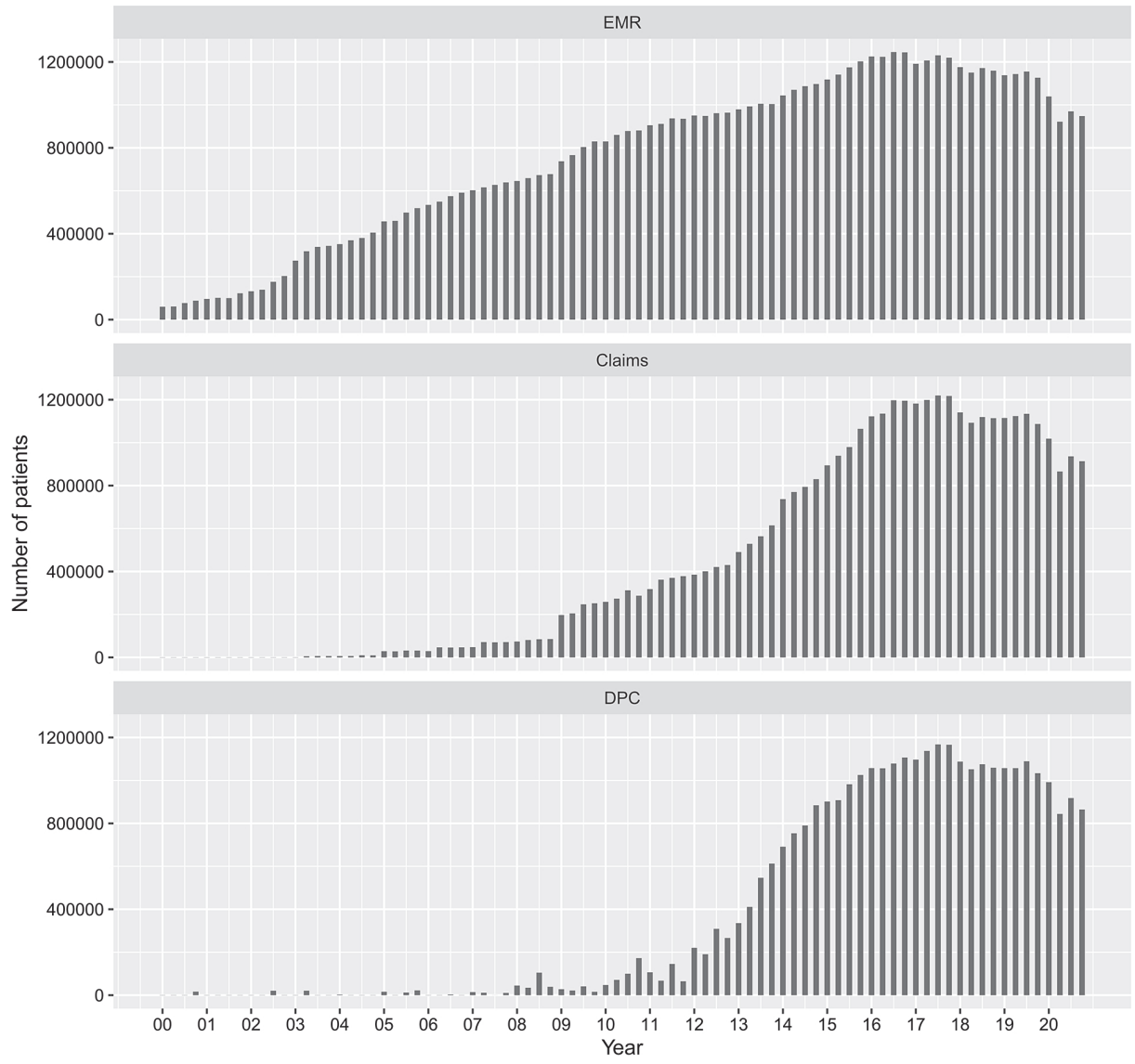

Fig. 1 **Number of patients included in the RWD database**

DPC, Diagnosis Procedure Combination; EMR, electronic medical record; RWD, Real World Database.

RESULTS

OVERVIEW

The HCEI-RWD agreed to collect data from 218 medical institutions for the period up to June 2021, and 90 of these institutions provided HCEI-RWD data for at least one quarter between 2000 and 2020. The number of patients increased annually, with the exception of 2020, owing to the influence of the coronavirus 2019 pandemic. Claims data and DPC data were not recorded during the early phases of data collection (**Fig. 1**). After 2017, most medical institutions with EMRs had claims data.

CHARACTERISTICS OF MEDICAL INSTITUTIONS

The characteristics of medical institutions distributed across all regions of Japan in 2019 are shown in **Table 2**. There were 82 medical institutions in the RWD database, consisting of eight clinics and 74 hospitals, whereas 69 of all the medical institutions were DPC hospitals.

Table 2 Number of medical institutions included in the Real World Database in 2019

Characteristics	Number of medical institutions
Total	82
Region	
Hokkaido	4
Tohoku	5
Kanto	14
Chubu	10
Kinki	34
Chugoku	4
Shikoku	2
Kyushu	9
Number of beds	
0–19	8
20–99	8
100–299	29
300–499	23
≥500	14
DPC hospital	69

DPC, Diagnosis Procedure Combination.

PATIENT CHARACTERISTICS

The characteristics of 2,184,666 patients who received treatment from the participating medical institutions in 2019 are shown in **Table 3**. There were 410,112 and 424,727 newly hospitalized and discharged patients, respectively, whereas 334,437 inpatients and 2,011,628 outpatients had at least one hospital stay or visit, respectively. The laboratory testing analysis revealed that 238 results were recorded in 2019, and **eTable 2** shows the number of inpatients and outpatients with laboratory test results. The most prevalent laboratory test for inpatients was hemoglobin level (77.8%), followed by hematocrit level (77.7%) and erythrocyte count (77.6%). The most prevalent laboratory test for outpatients was erythrocyte count (52.0%), followed by hemoglobin (52.0%) and hematocrit (51.5%) levels.

eTable 3 shows the number of inpatients and outpatients according to diagnosis category. The most prevalent diagnoses for inpatients were digestive system diseases (32.7%); circulatory system diseases (29.1%); and endocrine, nutritional, and metabolic diseases (25.6%). The most prevalent diagnosis for outpatients was diseases of the respiratory system (14.0%), followed by diseases of the digestive system (12.7%) and symptoms, signs, and abnormal clinical and laboratory findings not elsewhere classified (11.6%).

DISCUSSION

In this study, we introduced the RWD database, an EMR database linked to administrative data derived from 82 medical institutions. The RWD database is a potential data source for producing descriptive studies, comparative effectiveness studies, studies of adverse effects, and prediction studies in Japan. Studies using the RWD

Table 3 Characteristics of patients who received any treatments in 2019

Characteristics	n
Patients who received any treatment	2,184,666
Newly hospitalized cases	410,112
Newly discharged cases	424,727
Inpatients with at least one hospital stay	334,437
Outpatients with at least one visit	2,011,628
Number of in-hospital mortality	20,448
Number of out-hospital mortality	3,020

database have already been published in the fields of infectious diseases [15], cancer [16], endocrine diseases [17], cardiovascular diseases [18], rheumatoid arthritis [19], hemophilia [20], idiopathic pulmonary fibrosis [21], and renal disease [22].

The RWD has several advantages. First, the availability of laboratory test results enables researchers to use these data as outcome variables, such as blood glucose levels for diabetes care. Second, the two files for diagnosis information will assist researchers in conducting validation studies using claim-based diagnoses as index tests and EMR-based diagnosis as reference standards. Third, drug information derived from EMRs would allow researchers to investigate real-world prescription practices even in certain bundled payment plans, such as hospital fees for long-term care wards.

There are also some limitations to the RWD database that are worth mentioning. First, some medical institutions with EMRs do not have claims or DPC data, especially during the early phase of data collection. This is because medical institutions have no obligation to store administrative data for long-term. This limitation raises concerns about monitoring long-term patient follow-up. Second, no information was available for patients who received treatment outside the participating medical institutions because this was an institution-based study. Third, physiological function and imaging tests are not recorded in current stage of the project. Fourth, the severity of most disease conditions could not be identified using laboratory test results alone. Fifth, the participating medical institutions were mainly DPC hospitals, which limited the representativeness of the data for all medical institutions.

CONCLUSION

The established RWD database provides an opportunity and strategy to produce real-world clinical evidence for Japan.

CONFLICTS OF INTEREST

YO is a current employee, TF was an employee, HT is the former president, and TK is the current president of Real World Data Co., Ltd. This study was funded by the Real World Data Co., Ltd.

ACKNOWLEDGMENTS

We thank Editage (www.editage.jp) for the English language editing.

AUTHOR CONTRIBUTIONS

YO, TF, HT conceptualised the study. YO performed statistical analyses. YO and TF wrote the first draft of the paper. HT, TT, SH contributed to interpreting the data and to the writing and revising of the manuscript.

DATA AVAILABILITY STATEMENT

Data cannot be publicly shared due to privacy and ethical reasons, as stipulated in contracts with participating medical institutions.

REFERENCES

1. US Food and Drug Administration. *Submitting documents using real-world data and real-world evidence to FDA for drugs and biologics: guidance for industry*. https://www.fda.gov/media/124795/download. Accessed 2024 April 1.
2. Ministry of Health, Labour and Welfare. *Revised GPSP Ordinance (in Japanese)*. https://www.pmda.go.jp/files/000220766.pdf. Accessed 2024 April 1.
3. Corrigan-Curay J, Sacks L,Woodcock J. Real-world evidence and real-world data for evaluating drug safety and effectiveness. *JAMA* 2018;320:867–8.
4. Jarow JP, LaVange L,Woodcock J. Multidimensional evidence generation and FDA regulatory decision making: defining and using "Real-World" data. *JAMA* 2017;318:703–4.
5. Suto M, Iba A, Sugiyama T, Kodama T, Takegami M, Taguchi R, et al. Literature Review of Studies Using the National Database of the Health Insurance Claims of Japan (NDB): Limitations and Strategies in Using the NDB for Research. *JMA J* 2024;7:10–20.
6. Yasunaga H, Matsui H, Horiguchi H, Fushimi K,Matsuda S. Clinical epidemiology and health services research using the Diagnosis Procedure Combination Database in Japan. *Asian Pac J Dis Manag* 2015;7:19–24.
7. Hayashida K, Murakami G, Matsuda S,Fushimi K. History and profile of Diagnosis Procedure Combination (DPC): development of a real data collection system for acute inpatient care in Japan. *J Epidemiol* 2021;31:1–11.
8. National Hospital Organization. *NHO Clinical Data Archives (in Japanese)*. https://nho.hosp.go.jp/files/000145085.pdf. Accessed 2024 April 1.
9. Yamaguchi M, Inomata S, Harada S, Matsuzaki Y, Kawaguchi M, Ujibe M, et al. Establishment of the MID-NET((R)) medical information database network as a reliable and valuable database for drug safety assessments in Japan. *Pharmacoepidemiol Drug Saf* 2019;28:1395–404.
10. Yamada K, Itoh M, Fujimura Y, Kimura M, Murata K, Nakashima N, et al. The utilization and challenges of Japan's MID-NET® medical information database network in postmarketing drug safety assessments: a summary of pilot pharmacoepidemiological studies. *Pharmacoepidemiol Drug Saf* 2019; 28:601–8.
11. Ministry of Health, Labour and Welfare. *Survey of medical institutions in 2019 (in Japanese)*. https://www.mhlw.go.jp/toukei/saikin/hw/iryosd/16/dl/02_01.pdf. Accessed 2024 April 1.
12. Ministry of Health, Labour and Welfare. *Survey of medical institutions in 2017 (in Japanese)*. https://www.mhlw.go.jp/toukei/saikin/hw/iryosd/16/dl/02_01.pdf. Accessed 2024 April 1.
13. Ministry of Health, Labour and Welfare. *Hospital ward functioning report in fiscal 2018 (in Japanese)*. https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/open_data_00005.html. Accessed 2024 April 1.

14. Kimura M, Kanno T, Tani S, Satomura Y. Standardizations of clinical laboratory examinations in Japan. *Int J Med Inform* 1998;48:239–46.

15. Takeuchi M, Inokuchi S, Kimura T, Eguchi N, Kawakami K,Takahashi T. Descriptive epidemiology of COVID-19 in Japan 2020: insights from a multihospital database. *Annals of Clinical Epidemiology* 2023;5:5–12.

16. Fujiwara T, Kanemitsu T, Tajima K, Yuri A, Iwasaku M, Okumura Y, et al. Accuracy of algorithms to identify patients with a diagnosis of major cancers and cancer-related adverse events in an administrative database: a validation study in an acute care hospital in Japan. *BMJ Open* 2022;12:e055459.

17. Takeuchi M, Ogura M, Inagaki N, Kawakami K. Initiating SGLT2 inhibitor therapy to improve renal outcomes for persons with diabetes eligible for an intensified glucose-lowering regimen: hypothetical intervention using parametric g-formula modeling. *BMJ Open Diabetes Research & Care* 2022;10:e002636.

18. Fukasawa T, Seki T, Nakashima M,Kawakami K. Comparative effectiveness and safety of edoxaban, rivaroxaban, and apixaban in patients with venous thromboembolism: a cohort study. *J Thromb Haemost* 2022x;20:2083–97.

19. Yokoyama S, Ishii Y,Masuda J. Persistence and Safety of Golimumab in Elderly Patients with Rheumatoid Arthritis and Renal Dysfunction in a Real-World Setting. *Drugs - Real World Outcomes* 2022;10:51–60.

20. Fujiwara T, Miyakoshi C, Kanemitsu T, Okumura Y,Tokumasu H. Identification and validation of hemophilia-related outcomes on Japanese electronic medical record database (Hemophilia-REAL V Study). *J Blood Med* 2021;12:571–80.

21. Anan K, Kataoka Y, Ichikado K, Kawamura K, Johkoh T, Fujimoto K, et al. Early corticosteroid dose tapering in patients with acute exacerbation of idiopathic pulmonary fibrosis. *Respir Res* 2022;23:291.

22. Ide K, Fujiwara T, Shimada N, Tokumasu H. Influence of acetaminophen on renal function: a longitudinal descriptive study using a real-world database. *Int Urol Nephrol* 2021;53:129–35.