Millennial Medical Record Data Profile: A Japanese Electronic Medical Records Database Utilising Unstructured Data for Lung Cancer Research

RWD125

Ayumi Hamaguchi,[1] Kenichiro Shiba,[2] Mari Kato,[3] Daisuke Wakasugi,[4] Tadashi Koga,[3] Suguru Nozue,[4] Takayuki Sawada,[3] Tomoyo Morita,[4] Amanda Pulfer,[1] Dimitra Lambrelli[1]

[1]Thermo Fisher Scientific, London, UK; [2]General Incorporated Association Life Data Initiative, Kyoto, Japan; [3]Clinical Study Support, Inc., Nagoya, Japan; [4]NTT DATA Japan Corporation, Tokyo, Japan

Background

- Real-world data sources, such as health insurance claims and Diagnosis Procedure Combination (DPC) databases, are commonly used in Japanese research because of their broad coverage and suitability for longitudinal analyses,[1,2] but these sources often lack detailed clinical information.
- Electronic medical records (EMRs) offer rich clinical detail, including laboratory values and indicators of disease severity and progression,[3] but access to EMR-based data—particularly for commercial research—remains limited.
- The Millennial Medical Record (MMR) database includes data from EMRs, claims, and DPC sources. It has the potential to be leveraged for real-world research, including lung cancer and other disease areas, using its comprehensive clinical and longitudinal data.
- To describe the characteristics of the MMR database, lung cancer was selected due to its clinical relevance and its status in Japan as a leading cause of death,[4] with its rising age-adjusted incidence rate since 2019.[5]

Objectives

- **The goal of this study was to describe the characteristics of MMR data, illustrating its utility for lung cancer research.**

Methods

MMR Overview

- The MMR database includes structured and unstructured data from EMRs (**Table 1**) that complies with the Medical Markup Language 4.2.0 standard,[6] claims data, DPC data, and imaging data.
- Data are collected under the framework of the Next Generation Medical Infrastructure Act,[7] which allows only certified entities to manage and anonymise medical information for research use. Processing staff complete mandatory training in anonymisation, data security, and compliance (**Figure 1**). The MMR database is operated by the Life Data Initiative and NTT DATA Japan Corporation.
- As MMR's coverage expands throughout Japan, the number of participating hospitals increased from 43 in 2020 to 67 in 2025 (**Figure 2**). Of these, the number contributing EMR, DPC, and claims data usable for research increased from 10 to 26 over the same period.
- **Figure 3** provides an overview of data availability in the MMR database.

Table 1. MMR Overview

Participating hospitals	**67** secondary/tertiary care hospitals with an average of 600 beds, mostly university and regional core hospitals under secondary use agreements with LDI, as of Aug 2025
Patient coverage	1.9 million patients (approx. 1.5% of Japanese population) have complete EMR + DPC + claims data, as of Feb 2025
Data collection	Started in 2016 and is ongoing[a]
Update frequency	Quarterly updates[b]
Data access	Review required by the LDI 'Purpose of Data Use Review Board'

Abbreviations: DPC = Diagnosis Procedure Combination; EMR = electronic medical record; LDI = Life Data Initiative; MMR = Millennial Medical Record
[a]The latest data available as of September 2025 is from March 2025. [b]Data updated via automatic transfer or secure file upload from participating hospitals.

Figure 1. Data Processing and Anonymisation

Patient consent
- Opt-out procedures managed by contributing hospitals

Data anonymisation and linkage[a]
- Removal of direct identifiers
- Creation of pseudonymised patient IDs for cross-source linkage between EMR, DPC, and claims
- Re-identification risk assessed; low-frequency or potentially identifying values suppressed if needed

Three different options for data outputs
1. Anonymised patient-level datasets
2. Aggregated statistical datasets
3. Pseudonymised datasets available to a limited group of users, subject to additional regulatory requirements

Abbreviations: DPC = Diagnosis Procedure Combination; EMR = electronic medical record; ID = identification
[a]Data anonymisation must be performed in a secure workspace (e.g., fingerprint access, camera monitoring, no personal devices).

Figure 2. Hospitals Contributing to MMR, by Region

Total	**67**
Hokkaido	4
Tohoku	0
Kanto	12
Hokuriku	6
Tokai	3
Kinki (Kansai)	17
Chugoku	1
Shikoku	4
Kyushu	20
Okinawa	0

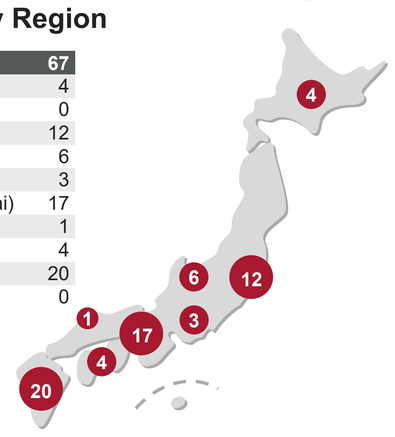


Figure 3. MMR Data Overview

EMR data

Structured data
- Demographics (sex, date of birth, nationality, insurance information, etc.)
- Diagnosis history (clinical outcomes/comorbidities using ICD-10 codes, disease onset and end date, outcome, etc.)
- Prescriptions and injections (drug code, daily dosage, route, start date, end date, etc.)
- Laboratory tests (specimen type, JLAC10 code, test result, reference range, flag for abnormal value, etc.)
- Vital signs (temperature, blood pressure, pulse, oxygen saturation)

Unstructured data
- Clinical notes (progress notes, physician comments)
- Discharge summaries (admission course, findings, outcomes)
- Radiology reports
- Referral letters and consultation notes

DPC data
- Admission and discharge information
- DPC group information

Claims data
- Demographics
- Insurance information
- Diagnoses
- Medical practice
- Prescriptions
- Hospital costs

Imaging data
- X-ray, CT, MRI

Abbreviations: DPC = Diagnosis Procedure Combination; EMR = electronic medical record; JLAC10 = Japan Laboratory Code, version 10; MMR = Millennial Medical Record

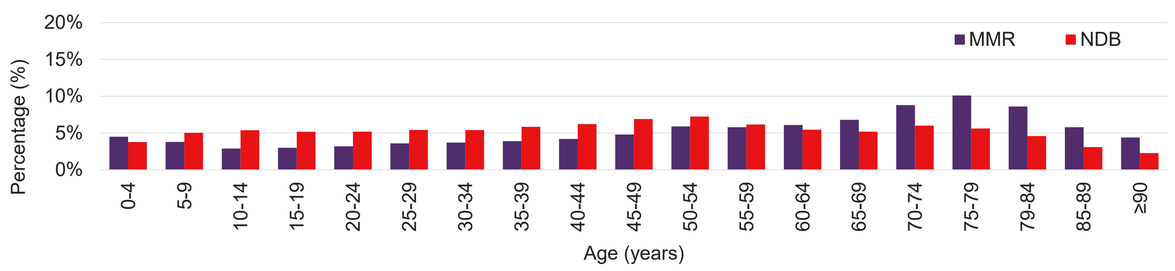
Statistical Analysis

- To describe MMR data characteristics, age distribution as of 2024 was calculated and compared with the National Database of Health Insurance Claims and Specific Health Checkups of Japan (NDB),[8] using patients with a first-visit claim between April 2023 and March 2024.
- Patients with lung cancer were identified using ICD-10 code a C34* as of February 2025, with counts stratified by sex and age.
- To demonstrate the value of EMR data, keywords were grouped into five domains: histology, genetic tests, disease scores, metastasis sites, and tumour marker tests. A rule-based keyword search was applied to assess data availability, regardless of confirmed presence. Keywords for histology, genetic tests, and disease scores were extracted from unstructured data using SQL scripts program supported by Generative AI. Analyses were descriptive.

Results

- As of 2024, the MMR database included 1.6 million individuals from 25 hospitals contributing EMR, DPC, and claims data across Japan, while the NDB database included 88 million individuals. Overall age distributions were similar, with all proportional differences under 5%, supporting the MMR database's representativeness. The largest gap was 4.49% in the 75–79 age group (**Figure 4**).

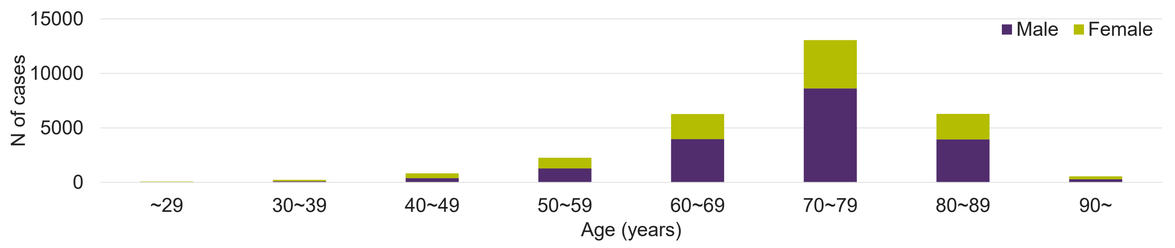
Figure 4. Age Distribution Among MMR and NDB



Abbreviations: MMR = Millennial Medical Record; NDB = National Database of Health Insurance Claims and Specific Health Checkups of Japan
Note: The NDB was used as a reference due to its comprehensive nationwide coverage.

- Among 29,545 patients with lung cancer, 63.1% were male (**Figure 5**). The largest age group was 70–79 years old (44.2%), consistent with national incidence data showing higher occurrence in adults aged 69 and older than in younger age groups.[9]

Figure 5. Age and Sex Distribution of Patients With Lung Cancer in MMR



Abbreviation: MMR = Millennial Medical Record

- Among 29,545 patients with lung cancer, keywords indicating tissue types—adenocarcinoma, small cell, and squamous cell carcinoma—were identified in 10%–40% of cases from unstructured data (**Table 2**). The keyword list was indicative; clinician validation would expand data identification.

Table 2. Availability of Clinical Data in Unstructured and Structured EMR

Variables	Keywords Lists/Description	Percentage (%) of Patients With Keywords/Codes
Keywords in unstructured EMR		
Histology		
Adenocarcinoma	(Adenocarcinoma or Ade)+(肺 or Lung)+without negative expressions[a]; OR (肺)+(腺癌 or 腺がん)+without negative expressions[a]	35~40%
Small cell carcinoma	(Small Cell Lung Cancer or SCLC)+without negative expressions[a]; OR (肺)+([^非]小細胞)+without negative expressions[a]; OR ([^非]小細胞)+(肺)+without negative expressions[a]	10~15%
Squamous cell carcinoma	(Squamous Cell Carcinoma or SCC)+(肺 or Lung); OR (SCCLC)+without negative expressions[a]; OR (肺)+(扁平上皮)+without negative expressions[a]; OR (扁平上皮)+(肺)+without negative expressions[a]	10~15%
Genetic test		
EGFR	(EGFR)[b]	20~25%
ALK	(ALK)[b]	15~20%
ROS1	(ROS1)[b]	5~10%
Disease score		
Performance status	(PS)+Any of the numbers 0–9; OR (Performance)+(Status)+any of the numbers 0–9	55~60%
Metastasis site		
Bone	(骨 or Bone)+(転移 or 病変 or Meta)+without negative expressions[a]	Keywords shown for reference purposes
Lymph node	(リンパ)+(転移 or 病変 or Meta)+without negative expressions[a]	
Brain	(脳 or Brain)+(転移 or 病変 or Meta)+without negative expressions[a]	
Data from structured EMR		
Tumour marker tests		
CEA	–	70~75%
SCC	–	35~40%
CYFRA21-1	–	15~20%

Abbreviations: ALK = anaplastic lymphoma kinase; CEA = carcinoembryonic antigen; CYFRA21-1 = cytokeratin 19 fragment; EGFR = epidermal growth factor receptor; PS = performance status; SCC = squamous cell carcinoma; SCCLC = squamous cell carcinoma lung cancer; SCLC = small-cell lung cancer. [a]Negative expressions indicating absence or uncertainty of a specific condition in Japanese include なし, ない, 無, 疑, ません, or 難し.
[b]Any alphabet before EGFR/ALK/ROS1 are excluded.

Conclusions

- **Age distributions were generally similar between the MMR and NDB databases, supporting its representativeness.**
- **By integrating complete EMR data—including histological type and biomarkers from unstructured records—with claims and DPC data, the MMR database enables research closely aligned with routine clinical practice, creating opportunities for real-world studies that complement clinical trials.**
- **The growth of the MMR database demonstrates the feasibility of harnessing EMR data for large-scale research in Japan. As coverage expands, the MMR will support innovative real-world evidence generation—including external control arms and longitudinal studies—ultimately accelerating advances in treatment and health policy.**

References

1. Sato S, Yasunaga H. *Ann Clin Epidemiol*. 2023;5(2):58-64. **2.** Fujinaga J, Fukuoka T. *Drugs - Real World Outcomes*. 2022;9(4):543-550. **3.** Zhao Y, Tsubota T. *JMIR Med Inform*. 2023;11:e39876. **4.** World Health Organization. Global Cancer Observatory - Japan. 2022.https://gco.iarc.who.int/media/globocan/factsheets/populations/392-japan-fact-sheet.pdf. **5.** Yamamoto H, et al. *Int J Clin Oncol*. 2025;30(2):199-209. **7.** Kobayashi S, Kume N. *J Med Syst*. 2020;44(4):69. **7.** Ministry of Health, Labour, and Welfare. Next-generation medical infrastructure law. 2023. https://www.mhlw.go.jp/content/10808000/001166476.pdf. In Japanese. **8.** Ministry of Health, Labour, and Welfare. The 10th NDB Open Data. https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000177221_00016.html. **9.** Foundation for Promotion of Cancer Research. Cancer Statistics in Japan. 2025. https://ganjoho.jp/public/qa_links/report/statistics/pdf/cancer_statistics_2025_fig_E.pdf.

Disclosures

AH, **AP**, and **DL** are employees of PPD™ Evidera™ Real-World Data & Scientific Solutions, Thermo Fisher Scientific. **KS** is an employee of the General Incorporated Association Life Data Initiative. **DW**, **SN**, and **TM** are employees of NTT DATA Japan Corporation. **MK**, **TS**, and **TK** are employees of Clinical Study Support, Inc. Funding was provided by Thermo Fisher Scientific.

Presented at ISPOR —The Professional Society for Health Economics and Outcomes Research—Europe | 9–12 November 2025 | Glasgow, Scotland, UK