Evaluating the Representativeness of a **Real-World Oncology Database** Developed by the **J-CONNECT** Consortium: Comparison With Japan's National Cancer Registry

RWD40

Masafumi Okada, MD, PhD[1], Shigemi Matsumoto, MD, PhD[2].

1. Prime Research Institute for Medical RWD, Inc., Kyoto, Japan,
2. Department of Real World Data R&D, Kyoto University, Kyoto, Japan.

Introduction

The J-CONNECT Consortium, a collaborative academic network in Japan, has developed a real-world pseudonymized database of chemotherapy-treated solid cancer patients using electronic medical records (EMRs) from 12 hospitals, with more institutions having already joined (Figure 1). The database includes core cancer registration data such as patient characteristics, pathological diagnosis, and clinical stage, as well as comprehensive prescription, injection, and laboratory test records extracted from EMRs. Gene mutation information is also extracted from pathological reports using natural language processing (Table 1). The database is constructed under an academic framework with an opt-out informed consent process. Researchers can access additional data for more detailed analyses through a federated model or by directly collaborating with individual hospitals. The representativeness of the registry was evaluated by comparing the number of cases by cancer type and patient demographic distributions with those from the national cancer registry.

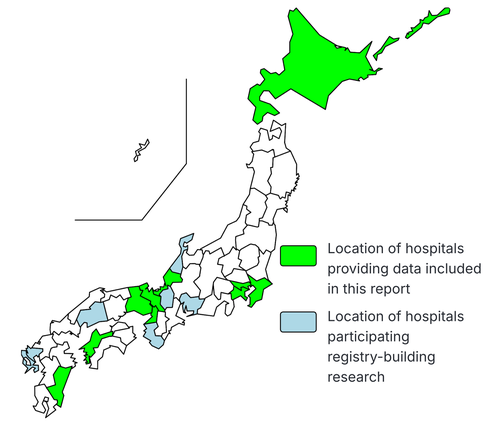


Figure 1: Location of hospitals participating J-CONNECT registry project.

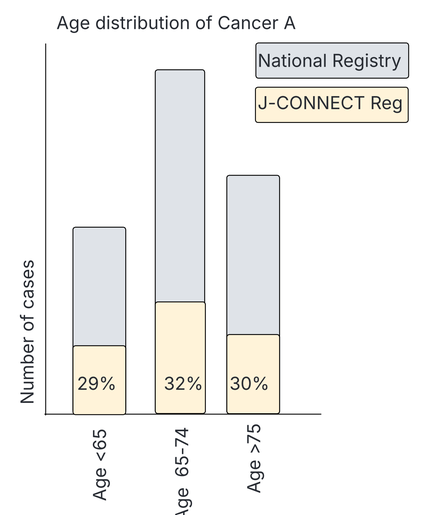
Table 1: Data collection items of J-CONNECT database

Category	Data Source	Data Item	Remarks
Patient Information	Hospital-based Cancer Registry	Age	Calculated from date of birth
	Hospital-based Cancer Registry	Sex	
Disease Information	Hospital-based Cancer Registry	Date of Diagnosis	
	Hospital-based Cancer Registry	Cancer type (detailed location/pathological diagnosis)	ICD-O-3
	Hospital-based Cancer Registry	Clinical stage	UICC TNM classification
Treatment Information	Hospital-based Cancer Registry	Date of initial chemotherapy	
	Electronic Medical Record	Drug data (prescription date / drug code / dosage / unit / duration)	YJ code / ATC code
	Electronic Medical Record	Laboratory test results (test date / test code / result)	
	Pathology Report	Biomarker test data (biomarker name / specimen collection date / testing method / result) incl. gene mutation	Extracted by natural language processing
Outcome	Hospital-based Cancer Registry	Survival and death information (date of death / last confirmation date of survival)	In-hospital death only

Table 2: Case coverage rate by cancer site

Cancer Site	J-CONNECT Count 2018-2023	National Count* 2018-2023	J-CONNECT Coverage Rate
Liver	1731	50547	0.03425
Esophagus	1960	58804	0.03333
Breast	12363	374907	0.03298
Pancreas	3720	113436	0.03279
Lung	6743	231478	0.02913
Prostate	5235	232128	0.02255
Stomach	2383	107623	0.02214
Colorectal	4012	200730	0.01999

* # of patients who received chemotherapy as the first treatment



Coefficient of variation (C.V.) of coverage rates (29%, 32%, 30%) = 0.05
Lower C.V. values indicate more consistent coverage across age groups and thus a more similar age distribution between the registries.

Figure 2: Schematic illustration of representativeness evaluation using subgroup-specific coverage rates.

Methods

In Japan, Designated Cancer Care Hospitals are required to maintain hospital-based cancer registries, making the registry data a reliable indicator of the total number of cancer cases nationwide. We compared the number of cases in the J-CONNECT registry to publicly available summary reports from the national cancer registry data. We evaluated the coverage stratified by age and gender for liver, esophageal, breast, pancreatic, lung, prostate, stomach, and colorectal cancers, and then calculated the coefficient of variation (CV) to assess the variability in coverage across strata (Figure 2). Lower CV values indicate more consistent coverage across strata, suggesting a more similar distribution between the registries.

Results

Cancer site-specific coverage ranged from 0.20% to 3.42%, CV = 0.208 (Table 2). Sex-stratified CV (excluding breast and prostate cancers) ranged from 0.003 (stomach) to 0.150 (biliary tract and gallbladder) (Table 3). Age-stratified CV ranged from 0.203 (liver) to 0.511 (stomach) (Table 4).

Table 3: Case coverage rate by sex

Cancer Site	Sex	National Count* 2018-2023	J-CONNECT Count 2018-2023	J-CONNECT Coverage Rate	Coefficient of Variation**
Stomach	Male	279072	965	0.003	0.0026
	Female	121309	421	0.003	-
Esophagus	Male	99800	869	0.009	0.0148
	Female	21256	189	0.009	-
Lung	Male	321425	3318	0.010	0.0389
	Female	158646	1550	0.010	-
Pancreas	Male	84846	969	0.011	0.0466
	Female	74920	801	0.011	-
Liver	Male	89728	848	0.009	0.0695
	Female	35483	370	0.010	-
Colorectal	Male	366207	1600	0.004	0.0823
	Female	252792	1241	0.005	-
Biliary tract	Male	43994	205	0.005	0.1495
(incl. Gallbladder)	Female	31843	120	0.004	-

* # of patients regardless of initial treatment, since there is no public report stratified by sex and treatment.

** Lower C.V. of coverage rate indicates more similar sex distribution

Table 4: Case coverage rate by age group

Cancer Site	Age Group (years)	National Count* 2018-2023	J-CONNECT Count 2018-2023	J-CONNECT Coverage Rate	Coefficient of Variation**
Liver	<65	21930	274	0.012	0.2032
	65-74	41385	408	0.010	-
	≥75	61896	519	0.008	-
Breast	<65	240654	5278	0.022	0.2033
	65-74	104673	1987	0.019	-
	≥75	87504	1267	0.014	-
Prostate	<65	45539	273	0.006	0.2880
	65-74	144518	1037	0.007	-
	≥75	146208	1515	0.010	-
Pancreas	<65	33266	485	0.015	0.3176
	65-74	53538	726	0.014	-
	≥75	72962	553	0.008	-
Esophagus	<65	28665	322	0.011	0.3414
	65-74	47725	486	0.010	-
	≥75	44667	245	0.005	-
Lung	<65	84173	1293	0.015	0.4055
	65-74	180774	2195	0.012	-
	≥75	215124	1363	0.006	-
Biliary tract	<65	9265	46	0.005	0.4060
(incl. Gallbladder)	65-74	21680	136	0.006	-
	≥75	44892	116	0.003	-
Colorectal	<65	164003	1127	0.007	0.4698
	65-74	207764	1120	0.005	-
	≥75	247232	587	0.002	-
Stomach	<65	68808	415	0.006	0.5110
	65-74	136800	600	0.004	-
	≥75	194773	365	0.002	-

* # of patients regardless of initial treatment, since there is no public report stratified by age and treatment.

* Lower C.V. of coverage rate indicates more similar age distribution

Conclusion

Although the database includes only chemotherapy-treated patients, sex distributions were largely consistent with national data. While variability existed in age and cancer site distributions, focusing on cancer types with less age-related variation suggests this real-world oncology registry may serve as a representative resource for Japanese solid cancer populations with rich background characteristics information including genotypes and complete history of prescriptions.

For details including availability of the registry data for commercial usage, please contact to the author.