

# Characterising Cancer Patients in Taiwan Using Real-World Data: Contributions to the OHDSI Network Study

Nguyen Thi Kim Hien<sup>1</sup>, Phan Thanh-Phuc<sup>2</sup>, Christianus Heru Setiawan<sup>3,4</sup>, Truong Thao Nguyen<sup>5</sup>, Irene López-Sánchez<sup>6,7</sup>, Talita Duarte-Salles<sup>8</sup>, Nguyen Phung-Anh<sup>9</sup>, Jason C. Hsu<sup>10,11,12,13\*</sup>

1. PhD Program in School of Nutrition and Health Sciences, College of Nutrition, Taipei Medical University, Taipei City, Taiwan.
2. Science and Training Department, University Medical Center Ho Chi Minh City, Ho Chi Minh City, Viet Nam.
3. Ph.D. Program, School of Pharmacy, Taipei Medical University, Taipei, Taiwan.
4. Faculty of Pharmacy, Sanata Dharma University, Yogyakarta, Indonesia.
5. PhD Program in Global Health and Health Security, College of Public Health, Taipei Medical University, Taipei City, Taiwan.
6. IDIAP Jordi Gol, Barcelona, Spain.
7. Programa de Doctorat en Metodologia de la Recerca Biomèdica i Salut Pública, Universitat Autònoma de Barcelona, Barcelona, Spain.
8. Erasmus MC University Medical Center, Rotterdam, The Netherlands.
9. Graduate Institute of Data Science, College of Management, Taipei Medical University, Taipei, Taiwan.
10. International Ph.D. Program in Biotech and Healthcare Management, College of Management, Taipei Medical University, Taipei, Taiwan.
11. Clinical Data Center, Office of Data Science, Taipei Medical University, Taipei, Taiwan.
12. Research Center of Data Science on Healthcare Industry, College of Management, Taipei Medical University, Taipei, Taiwan.
13. Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei Medical University, Taipei, Taiwan.

\* Corresponding author

**Background:** Cancer remains the leading cause of death in Taiwan, with the Taiwan Cancer Registry reporting over 121,000 new cases in 2021<sup>1</sup>. Timely monitoring of cancer occurrence and its demographic patterns, including age and sex, is essential for health policy and resource allocation<sup>2</sup>. Hospital electronic health record (EHR) databases harmonised to the OMOP Common Data Model (CDM) can complement cancer registries with longitudinal clinical detail and allow standardised comparison across the OHDSI network<sup>3</sup>. The Taipei Medical University Clinical Research Database (TMUCRD) is the Taiwanese data partner in OMOPCAN, a multinational OHDSI network study of the epidemiology and patient characteristics of 36 cancers<sup>4</sup>. This brief report presents a preliminary characterisation of cancer patients in TMUCRD as part of that collaboration.

## Methods

### *Data source*

TMUCRD integrates EHRs from three hospitals affiliated with Taipei Medical University (Taipei Medical University Hospital, Wan Fang Hospital and Shuang Ho Hospital). It was established in 2015, contained records for over 4.3 million patients by 2022, and has been mapped to the OMOP CDM. TMUCRD is a hospital-based (secondary-care) database, not a nationwide or population-based one: patients are observed only while they receive care at these hospitals, and the catchment population is not defined. The OMOPCAN protocol therefore classifies TMUCRD as a secondary-care source that contributes to characterisation and survival analyses, but not to population incidence or prevalence estimates<sup>4</sup>.

## Cohort definitions

Cancer cohorts used the OMOPCAN phenotypes, which were developed centrally for the network study. Condition concepts from SNOMED CT and ICD-O-3 were aligned with GLOBOCAN (ICD-10) cancer categories using the CodelistGenerator R package, implemented in OHDSI ATLAS, and reviewed by clinicians. In situ, benign, uncertain-behaviour and secondary (metastatic) codes were excluded<sup>4</sup>. Results are reported here as 34 cancer categories: colon and rectum are combined as colorectal cancer, and anal cancer is not reported. The index date was the first recorded qualifying diagnosis in TMUCRD between 2002 and 2022. Each cancer type was assessed separately, so a patient with more than one primary cancer could enter more than one cohort, and only the first diagnosis of each type was counted. Cases were identified from recorded diagnosis codes; pathology confirmation was not an inclusion criterion. Cohort definitions and study code are publicly available at <https://github.com/ohdsi-studies/OMOPCAN>.

## Analyses

We described the number of patients by cancer type, sex distribution and age at diagnosis. Rates by calendar year, sex and 10-year age group were obtained from the incidence-rate module of OHDSI CohortDiagnostics<sup>5</sup>. The numerator was the number of people with a first cohort entry in the stratum who had at least 365 days of prior observation in TMUCRD (washout). The denominator was the person-years observed in TMUCRD among people with at least 365 days of prior observation and no earlier occurrence of that cancer, ending at diagnosis or end of observation. Rates are expressed per 1,000 person-years and are crude (not age-standardised). Because the denominator is the hospital population rather than the general population, these are hospital-based rates and should not be read as population incidence. CohortDiagnostics was also used to review cohort construction, including included and orphan concepts, the index-event breakdown, and rates by calendar year, age and sex.

## Results

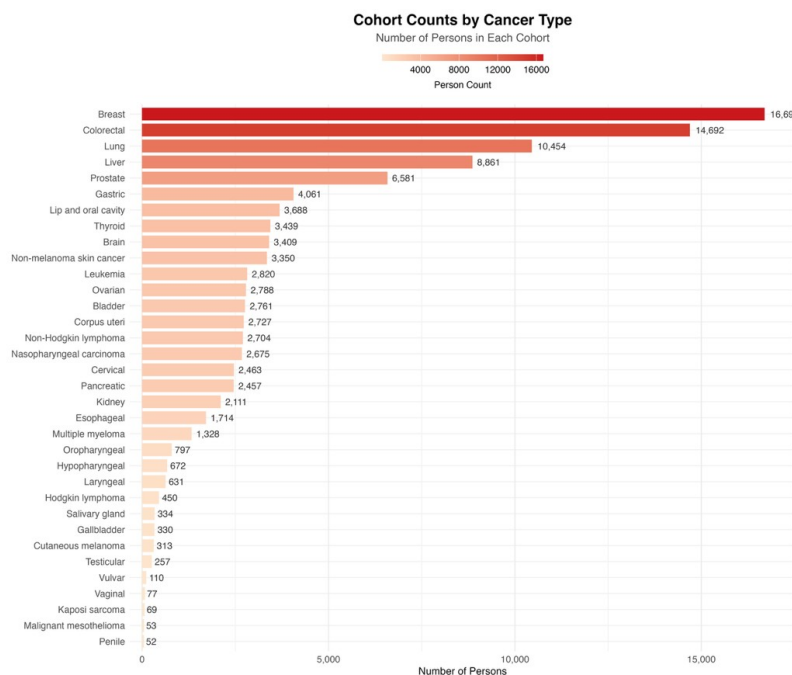


Figure 1. Number of cancer patients by cancer type in TMUCRD, 2002–2022

Between 2002 and 2022, nearly 106,000 patients with one of 34 cancer types were identified in TMUCRD. The most common were breast (16,698 patients), colorectal (14,692), lung (10,454), liver (8,861) and prostate cancer (6,581). Rare cancers were scarcely represented: penile (52), malignant mesothelioma (53), Kaposi sarcoma (69), vaginal (77) and vulvar cancer (110).

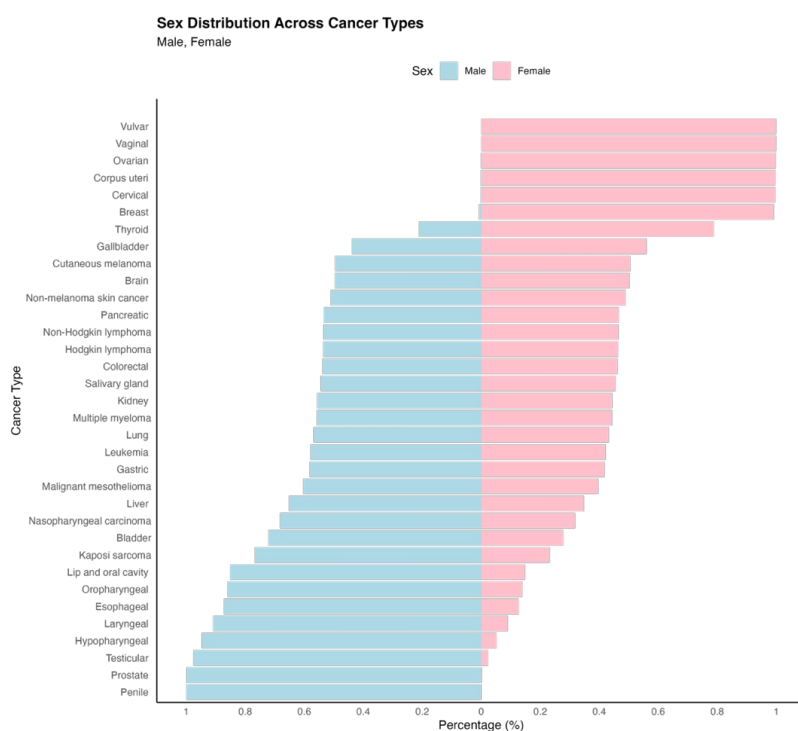


*Figure 2. Sex distribution of cancer patients by cancer type*

Sex distributions followed expected epidemiological patterns. Females predominated in breast, cervical, ovarian, corpus uteri, thyroid, vaginal and vulvar cancers, and males in lung, liver, colorectal, oral cavity, oesophageal, bladder, nasopharyngeal and prostate cancers. Leukaemia, lymphoma and multiple myeloma showed broadly balanced distributions with a slight male excess.

Rates of colorectal and lung cancer were highest in adults aged over 65 years, and prostate cancer rates rose steeply with age. Breast and cervical cancers occurred at younger ages, consistent with screening and reproductive risk profiles. Liver cancer spanned middle-aged and older groups, reflecting the combined impact of chronic viral hepatitis and metabolic risk factors.

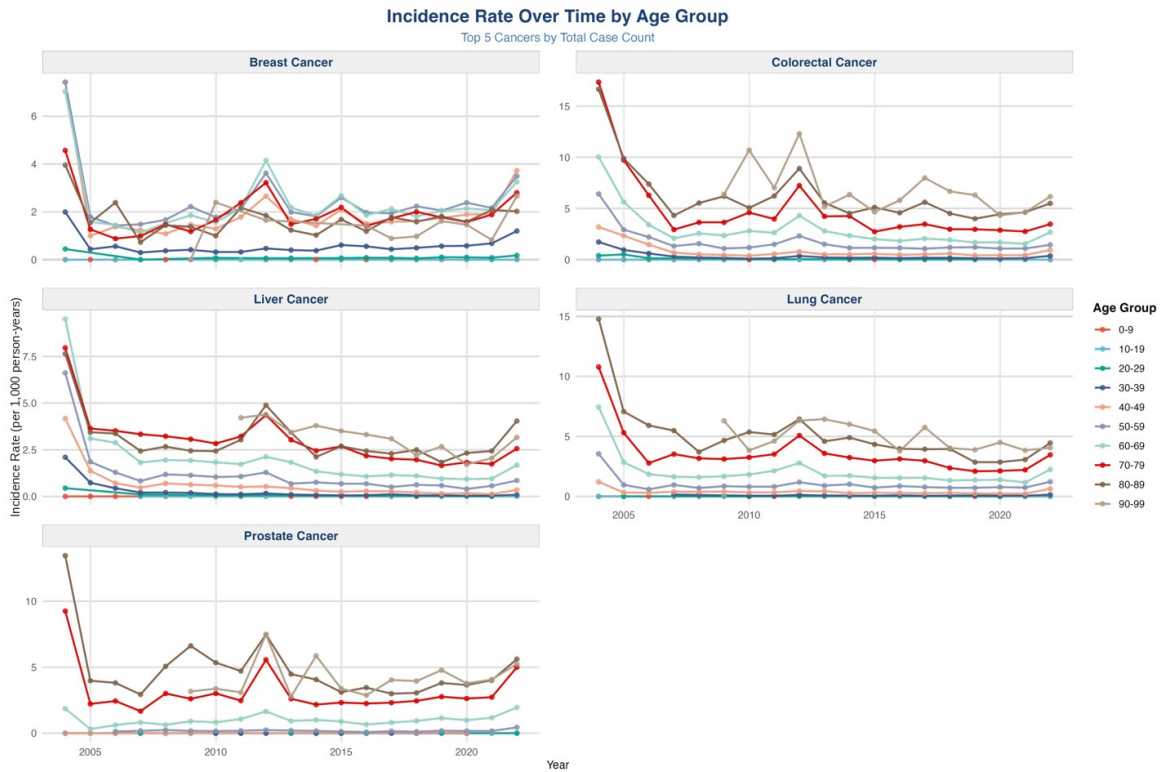
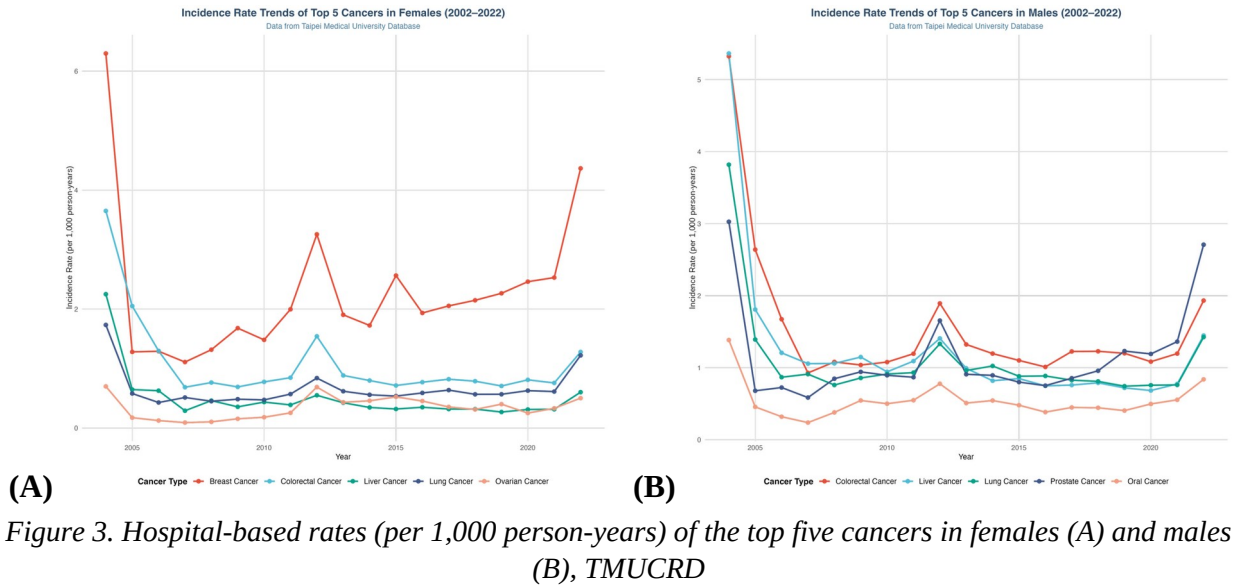


Figure 4. Age-specific hospital-based rates (per 1,000 person-years) for the top five cancers, TMUCRD

Three features of the rate series reflect data capture rather than epidemiology and should not be interpreted as trends:

- Rates in the first years shown (2004–2005) were markedly higher than in later years for every cancer, sex and age group. This pattern is consistent with prevalent-case contamination: early in the database, a 365-day lookback cannot distinguish patients diagnosed before their records began from genuinely new cases.

- All five cancers in both sexes peaked simultaneously in 2012, and rates rose again in 2022. These shifts coincide with changes in data capture in TMUCRD, namely the addition of hospital records and an EHR system migration.
- Absolute levels reflect the hospital population, which includes patients referred for cancer care, and are therefore not comparable with registry incidence.

Outside these periods, lung and liver cancer had the highest rates among men, with a modest rise in colorectal cancer. Among women, breast cancer rates increased gradually after 2010, and thyroid cancer showed a consistent upward pattern (data not shown).

## Discussion and limitations

This analysis shows that TMUCRD, mapped to the OMOP CDM, captures a broad range of cancers, including rare malignancies that are seldom available in Asian datasets, and can run standardised OHDSI phenotypes. Several limitations apply. First, TMUCRD is hospital-based; its rates describe the population attending three hospitals and cannot be generalised to Taiwan as a whole. Second, the index date is the first diagnosis recorded in TMUCRD, which may be later than the true date of diagnosis, particularly for patients diagnosed elsewhere and later referred. Third, case ascertainment relied on diagnosis codes without pathology confirmation, so some misclassification is likely. Fourth, the rate series contain data-capture artifacts (early-year inflation, the 2012 peak and the 2022 rise), and rates are crude rather than age-standardised. Consistent with the OMOPCAN protocol<sup>4</sup>, TMUCRD is best suited to characterisation and survival analyses. Linking to the Taiwan Cancer Registry, restricting trend analyses to periods of stable data capture, and age-standardisation would strengthen future estimates.

## Conclusion

This preliminary analysis characterises cancer patients across 34 cancer types in TMUCRD, a multi-hospital EHR database in Taiwan harmonised to the OMOP CDM. Its demographic patterns are consistent with Taiwan's known cancer burden, and it extends OHDSI oncology coverage in the Asia-Pacific region. Because TMUCRD is hospital-based, its main value lies in standardised characterisation and survival analyses within OMOPCAN, rather than population incidence. Broader participation from additional countries will enable more comprehensive cross-national assessments to support cancer research, healthcare planning and policy development.

## References

1. Liu PC, Huang TH, Wu YC, et al. Disease burden of 30 cancer groups in Taiwan from 2000 to 2021. *J Epidemiol Glob Health*. 2025;15(1):62. doi:10.1007/s44197-025-00406-w
2. Penberthy LT, Rivera DR, Lund JL, Bruno MA, Meyer AM. An overview of real-world data sources for oncology and considerations for research. *CA Cancer J Clin*. 2022;72(3):287-300.
3. Wang L, Wen A, Fu S, et al. A scoping review of OMOP CDM adoption for cancer research using real world data. *NPJ Digit Med*. 2025;8(1):189. doi:10.1038/s41746-025-01581-7
4. López-Sánchez I, Palomar-Cros A, Giuliadori A, et al. Temporal trends in epidemiology and patient characteristics of 36 cancers: a protocol for a multinational population-based cohort study using OMOP-standardised databases to investigate CANcer (OMOPCAN). *BMJ Open*. 2026. doi:10.1136/bmjopen-2026-119069
5. Observational Health Data Sciences and Informatics. CohortDiagnostics: diagnostics for OHDSI cohorts [R package]. <https://github.com/OHDSI/CohortDiagnostics>