Identification and Assessment of Real-World Data Sources for Oncology in China: A Data Landscaping Study

Lu Ban[1]; Jiang Li[2]; Amanda Pulfer[3*]

[1]Evidera, Beijing, China; [2]Evidera, Shanghai, China; [3]Evidera, London, UK

* Presenting author

RWD48

Background

- In China, cancer is among the top five leading causes of death, accounting for more than 2.4 million deaths in 2016.[1] The standardized incidence rate of cancer was 186.46 per 100,000 population in 2016, with the top five cancers reported as lung, colorectal, stomach, liver, and breast cancer.
- Despite this significant disease burden, very few observational studies using secondary data have been published in a representative Chinese population to understand the patient profile, treatment patterns, and clinical outcomes in a real-world setting.
- Worldwide, real-world data has increasingly been used to assess the disease burden in oncology.
- Due to the rapid development of information technology and artificial intelligence in China, a large amount of real-world healthcare data has been accumulated, but little is known about the current data landscaping and the potentials for research in oncology.

Objectives

The aim of the study was to identify real-world data sources for oncology in China and assess their strengths and limitations for conducting observational research.

Methods

A data landscaping exercise was conducted to identify existing data sources for oncology that can be used to assess patient profile, treatment, and clinical outcomes of adult patients with cancer in Mainland China. Specifically, a targeted literature review of studies supported by supplementary web-based searches was performed.

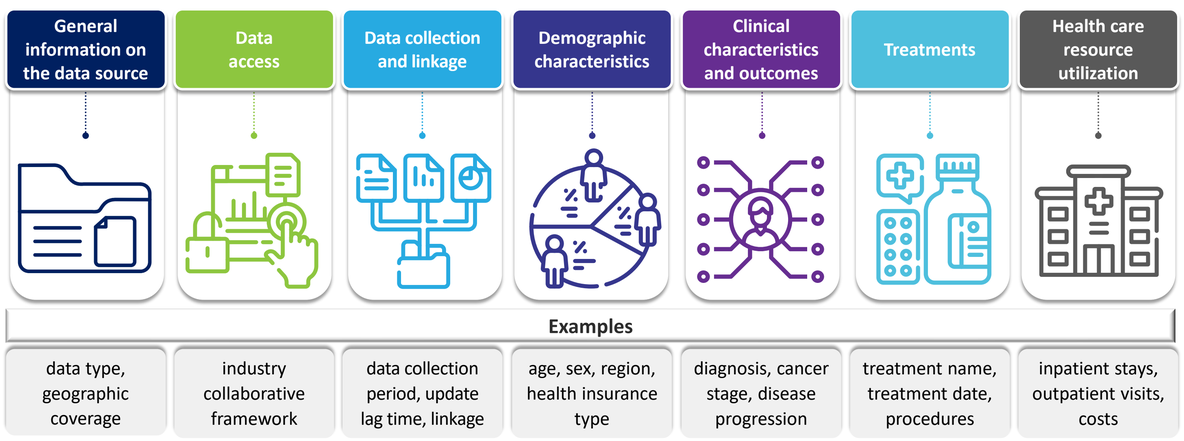
Screening

- Searches of studies including conference abstracts published between 2018 and 2023 were conducted in both English (i.e., PubMed/Embase) and Chinese search engines (i.e., Wanfang Data, and CNKI).
- Title and abstracts were screened for identification of existing data sources for oncology.
- Names and types of data sources, e.g., disease registries, claims, electronic health records (EHRs), were extracted.
- Full-text articles were retrieved if insufficient information was provided about a data source in the included abstracts.
- To identify data sources that can be used to study multiple indications, data sources including data for only one tumor type, from single centres, or only in paediatric populations were excluded.

Evaluation

Data elements relevant to real-world studies in oncology were evaluated based on publicly available information and (wherever possible) direct contacts with data custodians. A list of key criteria assessed is provided in the figure below.

Figure 1. Key evaluation criteria



Assessment

The strengths and limitations (including the accessibility) of each identified data source were assessed and summarized.

Results

Overall summary of results

- 2,918 citations were screened (1,278 from English search engines and 1,640 from Chinese ones).
- Over 30 real-world data sources for oncology containing individual-level patient data were identified.
- Five of them curated data from multiple geographic regions in China and were evaluated.
  - Disease registry: National Central Cancer Registry of China (NCCRC)
  - Claims: China Health Insurance Association (CHIRA) data
  - EHR data platforms/ databases:
    - Yidu Cloud (HLT)
    - Large-scale Data Analysis Center of Cancer Precision Medicine-LinkDoc database
    - National Anti-Tumor Drug Surveillance System (NATDSS)/National Cancer Center Oncology Information Database (NCCOID)
- The remaining data sources are regional (e.g., province- or citywide) and included regional cancer registries, claims, and local EHR data platforms (e.g., Tianjin citywide EHR database, Medical Database of Shanghai Hospital Development Center).
- The below section describes detailed data evaluation results for the five nationwide data sources and the evaluation results for the regional data sources are not described in this poster due to space limit.

Summary of key data evaluation results for nationwide data sources

Results are presented in Table 1 for the five nationwide data sources (one disease registry, one claims, and three EHR data platforms).

Results (Cont'd)

Table 1. Summary of the key data evaluation results

	Registry	*Claims*	*Hospital EHRs*		
	National Central Cancer Registry of China (NCCRC)	China Health Insurance Association (CHIRA) data	Yidu Cloud/HLT	Large-scale Data Analysis Center of Cancer Precision Medicine-LinkDoc database	National Anti-Tumor Drug Surveillance System (NATDSS)/ National Cancer Center Oncology Information Database (NCCOID)
General information on the data source	• Population-based cancer registries • Containing newly diagnosed cancer cases from over 500 local cancer registries in all 31 provinces • Covering about 31% of the national population in 2019	• Nationwide claims database of urban basic medical insurance • Containing data from the local insurance offices of selected areas of 66 sampling cities	• Data platform containing hospital EHRs of 30 to 40 hospitals • Mostly tier 3 hospitals from east and north parts of China • Non-cancer specific data platform	• Database for solid tumors with data extracted from EHRs of the oncology departments • The number of hospitals is unclear	• Database for solid tumors with data extracted from EHRs of the oncology departments • Over 2000 hospitals contributing data, with about 50 having good data quality in general
Data access	• Not open for collaborations with commercial entities	• Indirect access • Ethical approval is required • The study results must be published	• Indirect access • Data are owned by each participating hospital, and ethical approval and site contract are required for each site	• Indirect access • Data are owned by each participating hospital, and ethical approval and site contract are required for each site	• Indirect access • Data owned by National Cancer Center, and ethical approval and site contract are required for one site • The study results must be published
Data collection and linkage	• Established in 2002 • Data lag up to 3 years • Can be linked to the local and national death registries	• Data resampled annually from 2014 to 2017 • The data have not been updated since 2017 • No linkage to other data	• Some hospitals have data from 2010 onwards • Data lag 1-6 months • Data across participating hospitals are not linked and no linkage to other data either	• Started in 2014 • Data updates are not regular, depending on request of individual studies • Data across participating hospitals are not linked and no linkage to other data either	• Some hospitals have data from 2015 onwards • Data lag 5-6 months • Can be linked to the national Death Registry
Demographic characteristics	🟢	🟢	🟢	🟢	🟢
Clinical characteristics and outcomes	🟡	🔴	🟡	🟡	🟢
Treatments	🔴	🟢	🟡	🟡	🟡
HCRU	🔴	🟢	🟡	🟡	🟡
Recent publications in oncology	Chen R, et al. *J Natl Cancer Cent*. 2023;3(1)21-27.	Yang Y, et al. *J Glob Health*. 2023;13.04083.	Wang Y, et al. *Value Health*. 26(6)S403-S404.	Xu J, et al. *Transl Lung Cancer Res*. 2021; 10(1)402-414.	Qiu L, et al. *Front Oncol*. 29(12).

🟢 High availability 🟡 Medium availability 🔴 Low availability

Abbreviations: EHR = electronic health record; HCRU = healthcare resource utilization.

Summary of strengths and limitations

The overall strengths and limitations of the identified nationwide and regional data sources are summarized in Table 2 (by data source type).

Table 2. Strengths and limitations of the identified data sources by type

Data source type	Strengths	Limitations
Disease registry	• Good representativeness of the general population (nationwide or in specific geographic areas of interest) • Good cancer diagnosis information • Potential linkage to national or local death registries	• Lack of clinical follow-up data • Not open for collaborations with commercial entities • Long data lag
Claims	• Good representativeness of the general population (nationwide or in specific geographic areas of interest) • Good capture of health resource utilization	• Lacking clinical data • Some lacking follow-up data, e.g., cannot follow up the same individuals across multiple calendar years • Long data lag
EHR platforms/ databases	• Rich clinical data often with free texts from medical notes • Often having short data lag with more frequent data updates • Some can be linked to national or local death registries • Regional data often have longer follow-up periods with more data points and data linkage across different healthcare institutes	• Data are normally unstructured and unstandardized, which often require advanced data processing techniques like natural language processing • Data quality and completeness often vary by site • Some having incomplete follow-up data, e.g., diagnosis and treatment information from non-participating sites are not included • Incomplete clinical outcome data (e.g., tumor response, community deaths)

Abbreviation: EHR = electronic health record.

Conclusions

- Real-world data sources for oncology are emerging in China.
- Access to such data for commercial use often requires administrative, ethical, and/or regulatory approval and collaboration with academic or clinical institutes.
- Understanding the strengths, limitations, and practical requirements of each data source is crucial for appropriate study design and execution.
- An in-depth feasibility assessment (of data quality and completeness, accessibility, and operational considerations in addition to population coverage and representativeness and availability of key data elements) is thus recommended for selection of a data source for a study using real-world data for oncology in China.

References

Zheng R et al. Cancer incidence and mortality in China, 2016. *Journal of the National Cancer Center*. 2022; 2(1) 1-9. doi: 10.1016/j.jncc.2022.02.002

Presented at the ISPOR Europe Conference ● 12–15 November 2023 ● Copenhagen, Denmark

Funding provided by Evidera Inc., a business unit of PPD, part of Thermo Fisher Scientific.