Data Needs and Challenges in Cancer Epidemiology: A U.S. Real-World Database Perspective

Kathy H. Fraeman, SB, SM
Director, Data Analytics, Real-World Evidence, Evidera

Beth L. Nordstrom, PhD
Senior Research Scientist, Real-World Evidence and Executive Director, Center of Excellence for Epidemiology, Evidera

Kathy Fraeman

Beth Nordstrom

Introduction

With treatment options for many types of cancer increasing, there is an escalating demand for real-world evidence in oncology. The safety and efficacy of new antineoplastic drugs are demonstrated in clinical trials before U.S. Food and Drug Administration (FDA) approval, but are these treatments safe and effective for the patients prescribed these drugs in medical practice outside of trials? How are these drugs being prescribed by oncologists? Which patients receive which drugs? How long do they stay on treatment? Questions such as these can be answered only through real-world observational data.[1]

Evidence-based cancer epidemiology research using observational real-world data poses special challenges seldom found in other therapeutic areas. Treatment for cancer is often very complex, with multiple drugs given in combination regimens that frequently change over time. The course of cancer treatment may span many years, much longer than the average time an individual patient is tracked in many data sources. Progression-free survival is a high-priority target outcome, but can be exceedingly difficult to ascertain without the close, regular monitoring that occurs in clinical trials. Adverse events can be difficult or impossible to attribute to any particular treatment, given the treatment combinations used (both antineoplastic and as supportive care), and many adverse effects may be brought about by the disease itself and unrelated to treatment. Some studies examine cancer as an adverse outcome to treatment for non-cancer-related conditions; for these, the association between drug use and cancer may be difficult to assess due to long latency periods and the potential for unmanageable degrees of bias and confounding.[2]

> **"Evidence-based cancer epidemiology research using observational real-world data poses special challenges seldom found in other therapeutic areas."**

In the United States, there are two primary types of real-world databases available for oncology research: electronic medical record (EMR) data and administrative claims data containing medical and pharmacy claims information. The advantages of using these types of electronic databases for research are typically large patient population sizes, relatively timely updates to and availability of the data, and inclusion of many required data elements, such as patient diagnoses, medical procedures performed, inpatient admissions, and drug prescribing or dispensing. EMR databases often contain additional data elements relevant to oncology research, such as laboratory test results and detailed clinical information. In some cases, data from an EMR or claims database can even be supplemented with linked data from other sources, such as chart reviews, primary data collection such as patient or provider surveys, or registry data.

Cancer Epidemiology Database Study Types

Numerous oncology topics can be investigated using real-world databases. Incidence and prevalence studies look at rates of cancer relative to the general population, or to subgroups of the population with a particular disease or set of clinical or demographic characteristics. Patients with cancer can be followed for health outcomes such as disease progression, remission, or complications in studies that focus on the natural history of disease rather than on the effects of treatment. Treatment pattern studies examine the various antineoplastic or supportive care agents used to treat cancer patients in real-world settings and can identify the characteristics of patients prescribed each drug or regimen, their use across lines of therapy, and drug utilization measures such as adherence and persistence.

Drug safety and effectiveness are often investigated using real-world databases. Although treatment effectiveness can be very difficult to measure using real-world data, outcomes such as overall survival and, for hematologic cancers, key lab values indicating the likely effect of treatment can be studied. Many antineoplastic drugs carry a high burden of adverse events, and even supportive care oncology drugs have been associated with adverse outcomes. The incidence rates of these adverse events can be examined in databases. Finally, safety studies may be conducted to look for new-onset cancer as a safety outcome from the use of drugs intended as treatment for other diseases.

Real-World Databases for Oncology: What Is Available?

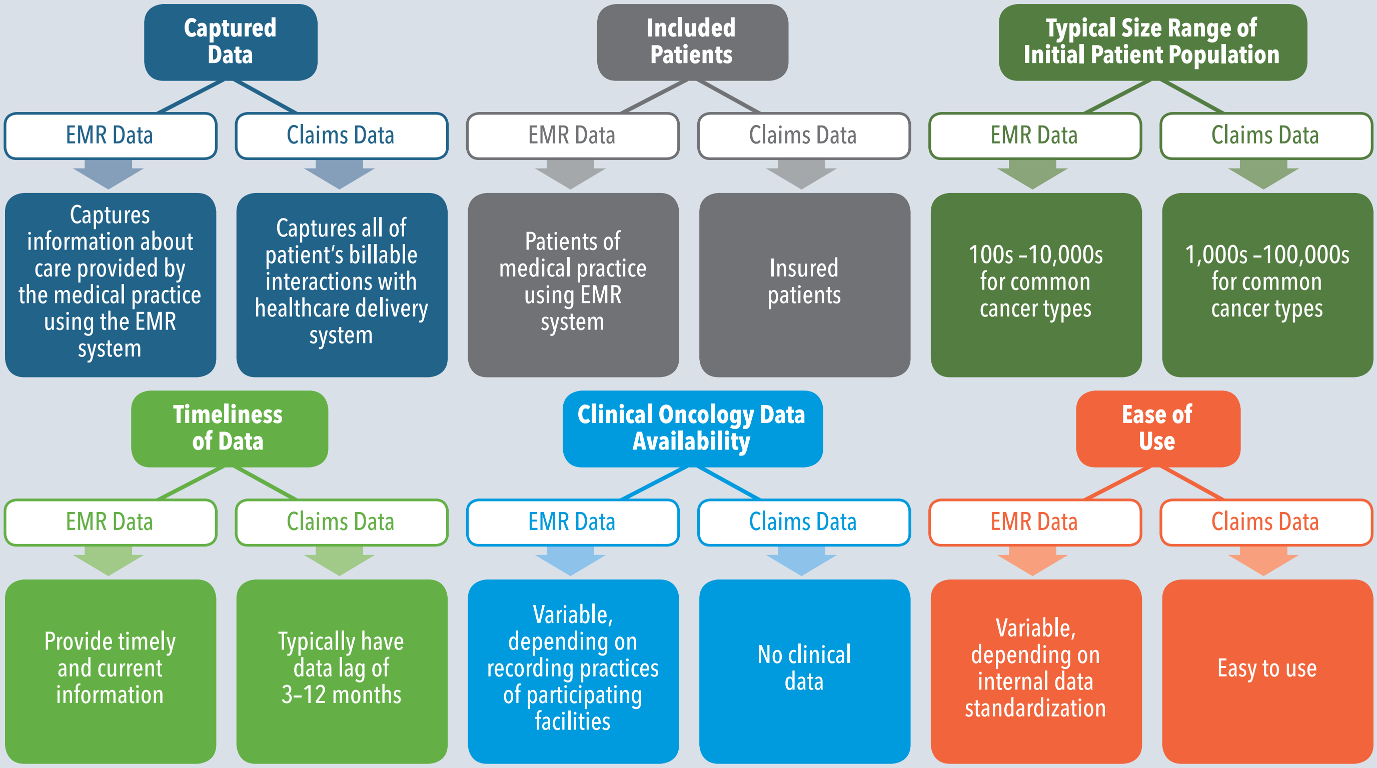
Insurance claims data summarize all of the billable interactions of an insured patient with the healthcare delivery system. These data include the dates corresponding to a variety of billing codes submitted to payers, including codes representing disease diagnoses (ICD-9, ICD-10), medical procedures (HCPCS, CPT4), and pharmacy drugs (NDC). In a closed system that contains data from payers, the claims data provide a complete picture of all covered medical and pharmacy services received by a patient in a clear, standardized format for a large number of insured patients. Open claims systems, which contain data from providers rather than payers, can be even larger than closed systems but are not complete for all patients, as not all providers caring for a given patient may submit claims to the same system.

EMR databases have many of the elements of claims data but also contain additional clinical information that is highly relevant to oncology studies. Some EMRs are designed to be used specifically in outpatient oncology clinics that provide treatment to cancer patients, making them a valuable real-world evidence data resource specifically for oncology studies. Other EMR databases not specific to oncology clinics may also be used for cancer epidemiology studies if the particular practice using the EMR system provides care for cancer patients. Some of these more general EMR databases have developed their own cancer "registries" containing in-depth information on histology, staging, treatment, and progression derived from progress notes and other data not typically included in an EMR extract.

Figure 1. Types of evidence-based cancer epidemiology studies

Incidence/Prevalence	**Treatment Patterns**	**Drug Safety of Antineoplastic Agents/Supportive Care**
Derive incidence or prevalence of type(s) of cancer relative to population or relative to patient populations with different demographic characteristics	Identify drugs and combinations of drugs being used to treat different types of cancer, during different lines of treatment	Examine adverse events associated with the use of antineoplastic drugs or supportive care agents prescribed to cancer patients
Natural History of Disease	**Treatment Effectiveness**	**Drug Safety with Cancer as Adverse Event**
Estimate the incidence of disease outcomes and complications among cancer patients, irrespective of treatment	Find the incidence of beneficial health outcomes associated with antineoplastic or supportive care treatment	Estimate the incidence of cancer as an adverse event associated with the use of drugs not given to treat cancer

Figure 2. Comparison of data sources used for oncology studies


Captured Data
EMR Data
Claims Data
Captures information about care provided by the medical practice using the EMR system
Captures all of patient's billable interactions with healthcare delivery system
Included Patients
EMR Data
Claims Data
Patients of medical practice using EMR system
Insured patients
Typical Size Range of Initial Patient Population
EMR Data
Claims Data
100s -10,000s for common cancer types
1,000s -100,000s for common cancer types
Timeliness of Data
EMR Data
Claims Data
Provide timely and current information
Typically have data lag of 3-12 months
Clinical Oncology Data Availability
EMR Data
Claims Data
Variable, depending on recording practices of participating facilities
No clinical data
Ease of Use
EMR Data
Claims Data
Variable, depending on internal data standardization
Easy to use

Data Needs for Oncology Studies: How Can We Fill the Gaps?

While insurance claims databases provide a comprehensive picture of a patient's medical care, they lack the clinical detail needed for many oncology studies. For example, claims data can indicate whether a medical test was conducted, but in general, the results of the test are not available. Some claims databases have linked laboratory results available, but usually for only a subset of patients and tests, so that the available data may be highly non-representative of lab results for the full patient population in a study. Diagnosis codes found in claims data are not confirmed and may indicate a diagnosis that was suspected but then ruled out by a given diagnostic test. Claims databases also lack clinical details such as cancer staging at initial diagnosis or progression over time. Metastatic cancer can in some cases be identified through diagnosis codes indicating a secondary tumor and/or treatment specific to metastatic cancer, but this approach is imperfect at best, and distinguishing among earlier stages in claims data may be even more difficult.[3]

EMR databases can help fill some of these gaps, as discussed above, but they have their own limitations. EMR systems are designed to help medical providers manage patient care and the business aspects of their practices, such as billing and scheduling. Diagnoses entered into an EMR may be no more valid than in a claims database, with rule-out codes and other erroneous diagnoses that do not reflect the patient's true medical conditions. While an EMR database may provide the opportunity to include data elements important for research – such as disease progression, comprehensive medical histories, and additional treatments administered outside of the practice – the availability and completeness of these data elements varies both across and within EMRs, depending on how each practice choses to enter data and to use the EMR for their own purposes. Information is often entered into an EMR as unstandardized free text, which then needs extensive cleaning and standardizing prior to initiating data analyses.

Despite these limitations, some cancer epidemiology studies can be conducted within a claims or EMR database and still produce valid results, as long as the needs of the study make use of the data source's strengths and do not rely on data elements that are absent or incomplete. For example, studies examining outpatient cancer treatment patterns or incidence of adverse events measured through validated coding algorithms or outpatient lab tests can be completed in an appropriate database. Yet many important research questions in cancer epidemiology cannot be answered through claims or EMR data alone. Many drug safety studies, for example, require detail from both inpatient and outpatient settings, where the adverse events under investigation are not reliably identified through ICD-9 or ICD-10 codes. Additional data gaps may include insufficient depth of clinical detail around the cancer at the start of follow-up, or around changes over time such as tumor size or response to treatment.

Approaches to filling these gaps may include linking to external data sources that contain the missing information or collecting data either retrospectively or prospectively. One commonly used linked database for oncology research is the SEER-Medicare database[4], which contains Medicare claims data combined with the cancer registry information collected by SEER (the Surveillance, Epidemiology, and End Results Program of the National Cancer Institute). The SEER data provide important clinical information from the time of initial cancer diagnosis that is missing from the claims data, including records of cancer type and stage, while the Medicare data for the subset of patients with drug coverage should be complete for services covered by Medicare, including cancer treatments and outcomes. This database is limited, however, by having a several-year lag time for the SEER data, as well as lacking follow-up registry or EMR-level data (e.g., lab results or disease progression).

Retrospective data collection typically involves chart review, which can be performed through text searches in electronic data if the information sought is recorded electronically (e.g., progress notes, radiology reports), or via manual review of paper charts. Even in pure EMR databases, where all records are kept electronically, data extracts generally cannot include free-text information because of concerns for patient privacy, and hence require an electronic chart review. Chart reviews can be used to validate diagnoses that qualify patients for the analysis or that occur as outcomes during follow-up, or to pull information that is missing in the data extract, such as results of a lab test that were not entered into the database. The chart review targets only the specific information that is needed, which can make it much more focused and study-appropriate than a database extract, but it can be time-consuming and expensive. Additionally, in many cases the required records for some patients are not available for review, leading to problems with missing data that need to be addressed.

Some topics in cancer epidemiology, such as assessing treatment response when the data needed to evaluate it are not usually measured in real-world clinical practice, require prospective data collection. Patients qualifying for the analysis are identified through a claims or EMR database, and the patients and/or their physicians are contacted to request enrollment in a prospective study. These endeavors may involve patient surveys to examine self-reported information from qualifying patients such as patient-reported outcomes, physician or caregiver surveys that inquire about their perspective on the patient's treatment or condition, blood draws or collection of tissue samples from patients to measure outcomes or biomarkers not assessed in the course of their medical care, or even enrollment into a registry with scheduled visits and examination of many follow-up characteristics and outcomes. Although these studies are by far the most expensive and time-consuming of the observational study types, they have an unsurpassed advantage in allowing investigation of exactly the information needed for the study.

Conclusion

Although many sources of real-world evidence are available to conduct cancer epidemiology studies, the data needs of these studies are not always fully met by a single data source. EMR databases lack complete information about diagnoses and treatments from outside the EMR practice, and the data entry can be highly idiosyncratic. Clinical details such as cancer staging and progression may be present for some patients but missing for many. While insurance claims data cover large patient populations, give complete data on all of a patient's billable medical care, and are easy to use, they are usually inadequate for many oncology studies due to their lack of clinical data. Data collection can help to overcome many of the shortcomings of these databases, but require markedly greater time and expense, as well as permission to collect the additional data. Ideally, more comprehensive oncology datasets could be constructed by linking together existing databases.

For more information, please contact Kathy.Fraeman@evidera.com or Beth.Nordstrom@evidera.com.

REFERENCES

[1] Booth CM, Tannock IF. Randomised Controlled Trials and Population-based Observational Research: Partners in the Evolution of Medical Evidence. *Br J Cancer.* 2014 Feb 4;110(3):551-5. doi: 10.1038/bjc.2013.725.

[2] Haynes K, Beukelman T, Curtis JR, et al. Tumor Necrosis Factor α Inhibitor Therapy and Cancer Risk in Chronic Immune-mediated Diseases. *Arthritis Rheum.* 2013 Jan;65(1):48-58. doi: 10.1002/art.37740.

[3] Chawla N, Yabroff KR, Mariotto A, McNeel TS, Schrag D, Warren JL. Limited Validity of Diagnosis Codes in Medicare Claims for Identifying Cancer Metastases and Inferring Stage. *Ann Epidemiol.* 2014 Sep;24(9):666-72, 672.e1-2. doi: 10.1016/j.annepidem.2014.06.099.

[4] SEER-Medicare Linked Database. Available at http://healthcaredelivery.cancer.gov/seermedicare. Accessed April 5, 2016.