Comparative Effectiveness Research in Cancer with Observational Data

Sharon H. Giordano, MD, MPH

OVERVIEW

Observational studies are increasingly being used for comparative effectiveness research. These studies can have the greatest impact when randomized trials are not feasible or when randomized studies have not included the population or outcomes of interest. However, careful attention must be paid to study design to minimize the likelihood of selection biases. Analytic techniques, such as multivariable regression modeling, propensity score analysis, and instrumental variable analysis, also can also be used to help address confounding. Oncology has many existing large and clinically rich observational databases that can be used for comparative effectiveness research. With careful study design, observational studies can produce valid results to assess the benefits and harms of a treatment or intervention in representative real-world populations.

Comparative effectiveness research increasingly is being recognized as a priority area in cancer research. Although researchers have been conducting comparative effectiveness research for decades, this area of research achieved national prominence when, as part of the American Recovery and Reinvestment Act of 2009, $1.1 billion was allocated to comparative effectiveness research. Shortly thereafter, the Patient Protection and Affordable Care Act of 2010 established the Patient-Centered Outcomes Research Institute (PCORI). The mandate of PCORI is “to improve the quality and relevance of evidence available to help patients, caregivers, clinicians, employers, insurers, and policy makers make informed health decisions. Specifically, we fund comparative clinical effectiveness research, or CER, as well as support work that will improve the methods used to conduct such studies.” These initiatives have stimulated interest and provided new funding opportunities in comparative effectiveness research.

The Institute of Medicine has formulated a definition of comparative effectiveness research as follows: “CER is the generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat and monitor a clinical condition or to improve the delivery of care. The purpose of CER is to assist consumers, clinicians, purchasers, and policy makers to make informed decisions that will improve health care at both the individual and population levels.” Thus, comparative effectiveness research is meant to compare the risks and benefits of a given approach in a real-world heterogeneous population with the ultimate goal of improving health care decisions and outcomes. Comparative effectiveness research encompasses a broad array of study types, including randomized trials (pragmatic clinical trials, cluster-randomized trials, adaptive trials), research synthesis (systematic reviews, meta-analysis, decision analysis), and observational studies (cohort, case-control, cross-sectional designs).

Randomized trials have long been considered the gold standard when evaluating the efficacy of a particular intervention. These studies test the effect of an intervention under carefully controlled conditions and are the gold standard to evaluate the efficacy of a particular intervention. Although randomized trials have the best internal validity, the generalizability of results can be limited because of strict eligibility criteria. Clinical trial participants tend to be systematically different than patients in the general population, with younger ages and fewer concurrent medical conditions. Pragmatic clinical trial designs address some of these issues, but randomized trials remain expensive and time consuming. Observational studies increasingly are being used to evaluate the effectiveness of treatments in real-world populations, because they can address the gaps in randomized studies and the benefits in patient populations who were not included in the clinical trials. In addition, observational studies can be conducted with a relatively low cost and often include large numbers of representative populations.

ROLE OF OBSERVATIONAL STUDIES

The Effective Health Care website has published a guide for methods in comparative effectiveness research. This article...
provides a conceptual framework for the use of observational studies in comparative effectiveness research. The authors note that researchers should ask two key questions before undertaking an observational CER study: (1) Are there gaps in the evidence from randomized controlled trials? and (2) Will observational studies provide valid and useful information?4

When data are available from randomized clinical trials, observational studies may provide little additional information. However, at times, randomized clinical trials may not be possible. Studies of medications, particularly if they are off patent, may have no financial support. In some situations, random assignment may not be feasible or ethical. For instance, studies of a rare cancer may never be able to accrue a sufficient number of patients to obtain statistically valid results. Trials may be unethical to conduct in situations when there is not clinical equipoise between the two therapeutic options. Even in situations when clinical trials may be possible, existing clinical trials may not be relevant to the population of interest or may not include the outcomes of interest. This situation often arises when evaluating the benefit of a treatment for patients who are older or who have many comorbid conditions. Randomized trials often focus on short-term outcomes and may not have sufficient follow-up times to evaluate late outcomes of interest. For example, clinical trials for patients with breast cancer typically have not observed patients to assess the risks of long-term outcomes, such as congestive heart failure related to anthracyclines or myocardial infarction as a result of radiation therapy, both of which may occur decades after treatment. In these situations, when no randomized studies exist, the trials were not conducted in the population of interest, or the studies do not capture the endpoint of interest, observational studies can be used to fill in the gaps in data.

The second question to ask before undertaking an observational study is whether an observational study can provide valid information.4 The major challenges in observational studies are assessing and addressing the risk of bias, particularly the risk of selection bias. Selection bias refers to systematic differences in the risk of the outcome between the two compared groups of patients. For instance, in a comparison of the effectiveness of chemotherapy, selection biases could work in several ways. First, patients who have high-risk disease features, such as higher-stage disease or high-grade tumors, would be more likely to be treated with chemotherapy. This selection bias could cause the chemotherapy-treated group to appear to have a worse outcome, simply because they had higher-risk disease. Many of these high-risk features can be measured and adjusted for in analyses, but some high-risk features, such as positive margins or lymphovascular invasion, may not be routinely captured or reported in national databases. A second way in which selection biases can work is that patients who are older or who have poorer performance statuses and more comorbidities will be less likely to be treated. This selection bias would result in the untreated patients appearing to have worse outcomes, simply because of poorer health at diagnosis. Again, some of these predictors of treatment, like age, can be measured and adjusted for in statistical models. However, other important factors that clinicians use to decide treatment, like performance status or frailty, are not routinely captured in databases. These factors can act as unmeasured confounders, which are extraneous variables that are associated with both the independent variable and dependent variable. For instance, performance status is associated with both the likelihood of treatment and survival and could lead to biased estimates of the efficacy of treatment. Another illustration of selection biases can be seen in an example of evaluating the benefit of treatment for patients with prostate cancer.5 In an analysis of older men with early-stage prostate cancer, the use of any active treatment (radiation or surgery) was associated with better survival after adjustment for all measured covariates.5 When evaluated by cause of death, treatment was associated with improved overall survival and improved prostate cancer–specific survival. However, prostate cancer treatment also was associated with better cause-specific survival from cardiovascular disease, pneumonia, and diabetes. There are no plausible mechanisms by which treatment for prostate cancer could affect survival from diabetes, other than through confounding from underlying health. This example illustrates the challenges of identifying and accounting for selection biases. Selection biases also can result in confounding by indication. Confounding by indication can occur when the patients who are prescribed a particular treatment differ from those who are not prescribed a given treatment because of the medical indication for which the drug is prescribed. Many other biases can occur in observational studies, such as performance bias (e.g., difference in adherence), detection bias (e.g., differential assessment of outcomes), and outcome-reporting bias.4 Before undertaking an observational analysis, careful consideration must be given to the possibility of bias and how potential biases could affect the results.

**KEY POINTS**

- Observational studies are particularly useful when randomized studies are not feasible or have not included the population or outcome of interest.
- Selection biases are the primary threat to validity of observational studies.
- Multivariable regression models, propensity score analyses, and instrumental variable analyses can be used to help address selection biases and confounding.

**ANALYTIC TECHNIQUES**

Several approaches can be taken to address bias in observational studies. Most studies will use multivariable regression analyses to adjust for confounders and to address issues related to nonrandom treatment assignment. This method can provide statistical adjustment for all measured variables but cannot adjust for unmeasured confounders. Propensity score analysis is another approach to improve the balance of un-
measured confounders between the two experimental groups. In propensity score analysis, multivariable regression is used to calculate the propensity of receiving the intervention of interest. This propensity score can then be entered as an explanatory variable in the regression models or can be used for matching. Some studies have suggested that propensity score stratification can remove more than 90% of the bias in observational studies. Another approach is the use of instrumental variable analysis, which is a method used in econometrics. An instrumental variable is a variable that is strongly correlated with the treatment or intervention but without any independent effect on the outcome. An appropriate instrumental variable can help address unmeasured confounding. In oncology studies, distance to the treatment center and regional treatment rates have been used as instruments. However, finding an appropriate instrumental variable can be challenging.

**TYPES OF OBSERVATIONAL STUDIES**

Several different study designs, such as cross-sectional, cohort, and case-control designs, can be used for observational research. A cross-sectional study is a study in which the prevalence of the exposure and outcome are measured at the same point in time. In this design, participants are identified independent of exposure and outcome. Cross-sectional studies can evaluate prevalence but cannot measure disease incidence. In a cohort study, participants are identified by their exposure and then observed over time for the outcome of interest. For instance, patients with breast cancer who were and were not treated with trastuzumab could be observed for the risk of congestive heart failure. A cohort study allows for the calculation of incidence in treatment groups. When they are prospectively designed, these studies may need an impractically large sample size if the outcome is not common. Large, retrospective, cohort studies using existing national databases get around this limitation with their large size and lengthy follow-up times. A third common study design is a case-control study. In this study design, cases and controls are identified by the outcome of interest, and then the prevalence of exposure is compared. In the example of trastuzumab cardiotoxicity above, patients with heart failure would be identified and then compared by exposure to trastuzumab. Other study designs, such as case cohorts or case crossovers, also may be used but are beyond the scope of this review.

**SAMPLE DATA SOURCES IN ONCOLOGY**

Oncology has a long track record of observational studies, in part because of the high-quality data provided by cancer registries. Incident cancer cases are reported to state cancer registries, which track cancer incidence, staging, first course of treatment, and outcomes. Cancer registries have provided data for numerous research studies and also have been linked with other databases. Commonly used observational data sets for cancer research are described below and are summarized in Table 1.

**Surveillance, Epidemiology, and End Results**

The Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute (NCI) collects and compiles data from cancer registries to provide information on cancer incidence and survival in the United States. This program was started in 1973 and has expanded over the years to include data from 10 states (Connecticut, Hawaii, Iowa, New Jersey, New Mexico, California, Virginia, Kentucky, Louisiana, and Utah), six metropolitan areas defined by county borders (Atlanta, Detroit, Los Angeles, San Francisco/Oakland, San Jose/Monterey, and Seattle/Puget Sound), the supplemental SEER registries of Alaska Natives, Arizona Native Americans, and Cherokee Nation, and 10 counties in rural Georgia. These SEER registries cover approximately 28% of the U.S. population. For each patient with an incident cancer, the following information is reported: (1) demographics: case number, age, sex, race/ethnicity, state and county of birth, state and county of residence, date and cause of death according to death certificate (of note, cancer recurrence is not captured); (2) cancer: type, month and year of diagnosis, diagnostic confirmation, and laterality; (3) extent of disease: American Joint Committee on Cancer (AJCC) stage, historic SEER stage, tumor size and extension, number of disease: American Joint Committee on Cancer (AJCC) stage, historic SEER stage, tumor size and extension, number

### TABLE 1. Data Sources

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Type of Data</th>
<th>Years Available with Incident Cases</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEER</td>
<td>Cancer registry</td>
<td>1973-2013</td>
<td>All incident cancer cases in SEER regions</td>
</tr>
<tr>
<td>100% Medicare</td>
<td>Health care claims</td>
<td>Through 2012</td>
<td>Age &gt; 65, disabled, end stage renal disease</td>
</tr>
<tr>
<td>SEER-Medicare</td>
<td>Registry-linked claims</td>
<td>1973-2011</td>
<td>Age &gt; 65, disabled, end stage renal disease</td>
</tr>
<tr>
<td>SEER-MHOS</td>
<td>Registry-linked to survey</td>
<td>1998-2011</td>
<td>Age &gt; 65</td>
</tr>
<tr>
<td>SEER-NLMS</td>
<td>Registry-linked to survey</td>
<td>1979-2011</td>
<td>All ages</td>
</tr>
<tr>
<td>HCCI</td>
<td>Health care claims from United, Aetna, Humana, and Kaiser</td>
<td>2009-2013</td>
<td>All; employees, spouses, dependents</td>
</tr>
<tr>
<td>Marketscan</td>
<td>Health care claims from 45 large employers</td>
<td>2003-2013</td>
<td>All; employees, spouses, dependents</td>
</tr>
<tr>
<td>NCDB</td>
<td>Clinical database</td>
<td>Since 1985</td>
<td>All ages; patients treated at CoC accredited facilities</td>
</tr>
</tbody>
</table>

Abbreviations: SEER, Surveillance, Epidemiology, and End Results; MHOS, Medicare Health Outcomes Survey; NLMS, National Longitudinal Mortality Study; HCCI, Health Care Cost Institute; NCDB, National Cancer Database; CoC, Commission on Cancer.
of lymph nodes examined, number of positive lymph nodes, histologic grade, histologic type; and (4) treatment: type of surgery and radiotherapy for first course of treatment. These data provide a definitive source of information on cancer incidence, staging, and survival. The information on treatment is less complete. In general, the ascertainment of surgery by registries is the highest-quality treatment variable. The use of radiation therapy may be under-ascertained by registries. In the case of breast cancer, the under-ascertainment has been reported to be differential by age, income, and other treatment.11,12 Complete information on chemotherapy administration is difficult to capture; thus, rates of chemotherapy use are not reported in the SEER database. The SEER database is de-identified, publically available, and free of charge.

Medicare
Medicare is the primary health insurer for 97% of the U.S. population age 65 and older. The Medicare Claims Data System collects information on all services provided to Medicare beneficiaries under its hospital (Part A), supplemental (Part B), and prescription drug (Part D) insurance plans. Part A covers inpatient hospitalizations and care in skilled nursing homes, whereas Part B covers physician services; hospital outpatient services; durable medical equipment; home health services; and other outpatient medical services, such as diagnostic x-rays and laboratory tests. Part D data includes information on prescription drug use and has been available since 2006. The unlinked Medicare data are comprised of enrollment files and health care claims. These data do not have information on the dates of diagnosis, stage, histology, or recurrence, and this lack of registry data has limited the use of the unlinked Medicare files for cancer research. For several cancer types, such as breast cancer, validated algorithms have been developed to identify incident cases.13 The strength of these data is in studies of patterns of care, geographic variation, and costs of therapy. To obtain the data set with identifiers, an application should be submitted to the Research Data Assistance Center (ResDAC). The cost is dependent on the specific data request.

SEER-Medicare Linked Database
Under an agreement between the NCI and the Centers for Medicaid & Medicare Services (CMS), patients in the SEER database who are eligible for Medicare have been linked with their Medicare records. Of persons who are reported by SEER as diagnosed with cancer at age 65 or older, 93% were matched with their Medicare enrollment records.14 At present, patients with cancer who were diagnosed through 2011 are linked, and their Medicare claims are available through 2012. For patients with cancer in the SEER database, Medicare enrollment and eligibility information and a subset of SEER data are in the Patient Entitlement and Diagnosis Summary File (PEDSF). The PEDSF also contains census-tract-level information, including ethnicity, education, income, percentage English-speaking residents, and population density. In addition to the registry data provided by SEER, these data include all billing claims for each patient. The claims can supplement data from the SEER registry and provide information on diagnostic testing and additional treatment information, such as chemotherapy use and patterns of surveillance. In addition, the data can be linked to the Area Resource File and the American Medical Association’s Physician Masterfile, which can provide additional information about health resources and physician characteristics, respectively. These data do have some substantial limitations. First and foremost, these data are administrative data generated for billing, so they lack the detail of clinical data. The claims are only complete for those patients in the fee-for-service plans (rather than in the HMO plans) and do not capture services provided by the U.S. Department of Veterans Affairs. Finally, as with the SEER data, cancer recurrences are not captured. To obtain these data, an application must be submitted to the NCI for review. Cost estimates can be found on the SEER-Medicare website.15 All publications resulting from these data must be reviewed before publication to ensure that patient confidentiality is protected.

SEER-Medicare Health Outcomes Survey Linked Database
The SEER-Medicare Health Outcomes Survey (MHOS) data set is a linkage of two large population-based databases. The SEER database, described above, has been linked with the MHOS, which is a large survey of health-related and quality-of-life information among participants in the Medicare Advantage Organization. These data include information on 12 cohorts of baseline and follow-up surveys at 2 years, which were conducted between 1998 and 2011.16 Collected data include patient demographics, marital status, income, smoking status, chronic conditions, and health-related quality of life. From 1998 to 2005, the MHOS used the 36-question short-form (SF-36) to collect data on health-related quality of life, but the MHOS switched to the Veterans RAND 12-item health survey (VR-12) in 2006. This data set provides additional patient-reported information to the SEER data. However, the survey was not limited to participants with cancer nor timed with a diagnosis of cancer, therefore the small sample sizes may limit potential analyses.

SEER-National Longitudinal Mortality Study Linked Database
The National Longitudinal Mortality Study (NLMS) is a data set that includes sociodemographic data collected by the Census Bureau by in-person and telephone interviews, and it includes cause of death information. The variables include race/ethnicity, education, income, employment, occupation, smoking, health status, and health insurance status. These data have been linked with SEER and with Medicare claims and can be used in studies evaluating socioeconomic determinants of cancer and cancer outcomes or outcomes related to income or employment, in addition to other uses.

Commercial Claims Databases
There is an increasing number of databases that can be licensed for research on the commercially insured population.
These databases are analogous to the unlinked Medicare data, in that they are health care claims without linkage to cancer registry data. They lack information on cancer incidence, stage, or survival, but they can provide detailed treatment information on younger, privately insured patients. One of the most commonly used data sets is Marketscan data, which consist of proprietary data sets licensed by Thomson Medstat Inc. The data undergo internal quality checks for reasonableness of data and validity before they are released, and they are HIPAA compliant. Marketscan is a large nationwide employment-based database that contains information on medical claims and outpatient prescription drug claims for employees and their spouses and dependents; the data represent claims from approximately 45 large employers, and Marketscan captures insurance claims data from over 100 payers. All files of claims data can be linked to the enrollment file via de-identified person identifiers. These data represent the medical experience of insured employees and their dependents for active employees, early retirees, COBRA continuers, and Medicare-eligible retirees. The specific demographic covariates available through Marketscan include patient age and sex, birth year, marital status, and three-digit zip codes. Information on race and ethnicity is not available.

The Health Care Cost Institute (HCCI) is another source of data on the commercially insured population. The HCCI is a nonprofit organization founded by academic economists to create a comprehensive source of information on health care activity and costs. Their research database covers about 40 million people, which represents approximately one-quarter of the population covered by employer-sponsored insurance. The data set contains claims from Aetna, Humana, Kaiser Permanente, and United Healthcare. The data capture complete payment information, including payments from both the benefit plan and patient. Pharmacy and mail-order prescriptions are included. Other proprietary claims data include IMS LifeLink data, Perspective, and Optum Insight. These data sets can provide comprehensive information on younger, privately insured patients. However, these data sets also have some limitations beyond the lack of linked registry data. The continuous coverage in these data sets tends to be short, so longitudinal studies can be challenging. In addition, no information on survival is available, and thus long-term outcomes cannot be assessed.

**National Cancer Database**

The National Cancer Database (NCDB) is a clinical oncology database jointly sponsored by the American College of Surgeons and the American Cancer Society. This database collects information from hospital registries that are Commission on Cancer (CoC)-accredited facilities. This includes more than 1,500 hospitals in the United States and Puerto Rico. The database is quite extensive and covers approximately 70% of newly diagnosed cancer cases in the United States. The data include information on patient characteristics, pathology, stage, prognostic factors, treatment, and outcomes. The strength of these data is the collection of clinically rich information, but the generalizability is somewhat limited because the data are not population based.

**CONCLUSION**

In summary, observational data can be a powerful resource for comparative effectiveness studies in cancer research. Observational studies are particularly useful to fill in gaps in data from randomized studies or when randomized studies cannot be conducted. Researchers must be cognizant of the risk of selection biases and confounding in observational research and must design studies carefully to minimize the risk of bias. The Agency for Healthcare Research and Quality’s Effective Health Care Program has provided a checklist for considerations in study design when observational comparative effectiveness studies are proposed. Given careful and thoughtful study design, observational studies can provide important and clinically relevant comparative effectiveness information in real-world populations. Finally, with the increasing use of electronic medical records and the ability to access large national data sets, this area of research is likely to continue to grow in scope and impact.

**Disclosures of Potential Conflicts of Interest**

The author(s) indicated no potential conflicts of interest.

**References**


