Journal of the National Comprehensive Cancer Network

The Last Word

Databases as Vehicles for Comparative Effectiveness Research

Christina A. Minami, MD, and Karl Y. Bilimoria, MD, MS

Although evidence-based medicine has grown over the past 2 decades and has become fundamental to oncologic practice, major questions remain. In an ideal world, clinical decisions would be based on data from exquisitely designed randomized controlled trials (RCTs). But RCTs are expensive, impractical, and can be unethical if they question entrenched practices that have become the standard of care. In addition, they are designed to answer questions of efficacy: the effect of an intervention compared with a placebo when all other variables are controlled. They poorly mimic oncologic practice in the real world, and as a result, investigators are turning to studies of effectiveness, which examine the effects of an intervention under real-world conditions.

In principle, comparative effectiveness research (CER) is not a new idea; the Institute of Medicine broadly defines it as the 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. New research tools and techniques, however, have made it possible to pursue CER in a more effective manner. One of the formidable resources that oncologic researchers now have to explore CER questions is large database analysis. With large amounts of patient data, databases allow researchers to conduct powerful retrospective studies and can offer a robust initial approach to certain questions of effectiveness. Researchers, however, must not only consider the strengths and weakness of a large database study in theory but also the various limitations of specific databases if they are to undertake a serious project in oncologic CER.

Commonly Used Cancer-Specific Nationwide Databases

National Cancer Data Base

Established in 1989, the National Cancer Data Base (NCDB) is a joint project between the Commission on Cancer (CoC) and the American Cancer Society. As the largest cancer registry in the world,[1] it contains data from 1985 to the present, with information from more than 1,500 CoC-accredited hospitals throughout the United States and Puerto Rico. The NCDB is a facility-based clinical surveillance resource that is also meant to be used in quality improvement (QI), allowing hospitals to benchmark their performance on process measures and outcomes against other CoC-accredited hospitals.[2] A prospective database collected by trained data abstractors, it houses reliable, high-quality data representing approximately 70% of all newly diagnosed cancers in the United States.[3] The participant user file, available to investigators associated with CoC-accredited programs by online application, contains deidentified HIPAA-compliant data regarding patient demographics and comorbidity score, hospital characteristics, disease stage, and treatment specifics. Up to 25 "site-specific factors" are listed in each participant user file and contain disease-specific information. For example, the breast cancer file contains "estrogen receptor assay" and "HER2: Immunohistochemistry test interpretation," while the melanoma file contains "vertical growth phase" and "serum [lactate dehydrogenase] LDH." Outcomes recorded include 30-day mortality, 90-day mortality, number of months from diagnosis to last contact or death, and overall "vital status," which conveys whether the patient is dead or alive.

With its emphasis on QI, the NCDB is a valuable tool in assessing hospital adherence to quality measures. For example, to evaluate the quality of lymph node examination

Christina A. Minami, MD

Christina A. Minami, MD, is a General Surgery resident at Northwestern Memorial Hospital and research fellow at the Center for Healthcare Studies at Northwestern University. She is also currently in the process of earning her Masters in Health Services and Outcomes Research at Northwestern and intends to pursue a career in surgical oncology.

The ideas and viewpoints expressed in this commentary are those of the author and do not necessarily represent any policy, position, or program of NCCN.

Karl Y. Bilimoria, MD, MS

Karl Y. Bilimoria, MD, MS, is a surgical oncologist at Northwestern Memorial Hospital and the Robert H. Lurie Comprehensive Cancer Center. He is the Vice Chair for Quality in the Department of Surgery at the Feinberg School of Medicine, Northwestern University and is the Director of the Surgical Outcomes and Quality Improvement Research Center. His research is funded by the National Institutes of Health, the Agency for Healthcare Research and Quality, the American Cancer Society, the National Comprehensive Cancer Network, the American College of Surgeons, the American Board of Surgery, the Accreditation Council for Graduate Medical Education, the California Health Care Foundation, the Health Care Services Corporation, and the Robert H. Lurie Comprehensive Cancer Center of Northwestern University. After completing his General Surgery residency training at Northwestern, he went on to a Surgical Oncology Fellowship at the M D Anderson Cancer Center.

after esophagectomy for cancer, Merkow et al[4] used the database to determine how many patients and hospitals between 1998 and 2007 met the benchmark of examining 15 lymph nodes. That this benchmark was only reached for 28.7% of patients and 7% of hospitals highlighted a national need for QI efforts in this arena.

SEER Program

Although the NCDB only houses data from CoC-accredited hospitals and is focused on oncologic surveillance and QI, the Surveillance, Epidemiology, and End Results (SEER) project was developed by the NCI with the goal of capturing the epidemiology of cancer. Widely available to investigators, SEER is a prospectively maintained region-based database that started with 7 regions in 1973. Since then, it has grown to 17 regional databases that represent large metropolitan areas and, in some cases, entire states. In total, it covers approximately 28% of the US population.[5] Because of this regional approach, SEER provides population-based cancer statistics and is a powerful tool to monitor cancer incidence. SEER also oversamples certain ethnic and racial minorities, allowing researchers to monitor the cancer incidence among specific groups. Patient demographics, disease variables, and treatment codes, all adhering to the same data definitions and coding manual as the NCDB, are available. In addition to date of last follow-up or death and vital status, SEER also provides the ICD-10 code for underlying cause of death, allowing researchers to explore questions regarding disease-specific survival in addition to overall survival.

An excellent use of SEER data can be seen in the recent study by Iqbal et al,[6] which was an important addition to the discussion regarding racial/ethnic differences in cancer biology and resource use. This study focused on the question of whether ethnic differences could be attributed to early detection or intrinsic biologic differences. They found that black women in all age groups were more likely to be diagnosed at a later stage than non-Hispanic white women, and their results suggested that this did not appear to be due to differences in screening trends, but rather due to biologic factors.

SEER-Medicare Data

Linking SEER registry data with Medicare enrollment and claims files, the SEER-Medicare database contains deidentified data on both individuals with cancer and a random 5% sample of Medicare beneficiaries who reside in a SEER area but do not have cancer.[7] Linkages between the 2 databases occur at discrete time points. At the last linkage, which took place in 2014, 93% of patients aged 65 years or older in the SEER database could be matched to the Medicare enrollment file.[8] This unique set of data allows for comparative studies on health disparities, quality of care, and cost of care across the spectrum of cancer diagnosis, treatment, recurrence, and mortality.

Although all data are deidentified, a remote chance of reidentification exists, which means that investigators must obtain approval from NCI to gain access to the dataset. In addition, a fee ($60 to $260) is charged for each file. The NCI tracks all publications using SEER-Medicare and posts these publication statistics online, allowing researchers to easily find studies already performed in their field.

This is a powerful dataset for CER. However, in combining all SEER data with the billing and claims data contained in the Medicare files—including use of hospital services and patters of care in the last year of life—researchers must consider that no data are included for services not covered by Medicare. Research questions must be appropriately designed to work around this limitation.

The potential of research using this database to affect policy was highlighted in a 2011 study that used cancer prevalence and survivor models in SEER and then analyzed net costs through the Medicare linkage data to project the burden of cancer care in the U.S. through 2020.9 The projected cost of care in 2020 was $173 billion,

representing a 39% increase from 2010 costs. This allows formulation of a possible scenario for policy makers focused on cancer funding and resource allocation.

Other Nationwide Databases for Possible Use in Oncologic CER

National Inpatient Sample

Sponsored by the Agency for Healthcare Research and Quality, the National Inpatient Sample (NIS) is the largest all-payer inpatient database, containing data from more than 7 million hospital stays annually.[10] Primary and secondary diagnosis codes, procedure codes, total charges, primary payer, and length of stay data are collected from administrative billing data retrospectively. Although the NIS is a good tool for investigators interested in hospital costs or trends in procedure use, these data are limited to a patient's discrete inpatient stay; thus, studies regarding patient outcomes cannot be performed. Patient-level risk adjustment is also largely impossible given that patient demographics and hospital characteristics are not contained in this dataset.

American College of Surgeons National Surgical Quality Improvement Program

All participating American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) hospitals report prospectively collected data on a random sample of patients undergoing certain operations. These data are abstracted by trained nurse registrars, making for a high-quality, standardized dataset. The ACS NSQIP contains data on patient demographics and comorbidities, operative and anesthesia details, patient laboratory values, and 30-day postoperative morbidity and mortality for all qualifying operations at ACS NSQIP hospitals. Fundamentally a project to improve surgical quality, this database is available to all investigators at ACS NSQIP institutions.

Strengths of Database Studies

These nationwide data sets are an effective way to approach many CER questions, facilitating powerful analyses pertaining to questions of health care delivery, cancer outcomes, and cost. They can also be used to create, evaluate, and refine existing cancer quality measures.

Time Trends

Given that many of the existing databases have reliable data from the 1990s through the present, the data sets provide ample opportunity to map trends over time, not only in cancer incidence and survival but also in treatment and procedural use. Subgroup analyses may guide future interventions targeting specific populations. These studies may have implications not only for data regarding guideline adherence but also for policymakers trying to curb increasing health care costs.

Practice Patterns

Physician practice patterns by region and/or hospital type may highlight disparities in care within oncologic communities. Gaps in quality identified by large database studies can in turn justify qualitative studies examining the barriers and facilitators of health care processes in certain regions or hospitals, and then spur institutional quality improvement projects. This approach represents the ideal way to identify and measure problems, analyze the specific areas of need, and implement a meaningful intervention.

CER of Outcomes

Tracking morbidity and mortality on a large scale is of utmost importance to tracking the quality of cancer care in the United States. Thirty-day complications and outcomes

can be tracked through NSQIP and the NCDB, and long-term survival data are available through SEER, SEER-Medicare, and the NCDB. Outcomes data generated through these retrospective database studies are an important complement to efficacy data generated by RCTs, as they are accessible measures of effectiveness of ever-evolving multidisciplinary treatment trends and practice patterns in the United States.

Limitations of Database Studies

Unfortunately, a number of limitations are inherent in database studies. First, because stringent quality checks are necessary before data release, there is often a time-lag of 1 to 2 years before nationwide data from a given year are widely available. Second, these large databases, although containing considerable study power due to the sheer amount of patient data, can only answer certain broad-based questions of cancer epidemiology, quality, and cost. That is, an investigator will necessarily be limited by the variables contained in a specific database. For example, although a database study may be able to show that specific cancer guidelines are not being followed in the community, the reasons behind this lack of compliance may only become apparent through smaller institutional-based prospective studies or qualitative studies. Third, a database study is only as good as its variables. No matter how well-versed one may be in statistics, each researcher must perform fundamental quality checks before analysis. Familiarizing oneself with the ins and outs of the variables in a given dataset is not simply a matter of reading the data dictionary but also being able to assess the accuracy and quality of data abstraction. Each database is in evolution; each year, variables may be added or taken away, or the coding may be changed. Knowing which version best suits a research question and how to appropriately modify the study period are important to a successful database study.

Conclusions

That databases can be powerful tools in oncologic CER is certain. However, as these types of studies continue to grow in the literature, researchers and clinicians must be aware of their limitations. Although researchers must intimately know their dataset for the purposes of performing high-quality projects, clinicians should also become familiar with the strengths and weakness of these registries to appropriately evaluate and apply the findings of published studies.

References

1. Bilimoria KY, Stewart AK, Winchester DP, et al. The National Cancer Data Base: a powerful initiative to improve cancer care in the United States. Ann Surg Oncol 2009;15:683–690.
2. Mohanty S, Bilimoria KY. Comparing national cancer registries: the National Cancer Data Base (NCDB) and the Surveillance, Epidemiology, and End Results (SEER) program. J Surg Oncol 2014;109:629–630.
3. Cancer Programs, American Cancer Society. National Cancer Data Base. Available at: www.facs.org/cancer/ncdb. Accessed May 13, 2015.
4. Merkow RP, Bilimoria KY, Chow WB, et al. Variation in lymph node examination after esophagectomy for cancer in the United States. Arch Surg 2012;147:505–511.
5. Surveillance Research Program, NCI: Surveillance, Epidemiology, and End Results Program. Available at: seer.cancer.gov. Accessed May 13, 2015.
6. Iqbal J, Ginsburg O, Rochon PA, et al. Differences in breast cancer stage at diagnosis and cancer-specific survival by race and ethnicity in the United States. JAMA 2015;313:165–173.
7. NCI: SEER-Medicare data fact sheet. Available at: http://healthcaredelivery.cancer.gov/seermedicare/overview/seermed_fact_sheet.pdf. Accessed May 13, 2015.
8. Warren JL, Klabunde CN, Schrag D, et al. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. Med Care 2002;40(8 Suppl):IV-3–18.
9. Mariotto AB, Yabroff KR, Shao Y, et al. Projections of the cost of cancer care in the United States: 2010-2020. J Natl Cancer Inst 2011;103:117–128.
10. AHRQ. Overview of the National (Nationwide) Inpatient Sample. Available at: www.hcup-us.ahrq.gov/nisoverview.jsp. Accessed May 13, 2015.