NCI SEER Public-Use Data: Applications and Limitations in Oncology Research

March 18, 2009
By James B. Yu, MD, MHS [1] and Benjamin D. Smith, MD [2]

The Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute (NCI) collects cancer survival and incidence information from population-based cancer registries, encompassing 26% of the US population.[1]

The Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute (NCI) collects cancer survival and incidence information from population-based cancer registries, encompassing 26% of the US population.[1] Over the past 3 decades, the SEER program database has become an internationally recognized cancer resource, increasingly utilized in the study of cancer epidemiology and outcomes. In this commentary, we will discuss applications and limitations of the SEER public-use database, to help clinicians interpret the many studies that are generated from this database, and to help clinical investigators implement future studies using this valuable national resource.

Background
Originally created in 1973 from two earlier NCI programs—the End Results Program and the Third National Cancer Survey,[2] the SEER program began collecting data from the states of Connecticut, Iowa, New Mexico, Utah, and Hawaii, and the metropolitan areas of San Francisco/Oakland, Seattle, Detroit, and Atlanta. These are considered the “original 9” SEER registries. In 1978, 10 predominantly black counties in rural Georgia were added, and American Indians in Arizona were added in 1980. New Orleans, Louisiana, (1974–1977, rejoined 2001), New Jersey (1979–1989, rejoined 2001), and Puerto Rico (1973–1989) were added prior to 1990.[3] Metropolitan Los Angeles County, and four counties in the San Jose/Monterey area were added in 1992, to increase coverage of minority populations, especially individuals of Hispanic origin. In 2001, the SEER program expanded to include Kentucky and the remainder of California, and New Jersey and Louisiana rejoined the registry.[3]
The SEER data are broadly representative of the US population, although there are some differences. Demographically, the population of patients recorded in the SEER database are more likely to be foreign born compared to the standard US 2000 population (17.3% vs 11.3%), and are more often urban inhabitants (88.2% vs 79%). There is also a higher proportion of the US Native Hawaiian/Pacific Islander (69.8%), Asian (53.3%), American Indian/Alaska Native (42.2%), and Hispanic populations (40.4%) covered compared to US white (23.4%) and US black (22.7%) populations in the SEER database.[3] Nonetheless, due to its large size and long follow-up, the SEER program database continues to be studied as an accurate representation of the US cancer population as a whole.

Table 1

General Description of SEER-Recorded Variables

<table>
<thead>
<tr>
<th>Registry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marital status</td>
</tr>
<tr>
<td>Race/Ethnicity</td>
</tr>
<tr>
<td>Spanish/Hispanic origin</td>
</tr>
<tr>
<td>Gender</td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Birthdate and birthplace</td>
</tr>
<tr>
<td>Date of diagnosis</td>
</tr>
<tr>
<td>Primary site, grade, and histology</td>
</tr>
<tr>
<td>Whether first malignancy, and sequence number of reported malignancy</td>
</tr>
<tr>
<td>Extent of disease at the time of diagnosis (variable detail)</td>
</tr>
<tr>
<td>Extent of lymph node dissection and pathology (variable detail)</td>
</tr>
<tr>
<td>Tumor markers and additional pathologic and clinical information (variable detail)</td>
</tr>
<tr>
<td>Staging (variable detail)</td>
</tr>
<tr>
<td>Extent of surgery (variable detail)</td>
</tr>
<tr>
<td>Radiation (broad terms)</td>
</tr>
<tr>
<td>Radiation (CNS: yes/no)</td>
</tr>
<tr>
<td>Date of death, and cause of death</td>
</tr>
</tbody>
</table>

CNS = central nervous system;
SEER = Surveillance, Epidemiology, and End Results.
Quality control is an important aspect of the SEER program. Registries are routinely audited for data accuracy, and a Data Quality Profile (DQP) is generated for each SEER registry. Individuals and registries that do well in reliability studies are identified and rewarded.[3] In addition, the NCI SEER program performs regular education and training programs in coordination with the National Cancer Registrars Association annual meeting. Registrars are tested via Web-based reliability studies, and audits of high-volume facilities are performed, to ensure that case ascertainment is complete and timely. As a result of these efforts, the SEER program has become the standard for data quality among international cancer registries.[3]

The SEER program regularly publishes national cancer statistics reviews and monographs as a part of its mission. Selected SEER publications can be accessed at [http://seer.cancer.gov/publications](http://seer.cancer.gov/publications), and include the Annual Report to the Nation on the Status of Cancer as well as Racial/Ethnic Patterns of Cancer in the United States. In addition to these statistical reviews, the SEER database is a rich resource for independent clinical researchers. With the increasing availability of easy-to-use statistical software and awareness of the SEER database, an increasing number of SEER-based reports have been published over the past decade (Figure 1), with over 400 peer-reviewed publications issued in 2008 alone.

**Recorded Variables**

Recorded in the SEER database are demographic variables and information describing the stage, extent of surgery, pathologic findings, whether or not radiation therapy was given, and the cause of death of patients with cancer (Table 1). A full list of variables can be found online.[4] More detail has been recorded in recent years than in the past. For example, for all years (1973+), stage at diagnosis is broken up into five main categories—in situ, localized, regional, distant, or unstaged. However, since 2004, most primary cancer sites have additional TNM staging data based on the AJCC Cancer Staging Manual, 6th edition.

In addition, since 2004, the Collaborative Staging Codes have been reported, offering many useful clinical and pathologic details. For example, reports since 2004 include information such as whether extracapsular extension is present pathologically or if nodes are described as “fixed” for head and neck cancers, as well as specific cervical lymph node levels of involvement.[5] Since 2004, estrogen receptor (ER), progesterone receptor (PR), and HER2 receptor status are reported for breast cancer, and pretreatment clinical variables such as prostate-specific antigen (PSA) and clinical stage are also reported for prostate cancer. Unfortunately, pathologic margins of resection continue to not be
Surgical detail also varies, with more recent years including more specific details. For example, records of patients who underwent skin surgery from 1983 to 1999 do not distinguish between Mohs surgery and other types of wide excision, whereas since 2004, data distinguishing Mohs surgery with a 1-cm margin or less vs Mohs surgery with a greater than 1-cm margin have been included. Although specific details about the surgery are still lacking (for example, whether a prostatectomy was laparoscopic or open), this additional level of detail is helpful in improving the fidelity of clinical observations. Unfortunately, the intent of the surgeon is not recorded, and therefore whether a patient receives surgery for palliative intent as opposed to curative intent is at times unclear.

The SEER database has been linked to Medicare billing data provided by the Centers for Medicare and Medicaid Services (CMS). As a result, much more detailed patient information is found in this much larger “SEER-Medicare” database. The addition of Medicare claims data allows for a more in-depth study of investigating cancer outcomes in patients age 65 and over (and in Medicare beneficiaries with disability or end-stage renal disease). A review of the applications and limitations of the SEER-Medicare linked data is beyond the scope of this commentary, and we will focus on the SEER “public-use” data for the remainder of this discussion. The methodology of studies involving SEER-Medicare linked data has been previously discussed in depth in an excellent supplement to the journal Medical Care in 2002.[6-18]

**Potential Research and Applications**

Many SEER investigations have centered around the demographic and epidemiologic description of rare malignancies.[19-22] Given the large size of the database and long period of record, the SEER data are very useful for describing the demographics and outcomes of rare malignancies. This gives valuable information about patients with rare diseases to clinicians, who would otherwise have to rely on case reports and small single-institution series. Temporal and regional trends in diagnosis, stage migration, and survival, and the frequency of different histologies, stages, interventions, and demographics are common avenues of study. For example, investigators were able to discover that ER/PR status is not related to overall survival in male breast cancer patients.[22] This finding would otherwise have been impossible to detect with smaller registries or retrospective series. Other examples of such studies are shown in Table 2.
Second malignancies have also been a rich area of investigation.[23-28] Because of the SEER database's large size and long follow-up, sufficient numbers of patients are included for investigating these otherwise rare diseases. For example, patients who have undergone radiotherapy for their first cancer of the head and neck (vs no radiotherapy as a part of initial cancer care) are less likely to have a subsequent head and neck cancer.[28] Further examples of such work are shown in Table 3.

Another important area of study using the SEER database is the investigation of national and regional trends in the delivery and outcomes of cancer care.[29-32] State registries can be compared to the national SEER cohort to benchmark a state’s cancer incidence rate or stage at presentation, and highlight areas where statewide intervention is needed. National trends can be compared to international registries to highlight US cancer rates in generalized ecologic studies. Whether progress is being made in the overall cure and early diagnosis of cancer over the years for different disease sites and demographics can be investigated. One example of such a study investigated Floridians compared to non-Floridians, and found that Floridians had higher lung cancer rates compared to the US population.[31] This information could potentially be used to highlight a need in Florida for improved public health awareness regarding tobacco. Examples of other studies are shown in Table 4.

The development of prognostic models predicting a person’s overall survival based on initial demographic and clinicopathologic variables are an active area of study. The internal and external validity of these models require robust numbers of patients across the spectrum of presentations for the disease being investigated. Because of its ability to provide large sample sizes, the creation and validation of clinical models is possible in a national setting.[33-35] For example, investigators from the M.D. Anderson Cancer Center were able to construct a useful nomogram predicting breast cancer mortality based on age, hormone-receptor status, tumor grade and size, and the number of nodes harvested.[35] Three examples of these studies are also shown at the bottom of Table 4.
Hypothesis generation, for example, regarding survival outcomes after treatment, is an active area of research. Where clinical trials are not forthcoming, either because of the rarity of the tumor, or when logistical reasons or patient and clinician bias preclude the ability to mount an effective clinical trial, the SEER database has been used to generate hypotheses regarding the overall survival efficacy of one treatment over another. Examples include investigation of the use of radiation[36-38] or radical surgery.[39,40] For example, Kapp et al found no association between oophorectomy and an improvement in overall survival in patients with uterine leiomyosarcoma.[40] The hypothesis generated is that perhaps ovarian preservation could be appropriate in patients with uterine leiomyosarcoma.[40] Examples of hypothesis-generating studies are listed in Table 5.

**Limitations of SEER-Based Research**

While the SEER database is an extremely valuable tool for clinical cancer research, several limitations should be taken into account when interpreting results from a SEER observational study, especially when there is an attempt to generate hypotheses regarding adjuvant therapy. Many of the limitations revolve around underreported and incomplete data regarding adjuvant therapy,
unrecorded variables, variations in data reporting, migration of patients in and out of SEER registry areas, and selection bias.

**Missing Data**

First, there appears to be underreporting of the use of radiation therapy for some disease sites. An investigation of the California Cancer Registry, which is one of the state registries that contributes to the national SEER data, revealed that there was significant underreporting of radiation therapy for breast cancer, with a registry sensitivity of 72.2%.[41] The authors suggested augmenting registry data with Medicare claims data as a method for improving the accuracy of observations. The underreporting of radiation therapy is partly due to the delivery of radiation as an outpatient therapy, which means that it is less likely to be captured by hospital-based cancer registries than inpatient therapy. In addition, for disease sites where prolonged chemotherapy is often delivered prior to radiotherapy (as in breast cancer), radiotherapy can be delivered beyond the time period of initial data abstraction, causing further underreporting. This underreporting of radiation therapy potentially increases the uncertainty surrounding hypotheses regarding radiotherapy generated from the SEER database.

Second, there is a lack of information regarding radiation fields, doses, and intent. The quality of radiation therapy has an impact on overall treatment outcomes, and inadequate radiation dose and treatment fields, or inappropriately administered radiotherapy, could reduce the apparent efficacy of radiation overall.

Third, information regarding some important aspects of cancer therapy and prognosis is lacking. There is no information regarding hormonal therapy for breast and prostate cancer, and no information on chemotherapy contained in the SEER database. There is also a lack of information on comorbidity. Receipt of chemotherapy and the presence of comorbid illness may be determining using billing claims included in the SEER-Medicare linked database,[13,48] but a thorough description of these techniques is beyond the scope of this paper. Whether positive or negative margins were obtained at the time of surgery is unrecorded. The presence of positive pathologic margins after attempted complete surgical resection has been shown to be an important prognostic factor in many cancers, and is an important consideration in investigating the benefit of any adjuvant therapy.

**Coding Reliability**

Reliability in coding for rare histologies is variable. For example, in an expert review of cases of non-Hodgkin lymphomas, agreement in the subclassification of histologies between the expert review and the SEER registry record varied from 5% to 100%.[42] In comparison, the reliability of Hodgkin disease diagnosis appeared good overall, though for older women, the SEER data slightly overestimated incidence in comparison to expert review.[43] Independent review of large-cell carcinoma of the lung in the Iowa Cancer Registry (one of the SEER cancer registries) noted a low sensitivity (21.9%, 95% confidence interval [CI] = 9.3%-40.0%) and positive predictive value (23.3%,
95% CI = 9.9%–42.3%) using consensus independent review as the reference diagnosis for analysis.[44] In comparison, small-cell carcinoma of the lung had excellent agreement.

**Patient Migration**

The migration of patients in and out of SEER registry geographic catchment areas is an important limitation, especially with regard to the investigation of second malignancies. A patient moving from a SEER registry region (for example from Connecticut to Florida), would be lost from the database, and any subsequent second malignancies would be unrecorded. This migration leads to a relative undercounting of second malignancies. In addition, the measurement of other outcomes dependent on long-term follow-up is less reliable because of the increasing likelihood of migration or losing the patient to follow-up.

**Selection Bias**

Finally, selection bias plays a large role in the evaluation of overall survival after cancer-directed therapy. A recent study investigating the association of adjuvant radiation therapy with survival in Merkel cell carcinoma of the skin found that there was a longer overall survival associated with the use of adjuvant radiation therapy vs no radiation therapy.[45] This study was criticized because it included patients who survived less than 4 months, potentially biasing results in favor of radiotherapy. Patients who die soon after cancer-directed surgery are less likely to receive radiation therapy. These immediate postsurgical deaths could potentially cause the nonradiation reference arm to have a lower relative survival in comparison to the radiation arm of an observational study.[46]

A similar discussion occurred regarding the association of adjuvant radiation therapy after surgery with improved survival for patients with gallbladder cancer.[47,48] The optimal cutoff for exclusion of perioperative death is unclear, and this remains one of the uncertainties that need to be wrestled with in past and future investigations of adjuvant therapy.

**An Illustrative Study**

An excellent illustrative study by Giordano et al,[49] highlights the inherent limitations found in any survival analysis of observational data. Although this study used the SEER-Medicare linked database, its conclusions extend to the SEER public-use data as well. The authors found that improbable results could be produced when investigating the SEER-Medicare data because elderly patients who underwent adjuvant therapy for prostate cancer and colon cancer were healthier than those who did not receive adjuvant therapy, and were even healthier than members of the general US noncancer population.

For example, as expected, active therapy for patients with prostate cancer (compared to observation) was associated with an improved cause-specific survival. However, active therapy was also associated with an improved other-cause mortality improvement. Patients who received active prostate cancer therapy were less likely to also die of cardiovascular disease, chronic obstructive pulmonary disease, and diabetes compared to patients who underwent observation, and also compared to patients from the general US noncancer population. Obviously, active prostate cancer therapy is not also a therapy for diabetes and the difference in other-cause survival is due to the selection bias of treating healthier patients with active therapy. The authors concluded that any observational studies that compare survival outcomes of different therapies (and with observation) should be viewed with caution, and that noncancer mortality, in addition to cancer-specific and all-cause mortality, should be reported.

Selection bias cuts both ways, of course, and perhaps in some disease groups, a patient selected for adjuvant therapy has a worse prognosis. For example, a patient undergoing radiotherapy within 2 months after resection of a pancreatic head mass for palliative intent because of rapidly recurrent disease would theoretically be placed in the same treatment group as a patient who underwent a resection with negative margins and subsequently underwent radiation therapy because of that patient's otherwise excellent performance status and desire for aggressive therapy.

**Conclusions**

The NCI SEER database is an excellent, increasingly utilized resource for clinical cancer investigation. The database is useful in the investigation of rare diseases, second malignancies, and national demographic differences in the diagnosis, treatment, and outcome of cancer. Regional trends in the diagnosis and treatment of cancer are other areas that can be studied using the SEER database, and hypotheses regarding national trends can be developed. Hypotheses regarding the benefits of cancer treatment can be generated, with the understanding that inherent limitations in the SEER database should be acknowledged.

The issue of selection bias is a difficult subject to negotiate in the study of any observational database. Statistical techniques such as propensity score analysis[50,51] and instrumental variable...
analysis[52,53] have been used in attempts to overcome this selection bias, but the gold standard for clinical evidence remains the double-blinded, randomized controlled trial. Nevertheless, there are clearly clinical situations and cancers that are logistically impossible to study in randomized clinical trials. Through judicious and careful investigation, new clinical advances and insights continue to be made using this important national resource.


Source URL:

Links: