Guide to Statistics and Methods

Practical Guide to Surgical Data Sets: Surveillance, Epidemiology, and End Results (SEER) Database

Kemi M. Doll, MD, MSCR; Alfred Rademaker, PhD; Julie A. Sosa, MD

Introduction

The Surveillance, Epidemiology, and End Results (SEER) database is a publicly available, federally funded cancer reporting system that represents a collaboration between the US Centers for Disease Control and Prevention, the National Cancer Institute, and regional and state cancer registries.[1] SEER data are national, with information from 18 states that represent all regions of the country. In contrast to other commonly used data sets (eg, the National Cancer Data Base), SEER is population-based, because local registries report information for all cancer cases within a specific region and/or defined racial/ethnic population. Given that SEER data is both a cancer reporting system and a research tool, we aim to present salient aspects of these data, strengths and limitations for analyses, and important statistical considerations.

Editorial pages 505 and 508

Data Considerations

Data Sources

SEER data are gathered at the local level. Trained registrars collect data from all clinical settings that diagnose or treat cancer and include patients of all ages, regardless of insurance status. Dates and causes of death come from death certificates, and mortality statistics are calculated using data from the US Census Bureau (**Table**). SEER data captures 28% of the US population; because of its targeted sampling strategy, it includes a high proportion of racial/ethnic minorities, foreign-born individuals, and those with income below the federal poverty line.

Time Trend Data

The SEER program originated in 1974, so it can be used to study trends in cancer incidence, prevalence, and survival in the United States over time. The addition of SEER registries since 1974 has resulted in numbered cohorts (eg, SEER-9 from 1974, SEER-13 from 1992, and SEER-18 from 2000). Trend studies should be restricted to a consistent SEER cohort for all years of the analysis to avoid shifts in base populations that create erroneous findings.

Cancer Data

Stage and histologic details are reported for all cancers, allowing for specific subpopulations and rare cancers to be studied. Unique to SEER is a variable termed *Summary Stage*, defined for each cancer site (local, regional, distant, and unknown) in manuals published online. Given the longevity of SEER data collection, shifts in stage classifications over time should be accounted for in time trend studies, using stratification or manually recoding for consistency. American Joint Committee on Cancer stage is available, usually for patients with summary stages reported. The *Collaborative Stage* variables (**Box**) for each cancer are site-specific factors that range from serum tumor markers (eg, cancer antigen 19-9) to diagnostic details (eg, number of prostate biopsy cores). Missingness, quality, and the time when each variable was introduced into the data set vary. (For example, in breast cancer, although HER2 laboratory test results are available for 76% of patients since 2010, they are often inconsistent with the HER2 status variable and therefore should not be used in analysis.[2]) Multiple imputation is a recommended method of accounting for variables with a high proportion of missing values, such as estrogen receptor status in breast cancer over time.[3]

Treatment Data

SEER data report receipt of surgery and radiation, and treatment sequence is captured such that analysis of treatment trends by specific histologic indications can be performed. For example, Ko et al[4] measured the use of adjuvant radiation therapy for high-to-

Table. Overview of the Surveillance, Epidemiology, and End Results Database

Type	Included in SEER	Not Included in SEER
Sociodemographic factors	Age at diagnosis, year of birth, race/ethnicity, sex, census tract education, census tract income, marital status, place of birth	Individual income, family income
Geographic variables	County and state of residence, originating SEER registry, urban/rural designation	Zip codes, site of treatment
Clinical factors	Prior cancer history	Comorbidity, functional status, medications
Cancer specific factors	Site, laterality, stage,[a] grade, lymph node status, extent of disease,[b] tumor markers[b]	Depending on the cancer site, information may be missing to varying degrees.
Pathologic variables	Lymphovascular invasion, perineural invasion, margin status	Pathologic variables collected vary by cancer site.
Treatment factors	Method of diagnostic confirmation, receipt of surgery, extent of surgery,[b] receipt of radiation, order of treatment	Clinician information, surgical approach, radiation dose, chemotherapy, hormonal therapy, immunotherapy
Outcomes	Date of death, cause of death	Cancer recurrence

Abbreviation: SEER, The Surveillance, Epidemiology, and End Results.

[a] These data points are SEER summary stages; American Joint Committee on Cancer Tumor, Nodes, and Metastases classification system was put in place starting in 2004.

[b] These data points are specific to certain cancer sites.

intermediate–risk endometrial cancer after 2 national clinical trials. It is more difficult to study treatment outcomes and perform comparative effectiveness research in SEER. Important details, such as comorbidity, intent of surgery (cure vs palliation), surgical route (minimally invasive vs open approaches), radiation dosing, and other treatments (eg, chemotherapy, hormonal therapy, or immunotherapy) are absent. The inability to address the influence of these missing variables on outcomes makes comparative effectiveness analyses prone to unmeasured confounding. Using the SEER-Medicare linked database can address this, but largely in adults 65 years and older.

Statistical Considerations

SEER data are available in 2 ways: (1) a binary format for which SEER*Stat software can be used to perform common but limited analyses; or (2) as text data that can be directly imported into external statistical software for more complex projects.[1] For incidence and mortality rates, results should be age-adjusted and reported as cases per 100 000 person-years. A trend analysis evaluates how rates change over time by comparing the annual percent change in rates using standard *t* or rank sum tests. A modeling strategy (eg, log-linear regression) can then be used to calculate the rate of change and generate illustrative graphics. The addition of joinpoint regression[5] can pinpoint years that demonstrate the most dramatic changes, as in a study by Lim et al[6] for thyroid cancer rates from 1974 through 2013.

Population-level survival statistics can be reported as relative survival (the ratio of overall survival of patients with the disease to the expected survival in a comparable cohort of the general population) or cancer-specific survival (the proportion of patients alive with a specific disease). Which to use depends on how best to limit bias for the population in question. Relative survival, which is based on the overall survival of patients with the disease, is less accurate for cancers for which patients commonly have other serious comorbidities (eg, lung cancer) because the competing mortality risks from these comorbidities are not taken into account. Cancer-specific survival is less reliable in cases of multiple primary cancers because of difficulty in identifying accurate causes of death from death certificates.[7] Cox proportional hazard models can be used to calculate how demographic factors and prognostic differences influence individual mortality. Overall, missing clinical data mean that comparative effectiveness research using SEER data alone should be undertaken with caution, given the limited ability to account for important clinical differences between treatment groups.

Box. Details of SEER Data

1. SEER is a nationally representative, population-based cancer reporting system that includes all cancer cases within specific US geographic regions.
2. Longitudinal trends in cancer diagnosis, treatment, and survival can be analyzed starting from 1974 to the present.
3. The SEER data are particularly well suited for longitudinal studies on specific subpopulations and rare or indolent cancer types.
4. The Collaborative Stage Data Collection system can be used to gather additional site-specific prognostic and treatment details for individual cancer sites.
5. Care should be taken to document and account for changes in staging classifications over time.
6. Comparative effectiveness analyses are limited by lack of information on comorbidity, recurrence, and chemotherapy.

Abbreviation: SEER, Surveillance, Epidemiology, and End Results.

Conclusions

SEER is a long-established resource that allows for population-based surveillance and analysis of all cancers in the United States. Excellent uses of SEER include epidemiologic studies of incidence, prevalence, and mortality rates over time, shifting treatment patterns between surgery and radiation, and quantifying diagnostic and treatment patterns by geographic and demographic factors.

ARTICLE INFORMATION

Author Affiliations: Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of Washington, Seattle (Doll); Seattle Cancer Care Alliance, Seattle, Washington (Doll); Department of Preventive Medicine, Northwestern University, Chicago, Illinois (Rademaker); Department of Surgery, Duke University Medical Center, Durham, North Carolina (Sosa); Department of Medicine, Duke University Medical Center, Durham, North Carolina (Sosa); Duke Cancer Institute, Durham, North Carolina (Sosa); Duke Clinical Research Institute, Durham, North Carolina (Sosa).

Corresponding Author: Kemi M. Doll, MD, MSCR, Department of Obstetrics and Gynecology, University of Washington, 1939 NE Pacific St, PO Box 356460, Seattle, WA 98195 (kdoll@uw.edu).

Published Online: April 4, 2018. doi:10.1001/jamasurg.2018.0501

Conflict of Interest Disclosures: Dr Doll receives research support from the National Comprehensive Cancer Network Foundation through a grant supported by Pfizer. Dr Sosa is a member of the data monitoring committee of the Medullary Thyroid Cancer Consortium Registry, which is supported by NovoNordisk, GlaxoSmithKline, AstraZeneca, and Eli Lilly. No other disclosures are reported.

REFERENCES

1. National Cancer Institute. Surveillance, Epidemiology, and End Results program website. http://www.seer.cancer.gov. Published 2018. Accessed February 26, 2018.

2. Howlader N, Chen VW, Ries LA, et al. Overview of breast cancer collaborative stage data items—their definitions, quality, usage, and clinical implications: a review of SEER data for 2004-2010. *Cancer*. 2014;120(suppl 23):3771-3780.

3. Krieger N, Jahn JL, Waterman PD. Jim Crow and estrogen-receptor-negative breast cancer: US-born black and white non-Hispanic women, 1992-2012. *Cancer Causes Control*. 2017;28(1):49-59.

4. Ko EM, Funk MJ, Clark LH, Brewster WR. Did GOG99 and PORTEC1 change clinical practice in the United States? *Gynecol Oncol*. 2013;129(1):12-17.

5. Dehkordi ZF, Tazhibi M, Babazade S. Application of joinpoint regression in determining breast cancer incidence rate change points by age and tumor characteristics in women aged 30-69 (years) and in Isfahan city from 2001 to 2010. *J Educ Health Promot*. 2014;3:115.

6. Lim H, Devesa SS, Sosa JA, Check D, Kitahara CM. Trends in thyroid cancer incidence and mortality in the United States, 1974-2013. *JAMA*. 2017;317(13):1338-1348.

7. Sarfati D, Blakely T, Pearce N. Measuring cancer survival in populations: relative survival vs cancer-specific survival. *Int J Epidemiol*. 2010;39(2):598-610.