Common Pitfalls of Head and Neck Research Using Cancer Registries

Evan A. Jones1, Andrew G. Shuman, MD2, Brian L. Egleston, MPP, PhD3, and Jeffrey C. Liu, MD1,4

Sponsorships or competing interests that may be relevant to content are disclosed at the end of this article.

Abstract
Objective. To highlight common pitfalls observed in scientific research derived from national cancer registries, predominantly the Survival, Epidemiology, and End Results Program and the National Cancer Database.

Data Sources. Literature review and expert opinion.

Review Methods. This state-of-the-art review consolidates the literature with editorial experiences describing how and why statistically flawed studies are usually rejected for publication, highlighting common errors in submitted articles employing national cancer registries.

Conclusions. Pitfalls were identified in 2 major areas—design and data analysis. Design pitfalls included unbalanced cohorts, uncontrolled covariates, and flawed oncologic variables. Analytical pitfalls included incorrect application of univariate analyses, inclusion of inaccurate data, and inclusion of stage IVc disease in curative survival analysis. Additional limitations of database studies were identified, including absence of patient-related outcomes, hypothesis-generating vs practice-changing implications, and inability to differentiate between overall survival and disease-specific survival.

Implications for Practice. Methodological strategies are suggested to ensure careful analytical design and appropriate interpretation. Although national cancer registries provide a wealth of data, researchers must remain vigilant when designing studies and analyzing these data sets. Inherent design flaws raise considerable problems with interpretation; however, when analyzed judiciously, registries can lead to a better understanding of cancer outcomes.

Keywords
NCDB, SEER, cancer registry, head and neck cancer, study design, pitfalls, data analysis, editorial experiences, statistical flaws, unbalanced, errors, submission, guidelines

Received June 12, 2018; accepted February 28, 2019.

National cancer registries afford researchers the ability to analyze large cohorts of populations with malignancies, an invaluable tool in examining patterns of cancer care delivery and outcomes.1 The 2 registries most frequently used to evaluate head and neck cancers in the United States are the National Cancer Database (NCDB) and the Survival, Epidemiology, and End Results Program (SEER).2,3 While these databases provide a wealth of information, meaningful analysis requires appropriate research study design and analysis. There has been significant growth in such cancer database studies in the past few decades given their ease of access.4,5 SEER-based publications have grown from the single digits per year in the 90s to now almost 200 per year (Figure 1). Similarly, there are now 50 to 100 publications per year using the NCDB.6

Given the large amount of data, a flawed analysis easily leads to statistically significant but misleading, inaccurate, or false conclusions. As prior training, mentorship, or specific qualifications are not explicitly required for access to NCDB or SEER, research with this resource by novice and inexperienced researchers occasionally generates flawed studies for consideration. This article serves to highlight some of the most common flaws seen in rejected manuscripts, indicating common blind spots in such database research. Because of the vast array of potential studies from these data sets, a single publication could not address all the potential challenges in an analysis plan. We hope this article will serve to highlight common pitfalls in such research and

1Department of Otolaryngology, Lewis Katz School of Medicine of Temple University, Philadelphia, Pennsylvania, USA
2Department of Otolaryngology–Head & Neck Surgery, University of Michigan, Ann Arbor, Michigan, USA
3Biostatistics and Bioinformatics Facility, Fox Chase Cancer Center, Temple University Health System, Philadelphia, Pennsylvania, USA
4Department of Surgical Oncology, Fox Chase Cancer Center, Philadelphia, Pennsylvania, USA

This article was presented at the AAO-HNSF 2018 Annual Meeting and OTO Experience; October 7-10, 2018; Atlanta, Georgia.

Corresponding Author:
Jeffrey C. Liu, MD, Department of Otolaryngology, Lewis Katz School of Medicine of Temple University, 3440 N. Broad Street, Kresge West, 3rd Floor, Philadelphia, PA 19140, USA.
Email: Jeffrey.Liu@temple.edu
the total US population, accounting for about a quarter of
lish cancer incidence and survival data representing 28% of
Twenty cancer registries across the nation collect and pub-
database to reduce the cancer burden of the US population.
Institute (NCI). Its purpose is to provide an epidemiologic
cancer registries founded in 1973 by the National Cancer
Registrars Association. Overall, the NCDB creates a power-
accuracy of abstracted data, CoC-accredited institutions
ment, adjuvant treatments, complications, and outcomes.9
Notably, the NCDB does not include cause of death, func-
tional outcomes, or patient-reported outcomes. To ensure
accuracy of abstracted data, CoC-accredited institutions
employ tumor registrars certified by the National Cancer
Registrars Association. Overall, the NCDB creates a power-
ful research tool for investigators and a quality improvement
gauge for CoC-accredited hospitals.

The SEER Program is a nationwide collaboration of
cancer registries founded in 1973 by the National Cancer
Institute (NCI). Its purpose is to provide an epidemiologic
database to reduce the cancer burden of the US population.
Twenty cancer registries across the nation collect and pub-
lish cancer incidence and survival data representing 28% of
the total US population, accounting for about a quarter of
newly diagnosed cancer cases.2 The state and regional
cancer registries were selected based on functional capacity
of operating a high-quality, population-based cancer report-
ing system in an area with significant racial and ethnic
minority subgroups. Racial and ethnic minorities are inten-
tionally overrepresented to gather more information on tra-
ditionally understudied subgroups; however, the sheer
power of the database enables the studied population to rep-
resent the country accurately.7,10 SEER data are reported by
certified cancer registrars into the SEER Data Management
System, which is a centrally designed system used by
cancer registries. Like the NCDB, the SEER database has
its shortfalls in completeness of data. Among the informa-
tion not captured by SEER is patient quality of life and per-
formance, health behaviors such as smoking or alcohol use,
adverse events without medical care, and explicit notation
of recurrence or metastasis after incident diagnosis.11

Although the NCDB and SEER are similar, the designs
of the databases dictate their intended utility. NCDB is
based on a convenience sample of patients at CoC-
accredited hospitals to be used as a surveillance tool and a
quality improvement resource, whereas SEER provides
population-based cancer statistics to more accurately mea-
sure cancer incidence and trends.7 While the NCDB and
SEER are the largest and most common cancer databases
used in cancer research and are the focus of this article,
other databases may also be relevant to this review.
Research using sufficiently large populations such as state,
payer, or other databases offers research opportunity but is
at similar risk for analytical pitfalls.12

Methods

While manuscripts submitted for publication in
Otolaryngology–Head and Neck Surgery are declined for
many reasons, it was noted by 1 associate editor that studies
employing large cancer databases were frequently rejected
for similar reasons. A list of common reasons for rejection
was composed and discussed with another associate editor,
who also regularly reviews oncologic manuscripts. Common
themes were developed. Comments from a biostatistician on
common statistical pitfalls were solicited.

Relevant publications were identified along these themes
from PubMed. Articles were then reviewed for relevance,
timeliness, and applicability to the review. Articles selected
for inclusion commented on the utility and characteristics of
using cancer databases in research, as well as potential
errors. The National Cancer Institute’s SEER website and
the American College of Surgeons’ NCDB website also
were reviewed for this study.2,3

Discussion

Design Pitfalls

A good study starts with thoughtful design and a cohesive
underlying hypothesis. Without an underlying hypothesis,
studies become descriptive and, occasionally, fishing stud-
ies. While descriptive studies may have value (eg, reporting
on rare diseases), a defined hypothesis significantly

Figure 1. Number of publications in the Survival, Epidemiology,
and End Results Program (SEER) by year (adapted from https://
improves a study by framing the data clearly, focusing the analysis, guiding direction for further research, and providing overall clarity.

“Fishing” studies are a common pitfall in large cancer database studies. The immense amount of data and variables facilitates extensive statistical analysis. Documentation of an a priori plan through predata analyses is often evidence of a well-thought-out study. A useful method of documenting that a hypothesis was specified a priori is including it in an institutional review board (IRB) approval document prior to conducting the study. While the NCDB database is de-identified, some SEER-related data releases might still contain some indirect identifiers (eg, individual ages above 90 years) and hence fall under the “limited data set” definition.\textsuperscript{13} Such SEER releases may require a data use agreement and IRB approval.

However, without a documented a priori plan, false or meaningless associations may arise. As is often said, correlation does not imply causation. Spurious relationships of data, many humorous, have been lampooned on the Internet.\textsuperscript{14} To resist the temptation of stretching associations in large cancer registries, defined hypotheses and appropriate statistical modeling should be established before analyzing data.

With a defined hypothesis, the next step is appropriate study design. Pitfalls of study design can occur in selecting patient cohorts, survival analysis, and introduction of bias from tumor biology. Careful selection of the cohort for analysis must be performed to limit sources of bias. In selecting the patient cohort, unbalanced demographics can introduce bias. For example, when studying outcomes of head and neck squamous cell carcinoma (HNSCC), older patients will invariably have worse overall survival than younger patients. Differences in male/female ratio in the cohort may affect measured end points. Unbalanced racial or socioeconomic variables may also introduce bias from differing access to care. When studying variable X in an HNSCC cohort, differences from these unbalanced demographics, rather than the actual variable of interest, may underlie a statistically significant finding.

Knowing subtleties of the data sets is important in cohort selection. SEER data contains information on multiple primaries from the same individual, for example. In conducting survival analyses, one might determine which record should be used for analysis so that the same individual does not provide multiple survival times. Cases diagnosed only at autopsy also might not be useful for many analyses.

As most studies evaluate survival as an end point, uncontrolled tumor or treatment factors can also introduce bias. Stage imbalance is a common pitfall seen when reporting survival. For example, in a study examining if variable X affects survival, 2 cohorts may be analyzed and show a survival difference. However, review of the cohort demographics will show a disproportionate proportion of early stage cancers in 1 cohort, thereby explaining the survival difference. Another example is imbalance in treatment. The 2 cohorts evaluated by variable X may have similar stages of cancer, but 1 group has a disproportionate percentage of adjuvant radiation after surgery. The survival difference may be due to differing adjuvant treatment frequency rather than variable X. The proper balancing of cohorts is paramount to the validity of good study design and reducing bias.

In these examples, the issue is summarized as a confounding variable problem (Figure 2). The independent variable, variable X, and dependent variable, survival, are not related but are influenced by the third variable (eg, race, stage, age, adjuvant treatment). The finding of a statistical association is spurious.

There are multiple strategies to address these problems. Good study design with stringent, detailed inclusion criteria will help balance variables. For example, in studying survival based on variable X, limiting the study to only advanced-stage patients eliminates bias from unbalanced early stage cancers. Another strategy is using matched analysis. By selecting variables for control in the experimental arm, matched controls can be selected and used for a summary analysis. Given the substantial number of cases often present in database studies, adequate cohort size is usually retained after applying stringent inclusion criteria. Further statistical approaches for appropriate bias reduction include propensity score–based weights, Bonferroni correction, and regression adjustment.

Another consideration when examining survival and curative intent outcomes is the appropriate exclusion of patients with stage IVc disease. These patients should be routinely excluded in curative intent survival studies as their management is typically palliative rather than curative.

Finally, an additional flaw in cancer registry use is inappropriate grouping of cancer pathologies. For example, a study may examine the association of sinonasal malignancies with sex and show that males have worse survival than females. On review of the cohort, there is a much higher

\begin{center}
\textbf{Figure 2.} Confounding variables lead to spurious association between independent and dependent variables.
\end{center}
percentage of esthesioneuroblastomas in the female cohort and a higher number of mucosal melanomas in the male cohort. Esthesioneuroblastoma has a 5-year disease-specific survival of almost 90%, whereas sinonasal melanoma is only about 37%. The survival differences observed are not due to sex but due to inappropriate cancer grouping. Grouping of heterogeneous cancers is often seen. For example, a study reporting survival differences as an end point will report all salivary cancers or all sinonasal malignancies as a large group. To avoid this problem, authors should be selective in the pathology groupings for survival analysis. Both SEER and NCDB use the International Classification of Diseases in Oncology (ICD-O-3) definitions. Familiarity with how cases are coded pathologically will help ensure appropriate case selection for study.

More recently, studies that evaluate head and neck squamous cell carcinoma sometimes do not control for the effects of human papillomavirus (HPV) oropharynx cancer. HPV status must be considered and controlled to prevent skewing of survival curves, as the incidence of HPV oropharynx cancer is on the rise and survival is significantly better for HPV than non-HPV oropharynx cancer. Studies submitted for publication have sometimes undergone multiple revisions, only to be rejected due to uncontrolled HPV oropharynx conditions as a cause of positive findings. The NCDB and SEER have included HPV status in some cases after 2010, allowing for some control in selecting recent cohorts. This is part of the CS Collaborative Staging System version 02.05 for head and neck sites.

Data Analysis Pitfalls

Even with employed certified cancer registrars, input and completeness of data in national registries are not perfect. The number of collected data fields per case is large, and medical records for submission may not contain all desired information or may not be coded accurately. For example, a recent review of the accuracy of SEER data on melanoma demonstrated the implications of incorrectly coded data. The article discovered that nearly 7% of thin melanomas (<1-mm lesions) in a single SEER registry were incorrectly coded by the misplacement of decimal points. This review invalidated prognostic conclusions based on SEER data, in turn changing therapeutic recommendations. Inconsistent data are fortunately rare; however, authors should drop cases with incomplete or inaccurate data. When assembling a cohort for study, a small percentage of cases will be missing key variables needed for analysis. These cases should be dropped from the study so the cohort is uniform and has all necessary data. In addition, T0 tumors (unknown primary) are reported with a site of tumor code and should be dropped from analysis unless the study is specifically of the unknown primary. As much as possible, investigators should use a single finalized data set for analysis. In addition, if many cases are dropped due to inconsistency, this may suggest an underlying issue with the data integrity that may need to be addressed.

Database subtleties exist that require consideration to avoid errors of case selection. In SEER, for example, cases introduced at autopsy or cases of multiple same-site primary malignancies yielding duplicate records may introduce inaccuracy and must be removed or consolidated, respectively. This is not an issue in the NCDB, which does not include cancer cases identified only at autopsy.

The considerable quantity of variables in cancer registries enables researchers to study the effects of many variables on cancer outcomes. However, caution must be taken when performing univariate analysis on multiple variables. Setting a P value of .05 for univariate analysis on average will yield a random positive result in 1 of every 20 analyses. Multiple hypothesis testing is common in otolaryngology. Remedies for this flaw include limiting the number of variables for univariate analysis or adjusting P values to account for multiple hypothesis testing using an appropriate correction method like the Bonferroni correction. Using a hypothesis-driven analysis also focuses the analysis and reduces unnecessary statistical testing. Also, it is often worthwhile to seek assistance from an individual with relevant data management and statistical training, such as an experienced biostatistician, for the project so that programming and inferential errors can be avoided.

Final Limitations

Along with statistical and analytical drawbacks, national cancer registries are limited in the type of data captured. Vital status (dead or alive) is captured, but the cause of death is not coded in the NCDB as it is in SEER. Therefore, disease-specific survival cannot be evaluated in some analyses and should be part of the discussion of any analysis. Inadequate database capture of chemotherapy, radiation, or multiple treatments skews treatment outcomes and treatment delay implications, a field of rising interest.

Coding of comorbidities is present albeit crude in the
NCDB, whereas it is absent in SEER, which limits evaluation. An additional limitation identified is that NCDB and SEER report only objective measures of outcomes in registry data. Patient-reported quality of life and function are not captured, which is an important part of cancer outcomes.

In summary, while major cancer registries provide a rich data set for analysis of cancer treatment and outcomes, obvious design and analytical pitfalls are to be avoided during study. These shortcomings have been observed on many submitted, and subsequently rejected, manuscripts to a major otolaryngology journal. With careful study design, planning, and awareness, investigators can avoid the painful experience of attempting to submit flawed research for publication.

Implications for Practice

This article presents common flaws observed by 2 associate editors in manuscripts employing cancer registries, which is an important undertaking because flawed studies are rarely published. By no means is this article an exhaustive collection of all possible shortcomings in database design or research development; however, it is a cautionary guide based on observations from editorial journal review. Herein, highlights are presented to identify heterogeneity and limitations of registry data, as well as recommendations provided for thoughtful analytical design.

Access to cancer databases like SEER and NCDB is free and easily available through online resources. Tremendous value can be mined from these databases that affects our understanding of the cancer epidemiology and biology, as well as its treatment. At the same time, the vast amount of data in these databases sets the stage for flawed design, analysis, and statistics.

While knowledgeable editorial staff and peer review often prevent flawed studies from being published, flawed studies sometimes proceed to print, even in reputable journals. Given the considerable time and effort involved in data analysis and manuscript writing, knowledge of these common pitfalls will greatly limit the wasted time and effort of researchers on a flawed study. In addition, this article provides insight for peer reviewers and journal subscribers. The examples provided will help clinicians reviewing manuscripts to better critically analyze cancer database research and judge interpretations of validity and meaning. For the novice database researcher, many introductory resources and publications are available.

We hope this review will provide a framework to help encourage quality, meaningful research with cancer databases.

Author Contributions

Evan A. Jones, development of concept, significant drafting and revision implementation from all authors to produce final manuscript, accountable for the manuscript; Jeffrey C. Liu, formulation of concept, significant drafting and critical revision to produce the final manuscript, accountable for the manuscript.

Disclosures

Competing interests: Brian L. Egleston, National Institutes of Health, National Cancer Institute P30CA006927. This grant subsidized time spent working on the project.

Sponsorships: None.

Funding source: None.

References


