Review Article

Strengths and limitations of large databases in lung cancer radiation oncology research

Vikram Jairam[1], **Henry S. Park**[1,2]

[1]Department of Therapeutic Radiology, Yale University School of Medicine, New Haven, CT, USA; [2]Cancer Outcomes, Public Policy, and Effectiveness Research (COPPER) Center, Yale School of Medicine, New Haven, CT, USA

Contributions: (I) Conception and design: All authors; (II) Administrative support: None; (III) Provision of study materials or patients: None; (IV) Collection and assembly of data: All authors; (V) Data analysis and interpretation: All authors; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Henry S. Park, MD, MPH. Department of Therapeutic Radiology, Yale University School of Medicine, 35 Park Street, Lower Level, New Haven, CT 06520, USA. Email: henry.park@yale.edu.

Abstract: There has been a substantial rise in the utilization of large databases in radiation oncology research. The advantages of these datasets include a large sample size and inclusion of a diverse population of patients in a real-world setting. Such observational studies hold promise in enhancing our understanding of questions for which evidence is conflicting or absent in lung cancer radiotherapy. However, it is critical that investigators understand the strengths and limitations of large databases in order to avoid the common pitfalls that beset observational analyses. This review begins by outlining the data variables available in major registries that are used most often in observational analyses. This is followed by a discussion of the type of radiotherapy-related questions that can be addressed using such datasets, accompanied by examples from the lung cancer literature. Finally, we describe some limitations of observational research and techniques to mitigate bias and confounding. We hope that clinicians and researchers find this review helpful for designing new research studies and interpreting published analyses in the literature.

Keywords: Lung cancer; large databases; big data; radiation oncology; radiotherapy; observational; limitations

Submitted May 01, 2019. Accepted for publication May 15, 2019.
doi: 10.21037/tlcr.2019.05.06
View this article at: http://dx.doi.org/10.21037/tlcr.2019.05.06

Introduction

The field of radiation oncology has witnessed a substantial increase in the publication of observational analyses using large healthcare databases. These datasets provide investigators with the tools to probe into questions that may not be feasible with prospective studies. The advent of drugs like immune checkpoint inhibitors and newer-generation targeted therapies, as well as enhanced radiotherapy and surgical techniques like intensity-modulated radiation therapy (IMRT), stereotactic body radiation therapy (SBRT), proton beam therapy, minimally invasive surgery, and robotic-assisted surgery, have led to a need to understand practice patterns, adherence to guidelines, healthcare disparities, and comparative effectiveness.

Common categories of databases include national cancer registries, claims-based datasets, national surveys, and hospital encounter data. A brief description of some of the major databases in the United States used in observational studies in lung radiation oncology follows below. *Tables 1* and *2* summarize some of the key similarities and differences among them.

National Cancer Data Base (NCDB)

The NCDB is a joint quality improvement program of the Commission on Cancer (CoC) between the American College of Surgeons and the American Cancer Society (1). This nationwide oncology outcomes database encompasses more than 1,500 CoC programs and approximately 70% of

Table 1 Comparison of NCDB, SEER, SEER-Medicare, and HCUP—descriptive and demographic information

Variables	NCDB	SEER	SEER-Medicare	HCUP (NEDS/NIS)
Number of patients	>34 million [2004–2015]	>10.4 million [1975–2016]	>2.3 million [1991–2015]	NEDS: >322 million [2006–2016]; NIS: >203 million [1988–2016]
Representation	Hospital-based	Population-based	Population-based	Population-based
Proportion of newly diagnosed cancers nationally	70%	35%	8%	N/A
Weighted	No	No	No	Yes
Ages represented	All ages	All ages	65+[a]	All ages
Demographic variables				
Comorbidities	Charlson-Deyo score	No	Medicare claims	Discharge diagnoses and Elixhauser score
Race	Yes	Yes	Yes	NEDS: no; NIS: yes
High school degree	Yes	Yes	Yes	No
Insurance	Yes	Yes	Yes	Yes
City size	Yes	Yes	Yes	Yes
Facility type	Yes	No	Yes	Yes
Hospital size	No	No	Yes	Yes
Hospital region	Yes	Yes	Yes	Yes

[a], also includes patients with end-stage renal disease and on disability. NCDB, National Cancer Data Base; SEER, Surveillance, Epidemiology, and End Results; HCUP, Healthcare Cost and Utilization Project; NEDS, Nationwide Emergency Department Sample; NIS, National Inpatient Sample.

all newly diagnosed cases of cancer in the U.S. (2). A portion of the dataset, known as the Participant User File, was made public in 2013 to researchers at programs accredited by the CoC, and is updated and released annually. Since then, the number of publications involving the NCDB has increased exponentially (3,4). In particular, the NCDB has been widely used to study patterns of radiation therapy as it includes granular data not encompassed other databases, including dose, fractionation, timing, modality, location, anatomical site, boost, and reason for no radiation. Extensive patient sociodemographic and clinical factors as well as facility characteristics are available. Variables like individual facility case volume can be extrapolated based on individual facility identification numbers.

Limitations of the NCDB include selection bias (5,6), lack of longitudinal treatment data, and lack of clinically relevant endpoints like complications, patient-reported outcomes (PROs), cause-of-death, local control, and disease-free survival. Furthermore, since the NCDB is hospital-based rather than population-based (not designed to be representative of the U.S. population overall), generalizability from this data is more limited than that of other databases.

Surveillance, Epidemiology, and End Results Program (SEER)

The SEER program of the National Cancer Institute is a widely used and authoritative source of information on cancer incidence, staging, treatment, demographics, and survival information (7). The program pools data from 19 geographic areas covering approximately 35 percent of the U.S. population. These areas are particularly chosen to be representative of the U.S. population and therefore have been used to answer important epidemiological questions such as cancer prevention and screening, healthcare disparities, effectiveness of public health interventions, and implementation of healthcare policy. Publications using the SEER database to study lung cancer have increased markedly from the 1980s to the 2010s (8).

Table 2 Comparison of NCDB, SEER, SEER-Medicare, and HCUP—clinical, treatment, and outcomes information

Variables	NCDB	SEER	SEER-Medicare	HCUP (NEDS/NIS)
Cancer data				
AJCC staging	Separate clinical and pathologic	Combined clinical and pathologic	Combined clinical and pathologic	No
Histology	Yes	Yes	Yes	No
Anatomical site	Yes	Yes	Yes	Yes
Genetic markers	Yes, limited (not lung)	Yes, limited (not lung)	Yes, limited (not lung)	No
Metastatic locations coded	Bone, brain, liver, lung (at diagnosis)	Bone, brain, liver, lung (at diagnosis)	Yes, multiple (SEER codes and Medicare claims)	Yes, multiple (discharge diagnoses)
Treatment				
Sequence	Yes	Yes	Yes	No
Surgery	Yes	Yes	Yes	No
Surgical margins	Yes	No	No	No
Number of lymph nodes dissected	Yes	Yes	Yes	No
Radiation therapy				
Received	Yes	Yes (upon request)	Yes	Yes
Modality	Yes	No	Yes	No
Dose	Yes	No	No	No
Fractionation	Yes	No	Yes[a]	No
Chemotherapy				
Received	Yes; includes single or multi-agent chemo	Yes (upon request)	Yes	Yes (very limited)
Type	No	No	Yes	No
Immunotherapy	Yes	No	No	Yes
Received	Yes	No	No	Yes
Type	No	No	No	No
Outcomes				
Overall survival	Yes	Yes	Yes	In-hospital mortality only
Disease-specific survival	No	Yes	Yes	No
Disease-free survival	No	No	Yes (limited[b])	No
Length of stay	Yes	No	Yes	Yes
Cost	No	No	Yes	Yes
Complications	No	No	Yes[c]	Yes[d]

[a], variable not collected but inferred through Medicare claims; [b], variable not collected but inferred through Medicare claims. This applies only for patients who receive treatment for their recurrence; [c], complications available via Medicare claims data; [d], complications available via discharge diagnoses, though it may be difficult to ascertain with certainty whether or not these diagnoses represent comorbidities or complications. NCDB, National Cancer Data Base; SEER, Surveillance, Epidemiology, and End Results; HCUP, Healthcare Cost and Utilization Project; NEDS, Nationwide Emergency Department Sample; NIS, National Inpatient Sample.

SEER data must be interpreted carefully as there are some important limitations (9). These include unrecorded variables, underreported and incomplete adjuvant treatment data, variations in coding and reporting, and migration of patients between SEER registry areas. Multiple analyses of local SEER registries in comparison to Medicare claims, medical record review, or patient self-report have demonstrated that the under-ascertainment of radiotherapy within those SEER registries may range from 10–30% (10-12). In 2016, the SEER program eliminated the routine reporting of radiotherapy. This is now available by request, although SEER requires the investigator sign a data use agreement acknowledging the limitations of the radiotherapy variable. Radiotherapy data in SEER is further limited to receipt, modality, and sequence with surgery.

SEER-Medicare

The SEER-Medicare dataset was created from the linkage of the SEER database to Medicare claims (13). The Medicare dataset contains diagnostic and billing codes for services covered by Medicare including tests, procedures, office visits, admissions, medical equipment, hospice care, and prescription drugs (14). This database has been primarily used to study patients age 65 or older, although Medicare does cover individuals who are disabled or with end-stage renal disease. For persons age 65 and older, 94% have been linked to their Medicare enrollment file. Claims data provides more granular comorbidity data based on International Classifications of Diseases (ICD) coding. Treatment-related complications and disease recurrence, while not coded, can be inferred from health claims (15). While SEER includes radiotherapy data from the first course of therapy, Medicare claims supplements this with information on type of modality administered, timing, fractionation, and anatomical site of treatment. A major additional limitation of this dataset beyond those associated with SEER would be the possibility of making inaccurate clinical inferences based on what ultimately is administrative and billing data (14). Generalizability to patients under the age of 65 is also highly limited.

Healthcare Cost and Utilization Project (HCUP)

The HCUP, which is sponsored by the Agency for Healthcare Research and Quality, includes the largest collection of longitudinal hospital data in the U.S., with all-payer and encounter level information. The HCUP includes healthcare databases such as the Nationwide Emergency Department Sample (NEDS) and the National Inpatient Sample (NIS), both representing a 20% stratified sample of U.S. hospital-based emergency departments (ED) and discharges, respectively. Like SEER, these are meant to be representative of the U.S. population, and so national estimates can be made. Discharge data is identifiable through ICD data, similar to most claims-based datasets. Cancer-related studies using these datasets have characterized general ED visits (16), complications of treatment (17), oncologic emergencies (18), and outcomes from oncologic surgeries (19-22). Since data is limited to the ED visit or inpatient encounter, few studies have examined radiotherapy utilization, although studies investigating patterns of emergent in-house radiotherapy would be feasible.

Survey data

Survey datasets have been gaining currency amongst health researchers due to their incorporation of PROs, which most large cancer databases lack. Examples of such datasets include the SEER-Medicare Health Outcomes Survey (SEER-MHOS), the National Health Interview Survey (NHIS), and the National Health and Nutrition Examination Survey (NHANES), with the latter two maintained by the Centers for Disease Control and Prevention. The strength of these studies lies in their ability to collect data on the patient experience, as well as demographic, financial, preventative care, and detailed social information like personal habits (alcohol, tobacco, drug use), diet, exercise, and reproductive health. These datasets have been used sparingly to answer radiotherapy-related questions (23), although a project to analyze quality-of-life after surgery or radiotherapy for stage I lung cancer patients is currently underway (24). The primary limitation of survey data would be non-response bias, with modern response levels hovering around 70% and trending downward (25).

Research questions using large datasets in lung cancer

Large databases have many advantages in epidemiological research, with their strength primarily resting on larger sample size, inclusion of more diverse subsets of patients, and completed outcomes. Below are some research questions related to lung cancer radiotherapy that have been addressed by large datasets.

Adoption of novel treatments or technology

One of the strongest advantages of large datasets lies in studying practice patterns across a disease site. This can be particularly useful in the field of lung radiotherapy, as differences in practice exist between providers based on the available evidence. One common application of this question is in the adoption of new techniques or technologies, including SBRT for early-stage non-small cell lung cancer (NSCLC) as well as metastatic disease to the lungs. Given the high rates of local control seen in prospective studies evaluating the efficacy of SBRT (26,27) in early-stage NSCLC, multiple groups have used the NCDB (28-30) to explore how this has changed practice patterns in the past 10–15 years. Another example has been the advent of IMRT in the treatment of locally advanced NSCLC. A SEER-Medicare analysis noted that IMRT utilization increased from 0.5% in 2001 to 14.7% in 2007 for patients with stage III NSCLC (31).

Healthcare disparities

Given that randomized controlled trials (RCTs) tend to enroll patients who are young, white, male, and healthy (32,33), understanding treatment variations in groups often excluded from RCTs is important. This is one manner by which investigators can learn about healthcare disparities based on age, sex, race/ethnicity, median household income, education, insurance status, and geographic region. One NCDB study examined barriers to chemotherapy or radiotherapy utilization in small cell lung cancer (SCLC) and found that patients on government insurance were less likely to receive radiotherapy (34). Other NCDB studies have found that black race was associated with lower likelihood of receiving SBRT or surgery in stage I NSCLC or in receiving cancer-directed care in stage III NSCLC (35,36). These studies can help inform future efforts to improve access to care and reduce inequality in our healthcare system.

Adherence to national guidelines

Large databases can be especially helpful for studying practice patterns with regard to adherence to national guidelines, such as those from the National Comprehensive Cancer Network (NCCN) (37) or professional societies like the American Society for Radiation Oncology (ASTRO) or the American Society of Clinical Oncology (ASCO). A SEER-Medicare study looking at guideline concordance in lung cancer in the elderly found that only 45% of patients received treatment consistent with guidelines, with non-concordant care being associated with worsened overall survival (38). Another study found that two thirds of potentially eligible patients for surgery in early-stage SCLC were not receiving surgery (39). A recent SEER-Medicare study investigating adherence to surveillance guidelines after curative therapy for NSCLC found that only 61% of patients received routine imaging surveillance, and that patients treated with SBRT were more likely to have undergone recommended imaging (40). These studies can enhance our understanding of current gaps in healthcare delivery and quality.

Complications of therapy and quality-of-life

In addition to expanding our repertoire of curative and life-prolonging therapies, there has been an increased emphasis on providing high-value care and improving the patient experience. The quality of PROs in clinical trial protocols is often poor, partly due to logistic challenges as well as limited resources available for monitoring these outcomes (41,42). However, a recent RCT comparing routine surveillance to electronic symptom monitoring in patients with lung cancer found an overall survival benefit in the monitoring group, demonstrating the importance of actively collecting PROs (43). Claims-based datasets such as SEER-Medicare and encounter-based hospital databases like HCUP-NEDS or HCUP-NIS can be useful for studying toxicities and other aspects of the patient experience. One SEER-Medicare study quantified the treatment burden of patients with stage I lung cancer (44). Another study using the NIS evaluated the utilization of intensive care during terminal hospitalizations in patients with metastatic lung cancer (45). Further study into PROs can be carried out with survey datasets.

Comparative effectiveness

Perhaps the most exciting but controversial application for large databases is in the area of comparative effectiveness research (CER). RCTs represent the gold standard in CER to evaluate the efficacy of an intervention (46). However, there are many instances in which a prospective RCT cannot be performed, often due to high costs, small sample sizes, long time to completion, strict enrollment criteria, and limited external validity. The reasons for the latter are multifactorial such as low inclusion of elderly, indigent, uninsured, underinsured, and racial ethnic minority patients (47). Clinical trial enrollment has been suggested as a particular problem in lung cancer (48), given the high likelihood of poor performance status, need for emergent radiation, and patient refusal (49). Patients with lung cancer also tend to have a smoking history that could lead to comorbidities that render them ineligible for clinical trials. Another concern is that lung cancer patients appear to be more likely to receive first-line treatment with their local oncologist, while referrals to the academic center may be more likely for second-line or non-therapeutic studies (50). These issues with trial recruitment or enrollment in lung cancer open a potential avenue for large observational datasets to address questions in specific underrepresented populations that may not be possible to answer in a prospective study. In addition, it is also important to understand the effectiveness of certain interventions in the "real world" among all populations in all settings, which may provide information that complements efficacy data derived from the RCT setting.

Below are some of the most common questions in lung radiotherapy that have begun to be addressed by CER studies.

Surgery *vs.* SBRT for early-stage NSCLC

One area of controversy in early-stage NSCLC is the efficacy of surgery compared to SBRT. While surgical resection with either lobectomy or sublobar resection has historically been the standard of care in operable patients, data from modern SBRT trials show local control rates above 90% (51) and comparable to that of surgery. A meta-analysis of the randomized STARS and ROSEL trials comparing surgery to SBRT for operable stage I NSCLC demonstrated a survival advantage to SBRT (52), but both trials closed early due to poor accrual, leaving this study to be criticized as an underpowered post-hoc analysis. Multiple database studies have attempted to answer this question, with nearly all of them demonstrating superiority of surgery over SBRT (53-55). However, the main limitation of these studies lies in the selection bias against SBRT as many of these patients receive SBRT due to poor performance status or high comorbidity burden. This was partially accounted for in an NCDB study examining patients who were surgical candidates and refused surgery for SBRT, thereby excluding patients who were not healthy enough for surgery (56). This study also observed a survival improvement for patients undergoing surgery compared to SBRT. Despite efforts to minimize bias, it is indeed impossible to account for all unknown confounders that may exist in large databases.

Post-operative radiotherapy (PORT) for locally advanced NSCLC

PORT has been another area of considerable controversy ever since the PORT meta-analysis showed no improvement and potentially a detriment in survival using older radiotherapy therapy techniques. Given that radiotherapy techniques have improved significantly since then, allowing for lower scatter doses to critical organs-at-risk like the lung, heart, and esophagus, more recent SEER and NCDB analyses have explored similar questions, demonstrating a potential survival improvement with PORT in patients with N2 disease or positive margins (57-59). Based largely on this observational data, current NCCN guidelines support the use of PORT in this population. Prospective trials like Lung-ART are aiming to answer this question more definitively, but the hypothesis generated by these large-database studies has been practice-changing.

Photon *vs.* proton beam therapy for locally advanced NSCLC

While the current standard-of-care for lung cancer involves photon therapy, proton beam therapy offers a distinct dosimetric advantage that could potentially improve the therapeutic ratio for radiotherapy by delivering higher doses to the target and sparing normal tissues. Current RCTs like RTOG 1308 are underway comparing photon to proton therapy in NSCLC, although we will likely have to wait years before learning these results. In the meantime, one NCDB study observed a 5-year overall survival improvement with proton therapy among NSCLC patients of all stages (60). Further database studies can potentially highlight whether these results bear out in other contexts such as the post-operative setting, in which the advantage of proton therapy in avoiding toxicity may be the greatest.

Chemoradiotherapy and prophylactic cranial irradiation (PCI) in limited-stage SCLC

The standard-of-care for limited-stage SCLC has been combined chemotherapy with radiotherapy (CRT), followed by PCI. However, one of the pivotal RCTs published in 1992 that demonstrated superiority of CRT over RT suggested that this benefit may be limited to patients younger than 70 years (61). Recent NCDB analyses demonstrated that this benefit may extend to highly elderly patients selected in the modern era given improvements in radiotherapy technique (62,63). This may be a result of lower toxicities to the normal lung and heart, but the hypothesis generated by these large-database projects support using aggressive therapy even in populations that were not well-represented in RCTs. Additionally, a SEER analysis noted that elderly patients may also continue to have a benefit to PCI, a finding which had been shown in RCTs that primarily included younger patients (64).

Radiotherapy dose and fractionation in NSCLC

One of the biggest advantages of the NCDB is in its detailed collection of dose and fractionation data for patients undergoing radiotherapy. This enabled multiple studies comparing survival outcomes for patients treated with differing radiation regimens. Dose escalation remains a controversial topic in NSCLC. For early-stage NSCLC, several single-institution and multi-institution studies have suggested that SBRT with a biologically effective dose using alpha/beta of 10 (BED_{10}) of at least 100 or 105 is necessary for optimal local control and overall survival, but NCDB studies have suggested a potential survival improvement with an even higher BED_{10} threshold (65,66). For locally advanced NSCLC, RTOG 0617 demonstrated inferior survival for patients receiving 74 Gy compared to 60 Gy (67), though many believe there still could be an advantage to some level of dose escalation when more stringent normal tissue constraints are met. This was explored by an NCDB study that found a survival benefit to dose escalation to 70 Gy but not beyond (68). There is certainly room for further study in this area such as for patients who are elderly or not candidates for chemotherapy.

Facility volume in NSCLC

Given that modern lung radiotherapy is commonly associated with significant technical demands in treatment planning, multidisciplinary coordination to ensure the reliability of concurrent treatment, severe toxicities, and substantial risk of recurrence and subsequent death, a high level of expertise may be necessary for providers treating lung cancer. Extensive literature shows that patients have improved outcomes when treated by high-volume surgeons and hospitals for oncologic resections, including video-assisted thoracoscopic lobectomy and robotic-assisted lobectomy (20,22). More recent data has examined the effect of the treating institution on outcomes following lung radiotherapy. A secondary analysis of the RTOG 0617 RCT in locally advanced NSCLC showed that receiving chemoradiotherapy at institutions with higher clinical trial accrual volume was associated with longer overall survival (69). NCDB analyses have also shown longer overall survival for patients receiving definitive chemo-RT for locally advanced NSCLC at facilities treating a high number of annual cases compared to those treated at lower-volume facilities (70), similar to findings noted regarding SBRT for stage I NSCLC (71).

Pitfalls and disadvantages of using large databases in lung cancer

Before embarking on a large database analysis, it is important to have a well-designed hypothesis and to be intimately aware of the strengths and limitations of the database at hand. For example, researchers interested in studying national disease incidence might choose to work with the SEER database, while others more interested in a radiotherapy dose escalation study may prefer the NCDB. The question should aim to fill a relevant gap in knowledge that prospective studies have not yet answered, and the researchers should consider what techniques they plan to use to minimize the effect of any potential inherent biases in the study question or the data itself prior to conducting the analysis.

Below are listed some of the most common design and analytical issues that befall such large database studies in lung cancer along with methods to address them.

Selection bias

Perhaps the most common and frequently cited criticism of large datasets is selection bias, an inherent result of non-random assignment. This stems from the idea that patients may be more likely to receive a particular treatment based on unmeasured confounders, such as performance status. Many patients who are elderly or medically unfit for surgery are more likely to be referred for radiotherapy. These patients may instead be more likely to die from comorbidities unrelated to their cancer. An additional missing variable in lung cancer is the location of lung tumors in relation to critical structures, which are often centrally located. Tumor location may influence which dose-fractionation schemes are chosen and therefore bias any comparisons of effectiveness among them. Therefore, in database comparisons of two management strategies or dose-fractionation regimens, investigators are often comparing two very different populations and may not be able to attribute the outcome of one group solely to the intervention.

Immortal time bias

In observational studies comparing adjuvant treatment to observation after surgery, patients who die soon after surgery will not have the opportunity to enter the adjuvant therapy cohort and will be much more likely to be included in the observation cohort. This would guarantee worse comparative survival for the observation arm than would otherwise be expected on an intention-to-treat analysis of an RCT. This is because the adjuvant therapy cohort has a certain time period of being "immortal" from early death after surgery, since those patients (by definition) must have lived a certain length of time to be included in that cohort. This type of bias is called immortal time bias (72), also known as guarantee time bias (73,74).

To account for this bias regarding time-dependent variables like survival or tumor response, investigators will often incorporate a landmark analysis, in which the patients who die a certain period soon after surgery are excluded from analysis. The optimal landmark to use for these analyses has remained elusive, as there is no standard definition for the period during which deaths would be classified as immediate postoperative mortality. Critics of stringent landmark corrections argue that disregarding the potential short-term survival benefit of PORT, particularly in aggressive cancers with poor prognosis, could cause a bias in the different direction. Therefore, a sequential landmark analysis method (selecting monthly landmarks from 1-6 months postoperatively) has been proposed as a sensitivity analysis to measure the robustness of the results from the primary analysis (72). Examples of landmark analyses in the lung radiotherapy literature are becoming increasingly abundant (58,75-78).

Strategies to address confounding

There are several techniques commonly employed to mitigate selection bias and address confounding in CER. One is stratification, in which patients can be grouped based on a covariate, such as comorbidity status. A second method is multivariable regression, in which all potential confounders are included in the multivariable model, and truly independent predictors will be revealed as statistically significant. A third method is propensity score matching (PSM), a statistical technique used to estimate the effect of an intervention or treatment by accounting for the propensity to receive a certain therapy based on the observed covariates (14). PSM has been applied often in the lung radiotherapy literature (53-56,60). Finally, instrumental variable analysis (IVA) is a technique originating from econometrics, but now appearing in epidemiological studies to account for confounding and mimic randomization. An instrumental variable is one that is strongly correlated with the treatment assignment but not with the outcome of interest. IVA has been used in multiple lung cancer database studies to account for geographic variation in radiotherapy utilization (79,80).

Even after employing techniques to reduce confounding, it is impossible to account for all potential confounders that might affect a CER study. One report found that even after multiple statistical adjustments it is still possible to obtain outcomes different from those of comparable RCTs (81). In a recent analysis, observational database studies from NCDB, SEER, and SEER-Medicare were matched with comparable RCTs and found that agreement between matched pairs occurred only 40% of the time without any variables that predicted stronger correlation (82). Criticisms of this study include the fact that the match criteria only included age and stage, and that CER studies are not designed to match inclusion and exclusion criteria RCTs perfectly, since they are typically intended to generate hypotheses that fill in gaps of knowledge that have not yet been answered by RCTs. While checklists and published guidelines for observational research, such as the STROBE guidelines (83), are helpful, it is imperative that researchers and journals continue to build and enforce rigorous standards for observational database studies.

Conclusions

Well-conducted observational studies using large databases hold much promise to enhance our understanding of lung cancer management. Given the need to deliver high-value care to patients from all segments of the population, observational analyses can help clarify ideal management strategies among certain subsets of patients as well as identify which patients may not be receiving optimal care. Furthermore, the advent of machine learning may help eliminate unknown confounders and improve the accuracy and predictions of models based off large datasets (84). While RCTs remain the gold standard by which we base our treatment decisions, observational analyses using large registries can provide important hypothesis-generating data, from which future practice-changing prospective trials can be built.

Acknowledgments

None.

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

References

1. Boffa DJ, Rosen JE, Mallin K, et al. Using the National Cancer Database for Outcomes Research: A Review. JAMA Oncol 2017;3:1722-8.
2. American College of Surgeons. About the National Cancer Database. 2019. Available online: https://www.facs.org/quality-programs/cancer/ncdb/about. Accessed April 9, 2019.
3. Blanchard P, Garden AS. Looking Beyond the Numbers: Highlighting the Challenges of Population-Based Studies in Cancer Research. J Clin Oncol 2016;34:2317-8.
4. Su C, Peng C, Agbodza E, et al. Publication trend, resource utilization, and impact of the US National Cancer Database: A systematic review. Medicine (Baltimore) 2018;97:e9823.
5. 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;109:629-30.
6. Mallin K, Browner A, Palis B, et al. Incident Cases Captured in the National Cancer Database Compared with Those in U.S. Population Based Central Cancer Registries in 2012–2014. Ann Surg Oncol 2019;26:1604-12.
7. Park HS, Lloyd S, Decker RH, et al. Overview of the Surveillance, Epidemiology, and End Results database: evolution, data variables, and quality assurance. Curr Probl Cancer 2012;36:183-90.
8. Komiya T, Guddati AK, Chaaya G. Overview of publications on lung cancer using the SEER database. Respir Investig 2018;56:424-6.
9. Park HS, Lloyd S, Decker RH, et al. Limitations and Biases of the Surveillance, Epidemiology, and End Results Database. Curr Probl Cancer 2012;36:216-24.
10. Malin JL, Adams J, Kahn KL, et al. Validity of Cancer Registry Data for Measuring the Quality of Breast Cancer Care. J Natl Cancer Inst 2002;94:835-44.
11. Jagsi R, Abrahamse P, Hawley ST, et al. Underascertainment of radiotherapy receipt in Surveillance, Epidemiology, and End Results registry data. Cancer 2012;118:333-41.
12. Walker GV, Giordano SH, Williams M, et al. Muddy water? Variation in reporting receipt of breast cancer radiation therapy by population-based tumor registries. Int J Radiat Oncol Biol Phys 2013;86:686-93.
13. 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:IV-3-18.
14. Jagsi R, Bekelman JE, Chen A, et al. Considerations for observational research using large data sets in radiation oncology. Int J Radiat Oncol Biol Phys 2014;90:11-24.
15. Warren JL, Mariotto A, Melbert D, et al. Sensitivity of Medicare Claims to Identify Cancer Recurrence in Elderly Colorectal and Breast Cancer Patients. Med Care 2016;54:e47-54.
16. Rivera DR, Gallicchio L, Brown J, et al. Trends in Adult Cancer-Related Emergency Department Utilization: An Analysis of Data From the Nationwide Emergency Department Sample. JAMA Oncol 2017;3:e172450.
17. Jairam V, Lee V, Park HS, et al. Treatment-Related Complications of Systemic Therapy and Radiotherapy. JAMA Oncol 2019. [Epub ahead of print].
18. Mak KS, Lee LK, Mak RH, et al. Incidence and treatment patterns in hospitalizations for malignant spinal cord compression in the United States, 1998-2006. Int J Radiat Oncol Biol Phys 2011;80:824-31.
19. Matsuo K, Mandelbaum RS, Adams CL, et al. Performance and outcome of pelvic exenteration for gynecologic malignancies: A population-based study. Gynecol Oncol 2019;153:368-75.
20. Park HS, Detterbeck FC, Boffa DJ, et al. Impact of hospital volume of thoracoscopic lobectomy on primary lung cancer outcomes. Ann Thorac Surg 2012;93:372-9.
21. Law TD, Boffa DJ, Detterbeck FC, et al. Lethality of cardiovascular events highlights the variable impact of complication type between thoracoscopic and open pulmonary lobectomies. Ann Thorac Surg 2014;97:993-9.
22. Tchouta LN, Park HS, Boffa DJ, et al. Hospital Volume and Outcomes of Robot-Assisted Lobectomies. Chest 2017;151:329-39.
23. Pearlstein KA, Basak R, Chen RC. Comparative Effectiveness of Prostate Cancer Treatment Options: Limitations of Retrospective Analysis of Cancer Registry

Data. Int J Radiat Oncol Biol Phys 2019;103:1053-7.

24. National Cancer Institute. Current Projects Using SEER-MHOS Data. 2018. Available online: https://healthcaredelivery.cancer.gov/seer-mhos/overview/current.html. Accessed April 12, 2019.
25. Johnson TP, Wislar JS. Response rates and nonresponse errors in surveys. JAMA 2012;307:1805-6.
26. Timmerman RD, Hu C, Michalski J, et al. Long-term Results of RTOG 0236: A Phase II Trial of Stereotactic Body Radiation Therapy (SBRT) in the Treatment of Patients with Medically Inoperable Stage I Non-Small Cell Lung Cancer. Int J Radiat Oncol Biol Phys 2014;90:S30.
27. Timmerman RD, Paulus R, Pass HI, et al. Stereotactic Body Radiation Therapy for Operable Early-Stage Lung Cancer: Findings From the NRG Oncology RTOG 0618 TrialStereotactic Body Radiation Therapy for Operable Early-Stage Lung CancerStereotactic Body Radiation Therapy for Operable Early-Stage Lung Cancer. JAMA Oncol 2018;4:1263-6.
28. Engelhardt KE, Feinglass JM, DeCamp MM, et al. Treatment trends in early-stage lung cancer in the United States, 2004 to 2013: A time-trend analysis of the National Cancer Data Base. J Thorac Cardiovasc Surg 2018;156:1233-46.e1.
29. McMurry TL, Shah PM, Samson P, et al. Treatment of stage I non-small cell lung cancer: What's trending? J Thorac Cardiovasc Surg 2017;154:1080-7.
30. Corso CD, Park HS, Moreno AC, et al. Stage I Lung SBRT Clinical Practice Patterns. Am J Clin Oncol 2017;40:358-61.
31. Shirvani SM, Jiang J, Gomez DR, et al. Intensity modulated radiotherapy for stage III non-small cell lung cancer in the United States: predictors of use and association with toxicities. Lung Cancer 2013;82:252-9.
32. Murthy VH, Krumholz HM, Gross CP. Participation in Cancer Clinical TrialsRace-, Sex-, and Age-Based Disparities. JAMA 2004;291:2720-6.
33. Unger JM, Hershman DL, Fleury ME, et al. Association of Patient Comorbid Conditions With Cancer Clinical Trial Participation. JAMA Oncol 2019;5:326-33.
34. Pezzi TA, Schwartz DL, Mohamed ASR, et al. Barriers to Combined-Modality Therapy for Limited-Stage Small Cell Lung Cancer. JAMA Oncol 2018;4:e174504.
35. Cassidy RJ, Zhang X, Switchenko JM, et al. Health care disparities among octogenarians and nonagenarians with stage III lung cancer. Cancer 2018;124:775-84.
36. Corso CD, Park HS, Kim AW, et al. Racial disparities in the use of SBRT for treating early-stage lung cancer. Lung Cancer 2015;89:133-8.
37. National Comprehensive Cancer Network. NCCN Guidelines. 2019. Available online: https://www.nccn.org/professionals/physician_gls/. Accessed April 12, 2019.
38. Nadpara PA, Madhavan SS, Tworek C, et al. Guideline-concordant lung cancer care and associated health outcomes among elderly patients in the United States. J Geriatr Oncol 2015;6:101-10.
39. Wakeam E, Varghese TK, Leighl NB, et al. Trends, practice patterns and underuse of surgery in the treatment of early stage small cell lung cancer. Lung Cancer 2017;109:117-23.
40. Erb CT, Su KW, Soulos PR, et al. Surveillance Practice Patterns after Curative Intent Therapy for Stage I Non-Small-Cell Lung Cancer in the Medicare Population. Lung Cancer 2016;99:200-7.
41. Kyte D. RA, Keely T, et al. Systematic evaluation of patient-reported outcome (PRO) protocol content and reporting in cancer clinical trials: the EPIC study. 25th Annual Conference of the International Society for Quality of Life Research; October 01; Dublin, Ireland: Springer Nature Switzerland, 2018:1-190.
42. Calvert M, Kyte D, Mercieca-Bebber R, et al. Guidelines for Inclusion of Patient-Reported Outcomes in Clinical Trial Protocols: The SPIRIT-PRO ExtensionGuidelines for Inclusion of Patient-Reported Outcomes in Clinical Trial ProtocolsGuidelines for Inclusion of Patient-Reported Outcomes in Clinical Trial Protocols. JAMA 2018;319:483-94.
43. Denis F, Basch E, Septans AL, et al. Two-Year Survival Comparing Web-Based Symptom Monitoring vs. Routine Surveillance Following Treatment for Lung CancerEffects on Survival of Web-Based Symptom Monitoring vs. Routine Surveillance After Lung Cancer TreatmentLetters. JAMA 2019;321:306-7.
44. Presley CJ, Soulos PR, Tinetti M, et al. Treatment Burden of Medicare Beneficiaries With Stage I Non-Small-Cell Lung Cancer. J Oncol Pract 2017;13:e98-107.
45. Mrad C, Abougergi MS, Daly B. One Step Forward, Two Steps Back: Trends in Aggressive Inpatient Care at the End of Life for Patients With Stage IV Lung Cancer. J Oncol Pract 2018:JOP1800515.
46. Lyman GH. Comparative effectiveness research in oncology. The oncologist 2013;18:752-9.
47. Hamel LM, Penner LA, Albrecht TL, et al. Barriers to Clinical Trial Enrollment in Racial and Ethnic Minority Patients With Cancer. Cancer Control 2016;23:327-37.
48. Yang CJ, Hartwig MG, D'Amico TA, et al. Large clinical

databases for the study of lung cancer: Making up for the failure of randomized trials. J Thorac Cardiovasc Surg 2016;151:626-8.

49. Baggstrom MQ, Waqar SN, Sezhiyan AK, et al. Barriers to enrollment in non-small cell lung cancer therapeutic clinical trials. J Thorac Oncol 2011;6:98-102.
50. Dubey S, Hammes L, Larson ML, et al. Changing patterns of patient accrual to lung cancer clinical trials in an academic center. J Clin Oncol 2005;23:7225.
51. Sebastian NT, Xu-Welliver M, Williams TM. Stereotactic body radiation therapy (SBRT) for early stage non-small cell lung cancer (NSCLC): contemporary insights and advances. J Thorac Dis 2018;10:S2451-64.
52. Chang JY, Senan S, Paul MA, et al. Stereotactic ablative radiotherapy versus lobectomy for operable stage I non-small-cell lung cancer: a pooled analysis of two randomised trials. Lancet Oncol 2015;16:630-7.
53. Yerokun BA, Yang CFJ, Gulack BC, et al. A national analysis of wedge resection versus stereotactic body radiation therapy for stage IA non-small cell lung cancer. J Thorac Cardiovasc Surg 2017;154:675-86.e4.
54. Yu JB, Soulos PR, Cramer LD, et al. Comparative effectiveness of surgery and radiosurgery for stage I non-small cell lung cancer. Cancer 2015;121:2341-9.
55. Puri V, Crabtree TD, Bell JM, et al. Treatment Outcomes in Stage I Lung Cancer: A Comparison of Surgery and Stereotactic Body Radiation Therapy. J Thorac Oncol 2015;10:1776-84.
56. Rosen JE, Salazar MC, Wang Z, et al. Lobectomy versus stereotactic body radiotherapy in healthy patients with stage I lung cancer. J Thorac Cardiovasc Surg 2016;152:44-54.e9.
57. Lally BE, Zelterman D, Colasanto JM, et al. Postoperative radiotherapy for stage II or III non-small-cell lung cancer using the surveillance, epidemiology, and end results database. J Clin Oncol 2006;24:2998-3006.
58. Wang EH, Corso CD, Rutter CE, et al. Postoperative Radiation Therapy Is Associated With Improved Overall Survival in Incompletely Resected Stage II and III Non-Small-Cell Lung Cancer. J Clin Oncol 2015;33:2727-34.
59. Robinson CG, Patel AP, Bradley JD, et al. Postoperative radiotherapy for pathologic N2 non-small-cell lung cancer treated with adjuvant chemotherapy: a review of the National Cancer Data Base. J Clin Oncol 2015;33:870-6.
60. Higgins KA, O'Connell K, Liu Y, et al. National Cancer Database Analysis of Proton Versus Photon Radiation Therapy in Non-Small Cell Lung Cancer. Int J Radiat Oncol Biol Phys 2017;97:128-37.
61. Pignon JP, Arriagada R, Ihde DC, et al. A meta-analysis of thoracic radiotherapy for small-cell lung cancer. N Engl J Med 1992;327:1618-24.
62. Corso CD, Rutter CE, Park HS, et al. Role of Chemoradiotherapy in Elderly Patients With Limited-Stage Small-Cell Lung Cancer. J Clin Oncol 2015;33:4240-6.
63. Miller ED, Fisher JL, Haglund KE, et al. The Addition of Chemotherapy to Radiation Therapy Improves Survival in Elderly Patients with Stage III Non-Small Cell Lung Cancer. J Thorac Oncol 2018;13:426-35.
64. Eaton BR, Kim S, Marcus DM, et al. Effect of prophylactic cranial irradiation on survival in elderly patients with limited-stage small cell lung cancer. Cancer 2013;119:3753-60.
65. Koshy M, Malik R, Weichselbaum RR, et al. Increasing radiation therapy dose is associated with improved survival in patients undergoing stereotactic body radiation therapy for stage I non-small-cell lung cancer. Int J Radiat Oncol Biol Phys 2015;91:344-50.
66. Yan SX, Qureshi MM, Dyer M, et al. Stereotactic body radiation therapy with higher biologically effective dose is associated with improved survival in stage II non-small cell lung cancer. Lung Cancer 2019;131:147-53.
67. Bradley JD, Paulus R, Komaki R, et al. Standard-dose versus high-dose conformal radiotherapy with concurrent and consolidation carboplatin plus paclitaxel with or without cetuximab for patients with stage IIIA or IIIB non-small-cell lung cancer (RTOG 0617): a randomised, two-by-two factorial phase 3 study. Lancet Oncol 2015;16:187-99.
68. Brower JV, Amini A, Chen S, et al. Improved survival with dose-escalated radiotherapy in stage III non-small-cell lung cancer: analysis of the National Cancer Database. Ann Oncol 2016;27:1887-94.
69. Eaton BR, Pugh SL, Bradley JD, et al. Institutional Enrollment and Survival Among NSCLC Patients Receiving Chemoradiation: NRG Oncology Radiation Therapy Oncology Group (RTOG) 0617. J Natl Cancer Inst 2016;108.
70. Wang EH, Rutter CE, Corso CD, et al. Patients Selected for Definitive Concurrent Chemoradiation at High-volume Facilities Achieve Improved Survival in Stage III Non-Small-Cell Lung Cancer. J Thorac Oncol 2015;10:937-43.
71. Koshy M, Malik R, Mahmood U, et al. Stereotactic body radiotherapy and treatment at a high volume facility is associated with improved survival in patients with

inoperable stage I non-small cell lung cancer. Radiother Oncol 2015;114:148-54.

72. Park HS, Gross CP, Makarov DV, et al. Immortal Time Bias: A Frequently Unrecognized Threat to Validity in the Evaluation of Postoperative Radiotherapy. Int J Radiat Oncol Biol Phys 2012;83:1365-73.
73. Giobbie-Hurder A, Gelber RD, Regan MM. Challenges of guarantee-time bias. J Clin Oncol 2013;31:2963-9.
74. Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response. J Clin Oncol 1983;1:710-9.
75. Koshy M, Fedewa SA, Malik R, et al. Improved survival associated with neoadjuvant chemoradiation in patients with clinical stage IIIA(N2) non-small-cell lung cancer. J Thorac Oncol 2013;8:915-22.
76. Lee HW, Noh OK, Oh YT, et al. Radiation Therapy-First Strategy After Surgery With or Without Adjuvant Chemotherapy in Stage IIIA-N2 Non-Small Cell Lung Cancer. Int J Radiat Oncol Biol Phys 2016;94:621-7.
77. Wong AT, Rineer J, Schwartz D, et al. Assessing the Impact of Postoperative Radiation Therapy for Completely Resected Limited-Stage Small Cell Lung Cancer Using the National Cancer Database. J Thorac Oncol 2016;11:242-8.
78. Corso CD, Rutter CE, Wilson LD, et al. Re-evaluation of the Role of Postoperative Radiotherapy and the Impact of Radiation Dose for Non-Small-Cell Lung Cancer Using the National Cancer Database. J Thorac Oncol 2015;10:148-55.
79. Wisnivesky JP, Halm EA, Bonomi M, et al. Postoperative radiotherapy for elderly patients with stage III lung cancer. Cancer 2012;118:4478-85.
80. Wisnivesky JP, Halm E, Bonomi M, et al. Effectiveness of radiation therapy for elderly patients with unresected stage I and II non-small cell lung cancer. Am J Respir Crit Care Med 2010;181:264-9.
81. Giordano SH, Kuo Y-F, Duan Z, et al. Limits of observational data in determining outcomes from cancer therapy. Cancer 2008;112:2456-66.
82. Soni PD, Hartman HE, Dess RT, et al. Comparison of Population-Based Observational Studies With Randomized Trials in Oncology. J Clin Oncol;0:JCO.18.01074.
83. von Elm E, Altman DG, Egger M, et al. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007;335:806-8.
84. Shew M, New J, Bur AM. Machine Learning to Predict Delays in Adjuvant Radiation following Surgery for Head and Neck Cancer. Otolaryngol Head Neck Surg 2019;160:1058-64.

Cite this article as: Jairam V, Park HS. Strengths and limitations of large databases in lung cancer radiation oncology research. Transl Lung Cancer Res 2019;8(Suppl 2):S172-S183. doi: 10.21037/tlcr.2019.05.06