Seminars in RADIATION ONCOLOGY

The Contribution of Cancer Surveillance Toward Real World Evidence in Oncology

Lynne Penberthy, MD, MPH,[*] Donna R. Rivera, PharmD, MSc,[*] and Kevin Ward, PhD, MPH[†]

Randomized control trials (RCTs) represent the gold standard by which new therapies are evaluated. However, there are many limitations to RCTs including biased populations enrolled and potential lack of generalizability. This has resulted in increasing interest in data that represent real world patients, who are not well represented in RCTs. These real world data (RWD) have the potential to provide data not captured in clinical trials such as longer term outcomes, sequelae and adverse events not identified in the sample size used for the RCT. There are many sources of RWD, all of which have strengths and limitations. This manuscript focuses on one source of RWD – the cancer registry. Cancer registries represent a set of consolidated data which are collected under state regulation in each state. As cancer surveillance expands the type and level of detail captured within registries, the potential value for complementing and supplementing clinical trials is increasing.

Semin Radiat Oncol 29:318–322 © 2019 Published by Elsevier Inc.

Randomized controlled trials (RCTs) represent the gold standard by which the efficacy of new cancer therapies is assessed. Efficacy of these treatments, however, is examined with strict inclusion criteria of participating patients to reduce confounding. Clinical trials include, at best, 5% of all cancer patients—a subset that is not representative of the typical cancer patient as clinical trial patients are generally younger and less likely to be nonwhite, and have fewer or no comorbid conditions.[1] This lack of representativeness to patients in the "real world" may limit the generalizability of the results reported from RCTs. With the advent of precision oncology, trials that focus on specific targets have further reduced sample sizes, widening the gap between patients enrolled in trials with those treated by clinicians in the real world. As a result, there is increasing interest and necessity to understand the effectiveness of new therapeutic modalities for the 95% of the cancer patient population who do not participate in RCTs. How do the outcomes of these treatments differ in patients with cardiovascular disease or declining renal function? Are the agents equally effective across differing population subgroups (eg, ethnicity and race) representing varying genetic composition? Acknowledging the importance of questions such as these has resulted in the use of real world data (RWD) to supplement or complement information obtained in clinical trials.

Strengths and Limitations of RWD

RWD have been defined as "information on health care that is derived from multiple sources outside typical clinical research settings." These sources include electronic health records, claims and billing data, product and disease registries, and data gathered through personal devices and health applications.[2] RWD can support research aimed at understanding and evaluating utilization patterns and outcomes of treatments among a heterogenous set of patients who may not be eligible for clinical trials. RWD could further supplement RCTs by providing information on longer term outcomes and possible adverse sequelae of treatments that are either not feasible to determine in RCTs due to limited duration of follow-up period or for which the sample size may be too small to permit identification of comorbidities or rare treatment-related events. The Food and Drug Administration is focusing on RWD to monitor postmarketing safety (pharmacosurveillance) and adverse drug events to inform regulatory decision-making. Various organizations are recognizing the value of RWD for patterns of care, comparative effectiveness, surveillance, and healthcare delivery research.

[*] Division of Cancer Control and Population Sciences, National Cancer Institute, Rockville, MD

[†] Georgia Center for Cancer Statistics, Rollins School of Public Health, Emory University, Atlanta, GA

Conflict of interest: None.

Address reprint requests to Lynne Penberthy, MD, MPH, Division of Cancer Control and Population Sciences, National Cancer Institute, 9609 Medical Center Drive, Rockville, MD 20850.
E-mail: lynnepenberthy.schumacher-penberthy@nih.gov

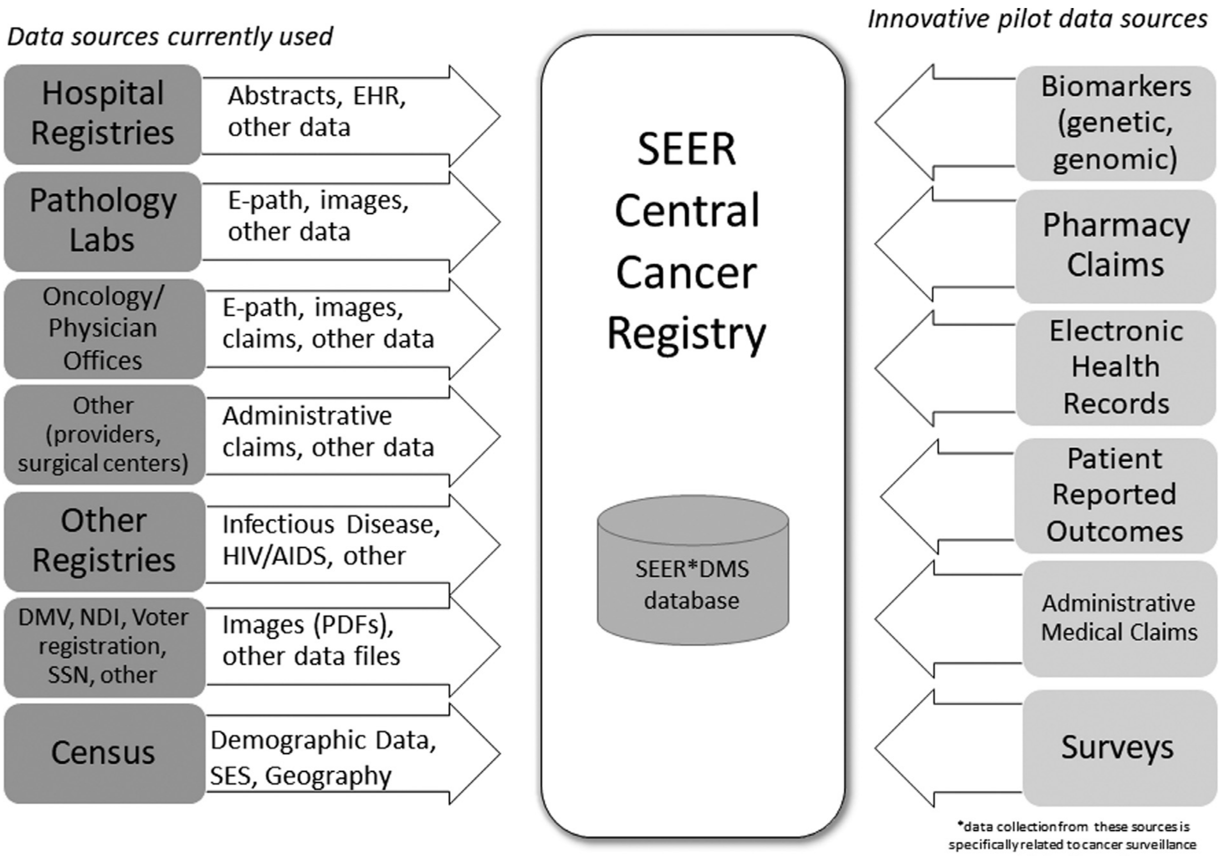
https://doi.org/10.1016/j.semradonc.2019.05.004
1053-4296/© 2019 Published by Elsevier Inc.

While RWD have substantial value in providing information not available through RCTs, there are also significant limitations to these types of data. First, RWD are largely observational, and when analyzing any observational data, it is impossible to control completely for confounding or identify all sources of bias. In addition, many RWD sources are convenience samples and thus suffer the same issues with external validity as RCTs. RWD often comprise a single data source, such as a single Electronic Medical Record system that may be used only in certain inpatient settings or specialty practices such as oncology. The single system RWD may not include all sources of information relevant to an analysis of a cancer patient who typically sees a broad spectrum of specialty providers. Furthermore, the quality of many RWD sources is unsubstantiated, as in many instances they may not be validated or curated. This presents concerns for errors in the data which, if nonrandom, could lead to erroneous or skewed results. Finally, RWD are often not collected for research, requiring caution in utilizing them for purposes other than their original intent.

Population-Based Cancer Registries

Cancer registries, including the National Cancer Institute's Surveillance, Epidemiology and End Results (SEER) Program, represent a RWD source that could supplement and complement RCTs as registries can address many of the shortcomings of other types of RWD. It could be argued, in fact, that a population-based cancer registry (PBCR) is the ideal source for linking, expanding, and optimizing RWD given the legal framework under which they operate. Cancer is a reportable disease in every US state and reporting to the central registry is Health Insurance Portability and Accountability Act of 1996 (HIPAA) exempt as reporting is done for the public health good. PBCRs have the authority to maintain patient identifiers, aggregate cancer data for each patient, and follow patients over time for cancer-related outcomes. These data can also be used as a foundation to support population-based cancer research activities, including those that involve contacting patients with appropriate approval. As such, a PBCR is unique in its ability to connect cancer patients across various RWD sources building upon its existing clinically rich data set that covers an entire population of cancer patients under the purview of the individual registry. Figure 1 illustrates the SEER registry with existing and proposed data linkages that might be used to support RWD analyses.

PBCRs operate under state regulations requiring all healthcare providers to report information on all patients with cancer. There is little opportunity for biased representation of patients within a PBCR because of its inherent nature of being population based (that is including all residents of


Data sources currently used
Innovative pilot data sources
Hospital Registries
Abstracts, EHR, other data
Pathology Labs
E-path, images, other data
Oncology/ Physician Offices
E-path, images, claims, other data
Other (providers, surgical centers)
Administrative claims, other data
Other Registries
Infectious Disease, HIV/AIDS, other
DMV, NDI, Voter registration, SSN, other
Images (PDFs), other data files
Census
Demographic Data, SES, Geography
SEER Central Cancer Registry
SEER*DMS database
Biomarkers (genetic, genomic)
Pharmacy Claims
Electronic Health Records
Patient Reported Outcomes
Administrative Medical Claims
Surveys

*data collection from these sources is specifically related to cancer surveillance public health reporting

Figure 1 SEER linkages to traditional and novel real world data sources.

the geographic regions included in the registry's catchment area). These registries include information that is captured from a variety of standardized data sources and then merged through a process known as consolidation. Many of these sources involve extensive manual curation (review and decision-making) of data. For example, one SEER registry reports that there is an average of 3.6 different data sources that are adjudicated and consolidated for each cancer case. In addition, many of these data sources are received at the registry in real time as electronic data streams. For example, SEER registries receive real time pathology reports for almost 90% of pathologically confirmed cancer cases. Pathology laboratories are required under state regulation to report cancer data. Further, as part of the quality control process, PBCRs, particularly SEER, use trained personnel (certified tumor registrars) to perform extensive visual editing and manual review of key variables. This manual review contributes to higher accuracy and validity of the data compared to many other RWD sources. The SEER registries also perform ongoing and active follow-up of all cancer patients to ensure complete, accurate capture of survival time and date of death, including information on cause of death − data frequently missing from many other commonly used RWD sources.

Despite these advantages, PBCR also have several key limitations that require caution in their utility for application to RWD analyses. There is limited information available on detailed longitudinal treatment. The first course of therapy is captured; however, specific systemic therapy agent information is not currently maintained in discrete fields, thus reducing the value in understanding differences in treatment-related outcomes across patient subsets. Registries do not currently collect information on recurrence or disease progression—both of which are key outcomes to understanding the effectiveness of cancer treatments. Lastly, registry data are often not current when made available at the national level: there is typically a 1- to 2-year lag in data availability which significantly limits their value in assessing uptake and outcomes for new therapies. The latter is growing in importance since there is an increasing number of new pharmacologic agents being approved monthly—in 2017 alone there were 19 newly approved agents.[3]

Current SEER Efforts to Address-Known Limitations

The SEER Program is working to enhance current data to include more clinically relevant information to better support analyses that impact clinical practice and patient care. Specifically, SEER is working to (1) capture detailed longitudinal treatment variables through multiple collaborative partnerships, (2) capture improved information on recurrence through innovative informatics work with the Department of Energy (DOE) Labs and other academic partners, and (3) provide closer to real time data by reducing the lag in reporting through 2 major efforts. The SEER program initiatives include the development of novel data linkages and of natural language processing and deep learning tools for automation and real time data capture; both of which are essential to gathering and sharing data within a shorter timeframe from actual patient care.

While linkages are often the most efficient and comprehensive source across many types of clinical data, there remain numerous data items of clinical importance that are only available from largely unstructured text in Electronic Medical Records or other clinical reports. The work with deep learning and natural language processing will enable automated and real time capture of key elements from unstructured documents (such as pathology reports) that are received in real time at each registry but currently require manual data extraction. This approach can facilitate early incidence reporting and for rapid screening of newly diagnosed patients for eligibility in clinical trials and other studies. As the data sources utilized by SEER are expanded to include radiology reports and other documents, the automation will be developed to capture other key information such as recurrence and progression that can be used not only to better understand outcomes, but also to identify patients who may be eligible for a trial at the time of their recurrence diagnosis. In this effort, SEER is working with several partners to develop, test, and implement a system for automatically extracting specific, targeted data elements from existing data sources such as pathology reports and radiology reports. The lead partner in this effort is the DOE National Laboratories, whose data science expertise and computational capacity are being leveraged. With close to 95% of cancer patients in the United States being pathologically confirmed at the time of diagnosis and real time pathology data flowing into PBCRs, the ability to greatly reduce the delay in registry data availability is rapidly approaching reality due to the development of this new system.[4-7] Furthermore, as both pathologic and radiologic confirmation are key sources of recurrence identification, the infrastructure being developed by the DOE is also focused on automated identification and capture of cancer recurrence.

SEER staff are also working to capture more detailed treatment information including additional information for radiation treatment as well as detailed systemic therapy agents. For radiation oncology, we are working with several radiation oncology practices and ASTRO to determine a minimal dataset that would provide clinically relevant and longitudinal data on radiation therapy. Variables to be included are: anatomic site, modality, technique, date, and dose.

For capturing detailed systemic therapy, SEER staff are working to utilize claims data from physician offices and pharmacies. Currently SEER is receiving real time claims data from oncology providers to capture detailed, longitudinal treatment data. SEER has extensive experience with utilizing registry-linked claims data from a long-standing linkage with Medicare data (SEER-Medicare).[8] Work is ongoing to expand linkages to commercial insurers to provide better population level longitudinal treatment information. An example of the types of analysis of PBCR-linked claims data analyzed is provided in Figure 2, which illustrates the dissemination over time for 4

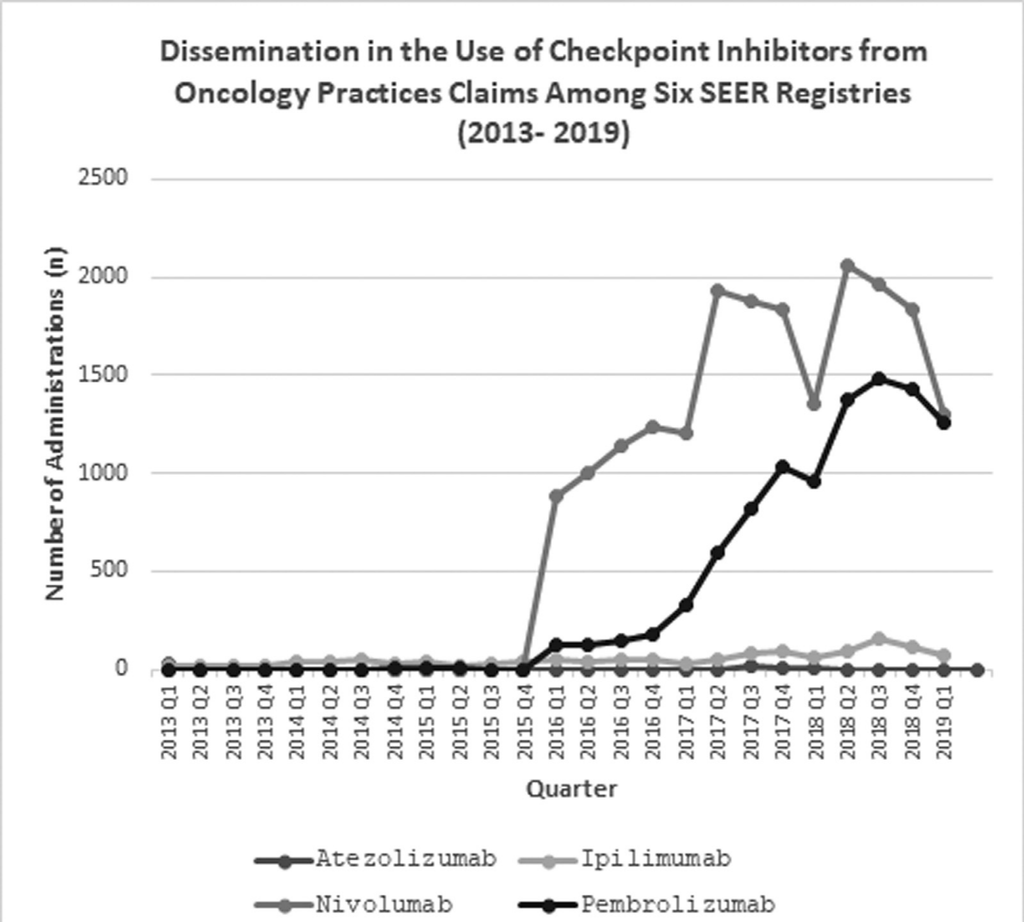


Figure 2 Dissemination in the use of checkpoint inhibitors from oncology practices claims (2013-March 2019).

checkpoint inhibitors in a subset of oncology practices in Georgia.

In addition to the data obtained from claims, there is a lack of RWD on the populations of cancer patients receiving oral antineoplastic medications. To address this need, SEER has been working with large commercial pharmacy organizations to link longitudinal oral antineoplastic medication claims with SEER to further enable research on medication utilization, disparities of care, and adherence to prescribed treatment. An example of the frequency of oral agents prescribed by class is illustrated in Table 1, which provides the number of patients receiving the 12 most

Table 1 The 12 Most Commonly Prescribed Oral Oncology Drug Classes from Georgia Administrative Pharmacy Claims (2013-2017)

Top 12 Drugs for Commercial Pharmacy Claims by Major Class	Number of Patients	Number of Prescription Fills
Alkylating agent	1820	8858
Antiandrogen	5161	35,075
Antimetabolite	15,146	136,479
Aromatase inhibitor	28,165	239,538
Cyclin dependent kinase inhibitor	661	4556
Enzyme Inhibitor	696	5561
Immunomodulator	1102	11,638
Monoclonal antibody	203	989
Plant alkaloid	183	529
Proteasome inhibitor	146	881
Selective estrogen receptor modulator (SERM)	16,393	187,006
Tyrosine kinase inhibitor	2706	19,928
Total	72,382	651,038

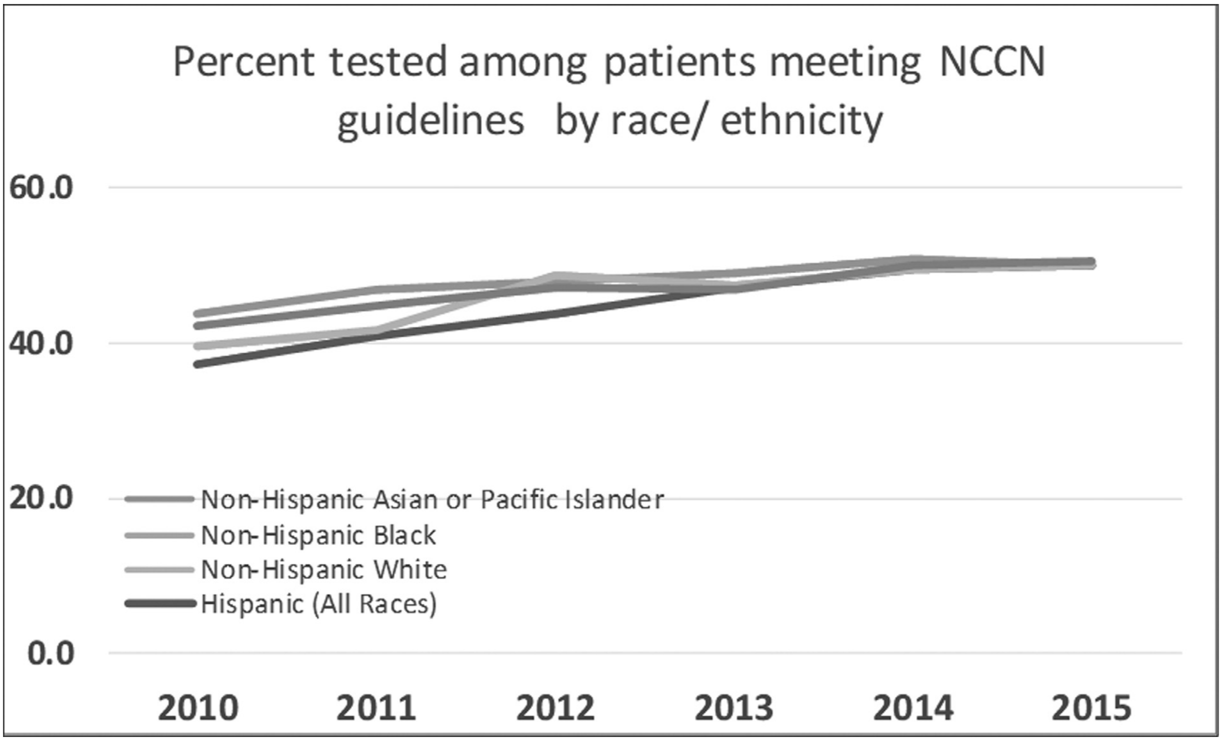


Figure 3 Oncotype DX testing over time by race/ethnicity in SEER from 2010 to 2015.

common oral antineoplastic treatments over a 4-year period in a single SEER registry. Even though this does not represent the entire US population, there were more than 100,000 unique patients who received >750,000 prescriptions filled, underscoring the importance of these data in understanding detailed treatment at the population level. While the data sources described above are not yet utilized for the entire SEER population, these efforts have demonstrated feasibility and the goal for SEER is to work to achieve such a level of completeness.

Genomic testing laboratories represent another key linkage. Genomic labs are also included in the public health reporting mandate. SEER is routinely linking with Genomic Health Incorporated to receive all Oncotype DX 21 and 16 gene assays since the test's approval.[9,10] An example of the use of such linked data is shown in Figure 3, which highlights the decrease over time in disparities of testing among women with node negative, HER2 negative early stage breast cancer by race and ethnicity.

In conclusion, PBCRs are in an unparalleled position to support the enhancement of existing RWD. Developing the types of activities described above at the population level, such as working across a myriad of external partnerships, while ensuring high-quality data capture is absolutely essential to enable us to understand the use of new therapies and the outcomes associated with these therapies outside the clinical trial setting.

References

1. Bleyer WA, Albritton K, et al: Lack of participation in clinical trials. In: Kufe DW, Pollock RE, Weichselbaum RR, eds. Holland-Frei Cancer Medicine (6th edition): HamiltonON: BC Decker; 2003
2. Sherman RE, Anderson SA, Dal Pan GJ: Real-world evidence – What is it and what can it tell us? N Engl J Med, 375:2293-2297, 2016
3. US Department of Health and Human Services Food And Drug Administration (FDA): Approved Drug Products. https://www.fda.gov/Drugs/InformationOnDrugs/ApprovedDrugs/default.htm. Accessed 21 February 2019
4. Qiu JX, Yoon HJ, Srivastava S: Scalable deep text comprehension for cancer surveillance on high-performance computing. BMC Bioinform 19:488, 2018
5. Rivera DR, Lee JSH, Hsu E, et al: Harnessing the power of collaboration and training within clinical data science to generate real-world evidence in the era of precision oncology. Clin Pharmacol Ther. 2019. http://dx.doi:10.1002/cpt.1459. [Epub ahead of print] PubMed PMID: 31166005.
6. Gao S, Qiu JX, Alawad M, et al: Classifying cancer pathology reports with hierarchical self-attention networks. Nature Methods (submitted for publication).
7. Alawad M, Gao S, Qiu JX, et al: Automatic extraction of cancer registry reportable information from free-text pathology reports using multi-task convolutional neural networks. JAMIA (pending NCI review - submitted for publication).
8. Warren JL, Klabunde CN, Schrag D: Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. Med Care 40: 2002(8 Suppl). p. Iv–3-18
9. Petkov VI, Miller DP, Howlader N, et al: Breast-cancer-specific mortality in patients treated based on the 21-gene assay: a SEER population-based study. NPJ Breast Cancer. 2:16017, 2016 http://dx.doi:10.1038/npjbcancer.2016.17
10. Lynch JA, Berse B, Petkov V: Implementation of the 21-gene recurrence score test in the United States in 2011. Genet Med 2016: 982-990. Oct 18