ELSEVIER

Seminars in
RADIATION
ONCOLOGY

Big Data in Oncology: Toward a Goal of Learning More From Every Patient

Ronald C. Chen, MD, MPH[*,†,‡]

The longstanding debate of randomized clinical trials (RCTs) vs "big data" research (a term often used synonymously with prospective or retrospective observational studies) goes something like this: proponents of RCTs emphasize that randomization is the most rigorous way to minimize observed and unobserved confounding among comparison groups, and is thus the research method most likely to produce internally valid results. On the other hand, proponents of observational research point to the well-known limitations of RCTs: long duration of time and high costs required for clinical trials to answer each question, as well as the limited participation of cancer patients on trials which reduce their external validity (generalizability).

This issue of *Seminars in Radiation Oncology* starts with writings by Soni et al[1] and Tsai et al[2] articulating this debate. As Soni et al[1] point out, current efforts to increase clinical trial participation and relax entry criteria will allow trial results to be more applicable to all patients, not just the younger and healthiest subgroup. They also raise concerns about the lack of agreement between results from RCTs and observational studies, hypothesizing that residual confounding in observational studies likely contribute to these result differences. Appropriately, the authors urge caution in interpreting comparative effectiveness studies using observational datasets.

The article by Tsai et al[2] describes some of the commonly used big datasets, and explains several misconceptions about big data research. These authors indicate that well-designed observational studies can indeed result in valid causal inferences, and provide results similar to clinical trials. When results from clinical trials and observational studies do not agree, Tsai et al point out the possibility that clinical trials could actually be wrong. Indeed, readers should be cautious when interpreting all studies including RCTs, and be cautious when applying RCT results to the clinical care of each patient, paying careful attention to whether specific patient groups are represented in the trials.

There are opportunities to learn from every cancer patient to drive more rapid advances in oncology. Currently, only 4% of cancer patients participate in clinical trials; even if this number increases, it is unlikely to ever approach near 100%. Big data bridge this important gap. In addition, as the article by DeWees et al[3] correctly indicates, not every important clinical question can be feasibly or ethically answered by RCTs. Thus, observational studies are an indispensable part of clinical cancer research. This article describes the strengths and limitations of common study designs, and analytic methodologies to reduce bias and confounding. Notably in observational studies, instrumental variable methodology has the potential to account for unobserved confounding, but finding a suitable instrument for each study is often difficult. DeWees et al also describe the issue of missing data and the importance of sensitivity analyses — both directly relevant to the validity of observational study results.

Population-based cancer registries represent existing efforts to learn from every cancer patient; and the Surveillance, Epidemiology, and End Results (SEER) registries are the best known examples. Within each SEER region, data are collected on nearly 100% of cancer patients, and SEER data have been used in thousands of published studies assessing patterns of cancer care and comparative effectiveness of different treatment options. Linkage of SEER with other datasets such as Medicare claims (SEER-Medicare) and patient-reported quality of life (SEER-MHOS) makes the datasets even richer.

Yet, as the article by Penberthy et al[4] acknowledges, current limitations in SEER data include the lack of information on comorbidities, radiation therapy details, systemic therapy agents, cancer recurrence, and subsequent treatments. However, the article also describes exciting current efforts by the National Cancer Institute to continue to enrich SEER data. These include the (1) use of natural language processing to abstract data from unstructured text in electronic health records (EHRs) and other clinical reports, which will allow collection of cancer recurrence data; (2) working with ASTRO to create a minimum dataset to collect data regarding radiation treatment; (3) linkage of SEER data with claims

[*]Department of Radiation Oncology, University of North Carolina, Chapel Hill, NC
[†]Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC
[‡]Cecil G. Sheps Center for Health Services Research, University of North Carolina, Chapel Hill, NC
Conflict of Interest: None.
Address reprint requests to Ronald C. Chen, MD, MPH, Department of Radiation Oncology, University of North Carolina — Chapel Hill, CB #7512, 101 Manning Drive, Chapel Hill, NC 27599.
E-mail: Ronald_chen@med.unc.edu

https://doi.org/10.1016/j.semradonc.2019.05.001

from pharmacies and insurers to collect details of systemic therapies (including oral agents) and longitudinal treatment data including for cancer recurrence; and (4) partnership with Genomic Health to link SEER with genomic testing data. Building upon the backbone of population-based registries and continuing to enrich the data collected for each patient is the best way toward realizing the goal of learning more from every patient.

There are many other exciting uses of cancer registry and other big data sources, including quality assessment and quality improvement, assessing cost and cost-effectiveness of care, and evaluating treatment-related side effects and patient health status. These data sources continue to evolve to become richer. The article by McCabe[5] describes another well-known large cancer registry, the National Cancer Database (NCDB), which pools data from approximately 1500 Commission on Cancer-accredited hospitals throughout the US and contains data on over 72% of newly-diagnosed cancer patients. The article describes the powerful ability of the Commission on Cancer to utilize its accrediting authority and the large NCDB infrastructure to assess and improve the quality of cancer care. The NCDB has a platform to assess National Quality Forum-endorsed quality measures, and provides annual performance reports to hospitals. Its Rapid Quality Reporting System has been developed to provide more timely feedback in order to detect and correct errors in patient care (such as patients receiving breast-conserving surgery but no adjuvant radiation therapy). During the pilot phase, one-third of participating hospitals prevented at least one patient error due to Rapid Quality Reporting System. As NCDB continues to move from retrospective data collection after a patient has completed primary treatment toward real-time data collection as patients are being diagnosed and treated, it will become an even more powerful tool for improving the quality of cancer care across the United States.

While NCDB can improve big-picture care quality based on data available in the cancer registry (eg guideline-adherence, timeliness of treatment), McNutt et al[6] describe the use of data to specifically improve the quality of radiation treatment. A compilation of a large amount of detailed patient data can be used to assist in accurate contouring of tumors and normal organs, and provide guidance on the optimal achievable radiation plan for each patient. The latter is currently being used in the cooperative group clinical trial NRG GY006; by reducing the variation in radiation plan quality across trial participating sites, this data-driven guidance is expected to significantly reduce radiation-associated toxicity. As this type of effort continues to mature, it is easy to envision a future where patients anywhere in the country can receive consistently high quality radiation treatment.

A pair of articles describe the evolving uses of big data to assess treatment-related toxicity. The article by Gross et al[7] describes the strengths and limitations of 3 different toxicity assessment methods: secondary data (including claims data) analysis, physician-graded toxicity (most commonly used in clinical trials), and patient-reported outcomes. Claims data, which can be linked with cancer registry data (eg, SEER-Medicare), represent a way to learn about treatment-related toxicity from every patient. As such, they can provide some types of information that are impractical to assess using clinical trials.

One example is in the assessment of a learning curve for new technologies. Gross et al cite a SEER-Medicare study which demonstrated relatively high genitourinary morbidity associated with robot-assisted radical prostatectomy when the technology was new; a repeat study several years later reported progressively declining complication rates over time. Claims data also continuously capture data from each patient, in contrast to discrete assessment time points in clinical trials, and also minimize the possibility of missing data especially for Medicare patients for whom loss or change of insurance is rare. However, claims data represent an indirect measure of toxicity, and rely upon clinicians to enter accurate diagnostic or procedure codes during the clinical encounter. As such, claims data are more likely to capture severe toxicities, especially those requiring procedures.

The article by Purswani et al[8] describes an exciting new way of assessing toxicity and patient health status. With a widespread use of cellular phones and continued popularization of smart wearables, it is now possible to use these devices to collect both subjective (patient-reported outcomes) and objective (step count, heart rate, and sleep) data from large numbers of patients. Multiple studies validating these measures continue, and clinical trials are starting to incorporate data from mobile devices. The potential to improve the care of cancer patients in real-time was demonstrated by a pilot study which used a smartphone app to monitor daily step counts of adult patients receiving chemotherapy.[9] For each patient, a decrease in step count by more than 15% triggered a phone call to the patient to assess and manage treatment-related toxicity; some of these phone calls resulted in sending the patient for urgent medical intervention.

With rising healthcare costs, it is increasingly important to evaluate the cost and cost-effectiveness of different cancer treatments, which are often best studied using administrative claims data. The article by Shih et al[10] describes commonly used data sources from Medicare and private insurers, and describes a study analyzing SEER-Medicare data to report national medical costs of cancer increasing from $124.6 billion in 2010 to $157.8 billion by 2020.[11] Another study used the Marketscan Commercial Claims and Encounters database to compare the costs associated with 3 different types of radiation technologies for privately insured prostate cancer patients: proton therapy, stereotactic body radiotherapy, and intensity-modulated radiotherapy.[12] Two-year total healthcare cost including treatment and subsequent medical management were $133,220 for proton therapy and $79,209 for intensity-modulated radiotherapy; stereotactic body radiotherapy costs were slightly lower. The cost data from claims can be used together with measures of effectiveness (eg, quality-adjusted life years) to model the cost-effectiveness of different treatments.

Finally, the widespread implementation of EHRs in almost every healthcare system across the United States has established the foundation for collecting data and learning from every patient. The challenge here is not a lack of data, but rather so much data (and often in unstructured formats) that more sophisticated analytic tools are needed to help make EHR data more meaningful. The article by Kim et al[13] describes evolving uses of machine learning and natural language processing methods for this purpose. As these methods continue to improve and as EHRs continue to be redesigned to better facilitate research (in addition to patient care and billing), these data will become increasingly useful toward the goal of learning more from every patient.

How exciting! If nearly all the medical care for nearly all cancer patients is captured in EHRs, then it is easy to envision a future where population-based cancer registries can collect more comprehensive data for each patient, and as a result reduce current concerns about unmeasured confounding in observational research. The ability exists to add further information regarding costs, toxicity, patient-reported outcomes, and even continuous monitoring of patient health status from mobile devices; as well as for real-time feedback to clinicians to reduce errors and improve the quality of cancer care. The articles in this issue of *Seminars in Radiation Oncology* together paint a picture where big data to will allow learning more from every patient, for the purpose of improving the care and outcomes of current and future cancer patients.

References

1. Soni PD, Spratt DE: Population-based observational studies in oncology: proceed with caution. Sem Radiat Oncol 2019
2. Tsai CJ, Riaz N, Gomez SL: Big data in cancer research - real-world resources for precision oncology to improve cancer care delivery. Sem Radiat Oncol 2019
3. DeWees TA, Vargas CE, Golafshar MA, et al: Analytical methods for observational data to generate hypotheses and inform clinical decisions. Sem Radiat Oncol 2019
4. Penberthy L, Rivera DR, Ward K: The contribution of cancer surveillance toward real world evidence in oncology. Sem Radiat Oncol 2019
5. McCabe RM: National Cancer Database - the past, present, and future of cancer registry and its efforts to improve the quality of cancer care. Sem Radiat Oncol 2019
6. McNutt T, Moore KL, Wu B, et al: Use of big data for quality assurance in radiation therapy. Sem Radiat Oncol 2019
7. Gross MD, Al Awamlh BAH, Hu JC: Assessing treatment-related toxicity using administrative data, patient-reported outcomes, or physician-graded toxicity - where is the truth? Sem Radiat Oncol 2019
8. Purswani JM, Dicker A, Champ C, et al: Big data from small devices: The future of smartphones in oncology. Sem Radiat Oncol 2019
9. Soto-Perez-De-Celis E, Kim H, Rojo-Castillo MP, et al: A pilot study of an accelerometer-equipped smartphone to monitor older adults with cancer receiving chemotherapy in Mexico. J Geriatr Oncol 9:145-151, 2018
10. Shih Y-CT, Liu L: Use of claims data for cost and cost-effectiveness research. Sem Radiat Oncol 2019
11. Mariotto AB, Yabroff KR, Shao Y, et al: Projections of the cost of cancer care in the United States: 2010-2020. J Natl Cancer Inst 103:117-128, 2011
12. Pan HY, Jiang J, Hoffman KE, et al: Comparative toxicities and cost of intensity-modulated radiotherapy, proton radiation, and stereotactic body radiotherapy among younger men with prostate cancer. J Clin Oncol 36:1823-1830, 2018
13. Kim E, Rubinstein S, Nead KT, et al: The evolving use of electronic health records (EHR) for research. Sem Radiat Oncol 2019