ELSEVIER

Seminars in
RADIATION
ONCOLOGY

Big Data in Cancer Research: Real-World Resources for Precision Oncology to Improve Cancer Care Delivery

Chiaojung Jillian Tsai, MD, PhD,* Nadeem Riaz, MD, MS,* and Scarlett Lin Gomez, MPH, PhD†

In oncology, the term "big data" broadly describes the rapid acquisition and generation of massive amounts of information, typically from population cancer registries, electronic health records, or large-scale genetic sequencing studies. The challenge of using big data in cancer research lies in interdisciplinary collaboration and information processing to unify diverse data sources and provide valid analytics to harness meaningful information. This article provides an overview of how big data approaches can be applied in cancer research, and how they can be used to translate information into new ways to ultimately make informed decisions that improve cancer care and delivery.
Semin Radiat Oncol 29:306–310 © 2019 Published by Elsevier Inc.

Introduction

In medicine, "big data" often refers to extensive records on a large number of patients, consisting of either structured or unstructured clinical information that can include patient characteristics, diagnostic and treatment history, genomic and molecular data, and sometimes billing history. The complexity and diversity of this vast amount of data require significant expertise to process. Proper study designs and thoughtful analytical strategies are necessary to turn big data into meaningful information for treatment decisions and patient care. Since big datasets usually contain real-world information, their use could bridge gaps between clinical and translational studies.[1] This may be especially important in some oncologic settings, where pharmaceutically funded trials often use suboptimal control arms and then leave an unclear decision between a new agent and the best existing standard of care regimen.[2] Moreover, randomized controlled trials include, on average, 4% of all cancer patients, and are known to underrepresent minorities and other underserved populations.[3,4] The other 96% of the cancer patient population that does not participate in trials looks very different from trial populations, not only in terms of sociodemographic factors, but also in terms of health status (ie, more comorbidities).[5,6] The lack of representativeness of trial populations thus makes clinical applicability of trial findings challenging.

At the 2018 American Society of Clinical Oncology Annual Meeting, the director of the National Cancer Institute (NCI), Dr. Norman E. Sharpless, identified big data as one of the 4 "areas of particular opportunity" worthy of focus for researchers. He also emphasized that the scientific community "must move from passive data sharing to data aggregation, establishing linkage and interoperability of diverse, complex data sets to understand cancer care, and provide real-world evidence."

In this article, we review the different types of big data and their potential applications in clinical oncology. We also address current trends in methods to improve data linkage and sharing. Furthermore, we dispel some common myths and misconceptions and lay out future directions for enhancing data access to enhance interdisciplinary collaborations and ultimately improve cancer care and delivery.

Sources and Forms of Big Data

National Population-Based Cancer Databases

The most notable population-based cancer databases are those assembled by NCI's Surveillance, Epidemiology, and

*Departement of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, NY
†Department of Epidemiology & Biostatistics, School of Medicine, University of California, San Francisco, CA
Supported in part by NIH/NCI Cancer Center Support Grant P30 CA008748
Conflict of Interest: None.
Address reprint requests to Chiaojung Jillian Tsai, MD, PhD, Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, NY 10065. E-mail: tsaic@mskcc.org

https://doi.org/10.1016/j.semradonc.2019.05.002
1053-4296/© 2019 Published by Elsevier Inc.

End Results (SEER) Program and the Centers for Disease Control and Prevention's National Program of Cancer Registries (NPCR). SEER currently collects and publishes cancer incidence and survival data from population-based cancer registries covering approximately 35% of the US population. SEER gathers patient demographics (age, gender, race, ethnicity, and birthplace); cancer characteristics (tumor cell types, biological and clinical aspects, and some biomarker and genomic information on tumors); stages of disease; treatment information (surgery, radiation, chemotherapy, hormone therapy, and immunotherapy); and patient outcomes (vital status and cause of death).[7]

NPCR was established by Congress in 1992 and is administered by the Centers for Disease Control and Prevention. It collects data on cancer occurrence, types of initial treatments, and outcomes. Together, NPCR and SEER capture data on 97% of the US population.[8] Such near-universal coverage allows researchers to evaluate the burden of cancer and identify needs for cancer prevention and control efforts at the national, state, and local levels.

Additionally, since 1990, the American College of Surgeons and the American Cancer Society have jointly sponsored the National Cancer Database (NCDB).[9,10] To date, NCDB covers more than 1500 Commission on Cancer-accredited facilities, representing more than 70% of newly diagnosed cancer cases nationwide and containing more than 34 million historical records.[11]

Although SEER, NPCR, and NCDB encompass data from most cancer patients in the United States, they do not capture the entire treatment course of each patient. Both longitudinal follow-up information and recurrence data are not included. Furthermore, often the available end point is overall survival, but no other details are given. There is no information beyond the first course of treatment and no second-line or salvage therapy recorded. Treatment information in the population-based cancer registries (SEER and NPCR) is very high level and specific protocols, drug names, and doses are not recorded. These limitations hinder the application of research efforts based solely on the current form of these national population databases. Recognizing these limitations, SEER has multiple ongoing efforts to enhance its current data (see article in this issue by Penberthy et al).

American Society of Clinical Oncology launched the CancerLinQ initiative, an attempt to assemble data from every cancer patient in the United States, and make them available for analyses.[12,13] CancerLinQ aggregates data from electronic health records (EHRs) via direct feeds without needing to reformat the data source. CancerLinQ then processes and transforms the datasets through cloud-based algorithms.[14] This is a challenging undertaking; since there is no common data format in real-world clinical settings or EHR systems, several computer algorithms are used to transform raw data into structured data suitable for future analyses. However, some unstructured clinician notes will require manual data abstraction and/or natural language processing. CancerLinQ now partners with SEER to facilitate information sharing, both to strengthen the nation's cancer surveillance efforts and to improve patient care outcomes.

Commercial and Private Cancer Databases

Several commercial big data companies have been launched to collect and compile real-world data from medical records, clinical notes, and billing information to give real-time feedback to cancer care stakeholders on treatments and outcomes. One notable example is Flatiron Health, which has created OncologyCloud for this purpose.[15] Flatiron's network comprises over 250 cancer clinics with 1.5 million active patients, compiled into a single data system via a cloud-based EHR platform. Their analytical algorithms can streamline the specific EMR system and analyze cost of individual patient care, assemble quality metrics, and identify potential clinical trial candidates. Flatiron's system has interestingly been both integrated with clinical genomics data from Foundation Medicine and been used by the Food and Drug Administration to evaluate the role of "real world evidence."[16,17] Other examples of health maintenance organization and private-sector databases include the Kaiser Permanente Clinical Research Networks and MarketScan by Truven Health.[18]

These types of big data differ from conventional large databases created solely to collect information for clinical trials or research purposes. A key distinction is that the data sources described above can be used to harvest unstructured data and provide real-time analysis and feedback, thereby increasing the value of the information gained for cancer care delivery.

Cancer Genomics and Other "-Omics" Databases

The Cancer Data Science Laboratory was established in 2018 by NCI's Center for Cancer Research. Its mission is to generate computational algorithms for analyzing and integrating cancer omics laboratory and patient data.[19] These computational methods can then be used to address fundamental research questions about genetic predisposition, susceptibility, and responsiveness to treatment of cancer.

The Cancer Genome Atlas (TCGA), supported by NCI and the National Human Genome Research Institute is one of the earliest large scale publicly accessible and collaborative genomic dataset.[20] As of March 2019, the TCGA has sequenced and molecularly profiled tumors from over 33,000 individuals with nearly 70 different types of cancer − analyzing over 22,000 genes, and discovering 3,140,000 mutations. Importantly the unique nature of the in-depth publically available molecular profiling generated by the TCGA has been reused by researcher all over the world, contributing to over 5000 studies.

The NCI's Clinical Proteomic Tumor Analysis Consortium is a consortium of institutions and investigators that uses pan-"omic" analyses to evaluate the molecular basis of cancer.[21,22] Proteomics data generated by Clinical Proteomic Tumor Analysis Consortium research projects are stored in a publicly available data repository accessible to researchers worldwide.

Additionally, the NCI's Therapeutically Applicable Research to Generate Effective Treatments (TARGET) program collaborates with different clinical trials cooperative groups and consortia to obtain clinical information and accrued tissue materials to generate, analyze, and interpret genomics data.[23] TARGET uses a "multiomic" approach and employs various sequencing and array-based methods to examine genomes and transcriptomes.

The development of genome-guided targeted therapies have led many academic cancer centers to develop their own panel-based DNA sequencing cancer (eg, Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets)[24] or use commercially available panels such as those from Foundation Medicine. The initial generation of this data has remained siloed; however, the American Association for Cancer Research has undertaken efforts to unify this clinically generated data centrally in the Genomics Evidence Neoplasia Information Exchange (American Association for Cancer Research-Genomics Evidence Neoplasia Information Exchange Project).

Data from these initiatives have helped researchers better understand the genetic compositions of different cancers and identify therapeutic targets. However, in some respects, we have only seen the tip of the iceberg and there remain vast amounts of data to be gathered and deciphered. As a result, there has been a new push in the research community to find ways to make these data, and the tools to analyze them, more widely accessible.

Common Myths and Misconceptions of Big Data Research

With increasing interest in, and use of, big data in clinical oncology research, some misconceptions have emerged. Below we list a few common ones and provide evidence to dispel such misconceptions. To clarify the nomenclature, all big data studies, whether genomic or registry-based, are observational studies.

1. *Databases containing information on molecular or other -omics data provide the greatest value and validity for big data research, whereas population registry databases containing primarily patient-level clinical information are more subject to bias.*

Some may believe that large-scale genomic databases do not face the same challenges as population-based registry databases. Yet all observational studies, regardless of their types of data, are subject to biases. For example, TCGA has characterized over 20,000 primary cancers and matched normal samples spanning 33 cancer types, an enormous undertaking that has proven extremely valuable in characterizing different tumors and potential therapeutic responses. However, tumor samples in TCGA were obtained from a highly selected population, many of whom were clinical trial participants. In addition, ethnic diversity is lacking in TCGA samples.[25] Similarly, in another recent analysis of SEER data from 4 laboratories that performed nearly all germline cancer genetic testing,[26] blacks had lower testing rates compared to whites. Therefore, genomic sequencing studies are subject to selection bias and may not represent the entire US population. Furthermore, if factors associated with receiving genetic testing are causally associated with prognosis/study outcomes, external validity of results could be compromised.

2. *Big observational databases are best suited for generating hypothesis, especially studies using population-based registry databases. Big data should not be used for assessing causal risk factors or patient treatment outcomes, or conducting comparative effectiveness studies.*

Observational studies, when *properly designed and conducted*, can yield valid causal inferences and provide results similar to that of clinical trials.[27] A prime example was the causal relation between cigarette smoking and lung cancer validated by several observational studies.[28] Such a causal relationship was not shown or easily studied in randomized clinical trials. Another example is the recent Food and Drug Administration approval of expended indications for palbociclib combination therapy to include men with certain advanced or metastatic breast cancer. The approval was based on data from EHRs and postmarketing reports of the real-world databases. The article in this issue by DeWees et al, and others, have addressed study design considerations for answering comparative effectiveness research questions using observational data.[29-32]

3. *When results from studies conducted using real-world populations differ from those of randomized clinical trials, the latter always trump real-world evidence.*

This idea is pervasive among clinicians, especially those with limited formal training in research methodology and study design. To elaborate on the example of smoking and cancer, a randomized intervention trial of smoking cessation found that those who received a smoking cessation intervention eventually had a higher incidence of lung cancer than those not urged to stop smoking.[33] These results did not negate prior findings from observational studies, and were attributed to inherent problems in the trial. Doubtless, most novel therapeutics or interventions are better tested in randomized controlled clinical trials. However, when properly designed, observational studies using real-world data can be as valid as clinical trials. It is easy to confuse internal and external validity. It is plausible that real-world observational studies, either genomic- or population-based, need not conform to results from randomized clinical trials, just as rigorously tested in vivo studies often do not agree with in vitro studies. Properly conducted randomized clinical trials can be internally valid in their conclusions within their study populations; however, since only about 4% of people participate in clinical trials, their characteristics differ from the general public, thereby

affecting the external validity (ie, generalizability) of trials in the real-world setting. Therefore, the type of study should not be taken as a guide to a study's validity[27] and failure to explain discrepancies should not automatically refute the validity of observational studies.[34]

4. *Bigger data are better than smaller data.*

Quality of the data is more important than quantity of the data, especially for certain questions — more is not necessarily better. Haphazardly collected and unprocessed data, even when analyzed with great accuracy, provide limited inherent value to the users. Clean and quality-controlled data are thus far more valuable and effective.

Future Directions of Big Data in Precision Oncology

Cultivating Cross Talks and Collaborations Between Researchers of Different Disciplines

Debates over the superiority of a study type to another are not always productive. Patients will derive the most benefits from interdisciplinary collaboration of researchers. Clinicians and data scientists will benefit from collaborations with genetic epidemiologists to properly design and carry out population-based genomic studies.[35,36] Additionally, more efforts should be put into improving the methodological rigor of clinical and observational studies rather than disputing the merits of various types of studies. In reality, very similar results can be obtained from both observational study designs and randomized clinical trials.[37] Clinicians wishing to conduct large-scale observational studies should consult population scientists to ensure meticulous scrutiny of patient identification, end-point selection, and analytical strategy. Large-scale observational studies will only be increasingly valuable for clinical situations when every single aspect of patient care and tumor characteristics will be collected in the future.

Advancing Analytical Strategies and Cancer Bioinformatics

Advancing data processing technologies and cancer bioinformatics is of paramount importance in deciphering big data. Bioinformatics uses advanced mathematical algorithms and technological platforms to store and transform data into an interpretable format. It is often difficult to integrate and analyze data from various platforms. Recently, there has been an increasing usage of virtual repositories, or "data clouds," to link and improve access to research data. Additionally, cognitive computing (ie, artificial intelligence) and machine learning are gaining popularity. With adequate "training," these new technologies will be able to identify differential therapeutic outcomes of a particular therapy, develop cancer treatment pathways, discover new cancer etiologies, and help deliver personalized interventions.[38]

Conclusions

As technology continues to progress in a fast pace, we may soon be able to obtain data on every treatment course, recurrence, genomic composition, and financial/quality measures from virtually every cancer patient. The general goal of big data is to link current population databases and gather comprehensive data on all cancer patients for more thorough analyses. These technologies will soon allow clinicians and researchers to learn from potentially all cancer patients, and no one can deny the incredible potential of this to rapidly improve learning and advances in clinical cancer care. Many of the current treatment paradigms are based on highly selected clinical trials encompassing only a small fraction of total patients with cancer. The advent of big data with massive clinic-genomic variables in large patient populations will be able to dissect common malignancies into distinct subtypes. Large prospective trials looking at a small subset of selected tumor subtypes will then be difficult to accomplish and less useful in this setting. Big data approaches will not only be used to bridge the gap between clinical trials; in many instances it will be the favored type of study in the future. With meticulous study design, data quality assurance, and sound analytical strategy, meaningful clinic-genomic information and treatment outcomes can be collected to aid clinical decisions in real time. Ultimately, big data will help deliver precision oncology treatment and patient care with greater accuracy and efficiency.

References

1. Dewdney SB, Lachance J: Electronic records, registries, and the development of "Big Data": Crowd-sourcing quality toward knowledge. Front Oncol 6:268, 2016
2. Hilal T, Sonbol MB, Prasad V: Analysis of control arm quality in randomized clinical trials leading to anticancer drug approval by the us food and drug administration. JAMA Oncol 5(6):887-892, 2019
3. Enrollment of Racial Minorities in Clinical Trials: Old Problem Assumes New Urgency in the Age of immunotherapy. Bassel Nazha, Manoj Mishra, Rebecca Pentz
4. Taofeek K. Owonikoko. American Society of Clinical Oncology Educational Book 39:3-10, 2019
5. Hutchins LF, Unger JM, Crowley JJ, Coltman Jr CA, Albain KS: Underrepresentation of patients 65 years of age or older in cancer-treatment trials. N Engl J Med 341:2061-2067, 1999. PMID: 10615079
6. Elting LS, Cooksley C, Bekele BN, et al: Generalizability of cancer clinical trial results: Prognostic differences between participants and nonparticipants. Cancer 106:2452-2458, 2006. PMID: 16639738
7. Surveillance, Epidemiology, and End Results (SEER) Program. 2019. https://seer.cancer.gov/ztml(Accessed 1 May 2019).
8. National Program of Cancer Registries (NPCR). 2019. https://www.cdc.gov/cancer/npcr/about.htm.
9. Garfinkel L: The National Cancer Data Base: A cancer treatment resource. CA Cancer J Clin 43:69-70, 1993
10. Murphy GP: The National Cancer Data Base. CA Cancer J Clin 41:5-6, 1991
11. National Cancer Database (NCDB). 2019. https://www.facs.org/quality-programs/cancer/ncdb (Accessed 1 May 2019).
12. ASCO forges ahead with CancerLinQ. *Cancer Discov*2014;4: OF4.
13. Sledge Jr. GW, Miller RS, Hauser R: CancerLinQ and the future of cancer care. Am Soc Clin Oncol Educ Book 2013: 430-434. PMID: 23714566
14. Miller RS, Wong JL: Using oncology real-world evidence for quality improvement and discovery: The case for ASCO's CancerLinQ. Future Oncol 14(1):5-8, 2018

15. Flatiron Health. 2019. https://flatiron.com/oncology/ (Accessed 1 May 2019).
16. Singal G, Miller PG, Agarwala V, et al: Association of patient characteristics and tumor genomics with clinical outcomes among patients with non-small cell lung cancer using a clinicogenomic database. JAMA 321:1391-1399, 2019
17. Khozin S, Abernethy AP, Nussbaum NC, et al: Characteristics of real-world metastatic non-small cell lung cancer patients treated with nivolumab and pembrolizumab during the year following approval. The Oncologist 23:328-336, 2018
18. Kulaylat AS, Schaefer EW, Messaris E, et al: Truven health analytics marketscan databases for clinical research in colon and rectal surgery. Clin Colon Rectal Surg 32:54-60, 2019
19. Ruppin E.. Cancer Data Science Laboratory. 2019. https://ccr.cancer.gov/cancer-data-science-laboratory (Accessed 1 May 2019).
20. Wang Z, Jensen MA, Zenklusen JC: A practical guide to The Cancer Genome Atlas (TCGA). Methods Mol Biol 1418:111-141, 2016
21. Edwards NJ, Oberti M, Thangudu RR, et al: The CPTAC data portal: A resource for cancer proteomics research. J Proteome Res 14:2707-2713, 2015
22. Rudnick PA, Markey SP, Roth J, et al: A description of the clinical proteomic tumor analysis consortium (CPTAC) common data analysis pipeline. J Proteome Res 15:1023-1032, 2016
23. TARGET: Therapeutically Applicable Research To Generate Effective Treatments. 2019.
24. Cheng DT, Mitchell TN, Zehir A, et al: Memorial sloan kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT): A hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. J Mol Diagn 17:251-264, 2015
25. Spratt DE, Chan T, Waldron L, et al: Racial/ethnic disparities in genomic sequencingracial/ethnic disparities in genomic sequencingracial/ethnic disparities in genomic sequencing. JAMA Oncol 2:1070-1074, 2016
26. Kurian A.W., Ward K.C., Howlader N., et al. Genetic testing and results in a population-based cohort of breast cancer patients and ovarian cancer patients. *J Clin Oncol*;0: JCO.18.01854.
27. Rothman KJ: Six persistent research misconceptions. J Gen Intern Med 29:1060-1064, 2014
28. Hammond EC, Horn D: The relationship between human smoking habits and death rates: A follow-up study of 187,766 men. JAMA 155:1316-1328, 1954
29. Bradley CJ, Penberthy L, Devers KJ, et al: Health services research and data linkages: issues, methods, and directions for the future. Health Serv Res 45(5 Pt 2):1468-1488, 2010
30. 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 90:11-24, 2014
31. Mathes T, Pieper D: Study design classification of registry-based studies in systematic reviews. J Clin Epidemiol 93:84-87, 2018
32. Meyer AM, Wheeler SB, Weinberger M, et al: An overview of methods for comparative effectiveness research. Semin Radiat Oncol 24:5-13, 2014
33. Shaten BJ, Kuller LH, Kjelsberg MO, et al: Lung cancer mortality after 16 years in MRFIT participants in intervention and usual-care groups. Multiple Risk Factor Intervention Trial. Ann Epidemiol 7:125-136, 1997
34. Soni PD, Hartman HE, Dess RT, et al: Comparison of population-based observational studies with randomized trials in oncology. J Clin Oncol 37:1209-1216, 2019
35. Whittemore AS, Nelson LM: Study design in genetic epidemiology: theoretical and practical considerations. J Natl Cancer Inst Monogr 1999: 61-69. PMID: 10854488
36. Wang Q, Lu Q, Zhao H: A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing. Front Genet 6:149, 2015
37. Anglemyer A, Horvath HT, Bero L: Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev 2014(6):Mr000034. https://doi.org/10.1002/14651858.MR000034.pub2
38. Kantarjian H, Yu PP: Artificial intelligence, big data, and cancer. JAMA Oncol 1:573-574, 2015