Translating “Big Data” in Oncology for Clinical Benefit: Progress or Paralysis

Anna D. Barker[1,2] and Jerry S.H. Lee[1,3]

ABSTRACT

The molecular characterization of cancer through genomics, data from multiomics technologies, molecular-driven clinical trials, and internet-enabled devices capturing patient context and real-world data are creating an unprecedented big data revolution across the cancer research-care continuum. While big data has translated to benefit for some patients, it has also created new problems. Our intent in this brief communication is to explore some examples of progress and key challenges that remain. The problems are not intractable, but success will require rethinking and rebuilding an information and evidence-based learning system that moves beyond paralysis to shape a better future for patients with cancer.

Introduction

Although there is no common definition of “big data,” it has been described as “large-scale datasets with complex organization that arise from different sources and fields (e.g., genomics, physiology, imaging, health informatics, real -world data, etc.).” Currently, a multiomics profile of a single cancer patient’s sample can produce from 2–4 terabytes of data or more; and when integrated with myriad other clinical and disease measures demonstrate that building a big data foundation for “precision oncology” is challenging but feasible. However, it is currently estimated that of the 8% of patients with cancer who qualify for big data–driven targeted therapies, only 5% will benefit (1). A major reason more patients are not benefitting from big data is because data analysis has lagged significantly behind data generation across the discovery to care continuum (2). Given the sheer volume and variety of data types, only a small percentage of oncology data has been analyzed to date and even the best algorithms are fraught with risk in oncology due to the difficulties of applying machine learning and artificial intelligence (AI) to a dynamic system such as cancer (3). Although big data may well prove to be transformative in diagnosing, treating, and preventing cancer, progress in converting data to information and using the information to realize patient benefit is slow. There is general acceptance that information from these massive data efforts can be transformative for patients, but issues must be addressed before most patients will benefit (**Table 1**).

How Genomics Came to Drive Big Data in Oncology

Cancer is an extraordinarily complex and heterogenous disease. Despite 50 years of concentrated research by thousands of scientists, understanding cancer at a fundamental level remains the greatest challenge of cancer research. Cancer is often described as a disease of the genome, reflecting the fact the disease results from aberrations in the normal genome. The long history of aberrant gene discovery in cancer is a major driver of today’s oncology’s big data “tsunami.” The early discoveries of key cancer genes (*SRC, RAS, TP53*, etc.) inevitably led to the concept of targeting proteins coded by these abnormal genes as a path to understand and control cancer. FDA approval in 1998 of trastuzumab for HER2-positive breast cancers and in 2002 of imatinib targeting BCR-ABL in patients with chronic myeloid leukemia provided momentum to the search for other targeted agents in oncology. The completion of the sequence of the normal human genome in 2004 was a watershed moment for all of biomedicine, especially cancer, that catalyzed a massive search for similarly effective cancer targets in the dysregulated cancer genome. Building on decades of genome-centric discoveries in cancer, finally knowing the normal genome sequence provided a strong rationale for applying sequencing technologies to systematically identify (all or most) aberrant genes in cancer. These events set the stage for large-scale sequencing projects in oncology including The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (4).

TCGA was the first “big science” project with an original goal to develop a comprehensive catalog of the genomic changes in 20 cancers. TCGA’s mission was to provide the high-quality curated data needed to discover new targets for cancer diagnosis, treatment, and prevention. Key goals for the pilot project included establishing the network of genome characterization and sequencing centers, a centralized biospecimen resource and data center to sequence up to three tumors. The project was scaled from pilot phase in 2008 to ultimately sequence and provide multiomics data on 33 different types of cancer including 10 rare tumors—producing 2.5 petabytes of publicly available data. TCGA set high standards for data quality, showed the value of machine-readable and computable data, identified new driver genes and demonstrated the value of multiomics data in establishing tumor subtypes. TCGA and other large scale genome sequencing projects have driven sequencing technology to where it is estimated that 100 million to 2 billion human genomes could be sequenced by 2025 producing up to 1 zettabase of sequence data per year (5).

Examples of Big Data–Driven Progress

As we reflect on lessons learned, data generated, and frameworks created by “the Gleevec moment” just 20 years ago and projects such as TCGA, clearly the “big data” revolution has impacted almost every

[1] Lawrence J. Ellison Institute for Transformative Medicine, Los Angeles, California. [2] Complex Adaptive Systems Initiative and School of Life Sciences, Arizona State University, Tempe, Arizona. [3] Keck School of Medicine and Viterbi School of Engineering, University of Southern California, Los Angeles, California.

Corresponding Author: Anna D Barker, Ellison Institute for Transformative Medicine, Los Angeles, CA 90064. Phone: 503-703-1945; E-mail: abarker@eitm.org

Cancer Res 2022;82:2072–5

doi: 10.1158/0008-5472.CAN-22-0100

Table 1. Select examples of Big Data's impact in oncology.

Clinical research phase	Areas of advancement	Synopsis of clinical impact from Big Data	Challenges ahead
(i) Discovery	Individual and pan-cancer subtypes (2, 6)	Molecular profiling of biospecimens has generated new molecular classifications within tissue types (e.g., primary and secondary GBMs) as well as finding molecular commonalities across tissue types that have advanced discovery and development of interventions (e.g., NTRK).	Need to keep pace in developing actionable targeted interventions as new cancer and pan-cancer subtypes are discovered and analytically validated using growing datasets to improve both decision science and drug discovery.
(ii) Translation	Multiomics approaches (3)	Reuse of retrospective cohorts to inform prospective multiomics has demonstrated the need for orthogonal measurements using biospecimens collected at the same time and longitudinally, as shown by genome to proteome studies.	As omics technologies continue to emerge and evolve, it is critical that these studies are conducted with high-quality biospecimens to avoid secondary analysis of pooled data, with new analytic techniques mistaking batch effects as real biological/clinical insights.
	Targeted therapies with companion diagnostics (2, 7)	Between 2000 and 2020, 124 anticancer agents were approved by FDA. In 2021, 15 were approved including the first KRAS inhibitor. Since 2016, there have been four biomarker-driven, tumor agnostic–approved indications of targeted therapies and immunotherapies across 20 cancers (e.g., MSI+, *NTRK*, and TMB-H).	Need an optimized solution to update treatments and feedbacks based on molecular profiles and actionable alterations as analysis across 11 countries between 2014 and 2020 reported only 20% of patients received treatment informed by molecular tumor board.
(iii) Clinical trials	Novel molecular biomarker-driven clinical trials (8)	Adaptive platform trials to stratify broad cohorts into different subtypes based on molecular profile and clinical phenotypes and to test different treatment strategies on the basis of a predefined decision algorithm.	Adaptive trials such as I-SPY 2, an adaptive platform trial for neoadjuvant breast cancer launched in 2009, demonstrate that these innovative trials must be comprehensive in collecting longitudinal data to establish subtypes for ever-increasing precision in therapy selection.
(iv) Delivery	Therapy selection—AI-based decision oncology (9)	Use of clinicogenomics and real-world evidence outcome datasets to determine benefit and to develop sequencing of combination regimens informed by molecular and clinical factors of individual patients.	As combination regimens continue to be approved across tumor types, how do we balance N-of-one "evidence" when an individual patient is choosing his/her initial treatment plan?

Abbreviation: GBM, glioblastoma.

aspect of cancer research from discovery to clinical application. Although early, the use of molecular profiling and other technologies such as advanced imaging, digital pathology, immunology, and innovative clinical trials have moved the pharmaceutical and biotechnology industries to shift their focus from cytotoxic drugs to more targeted agents—including immunotherapies. This big data focus combined with associated advanced analytics has led to approval of unprecedented numbers of new targeted drugs across several cancer types and numerous advances in diagnostics ranging from specific "omics" profile–directed therapies to companion diagnostics for drug development (2). The last few years have seen significant progress in applying big data to advance precision oncology concept to reality. Although far from exhaustive, the areas and highlights shown in **Table 1** reflect a breathtaking pace of change that hopefully portend progress versus paralysis:

(i) **Discovery.** TCGA and other big data projects drove development of targeted gene panels that support the identification of cancer subtypes significantly influencing clinical practices (2). For example, in 2016, the World Health Organization incorporated molecular criteria for the first time to redefine glioblastomas into primary and secondary diseases based on molecular differences, prevalence, and survival rate (6). Including these molecular characteristics as part of standard molecular work up is critical as the results from limited surgical samples used to improve decision-making after surgery will become clinical data elements for future pooled secondary analysis.

(ii) **Translation.** We have vast amounts of data in oncology that was developed either before or very early in the evolution of next-generation sequencing (NGS). Lessons learned in large-scale genomics (e.g., TCGA) and proteomics (e.g., CPTAC) programs have demonstrated the vital importance of determining fit-for-purpose quality of biospecimens to comprise true multiomics for individual patients. In addition, databases used to house multiomics data for future secondary analysis will require careful attention to context of use for each study, especially as the use of companion diagnostics to determine actionable treatments evolve within and across tumor types over time (7).

(iii) **Clinical trials.** Validating molecular signatures requires clinical trials that test multiple interventions against standard of care with longitudinal follow-up and designs that speed the testing of the effectiveness of new drugs and combinations. Launched in 2010, the I-SPY2 Bayesian statistics–driven adaptive platform trial has continuously enrolled patients on a multi-arm, master protocol to evaluate new neoadjuvant therapies in combination with standard-of-care chemotherapy for high-risk neoadjuvant breast cancer. Patients are evaluated on the basis of molecular markers and imaging for molecular subtyping and assigned to agent arms based on their subtype. Pathologic complete response (pCR) is employed to determine success or failure of a specific agent. Big data plays a key role in nearly all aspects of I-SPY2 as evident by the recent update demonstrating the association of event-free survival with individual-level pCR that could only be achieved through longitudinal monitoring and continued follow-up (8).

(iv) **Delivery.** Big data from NGS is rewriting the way oncologists make decisions on therapy recommendations for patients informed by clinicogenomic real-world evidence repositories (e.g., AACR's Project GENIE, ASCO TAPUR; ref. 2). Using cohorts of molecularly profiled patients with colorectal cancer given different first-line combination regimens, an aggregate biomarker signature was developed and trained to predict relative efficacy to inform future patients with colorectal cancer faced with similar choices of initial therapy (9). Clinical validation using an independent cohort demonstrated the signature not only predicted patients with decreased benefit from approved combinations for initial treatment of colorectal cancer, but also predicted those who would benefit from alternative strategies for initial treatment for esophageal, gastric, and pancreatic cancers. As AI-derived biosignatures like FOLFOXai continue to be developed and aid clinical decisions, it is essential to continue to augment genomic-only AI training with emerging molecular profiling datasets (e.g., microenvironment, proteome, epigenome).

Major Issues and Barriers

Although big data has enabled progress in several areas of the cancer research continuum, for the major goals of precision oncology to be achieved, there are both legacy and prospective barriers that must be understood and removed (see **Table 1**). For example, legacy (retrospective) data are often collected by individual investigators without the benefit of future secondary analysis revealing batch effects caused by either sample quality or legacy data standards (7). These data, sometimes referred to as "fuzzy data" are of varying quality and questionable value. One approach to increase the value of certain of these data sets is the creation of "data lakes," which could be structured in a manner agnostic to questions and analytics as described in Part 14 of Jaffee and colleagues (10).

Despite government mandates to make all data from government funded large-scale genomics studies public, data sharing continues to be a legacy problem as new multiomics platforms are introduced and mature (6). Solving this problem requires progress on developing and deploying common data models, data standards, and interoperability of systems, but of equal importance is changing career reward structures that currently select against data sharing before publication.

Beyond these legacy problems, as highlighted in **Table 1** and further explored below, there are more intractable problems that require creative solutions:

(i) **Poor data quality.** Data quality begins with the quality of biospecimens. Rigorous biospecimen standards exist and as molecular profiling becomes more powerful, the mandate to ensure reproducibility of data has become critical. Best practices for the acquisition of high-quality biospecimens are available from several sources including NCI's Best Practices (https://biospecimens.cancer.gov/bestpractices/). History has shown AI methods trained using "samples of convenience" have generated biosignatures that have little benefit for patients and sometimes even resulted in harm (2).

(ii) **Unstructured databases.** Prior to the big data "omics-centric" revolution, oncology data were collected with minimal attention paid to the structure of databases, creating a problem for secondary analysis using new machine learning and deep learning–based analytics (3). Oncology databases, especially housing "multiomics centric" big data, should require methods to harmonize legacy data with new emerging technologies.

(iii) **Inadequate analytics and lack of delivery.** Analysis of the ever-increasing data explosion in cancer research and care represents perhaps the field's biggest challenge in providing the right algorithm-based decision tools to oncologists. Analytics that mirror the dynamics of cancer require copious amounts of data for training and testing these algorithms; and there is a critical lack of trained technical computational professionals to solve these problems. Moreover, there is a dearth of robust processes to deliver these decision tools to oncologists. Creating information flows that sync with electronic medical and electronic health records are challenges that will likely only be addressed by the major hospital records companies.

Harnessing the power of big data will require the creation of prospective databases that address all of the issues above and more. In addition, there is a need for longitudinal databases to inform the dynamics of cancer in patients across the research-care continuum including: diagnostic staging, therapy selection, cancer subtyping, companion diagnostics for treatment selection and creating new innovative clinical trials.

The Future

Although the definition of "big data" is still evolving, as existing technologies advance and new technologies emerge, the questions that both oncologists and patients hope "big data" will answer remain relatively constant: (*oncologist*) what therapies to recommend and (*patient*) what therapy to choose. Shared decision making in the era of "big data" should be informed by the contributed knowledge and experience of every patient with cancer. Most patients would like to see their data shared in a continuous learning system.

The increase in our understanding of biological pathways, by TCGA and other large-scale genomic studies, must now be mirrored and enabled for clinical pathways (e.g., guidelines, reimbursement). How can we build an ecosystem where the utilization driven by clinical pathways dictated by payors and longitudinal outcomes are candidly shared so we can determine what test to give to whom and what treatment should be recommended per person on their cancer journey? Our current system of waiting for enough patients to change guidelines or using cancer registries as surrogate delayed indicators is not sustainable.

We find ourselves at the beginning of a "big data" revolution and are ill prepared to achieve progress without creating patient-centric systems to collect, manage, analyze and move these data into patients. Are we making progress? Yes, but it is too slow and the numbers of patients benefitting, especially in underserved populations, is disappointing. Change will require a combination of government enforcement of strict privacy laws, interoperability across research and care, common data standards and models, mandated data sharing, and strong engagement of the private sector to ensure delivery of the AI decision tools needed to deliver these new technologies at scale to maximize clinical benefit—to name a few actions.

Finally, as we contemplate how to proceed in an environment where big data is literally accumulating at petabytes per day, we will soon (or have) reach a point where individual investigators armed even with the most powerful AI tools will be unable to integrate and interpret the vast amounts of data required to produce the information needed to inform patient decisions. We are reaching an inflection point in the oncology big data revolution where we must learn from other disciplines and understand that data are entropy until it has context. We can only move to establish such context through the acceptance and implementation of appropriate theoretical constructs to drive making sense of big data. This will be a difficult but necessary frontier that must be crossed if we are to unleash the power of big data to inform and improve all aspects of the cancer patient’s journey diagnosis through survivorship.

Authors’ Disclosures

A.D. Barker reports personal fees from Caris Life Sciences and personal fees from Ellison Institute, LLC, outside the submitted work. J.S.H. Lee reports other support from AtlasXomics, Inc., Accelerating Therapeutics for Opportunities in Medicine Consortium, Health and Environmental Sciences Institute, and Blood-PAC; personal fees from Ellison Institute, LLC, and Henry M. Jackson Foundation outside the submitted work. No other disclosures were reported.

Received March 16, 2022; revised April 4, 2022; accepted April 6, 2022; published first April 13, 2022.

References

1. Marquart J, Chen EY, Prasad V. Estimation of the percentage of US patients with cancer who benefit from genome-driven oncology. JAMA Oncol 2018;4:1093–8.
2. Malone ER, Oliva M, Sabatini PJB, Stockley TL, Siu LL. Molecular profiling for precision cancer therapies. Genome Med 2020;12:8.
3. Howard FM, Dolezal J, Kochanny S, Schulte J, Chen H, Heij L, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nat Commun 2021;12:4423.
4. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature 2020;578:82–93.
5. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big data: astronomical or genomical?. PLoS Biol 2015;13:e1002195.
6. Louis DN, Perry A, Reifenberger G, Von Deimling A, Figarella-Branger D, Cavenee WK, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathol 2016;131:803–20.
7. Frost H, Graham DM, Carter L, O’Regan P, Landers D, Freiatas A, et al. Patient attrition in molecular tumour boards: a review. medRxiv. doi: https://doi.org/10.1101/2021.10.07.21264241.
8. I-SPY2 Trial Consortium, Yee D, DeMichele AM, Yau C, Isaacs C, Fraser Symmans W, et al. Association of event-free and distant recurrence-free survival with individual-level pathologic complete response in neoadjuvant treatment of stages 2 and 3 breast cancer: three-year follow-up analysis for the I-SPY2 Adaptively Randomized Clinical Trial. JAMA Oncol 2020;6:1355–62.
9. Abraham JP, Magee D, Cremolini C, Antoniotti C, Halbert DD, Xiu J, et al. Clinical validation of a machine-learning-derived signature predictive of outcomes from first-line oxaliplatin-based chemotherapy in advanced colorectal cancer. Clin Cancer Res 2021;27:1174–83.
10. Jaffee EM, Dang CV, Agus DB, Alexander BM, Anderson KC, Ashworth A, et al. Future cancer research priorities in the USA: a Lancet Oncology Commission. Lancet Oncol 2017;18:e653–706.