# Cancer Needs a Robust "Metadata Supply Chain" to Realize the Promise of Artificial Intelligence

Caroline Chung and David A. Jaffray

## ABSTRACT

Profound advances in computational methods, including artificial intelligence (AI), present the opportunity to use the exponentially growing volume and complexity of available cancer measurements toward data-driven personalized care. While exciting, this opportunity has highlighted the disconnect between the promise of compute and the supply of high-quality data. The current paradigm of ad-hoc aggregation and curation of data needs to be replaced with a "metadata supply chain" that provides robust data in context with known provenance, that is, lineage and comprehensive data governance that will allow the promise of AI technology to be realized to its full potential in clinical practice.

## Introduction

Hippocrates, considered by some as the "Father of Western Medicine," set us out on a lofty goal of achieving predictive, personalized medicine with a patient-centered approach when he stated "Declare the past, diagnose the present, foretell the future" while reminding us that "It's far more important to know what person the disease has than what disease the person has." Modern cancer medicine, particularly with the emergence of genomics, has set the field on a journey to seek a myriad of new measurements to describe both the disease and the individual (i.e., phenomics). This includes patient characteristics, stage, tumor, and germline genomics, imaging data, and a multitude of additional emerging measures such as the microbiome, circulating markers, and data streaming from personal health monitors. It has been argued that the current wealth of information available for decision making exceeds human cognitive capacity (1) such that new computational approaches such as artificial intelligence (AI) and machine learning are needed for meaningful integration of the growing rate and diversity of data while ensuring that we do not lose sight of the ultimate goal, which is to treat the patient and not the lab tests or investigations. While the promise of AI has driven a revived hype cycle in the tsunami of data of modern medicine, recognizing both the promise and current limitations of this technology and identifying the necessary changes required in healthcare's approach to managing data are critical steps towards accelerated realization of the benefits of AI.

## The Promise and Challenges of AI in Personalized Medicine

There are many examples where the use of AI is being pursued in cancer medicine today ranging from the introduction of automation to improve efficiency and consistency to the exploration of AI to provide data-driven decision support, but this growing experience also highlights the need to rethink how we manage the key ingredient for their utility–the data. For instance, the growing dependence of cancer management on imaging including cancer staging, image-guided therapies (e.g., surgery, radiation, interventional radiology) and response assessment has created a demand for AI-based approaches to improve the consistency and efficiency of image-acquisition and interpretation. Hickman and colleagues touch on the many benefits and challenges of AI adoption in breast cancer imaging but highlight the pressure that the development of these algorithms are putting on access and management to data (2). Similarly, the growing development of digital pathology platforms introduce the promise of integrated AI-based tools to assist with workflow but rely on expert human classification and curation for their development. Building upon large volumes of annotated data found in radiation oncology, one of the earliest applications of machine learning in cancer has been to improve the efficiency and consistent quality of radiation treatment planning through automation of currently manual processes of tumor and normal tissue segmentation and plan generation (3), thereby potentially enabling access to expert level radiation treatments in underserved populations globally (4). AI approaches are also transforming cancer research with algorithms that can automatically flag patients for eligibility to open clinical trials, the development of natural language processing (NLP) tools for mining published corpora to illuminate potential linkages between clinical observations and underlying biology, or the development of massive discovery machinery that bridges features from across the basic and clinical data domains using unified frameworks to predict time-to-event outcomes (5). In addition to these domains of impact in oncology, there are numerous applications of AI in more fundamental work such as the modeling of protein folding, improvements in sequencing analyses, and the generation of imaging data (6).

While the growing applications highlight the breadth of promise of AI, they reinforce that our current paradigm to managing data and models is not mature enough to support the confident deployment of AI technologies at scale. A recent example was published by Wong and colleagues that reported that while the initial performance leading to regulatory approval of an integrated sepsis prediction tool in the EMR reported area under the receiver operating characteristic curve (AUC) results of 0.76–0.83, an evaluation of over 27,000 patients across over 38,000 hospitalizations found the AUC of this proprietary sepsis model was only 0.63, raising caution around broad adoption of models

The University of Texas MD Anderson Cancer Center, Houston, Texas.

**Corresponding Author:** Caroline Chung, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77027. E-mail: cchung3@mdanderson.org

Cancer Res 2021;81:5810–2

**doi:** 10.1158/0008-5472.CAN-21-1929

without rigorous evaluation of the performance of these models in specific settings (7). It is not only the specific data feeding into the model but the contextual information around that data (e.g., patient population included in training, quality of the training data, quality of the annotations), which can impact the performance of the model in new settings and appropriate use of the model in similar yet unique clinical scenarios or populations. There is a growing awareness of this issue at the regulatory level with the FDA seeking input from the community on the management of AI/ML-based “Software-as-a-Medical Device” (SaMD) and establishing guidelines for use with a focus on understanding the source of the data to assure cohort-appropriate application. In addition, there is growing awareness of the need to establish frameworks for algorithm development, validation, clinical commissioning, and on-going monitoring and quality assurance in the clinical setting (8). These are needed to provide evidence in support of the safe and reliable use of machine-based decision-making, assure “human-in-the-loop” validation, manage the issue of interpretability, and attend to ethical aspects to protect against societal biases that are likely embedded in the data or in the processes they are replicating.

Far beyond the obvious supply and demand relationship of data and AI, it is becoming apparent that robust data governance will both support and drive AI model management and performance. The ability to draw together diverse data is critical to build an understanding of the complexity of cancer, but also raises important questions regarding access, curation, and governance of these data to assure both appropriate use and extraction of robust insights. The current frameworks for finding and accessing data are straining under the growing demand and are undermining our ability to collaborate effectively and efficiently at the institutional and individual investigator level. This will become more critical as the sources of data become more personal (e.g., personal health devices and voice to text and sentiment technologies capturing more accurate clinical notes and patient-reported outcomes; ref. 9) and expand to include patient-directed data contributions (10).

## How Do We Realize the Full Potential of AI?

Organizations that want to benefit from AI technologies need to pivot their focus from the “promise of compute” to investing in the “supply of data.” The reliable flow of high-quality data required to effectively develop and benefit from AI technologies needs a “metadata supply chain.” This chain links the explicit capture and transport of metadata (i.e., data that is descriptive of the data and its origin that provides the what, when, where, who, how, which, and why of the data that embodies the context of the data) from the point of collection to the point of insight extraction and engages all data stakeholders in the generation, maintenance, and appropriately use high-quality data throughout the data lifecycle. As highlighted above in the various opportunities and promising developments of AI-based tools along with their limitations, data needs its associated metadata to maintain the meaning, assure the quality, trace the provenance (i.e., lineage) of the data, and also engage appropriate individuals or groups in the governance of the data. We have captured these characteristics of the “metadata supply chain” in a set of principles grounded in the treatment of data points as observations within a scientific “observation paradigm” that is better aligned with the data science empowered future of medicine than the traditional “medical record paradigm.”

1. *Observations must be in context.* It is critical that the context surrounding every observation (i.e., measurement or data point) is captured in the form of the metadata associated with the observation. Metadata that provides context can include the device or person who has generated the observation, time of day, and circumstances of the observation. For example, a simple numeric measurement like a pain score can be impacted by how the measurement was taken (e.g., uncertainties associated with that measurement) and how that measurement will be used to generate insights can be impacted on the circumstances (e.g., patient reported or provider reported, how the measurement was requested–pain currently or within the past day/week or worst pain in the past day/week, what events preceded the pain measurement, emotional state when asked about pain). This is the start of the supply chain and investment here is critical to assure the data has value and utility and is used appropriately. The lack of this metadata in our current data management practices introduces risk of misinterpretation and this risk is realized and magnified when observations are blindly fed into algorithms without consideration of the context of those observations.
2. *The quality of an observation is captured in the context.* Contextualizing metadata that allows for an integrated, granular data quality monitoring approach will enable comparison of the robustness of algorithm performance to data quality and thereby evoke some level of confidence in the extracted insights. Moving forward, capturing comprehensive contextualizing metadata at the point of data generation will deliver a level of data quality measurement that no amount of *post hoc* data curation will ever achieve. For instance, tumor response assessment, which increasingly relies on image-based measurements can be impacted by the protocol used during the imaging session, the type of contrast agent used, the image processing completed prior to image interpretation and, as well as, the criteria used to determine response. The metadata around each of these aspects collectively characterizes the quality of the underlying observations and the certainty in the determination of tumor response. In the case of building an AI tool to achieve such a task, the lack of metadata to help measure and monitor data quality would leave many unanswered questions when the algorithm fails to perform.
3. *Provenance links insights to observations.* Provenance is crucial to making determinations about whether an insight can be trusted, how to track the integration of diverse observations, how to reveal potential biases in the analyses or observations, and how to verify the rights of use or, alternatively, give credit and attribution to the human or machine contributor. By capturing the contextualizing metadata around observations, the information necessary to track the provenance is captured at the time of data generation. With growing interest in data collaborations and team science, this lack of provenance has raised alarms around how institutions will appropriately manage the complex issues of encumbered data and academic credit and attribution.
4. *Data governance must be granular and consistent with the needs of the demand.* Precision and granularity in data governance that captures the data and contextualizing metadata will empower organizations to utilize their data with confidence and provide a level of transparency that will assure the trust of all their patients and stakeholders, including researchers, clinical teams, operations, finance, external partners, and collaborators. On the

basis of first principle, the context of the observations would capture the subject, observer, their roles, and the associated consented rights for use.

Establishing a set of principles is a proven method to support transformative change that requires broad engagement across diverse teams. Health care, in particular, will benefit from these principles as a robust metadata supply chain within healthcare requires understanding and cooperation across the entire organization including those who contribute (e.g., patients, frontline and clinical staff), manage (e.g., operational and research teams), and consume the data (e.g., researchers and administrators), as well as those responsible for overseeing the supply chain and building the underlying technological infrastructure. It can be argued that the only way to make real progress in medicine is by learning from the care we deliver each and every day and the lack of robust data with its contextualizing metadata is the single largest impediment to making this a reality. The presented principles provide a framework to create a sustainable solution that reinforces the scientific paradigm of rigorous observation, assures patient engagement in the governance of their data, and is necessary to realize the promise of technologies like AI. Looking to the future, the next generation of AI tools will strive to improve performance and will be built on a fusion of new methods and new data. This cycle will continue and will put immense pressure on our ability to understand where exactly the data that is “programming” our new AI-enabled world is coming from. It will also highlight the remarkably inefficient and irreproducible manual curation steps that are routinely employed today to improve data quality prior to analysis. Finally, a unique challenge in healthcare is the current design and operational paradigm of electronic medical record (EMR) technology, which remain highly fragmented and have not been engineered to support the “metadata supply chain” paradigm. There is a major need to develop and adopt new data governance and provenance maintaining technologies for healthcare and health more broadly as patients become an active participant in the metadata supply chain for the future of cancer medicine.

## Summary

In our increasingly digital world and explosion of AI, data has been recognized as “the most valuable resource” (The Economist, 2017) and as a potential “new currency” with great value; however it is being increasingly realized that some data with differing qualities and characteristics will have greater value than others. This quality and value measure will largely be reflected in the metadata. While healthcare is the domain where AI technologies could be of greatest value to humanity, we argue that a major change in approach to data is required within healthcare to set a new foundation that aligns the culture, processes and technology and enables a robust and trustworthy “metadata supply chain.” Organizations that decide to make the investment to steward their data comprehensively will reap the benefits promised by AI technologies and will be the best partners for their patients and for their collaborations with other organizations whether academic, industry, or governmental that have made similar decisions to invest in managing their data.

### Authors’ Disclosures

C. Chung reports grants from Siemens Healthineers and grants from RaySearch Laboratories outside the submitted work. No disclosures were reported by the other author.

### Acknowledgments

This work was supported by Data-Driven Determinants for COVID-19 Oncology Discovery Effort (D3CODE) Team, University of Texas, MD Anderson Cancer Center (Houston, TX).

Received June 29, 2021; revised August 4, 2021; accepted October 7, 2021; published first December 1, 2021.

### References

1. Abernethy AP, Etheredge LM, Ganz PA, Wallace P, German RR, Neti C, et al. Rapid-learning system for cancer care. J Clin Oncol 2010;28:4268–74.
2. Hickman SE, Baxter GC, Gilbert FJ. Adoption of artificial intelligence in breast imaging: evaluation, ethical constraints and limitations. Br J Cancer 2021;125:15–22.
3. McIntosh C, Conroy L, Tjong MC, Craig T, Bayley A, Catton C, et al. Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. Nat Med 2021;27:999–1005.
4. Lewis PJ, Amankwaa-Frempong E, Makwani H, Nsingo M, Addison ECDK, Acquah GF, et al. Radiotherapy planning and peer review in sub-saharan Africa: A needs assessment and feasibility study of cloud-based technology to enable remote peer review and training. JCO Global Oncol 2021;7:10–16.
5. Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Vega JEV, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. PNAS 2018;115:E2970–E9.
6. Chung C, Kalpathy-Cramer J, Knopp MV, Jaffray DA. In the era of deep learning, why reconstruct an image at all? J Am Coll Radiol 2021;18: 170–3.
7. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med 2021;181:1065–70.
8. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet 2019;393:1577–9.
9. Car LT, Dhinagaran DA, Kyaw BM, Kowatsch T, Joty S, Theng Y-L, et al. Conversational agents in health care: scoping review and conceptual analysis. J Med Intern Res 2020;22:e17158.
10. Dhruva SS, Ross JS, Akar JG, Caldwell B, Childers K, Chow W, et al. Aggregating multiple real-world data sources using a patient-centered health-data-sharing platform. NPJ Digit Med 2020;3:60.