



REVIEW

Artificial intelligence to unlock real-world evidence in clinical oncology: A primer on recent advances

Alex K. Bryant^{1,2}  | Rafael Zamora-Resendiz³ | Xin Dai⁴ | Destinee Morrow³ | Yuewei Lin⁴ | Cassidy M. Jungles⁵ | James M. Rae^{5,6} | Akshay Tate¹ | Ashley N. Pearson¹ | Ralph Jiang^{1,7} | Lars Fritsche⁷ | Theodore S. Lawrence¹ | Weiping Zou^{7,8,9,10} | Matthew Schipper^{1,5} | Nithya Ramnath^{11,12} | Shinjae Yoo⁴ | Silvia Crivelli³ | Michael D. Green^{1,2,10,13,14} 

¹Department of Radiation Oncology, University of Michigan School of Medicine, Ann Arbor, Michigan, USA

²Department of Radiation Oncology, Veterans Affairs Ann Arbor Healthcare System, Ann Arbor, Michigan, USA

³Applied Mathematics and Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA

⁴Computational Science Initiative, Brookhaven National Laboratory, Upton, New York, USA

⁵Department of Pharmacology, University of Michigan School of Medicine, Ann Arbor, Michigan, USA

⁶Department of Internal Medicine, University of Michigan School of Medicine, Ann Arbor, Michigan, USA

⁷Department of Statistics, University of Michigan, Ann Arbor, Michigan, USA

⁸Center of Excellence for Cancer Immunology and Immunotherapy, University of Michigan Rogel Cancer Center, Ann Arbor, Michigan, USA

⁹Department of Pathology, University of Michigan, Ann Arbor, Michigan, USA

¹⁰Graduate Program in Immunology, University of Michigan, Ann Arbor, Michigan, USA

¹¹Division of Hematology Oncology, Department of Medicine, University of Michigan School of Medicine, Ann Arbor, Michigan, USA

¹²Division of Hematology Oncology, Department of Medicine, Veterans Affairs Ann Arbor Healthcare System, Ann Arbor, Michigan, USA

¹³Graduate Program in Cancer Biology, University of Michigan, Ann Arbor, Michigan, USA

¹⁴Department of Microbiology and Immunology, University of Michigan School of Medicine, Ann Arbor, Michigan, USA

Correspondence

Michael D. Green, Department of Radiation Oncology, University of Michigan School of Medicine, Medical Science Building I Room 4424F, 1301 Catherine Street, Ann Arbor, MI 48109, USA.

Email: migr@med.umich.edu

Funding information

LUNgevity Foundation, Grant/Award Number: 2021-07; National Institute of Health, Grant/Award Number: R01CA276217; Melanoma Research Alliance, Grant/Award Number: 689853; Veterans Affairs, Grant/Award Number: MVP064, BX005267, 150CU000182 and VA 150CU000182

Abstract

Purpose: Real world evidence is crucial to understanding the diffusion of new oncologic therapies, monitoring cancer outcomes, and detecting unexpected toxicities. In practice, real world evidence is challenging to collect rapidly and comprehensively, often requiring expensive and time-consuming manual case-finding and annotation of clinical text. In this Review, we summarise recent developments in the use of artificial intelligence to collect and analyze real world evidence in oncology.

Methods: We performed a narrative review of the major current trends and recent literature in artificial intelligence applications in oncology.

Results: Artificial intelligence (AI) approaches are increasingly used to efficiently phenotype patients and tumors at large scale. These tools also may provide novel biological insights and improve risk prediction through multimodal integration

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *Cancer Medicine* published by John Wiley & Sons Ltd.

of radiographic, pathological, and genomic datasets. Custom language processing pipelines and large language models hold great promise for clinical prediction and phenotyping.

Conclusions: Despite rapid advances, continued progress in computation, generalizability, interpretability, and reliability as well as prospective validation are needed to integrate AI approaches into routine clinical care and real-time monitoring of novel therapies.

KEYWORDS

Artificial Intelligence, Cancer Outcomes Research, Large language models, Observational Data, prognostic factor

1 | INTRODUCTION

The daily practice of oncology is filled with uncertainty: the unclear real-world effectiveness of novel therapies; the lack of precise prognostic information relevant to individual patients; a plethora of management decisions for which there is no prospective evidence; and unknown real-world treatment toxicity rates. These gaps have increased the value of real world evidence that rigorously defines patient prognosis, treatment response, and toxicity outside of clinical trials, and of risk prediction models that can reduce diagnostic and prognostic uncertainty.^{1,2} However, much of the narrative experience of oncology exists within unstructured medical records such as oncologist notes, discharge summaries, and radiographic or pathologic reports. Critical prognostic information is also embedded within complex multimodal datasets such as radiographic images, tumor and germline sequencing, germline sequencing, and histology slides. The significant effort required for annotation, curation, and interpretation of these unstructured data sources prevents healthcare systems from rapidly learning from emerging experiences with novel therapies.³ There is a critical need for flexible approaches that capture and distill the complexities of real-world cancer care. Artificial intelligence (AI), which refers to the use of complex computer algorithms to provide solutions and inform decision making for unsolved problems, has shown significant potential in this area. Neural network-based AI approaches such as natural language processing (NLP) and deep learning (DL) are revolutionizing the collection and interpretation of real-world oncological data and are beginning to touch many aspects of cancer care.^{4,5} In this review, we highlight recent advances in NLP, DL, and other AI approaches being applied to routinely collected oncological data to advance our understanding of real-world cancer care and discover novel therapies (Figure 1).

2 | AI APPLICATIONS IN CANCER SCREENING, DIAGNOSIS, AND STAGING

Prospective trials have highlighted the benefits of screening for breast, colorectal, lung, cervical, and prostate cancer,^{6–8} but there is continuing controversy over which real-world patient subgroups benefit. There is also controversy whether the small absolute survival benefits seen in many screening trials is outweighed by the increased detection of indolent disease and unnecessary workup leading to emotional and financial stresses.⁹ As a result, uptake of some screening procedures remain low; for example, <15% of eligible patients undergo low-dose computed tomography screening for lung cancer¹⁰ and prostate-specific antigen screening rates have declined in recent years due in part to concerns of overdiagnosis of indolent cancers.^{11,12} Given these challenges, there is increasing interest in applying AI approaches to augment the interpretation of screening studies and maximize the potential benefits of screening while minimizing harms of overdiagnosis.

Screening approaches include direct visualization (cutaneous malignancies), radiographic evaluations (breast and lung malignancies), endoscopy (gastrointestinal malignancies), blood markers (prostate cancer), and tissue sampling (cervical cancer). Advances in image analysis have enabled DL approaches to contribute to the interpretation of most of these screening modalities. Convolutional neural networks (CNNs) use a grid-like topology using layers of filters or “convolutions” to model high dimensional, non-linear feature spaces such as those found in medical imaging datasets. CNNs have shown impressive accuracy in multiple studies for skin cancer detection,^{13,14} in some cases outperforming human raters^{15,16} with positive predictive values for melanoma in the 0.80–0.90 range and negative predictive values in the 0.90–0.95 range.¹⁷ There are significant challenges in applying these algorithms

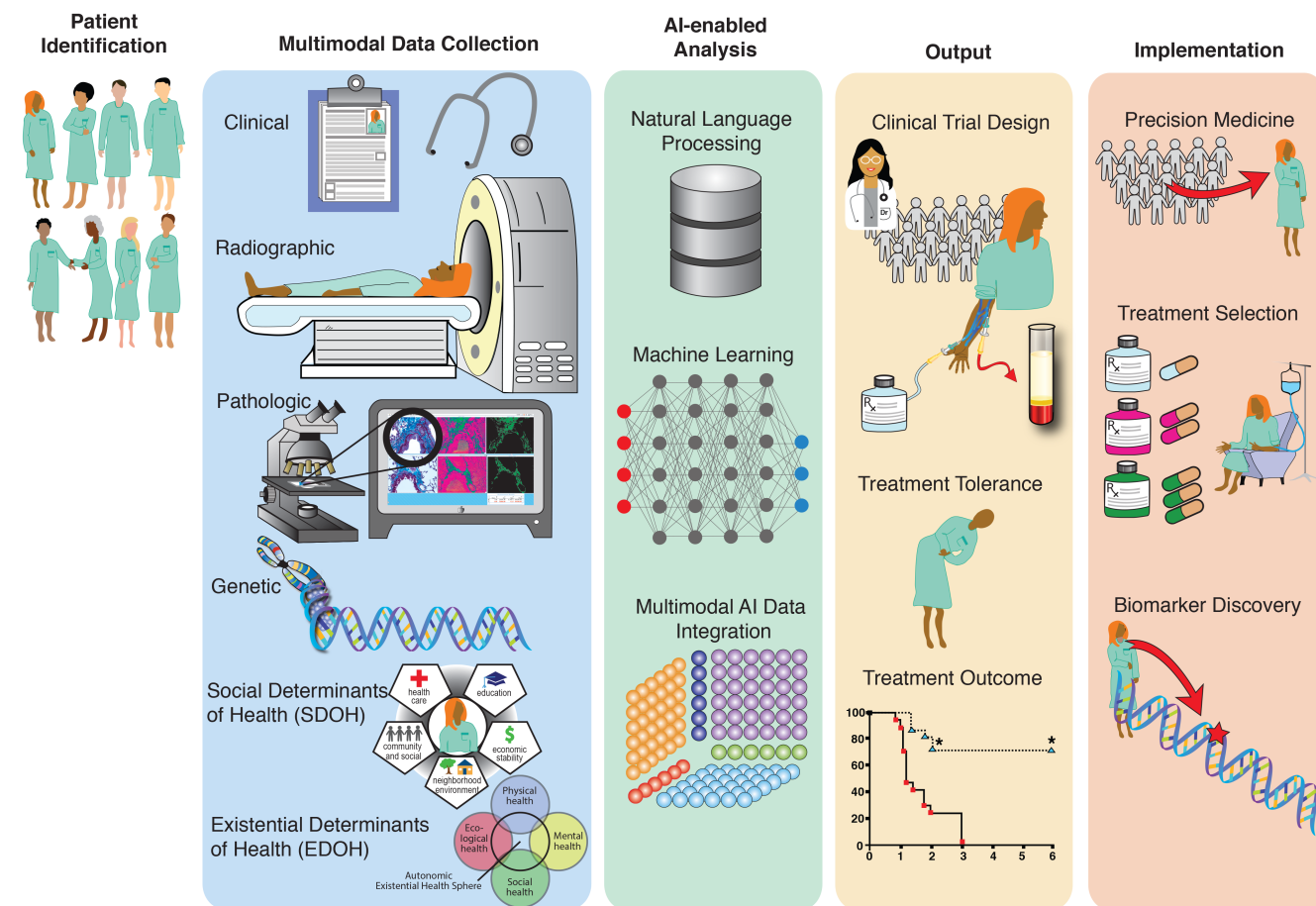


FIGURE 1 Ongoing and Potential Integrations of AI in Oncology. There is increasing efforts to combine clinicopathologic factors, social and existential determinants of health, and radiographic features using AI approaches to evaluate cancer therapy treatment outcomes and improve cancer care.

in routine care, where confounding visual features such as artifact from pen markings and crusted lesions can cause a substantial decrement in model accuracy.¹⁸ In colorectal cancer screening, CNNs have been shown in a randomized controlled trial to decrease the endoscopic miss rate of adenomas by half, primarily by increasing the detection of flat, small, and subtle neoplastic lesions.¹⁹ Moreover, CNNs also show promise in improving the cost-effectiveness of screening²⁰ and in mitigating quality disparities.²¹ CNNs can improve the efficiency and accuracy of cervical cytology, with one study showing higher specificity compared to conventional screening and lower rates of referral for colposcopy;²² such approaches may be especially relevant in resource-limited settings where cervical cancer incidence is highest.²³ While AI-assisted mammography interpretation holds promise for breast cancer detection,²⁴ studies have yet to conclusively establish the benefit over expert radiographic review and large prospective validation studies are needed.²⁵ In lung cancer, a three-dimensional CNN termed Sybil has shown impressive accuracy in detecting short-term risk of lung cancer from a single low-dose computed tomography scan of

the chest, with a 1-year area under the receiver operating characteristic curve of 0.86–0.94 in validation sets.²⁶ Sybil and similar models trained on computed tomography data may refine risk prediction in lung cancer screening and reduce false-positive rates seen in screening trials,⁶ particularly when combined with longitudinal clinical data and when incorporating repeated screens. A critical weakness of many imaging-based AI algorithms is the narrowness of training datasets, which are often drawn from European-centric patient populations, single medical centers, and a limited set of radiological scanner vendors, each of which can produce biased, poorly calibrated models which fail to generalize. As such, multi-institutional validation studies are needed to test models in the full range of scanning technologies and patient populations reflective of routine clinical care.

Many groups have studied the application of AI techniques to aid interpretation of digitized pathology slides of tumor biopsies and cytology.²⁷ The recent availability of whole slide imaging systems has produced a flood of histopathological imaging data which is increasingly used to train DL models.²⁸ Major tasks include tumor

segmentation and classification,^{29,30} automated characterization of the tumor microenvironment (such as tumor-infiltrating lymphocyte density, which may predict responsiveness to immunotherapies),³¹ detection of metastases in tumor draining lymph nodes,³² and prediction of tumor mutational status,^{33–35} among many others.²⁷ Serious challenges remain in ensuring model stability across disparate laboratories and preparation techniques, managing the massive quantity of imaging data produced by whole-slide scanners, and defining the clinical utility and cost-effectiveness of digitization efforts and AI model deployment. Clinical questions also remain about the treatment implications of highly sensitive AI algorithms. For example, deployment of a model that significantly increases detection of micro metastases may expose patients to intensified therapy and over-treatment without a clearly defined clinical benefit. Carefully controlled prospective studies are therefore needed to assess the patient-centered benefit of incorporating model outputs into cancer staging and management decisions.

Accurate interpretation of radiographic images is critical in assigning clinical stages to cancer patients, which largely dictates subsequent treatments. DL models are rapidly being applied to baseline staging images to provide more accurate staging and risk stratification. DL models have been applied in breast cancer to improve detection of subclinical axillary nodal metastases on ultrasound,³⁶ in head and neck cancer to predict the presence of extra nodal extension in pre-treatment computed tomography scans,³⁷ and in lung cancer to predict pathologic node positivity in the mediastinum,³⁸ among others.³⁹ While these tools can reduce the burden of labor-intensive cognitive tasks and improve diagnostic accuracy, it is unknown whether allowing clinical management to be influenced by the probabilistic outputs of DL models will improve patient-centered outcomes like quality of life or survival, and prospective studies are needed to more clearly define the benefits of these novel algorithms.

3 | AUTOMATED EXTRACTION OF PATIENT AND CANCER CHARACTERISTICS

Once a cancer is appropriately screened, diagnosed, and staged, this critical information is typically transcribed in free-text notes from oncologists or other front-line providers. Additional crucial data such as performance status, symptoms from medical comorbidities, specific pathological and radiological findings, medication compliance, and family history are also primarily contained in free-text form. While structured clinical data derived from electronic medical records—such as laboratory

results, diagnosis codes, and procedures—provide important prognostic information, much of the crucial patient and cancer features that dictate subsequent treatment are locked in unstructured text and invisible to traditional methods of analysis that require structured clinical data. As such, there have been major efforts to develop NLP approaches—both deterministic rule-based systems and probabilistic systems based on large language models (LLMs)—to facilitate analysis of free text and enable real-time case identification, patient profiling, and disease characterization.

Deterministic rule-based NLP systems have demonstrated excellent performance in the extraction of targeted features from unstructured text in scenarios where a limited lexicon is used to describe the features of interest. The Leo framework has been applied successfully to extract pathologic stage from surgical pathology reports⁴⁰ and the CLAMP (Clinical Language Annotation, Modeling, and Processing) framework for extraction of tumor size, stage, and biomarker results.⁴¹ Custom NLP pipelines engineered for extraction of specific features can include initial development of a lexicon reflecting the features of interest, followed by term expansion using computational methods such as word embeddings derived from statistical language models. Free-text extracted around the term list is then processed using standardized ontologies and rule-based NLP tools to produce the final feature vector. This process is exemplified in a recent report describing the development of a high-performance algorithm to extract new diagnoses of metastatic prostate cancer in the Veterans Affairs system based on a complex logic of anatomical and diagnostic text patterns.⁴² Such pipelines require substantial development time and may be applicable only to the healthcare setting in which they were developed, with idiosyncratic note types, data structures, and vocabularies. There have therefore been efforts to develop generalizable NLP algorithms that use rule-based systems like cTAKES to map concepts in unstructured free text to large cancer ontologies.⁴³ While promising, the results of these generalized systems have not been extensively validated and may not produce reliable outputs for all features.

A fundamental weakness of rule-based NLP systems is a lack of flexibility and context-dependence, weaknesses that may be overcome in the era of transformer-based LLM like GPT,⁴⁴ LLAMA,⁴⁵ and Claude⁴⁶ that model semantic context. While the training of massive language models has been limited by the dearth of large, diverse corpora of deidentified clinical text for training, early efforts have shown remarkable results in feature extraction even in clinically ambiguous contexts.^{47,48} LLMs may quickly subsume previous rule-based NLP systems as a “one size fits all” solution for feature extraction, document classification, and other traditional clinical NLP tasks.⁴⁸

Importantly, the optimal method of extracting structured variables from large language models is not yet clear, and the emerging field of prompt engineering has highlighted the sensitivity of generative LLM outputs to the precise wording and sequencing of prompts.^{47,49} While the application of these models is still in its infancy, the highly visible success of OpenAI's ChatGPT is producing a massive acceleration of effort in this area and LLMs will find immediate application in feature extraction tasks across medicine.

4 | DL FOR OUTCOME AND TOXICITY PREDICTION

Cancer researchers have made great strides in unraveling the major factors affecting long-term prognosis. These include a refined understanding of cancer staging and patterns of spread,⁵⁰ competing mortality prediction,⁵¹ molecular tumor subtyping,⁵² and heterogenous treatment effects,⁵³ in addition to the many novel therapies and treatment approaches entering routine practice each year. While our ability to use these factors to better predict outcomes of subgroups has improved dramatically, the clinical course for individual patients has remained stubbornly unpredictable. There is a need for new approaches to dynamically predict treatment efficacy, toxicity, and non-cancer mortality events throughout the patient's clinical course, enabling early interventions for supportive care and better personalizing treatment selection. The confluence of increasing centralization of multimodal clinical data and integrative AI approaches can bring the field closer to realizing the dream of personalized medicine.

The advancing flood of multimodal data—radiographic, genomic, pathomic, and clinical—has produced a flowering of AI approaches to integrate these fundamentally disparate data sources and identify novel connections across modalities. In rectal cancer, radiomic features of the pre-treatment pelvic MRI were combined with pathomic features from tumor biopsies to improve prediction of pathological complete response after neoadjuvant therapy, outperforming radiomic or pathomic features alone.⁵⁴ In breast cancer, whole-slide pathomic images and clinical variables were integrated to predict response to neoadjuvant chemotherapy in triple negative breast cancer through federated learning.⁵⁵ This is a novel example of edge computing which does not require central pooling of data for model training and therefore protects data privacy while improving feasibility.⁵⁵ In lung cancer, radiomic analysis of computerized tomography (CT) images could predict lung cancer EGFR genotype non-invasively with area under the curve (AUC) ranging from 0.75 to 0.81,⁵⁶ and integration of radiomic, pathomic,

and genomic features improved prediction of immunotherapy responses beyond unimodal measures alone.⁵⁷ Similar results have been found in serous ovarian cancer⁵⁸ and prostate cancer⁵⁹ among many others.⁶⁰ The optimal method of multimodal data fusion continues to evolve and can range from early fusion, in which raw features from each modality are fed into a single neural network with minimal pre-processing, to late fusion, in which separate networks are trained on each modality and then aggregated for a combined prediction.⁶⁰ Interpretability of these incredibly complex networks remains a challenge, though attentional heat maps can assist human observers in identifying the most salient features and produce new avenues for mechanistic research.^{60,61}

While many model development efforts have focused on using baseline clinical data for prediction, incorporating longitudinal data could increase predictive accuracy and relevance of model predictions to on-the-fly clinical decision making. Bayesian machine learning (ML) methods have been developed and applied to patients with diffuse large B-cell lymphoma, incorporating baseline clinical measures and longitudinal biomarker measurements to dynamically update disease progression and survival risks.⁶² Flexible longitudinal ML approaches are increasingly applied in other domains including Bayesian models for prediction of renal survival after kidney transplant using longitudinal laboratory values⁶³ and recurrent neural networks to predict acute kidney injury in the Veterans Affairs system, incorporating high-dimensional clinical text features.⁶⁴ DL approaches have been applied to infer formal tumor response criteria from radiology text reports^{65,66} as well as approaches like term frequency inverse documentation weighting and support vector machines, which is a supervised learning model that uses classification and regression analysis to define optimal hyperplanes for data discrimination.⁶⁷ Groups have also developed medical concept extraction approaches to infer therapeutic benefit from other sources including operative reports, pathology reports, and the narrative medical record.⁶⁸

LLMs trained on clinical text have also been proposed as general-purpose prediction engines that can quickly produce best-in-class risk predictions for virtually any clinical outcome.⁶⁹ Well-trained LLMs appear to produce large improvements in prediction accuracy over traditional ML approaches using structured clinical features (such as diagnosis codes and laboratory values) alone.⁶⁹ Important questions remain about the optimal model size, the need for a single foundational model versus multiple smaller models fine-tuned for distinct prediction tasks, model architecture, pre-training method, and need for site-specific fine-tuning of model parameters at individual hospitals in a network. Operationalized LLMs will also likely require

periodic re-tuning and vocabulary expansions to accommodate shifts in medical terminology usage and new drug names or procedures. Finally, LLM prediction models will likely benefit from incorporation of orthogonal datasets not directly present in the electronic medical record, such as three-dimensional radiographic images and genomic sequencing data. The optimal architecture for incorporating LLM predictions with other multimodal datasets remains unknown.

Even successful cancer therapies can have lifelong toxicities, some of which can be functionally devastating. Unfortunately, many acute and long-term toxicities remain largely unpredictable despite our increasing reliance on genomically-tailored therapies, immunotherapy, and precision radiotherapy. This problem is likely more acute in real-world settings compared to clinical trials, as patients enrolled on clinical trials are well known to have more extensive support networks and fewer medical comorbidities, which can improve treatment tolerance.⁷⁰ AI approaches are increasingly used to fill the knowledge gap by describing real-world, longitudinal patient experiences, such as capturing pain, fatigue, and other patient-reported side effects from the medical record.^{71–73} A challenge with many NLP approaches is the need to generate a lexicon of terms for a given toxicity; however, the use of weak labeling and large standardized ontologies may help overcome this challenge.^{74,75} Cancer patients can discontinue therapy due to symptom burden in a specific domain as well as the overall perception of diminished quality of life, and traditional NLP models have shown success in defining the clinical rationale for treatment discontinuation.⁷⁶ NLP monitoring of Twitter can even provide insight into real-world drug tolerance.^{77,78} These studies highlight that AI approaches may be able to monitor a variety of data sources to enable early warning of clinicians when patients are developing toxicities as well as aid in post-approval surveillance for cancer treatments.

5 | AI INTEGRATION INTO ROUTINE PRACTICE

There are emerging performance frameworks developed with the Food and Drug Administration (FDA) for implementation of AI as a medical device,⁷⁹ and AI is increasingly integrated into routine workflows in multiple specialties such as aiding radiologist interpretation and assisting in adaptive radiotherapy planning.⁸⁰ AI systems can even generate preliminary treatment recommendations, though these remain rudimentary and have yet to inform routine clinical practice.⁸¹ While the burgeoning AI literature is replete with model-building exercises and

impressive performance on test datasets, prospective validation of prediction algorithms in randomized controlled trials has been scarce. One recent trial in radiation oncology showed that an ML model predicting unplanned ED visits or hospitalizations could help direct intensive clinical monitoring during radiation and reduce acute care visits.⁸² Importantly, this trial did not randomize patients to ML-directed care versus usual care, but rather used ML to identify patients at risk of acute care visits who were then randomized to usual care versus intensive monitoring. Another recent trial used a stepped-wedge design to evaluate the impact of an ML algorithm to identify patients at high risk of death within 6 months, whose providers were then prompted to consider initiating a serious illness conversation with the patient.⁸³ This strategy led to an increase in serious illness conversations and reduction in end-of-life systemic therapy administration.⁸³ ML tools have been developed to identify patients eligible for enrollment on clinical trials of new cancer therapies^{84,85} and to explore inefficiencies in restrictive clinical trial enrollment criteria relative to real-world populations.⁸⁶ Despite these efforts, prospective validation of ML models and rigorous testing of their impact on clinical care remain disappointingly rare.

Recent advances in DL have significantly impacted radiation oncology, particularly in the automatic generation of organ-at-risk (OAR) and target contours on computed tomography radiation planning scans. These algorithms, primarily based on CNNs, have shown remarkable accuracy and efficiency in delineating OARs and tumor targets and have rapidly entered routine clinical practice.⁸⁷ A noteworthy development is the integration of 3D CNNs, which better capture the spatial relationships in volumetric data, leading to improved contour accuracy compared to traditional approaches.⁸⁸ Generative adversarial networks (GANs) have been employed to augment training datasets and to perform data normalization, enhancing model robustness across different imaging modalities and protocols.⁸⁹ AI applications in radiation treatment planning have also advanced notably. These models expedite the treatment planning process by automating dose distribution predictions and beam angle and intensity selections, tasks that are traditionally manual and time-consuming.⁹⁰ However, challenges such as ensuring model generalizability across different patient demographics and smooth integration into existing clinical workflows persist.

Technical challenges in training and validating complex ML models can also hinder performance and deployment efforts. ML models generally require many thousands of training examples to perform adequately, and performance should then be tested on large, preferably external validation datasets derived from an entirely different population than the training data to prove generalizability.⁹¹

Well-powered training data and external validation datasets may not be practically available for some prediction problems. Further, while external validation has typically been held as the gold-standard in testing model performance, future models may be increasing trained and deployed in only one hospital or healthcare system,⁶⁹ suggesting that rigorous, repeated internal validation may be more important than external generalizability.⁹² Some argue that there is no such thing as a truly validated prediction model due to the constant performance drift over time and practice settings that requires periodic model updating.⁹³ Technical aspects of the model training process can also affect performance and generalizability such as choosing among multiple architectures, tuning hyper parameters, and guarding against overfitting with methods such as cross-validation. AI implementation efforts need well-designed governance structures to provide oversight on these issues while also being flexible enough to rapidly identify, test, and deploy technical advancements in this fast-moving field.

A core challenge to implementing models remains physician and patient trust in model outputs. Both physicians and patients have interests in understanding the underlying principles of clinically deployed AI algorithms.^{94,95} While approaches like Local Interpretable Model-Agnostic Explanations (LIME),⁹⁶ Shapley additive explanation (SHAP),⁹⁷ and many others⁹⁸ provide frameworks for explainable AI,⁹⁹ there remain deep limitations to these approaches.¹⁰⁰ Further undermining trust is the well-documented biases apparent in models that have not been trained on appropriately diverse data sources, including racial biases.¹⁰¹ Equity must be a conscious design principle and training on diverse patient populations should be a requirement before model deployment. Finally, clinical care requires robust models that can be applied in diverse healthcare environments and which are temporally stable and continuously monitored.¹⁰² Thus, advances in reliability, stability, portability, adaptability, and equity are needed as AI continues to integrate into oncology practice.

6 | CONCLUSION

Cancer impacts more than 20 million new individuals each year. Understanding and integrating the totality of cancer experiences is needed to make improvements in cancer care, but the unstructured nature of the medical record has made large-scale data integration highly resource intensive under classical methods. AI approaches are filling this gap and have already revolutionized the collection, analysis, and interpretation of routinely collected unstructured health data. This includes tools

aiding in patient selection for cancer screening, interpreting radiographic studies, accurately staging new cancer patients, profiling patients using unstructured medical text, predicting treatment response and toxicity, generating novel connections between disparate data modalities, and directing clinical trial enrollment, among others. While there is great excitement surrounding these tools, their integration into routine care has been hampered by a relative lack of high-quality prospective validation studies and randomized clinical trial to prove the benefit of AI-assisted care. There are also ongoing concerns regarding the reliability, generalizability, accuracy, equity, and temporal stability of deployed prediction models. Despite the challenges, these methods hold incredible promise for improving the care of patients with cancer.

AUTHOR CONTRIBUTIONS

Alex K. Bryant: Conceptualization (equal); writing – original draft (equal). **Rafael Zamora-Resendiz:** Writing – original draft (equal). **Xin Dai:** Writing – original draft (equal). **Destinee Morrow:** Writing – original draft (equal). **Yuewei Lin:** Writing – original draft (equal). **Kassidy M. Jungles:** Writing – original draft (equal). **James M. Rae:** Writing – original draft (equal). **Akshay Tate:** Writing – original draft (equal). **Ashley N. Pearson:** Writing – original draft (equal). **Ralph Jiang:** Writing – original draft (equal). **Lars Fritsche:** Writing – original draft (equal). **Theodore S. Lawrence:** Writing – original draft (equal). **Weiping Zou:** Writing – original draft (equal). **Matthew Schipper:** Writing – original draft (equal). **Nithya Ramnath:** Writing – original draft (equal). **Shinjae Yoo:** Funding acquisition (equal); writing – original draft (equal). **Silvia Crivelli:** Funding acquisition (equal); writing – original draft (equal). **Michael D. Green:** Conceptualization (equal); funding acquisition (equal); supervision (lead); writing – original draft (equal).

FUNDING INFORMATION

Lung Precision Oncology Program (VA 150CU000182; PI Ramnath), LUNGeVity (2021–07, PI Green), NCI (R01CA276217, PI Green), Veterans Affairs (I01 BX005267; PI Green), Melanoma Research Alliance (MRA 689853; PI Green), Veterans Affairs (MVP064; PI Green, Crivelli, Yoo).

CONFLICT OF INTEREST STATEMENT

Authors report no conflicts of interest.

DATA AVAILABILITY STATEMENT

No new data sets were developed as part of this manuscript.

ORCID

Alex K. Bryant  <https://orcid.org/0000-0003-0194-8381>

Michael D. Green  <https://orcid.org/0000-0003-4951-7118>

REFERENCES

- Arondekar B, Duh MS, Bhak RH, et al. Real-world evidence in support of oncology product registration: a systematic review of new drug application and biologics license application approvals from 2015–2020. *Clin Cancer Res*. 2022;28(1):27–35.
- Raphael MJ, Gyawali B, Booth CM. Real-world evidence and regulatory drug approval. *Nat Rev Clin Oncol*. 2020;17(5):271–272.
- Xia F, Yetisgen-Yildiz M. *Clinical Corpus Annotation: Challenges and Strategies*. 2012.
- Yim W-W, Yetisgen M, Harris WP, Kwan SW. Natural language processing in oncology: a review. *JAMA Oncol*. 2016;2(6):797–804.
- Savova GK, Danciu I, Alamudun F, et al. Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. *Cancer Res*. 2019;79(21):5463–5470.
- National Lung Screening Trial Research Team, Aberle DR, Adams AM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*. 2011;365(5):395–409.
- Faivre J, Dancourt V, Lejeune C, et al. Reduction in colorectal cancer mortality by fecal occult blood screening in a French controlled study. *Gastroenterology*. 2004;126(7):1674–1680.
- Schröder FH, Hugosson J, Roobol MJ, et al. Screening and prostate cancer mortality: results of the European randomised study of screening for prostate cancer (ERSPC) at 13 years of follow-up. *Lancet*. 2014;384(9959):2027–2035.
- Gulati R, Morgan TM, Amar T, Psutka SP, Tosoian JJ, Etzioni R. Overdiagnosis and lives saved by reflex testing men with intermediate prostate-specific antigen levels. *J Natl Cancer Inst*. 2020;112(4):384–390.
- Jemal A, Fedewa SA. Lung cancer screening with Low-dose computed tomography in the United States-2010 to 2015. *JAMA Oncol*. 2017;3(9):1278–1281.
- Becker DJ, Rude T, Walter D, et al. The Association of Veterans' PSA screening rates with changes in USPSTF recommendations. *J Natl Cancer Inst*. 2021;113(5):626–631.
- Zeliadt SB, Hoffman RM, Etzioni R, Gore JL, Kessler LG, Lin DW. Influence of publication of US and European prostate cancer screening trials on PSA testing practices. *J Natl Cancer Inst*. 2011;103(6):520–523.
- Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115–118.
- Tschanidl P, Rinner C, Apalla Z, et al. Human-computer collaboration for skin cancer recognition. *Nat Med*. 2020;26(8):1229–1234.
- Maron RC, Weichenthal M, Utikal JS, et al. Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. *Eur J Cancer*. 2019;119:57–65.
- Tschanidl P, Codella N, Akay BN, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *Lancet Oncol*. 2019;20(7):938–947.
- Jones OT, Matin RN, van der Schaar M, et al. Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: a systematic review. *Lancet Digit Health*. 2022;4(6):e466–e476.
- Combalia M, Codella N, Rotemberg V, et al. Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 international skin imaging collaboration grand challenge. *Lancet Digit Health*. 2022;4(5):e330–e339.
- Wallace MB, Sharma P, Bhandari P, et al. Impact of artificial intelligence on miss rate of colorectal neoplasia. *Gastroenterology*. 2022;163(1):295–304.e5.
- Areia M, Mori Y, Correale L, et al. Cost-effectiveness of artificial intelligence for screening colonoscopy: a modelling study. *Lancet Digit Health*. 2022;4(6):e436–e444.
- Lu Z, Zhang L, Yao L, et al. Assessment of the role of artificial intelligence in the association between time of day and colonoscopy quality. *JAMA Netw Open*. 2023;6(1):e2253840.
- Wentzensen N, Lahrman B, Clarke MA, et al. Accuracy and efficiency of deep-learning-based automation of dual stain cytology in cervical cancer screening. *J Natl Cancer Inst*. 2021;113(1):72–79.
- Holmström O, Linder N, Kaingu H, et al. Point-of-care digital cytology with artificial intelligence for cervical cancer screening in a resource-limited setting. *JAMA Netw Open*. 2021;4(3):e211740.
- Lehman CD, Mercaldo S, Lamb LR, et al. Deep learning vs traditional breast cancer risk models to support risk-based mammography screening. *J Natl Cancer Inst*. 2022;114(10):1355–1363.
- Freeman K, Geppert J, Stinton C, et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. *BMJ*. 2021;374:n1872.
- Mikhael PG, Wohlwend J, Yala A, et al. Sybil: a validated deep learning model to predict future lung cancer risk from a single low-dose chest computed tomography. *J Clin Oncol*. 2023;41:JCO2021345–JCO2022000.
- Baxi V, Edwards R, Montalto M, Saha S. Digital pathology and artificial intelligence in translational medicine and clinical practice. *Mod Pathol*. 2022;35(1):23–32.
- Hanna MG, Parwani A, Sirintrapun SJ. Whole slide imaging: technology and applications. *Adv Anat Pathol*. 2020;27(4):251–259.
- Cruz-Roa A, Gilmore H, Basavanahally A, et al. Accurate and reproducible invasive breast cancer detection in whole-slide images: a deep learning approach for quantifying tumor extent. *Sci Rep*. 2017;7:46450.
- Teramoto A, Tsukamoto T, Kiriya Y, Fujita H. Automated classification of lung cancer types from cytological images using deep convolutional neural networks. *Biomed Res Int*. 2017;2017:4067832.
- Rakae M, Adib E, Ricciuti B, et al. Association of Machine Learning–Based Assessment of tumor-infiltrating lymphocytes on standard histologic images with outcomes of immunotherapy in patients with NSCLC. *JAMA Oncol*. 2023;9(1):51–60. Accessed November 17, 2022. <https://jamanetwork.com/journals/jamaoncology/fullarticle/2798850>
- Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*. 2017;318(22):2199–2210.
- Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med*. 2018;24(10):1559–1567.

34. Ding K, Zhou M, Wang H, Zhang S, Metaxas DN. Spatially aware graph neural networks and cross-level molecular profile prediction in colon cancer histopathology: a retrospective multi-cohort study. *Lancet Digit Health*. 2022;4(11):e787-e795.
35. Saldanha OL, Loeffler CML, Niehues JM, et al. Self-supervised attention-based deep learning for pan-cancer mutation prediction from histopathology. *NPJ Precis Oncol*. 2023;7(1):35.
36. Zhou L-Q, Wu X-L, Huang S-Y, et al. Lymph node metastasis prediction from primary breast cancer US images using deep learning. *Radiology*. 2020;294(1):19-28.
37. Kann BH, Hicks DF, Payabvash S, et al. Multi-institutional validation of deep learning for pretreatment identification of extranodal extension in head and neck squamous cell carcinoma. *J Clin Oncol*. 2020;38(12):1304-1311.
38. Shimada Y, Kudo Y, Maehara S, et al. Artificial intelligence-based radiomics for the prediction of nodal metastasis in early-stage lung cancer. *Sci Rep*. 2023;13(1):1028.
39. Court LE, Rao A, Krishnan S. Radiomics in cancer diagnosis, cancer staging, and prediction of response to treatment. *Transl Cancer Res*. 2016;5(4):337-339. Accessed February 26, 2023. <https://tcr.amegroups.com/article/view/8701>
40. Abedian S, Sholle ET, Adekkanattu PM, et al. Automated extraction of tumor staging and diagnosis information from surgical pathology reports. *JCO Clin Cancer Inform*. 2021;5:1054-1061.
41. Soysal E, Warner JL, Wang J, et al. Developing customizable cancer information extraction modules for pathology reports using CLAMP. *Stud Health Technol Inform*. 2019;264:1041.
42. Alba PR, Gao A, Lee KM, et al. Ascertainment of veterans with metastatic prostate cancer in electronic health records: demonstrating the case for natural language processing. *JCO Clin Cancer Inform*. 2021;5:1005-1014.
43. Savova GK, Tseytlin E, Finan S, et al. DeepPhe: a natural language processing system for extracting cancer phenotypes from clinical records. *Cancer Res*. 2017;77(21):e115-e118.
44. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Adv Neural Inf Proces Syst*. 2020;33:1877-1901.
45. Touvron H, Martin L, Stone K, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv [cs.CL]. 2023; <http://arxiv.org/abs/2307.09288>
46. Agarwal M, Goswami A, Sharma P. Evaluating ChatGPT-3.5 and Claude-2 in answering and explaining conceptual medical physiology multiple-choice questions. *Cureus*. 2023;15(9):e46222.
47. Agrawal M, Hagselmann S, Lang H, Kim Y, Sontag D. Large Language Models are Few-Shot Clinical Information Extractors. arXiv [cs.CL]. 2022; <http://arxiv.org/abs/2205.12689>
48. Gehrman S, Deroncourt F, Li Y, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PLoS One*. 2018;13(2):e0192360.
49. Sivarajkumar S, Wang Y. HealthPrompt: A Zero-shot Learning Paradigm for Clinical Natural Language Processing. arXiv [cs.CL]. 2022; <http://arxiv.org/abs/2203.05061>
50. Amin MB, Edge S, Greene F, et al. *AJCC Cancer Staging Manual*. 8th ed. Springer International Publishing; 2017.
51. Carmona R, Zakeri K, Green G, et al. Improved method to stratify elderly patients with cancer at risk for competing events. *J Clin Oncol*. 2016;34(11):1270-1277.
52. Chambers P, Man KKC, Lui VWY, et al. Understanding molecular testing uptake across tumor types in eight countries: results from a multinational cross-sectional survey. *JCO Oncol Pract*. 2020;16(8):e770-e778.
53. Angus DC, Chang C-CH. Heterogeneity of treatment effect: estimating how the effects of interventions vary across individuals. *JAMA*. 2021;326(22):2312-2313.
54. Feng L, Liu Z, Li C, et al. Development and validation of a radiopathomics model to predict pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer: a multicentre observational study. *Lancet Digit Health*. 2022;4(1):e8-e17.
55. Ogier du Terrail J, Leopold A, Joly C, et al. Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. *Nat Med*. 2023;29(1):135-146.
56. Wang S, Yu H, Gan Y, et al. Mining whole-lung information by artificial intelligence for predicting EGFR genotype and targeted therapy response in lung cancer: a multicohort study. *Lancet Digital Health*. 2022;4(5):e309-e319. <https://linkinghub.elsevier.com/retrieve/pii/S2589750022000243>
57. Vanguri RS, Luo J, Aukerman AT, et al. Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. *Nat Can*. 2022;3:1151-1164. doi:10.1038/s43018-022-00416-8
58. Boehm KM, Aherne EA, Ellenson L, et al. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nat Can*. 2022;3(6):723-733.
59. Esteva A, Feng J, van der Wal D, et al. Prostate cancer therapy personalization via multi-modal deep learning on randomized phase III clinical trials. *NPJ Digit Med*. 2022;5(1):81.
60. Lipkova J, Chen RJ, Chen B, et al. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell*. 2022;40(10):1095-1110.
61. Geessink OGF, Baidoshvili A, Klaase JM, et al. Computer aided quantification of intratumoral stroma yields an independent prognosticator in rectal cancer. *Cell Oncol*. 2019;42(3):331-341.
62. Kurtz DM, Esfahani MS, Scherer F, et al. Dynamic risk profiling using serial tumor biomarkers for personalized outcome prediction. *Cell*. 2019;178(3):699-713.e19.
63. Raynaud M, Aubert O, Divard G, et al. Dynamic prediction of renal survival among deeply phenotyped kidney transplant recipients using artificial intelligence: an observational, international, multicohort study. *Lancet Digit Health*. 2021;3(12):e795-e805.
64. Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*. 2019;572(7767):116-119.
65. Arbour KC, Luu AT, Luo J, et al. Deep learning to estimate RECIST in patients with NSCLC treated with PD-1 blockade. *Cancer Discov*. 2021;11(1):59-67.
66. Fink MA, Kades K, Bischoff A, et al. Deep learning-based assessment of oncologic outcomes from natural language processing of structured radiology reports. *Radiol Artif Intell*. 2022;4(5):e220055.
67. Chen PH, Zafar H, Galperin-Aizenberg M, Cook T. Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports. *J Digit Imaging*. 2018;31(2):178-184.
68. Ping XO, Tseng YJ, Chung Y, et al. Information extraction for tracking liver cancer patients' statuses: from mixture of clinical narrative report types. *Telemed J E Health*. 2013;19(9):704-710.
69. Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. *Nature*. 2023;619:357-362. doi:10.1038/s41586-023-06160-y

70. Antman K, Amato D, Wood W, et al. Selection bias in clinical trials. *J Clin Oncol*. 1985;3(8):1142-1147.
71. Lu Z, Sim J-A, Wang JX, et al. Natural language processing and machine learning methods to characterize unstructured patient-reported outcomes: validation study. *J Med Internet Res*. 2021;23(11):e26777.
72. Forsyth AW, Barzilay R, Hughes KS, et al. Machine learning methods to extract documentation of breast cancer symptoms from electronic health records. *J Pain Symptom Manag*. 2018;55(6):1492-1499.
73. Lindvall C, Deng C-Y, Agaronnik ND, et al. Deep learning for cancer symptoms monitoring on the basis of electronic health record unstructured clinical notes. *JCO Clin Cancer Inform*. 2022;6:e2100136.
74. Koleck TA, Tatonetti NP, Bakken S, et al. Identifying symptom information in clinical notes using natural language processing. *Nurs Res*. 2021;70(3):173-183.
75. Grossman Liu L, Grossman RH, Mitchell EG, et al. A deep database of medical abbreviations and acronyms for natural language processing. *Sci Data*. 2021;8(1):149.
76. Alkaitis MS, Agrawal MN, Riely GJ, Razavi P. Automated NLP extraction of clinical rationale for treatment discontinuation in breast cancer. *JCO Clin Cancer Inform*. 2021;5:550-560.
77. Yin Z, Harrell M, Warner JL, Chen Q, Fabbri D. The therapy is making me sick: how online portal communications between breast cancer patients and physicians indicate medication discontinuation. *J Am Med Inform Assoc*. 2018;25(11):1444-1451.
78. Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *J Biomed Inform*. 2015;53:196-207.
79. Administration, U.S. Food and Drug. Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). Action Plan. 2021.
80. Chen W, Wang C, Zhan W, et al. A comparative study of auto-contouring softwares in delineation of organs at risk in lung cancer and rectal cancer. *Sci Rep*. 2021;11(1):23002.
81. Jie Z, Zhiying Z, Li L. A meta-analysis of Watson for oncology in clinical application. *Sci Rep*. 2021;11(1):5792.
82. Hong JC, Eclow NCW, Dalal NH, et al. System for high-intensity evaluation during radiation therapy (SHIELD-RT): a prospective randomized study of machine learning-directed clinical evaluations during radiation and chemoradiation. *J Clin Oncol*. 2020;38:JCO2001688-JCO2003661.
83. Manz CR, Zhang Y, Chen K, et al. Long-term effect of machine learning-triggered behavioral nudges on serious illness conversations and end-of-life outcomes among patients with cancer: a randomized clinical trial. *JAMA Oncol*. 2023;9(3):414-418. doi:10.1001/jamaoncol.2022.6303
84. Kirshner J, Cohn K, Dunder S, et al. Automated electronic health record-based tool for identification of patients with metastatic disease to facilitate clinical trial patient ascertainment. *JCO Clin Cancer Inform*. 2021;5:719-727.
85. Chow R, Midroni J, Kaur J, et al. Use of artificial intelligence for cancer clinical trial enrolment: a systematic review and meta-analysis. *J Natl Cancer Inst*. 2023;115(4):365-374. doi:10.1093/jnci/djad013
86. Liu R, Rizzo S, Whipple S, et al. Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature*. 2021;592(7855):629-633.
87. Doolan PJ, Charalambous S, Roussakis Y, et al. A clinical evaluation of the performance of five commercial artificial intelligence contouring systems for radiotherapy. *Front Oncol*. 2023;13:1213068.
88. Henderson EGA, Vasquez Osorio EM, van Herk M, Brouwer CL, Steenbakkers RJHM, Green AF. Accurate segmentation of head and neck radiotherapy CT scans with 3D CNNs: consistency is key. *Phys Med Biol*. 2023;68(8):085003. doi:10.1088/1361-6560/acc309
89. Kawahara D, Tsuneda M, Ozawa S, et al. Deep learning-based auto segmentation using generative adversarial network on magnetic resonance images obtained for head and neck cancer patients. *J Appl Clin Med Phys*. 2022;23(5):e13579.
90. Jones S, Thompson K, Porter B, et al. Automation and artificial intelligence in radiation therapy treatment planning. *J Med Radiat Sci*. doi:10.1002/jmrs.729
91. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J*. 2021;14(1):49-58.
92. Youssef A, Pencina M, Thakur A, Zhu T, Clifton D, Shah NH. External validation of AI models in health should be replaced with recurring local validation. *Nat Med*. 2023;29(11):2686-2687.
93. Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. *BMC Med*. 2023;21(1):70.
94. Diprose WK, Buist N, Hua N, Thurier Q, Shand G, Robinson R. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *J Am Med Inform Assoc*. 2020;27(4):592-600.
95. Price WN. Big data and black-box medical algorithms. *Sci Transl Med*. 2018;10(471):eaa05333. doi:10.1126/scitranslmed.aao5333
96. Li R, Shinde A, Liu A, et al. Machine learning-based interpretation and visualization of nonlinear interactions in prostate cancer survival. *JCO Clin Cancer Inform*. 2020;4:637-646.
97. Lundberg SM, Lundberg SM, Lee SI. *Adv Neural Inf Proces Syst*. 2017;30.
98. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. *Entropy*. 2020;23(1). doi:10.3390/e23010018
99. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*. 2018;2(10):749-760.
100. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021;3(11):e745-e750.
101. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453.
102. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*. 2020;368:m689.

How to cite this article: Bryant AK, Zamora-Resendiz R, Dai X, et al. Artificial intelligence to unlock real-world evidence in clinical oncology: A primer on recent advances. *Cancer Med*. 2024;13:e7253. doi:10.1002/cam4.7253