Please cite this article in press as: Kann et al., Artificial intelligence for clinical oncology, Cancer Cell (2021), https://doi.org/10.1016/j.ccell.2021.04.002

Cancer Cell

CellPress

Perspective

# Artificial intelligence for clinical oncology

Benjamin H. Kann,[1,2] Ahmed Hosny,[1,2] and Hugo J.W.L. Aerts[1,2,3,4,*]
[1]Artificial Intelligence in Medicine (AIM) Program, Mass General Brigham, Harvard Medical School, Harvard Institutes of Medicine – HIM 343, 77 Avenue Louis Pasteur, Boston, MA 02115, USA
[2]Department of Radiation Oncology, Brigham and Women's Hospital, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA
[3]Department of Radiology, Brigham and Women's Hospital, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA
[4]Radiology and Nuclear Medicine, CARIM & GROW, Maastricht University, Maastricht, the Netherlands
*Correspondence: haerts@bwh.harvard.edu
https://doi.org/10.1016/j.ccell.2021.04.002

## SUMMARY

Clinical oncology is experiencing rapid growth in data that are collected to enhance cancer care. With recent advances in the field of artificial intelligence (AI), there is now a computational basis to integrate and synthesize this growing body of multi-dimensional data, deduce patterns, and predict outcomes to improve shared patient and clinician decision making. While there is high potential, significant challenges remain. In this perspective, we propose a pathway of clinical cancer care touchpoints for narrow-task AI applications and review a selection of applications. We describe the challenges faced in the clinical translation of AI and propose solutions. We also suggest paths forward in weaving AI into individualized patient care, with an emphasis on clinical validity, utility, and usability. By illuminating these issues in the context of current AI applications for clinical oncology, we hope to help advance meaningful investigations that will ultimately translate to real-world clinical use.

## INTRODUCTION

Over the last decade, there has been a resurgence of interest in artificial intelligence (AI) applications in medicine. This is driven by the advent of deep-learning algorithms, computing hardware advances, and the exponential growth of data that are being generated and used for clinical decision making (Esteva et al., 2019; Kann et al., 2020a; LeCun et al., 2015). Oncology is particularly poised for transformative changes brought on by AI, given the proven advantages of individualized care and recognition that tumors and their response rates differ vastly from person to person (Marusyk et al., 2012; Schilsky, 2010). In oncology, much like other medical fields, the overarching goal is to increase quantity and quality of life, which, from a practical standpoint, entails choosing the management strategy that optimizes cancer control and minimizes toxicity.

As multi-dimensional data are increasingly being generated in routine care, AI can support clinicians to form an individualized view of a patient along their care pathway and ultimately guide clinical decisions. These decisions rely on the incorporation of disparate, complex data streams, including clinical presentation, patient history, tumor pathology, and genomics, as well as medical imaging, and marrying these data to the findings of an ever-growing body of scientific literature. Furthermore, these data streams are in a constant state of flux over the course of a patient's trajectory. With the emergence of AI, specifically deep learning (LeCun et al., 2015), there is now a computational basis to integrate and synthesize these data to predict where the patient's care path is headed and ultimately improve management decisions.

While there is much reason to be hopeful, numerous challenges remain to the successful integration of AI in clinical oncology. In analyzing these challenges, it is critical to view the promise, success, and failure of AI not only in generalities but on a clinical case-by-case basis. Not every cancer problem is a nail to AI's hammer; its value is not universal, but inextricably linked to the clinical use case (Maddox et al., 2019). The current evidence suggests that clinical translation of the vast majority of published, high-performing AI algorithms remains in a nascent stage (Nagendran et al., 2020). Furthermore, we posit that the imminent value of AI in clinical oncology is in the aggregation of narrow-task-specific, clinically validated, and meaningful applications at clinical "touchpoints" along the cancer care pathway, rather than general, all-purpose AI for end-to-end decision making. As the global cancer incidence increases and the financial toxicity of cancer care is increasingly recognized, many societies are moving toward value-based care systems (Porter, 2009; Yousuf Zafar, 2016). With development of these systems, there will be increasing incentive for the adoption of data-driven tools—potentially powered by AI—that can lead to reduced patient morbidity, mortality, and healthcare costs (Kuznar, 2015).

Here, we describe the key concepts of AI in clinical oncology and review a selection of AI applications in oncology from the lens of a patient moving through clinical touchpoints along the cancer care path. We therein describe the challenges faced in the clinical translation of AI and propose solutions, and finally suggest paths forward in weaving AI into individualized patient cancer care. By illuminating these issues in the context of current AI applications for clinical oncology, we hope to provide concepts to help drive meaningful investigations that will ultimately translate to real-world clinical use.

Please cite this article in press as: Kann et al., Artificial intelligence for clinical oncology, Cancer Cell (2021), https://doi.org/10.1016/j.ccell.2021.04.002



## ARTIFICIAL INTELLIGENCE: FROM SHALLOW TO DEEP LEARNING

The concept of AI, formalized in the 1950s, was originally defined as the ability of a machine to perform a task normally associated with human performance (Russell and Haller, 2003). Within this field the concept of machine learning was born, which refers to an algorithm's ability to learn data and perform tasks without explicit programming (Samuel, 1959). Machine-learning research has led to development and use of a number of "shallow" learning algorithms, including earlier generalized linear models such as logistic regression, Bayesian algorithms, decision trees, and ensemble methods (Bhattacharyya et al., 2019; Richens et al., 2020). In the simplest of these models, such as logistic regression, input variables are assumed to be independent of one another, and individual weights are learned for each variable to determine a decision boundary that optimally separates classes of labeled data. More advanced shallow learning algorithms, such as random forests, allow for the characterization and weighting of input variable combinations and relationships, thus learning decision boundaries that can fit more complex data.

Deep learning is a newer subset of machine learning that has the ability to learn patterns from raw, unstructured input data by incorporating layered neural networks (LeCun et al., 2015). In supervised learning, which represents the most common form within medical AI, a neural network will generate a prediction from this input data and compare it with a "ground truth" annotation. This discrepancy between prediction and ground truth is encapsulated in a loss function, which is then propagated back through the neural network, and over numerous cycles the model is optimized to minimize this loss function.

For the purpose of clinical application, we can view AI as a spectrum of algorithms, the utility of which are inextricably linked to the characteristics of the task under investigation. Thorough understanding of the data stream is necessary to choose, develop, and optimize an algorithm. In general, deep-learning networks offer nearly limitless flexibility in input, output, and architectural and parameter design, and thus are able to fit vast quantities of heterogeneous and unstructured data never before possible (Esteva et al., 2017). Specifically, deep learning has a high propensity to learn non-linear and high-dimensional relationships in multi-modal data including time series data, pixel-by-pixel imaging data, unstructured text data, audio/video data, or biometric data. Data with significant spatial and temporal heterogeneity are particularly well suited for deep-learning neural networks (Zhong et al., 2019). On the other hand, this power comes at the expense of limited interpretability and a proclivity for overfitting data if not trained on a large, representative dataset (Zhu et al., 2015). While traditional machine learning and statistical modeling can perform quite well at certain predictive tasks, they generally struggle to fit unprocessed, unstructured, and high-dimensional data compared with deep learning. Therefore, despite its limitations, deep learning has opened the door to "big data" analysis in oncology and promises to advance clinical oncology, as long as certain pitfalls in development and implementation can be overcome.

## CANCER CARE AS A MATHEMATICAL OPTIMIZATION PROBLEM

To appreciate the promise surrounding AI applications for clinical oncology, it is essential to incorporate a mathematical lens to the patient care path through cancer risk prediction, screening, diagnosis, and treatment. From the AI perspective, the patient path is an optimization problem, wherein heterogeneous data streams converge as inputs into a mathematical scaffold (i.e., machine-learning algorithms) (Figure 1). This scaffold is iteratively adjusted during training until the desired output can be reliably predicted and an action can be taken. In this setting, an ever-growing list of inputs includes patient clinical presentation, past medical history, genomics, imaging, and biometrics, and can be roughly subdivided as tumor, host, or environmental factors. The complexity of the algorithms is often driven by the quantity, heterogeneity, and dimensionality of such data. Outputs are centered, most broadly, on increasing survival and/or quality of life, but are often evaluated by necessity as a series of more granular surrogate endpoints.

## DATA STREAMS FOR CLINICAL ONCOLOGY

The arc of research in oncology, increasing data generation, and advances in computational technology have collectively resulted in a frameshift from low-dimensional to increasingly high-dimensional patient data representation. Earlier data and computational limitations often necessitated reducing unstructured patient data (e.g., medical images and biopsies) into a set of human-digestible discrete measures of disease extent. One notable example of such simplification lies within cancer staging systems, most prominently the American Joint Committee on Cancer (AJCC) TNM classification (Amin et al., 2017). In 1977, with only three inputs commonly available—tumor size, nodal involvement, and presence of metastasis (TNM)—the first edition of AJCC TNM staging became the standard of care for risk stratification and decision management in oncology. Over the subsequent decades, with the incorporation of other discrete data points, predictive nomograms could be generated using simple linear models, which have found practical use in certain situations (Bari et al., 2010; Creutzberg et al., 2015; Mittendorf et al., 2012; Stephenson et al., 2007). More recently, improved methods to extract and analyze existing data coupled with new data streams and a growing understanding of inter- and intra-tumoral heterogeneity have all led to the development of increasingly complex and specific stratification models. Key examples of novel data streams introduced over the past two decades are the Electronic Health Record (EHR), The Cancer Genome Atlas (Weinstein et al., 2013), The Cancer Imaging Archive (Clark et al., 2013), and the Project GENIE initiative (AACR Project GENIE Consortium, 2017). Key examples of advanced risk stratification and prediction models are the prostate cancer Decipher score (Erho et al., 2013) and breast cancer OncotypeDx score (Paik et al., 2004), which utilize discrete genomic data and shallow machine-learning algorithms to form clinically validated predictive models. Useful oncology data streams, roughly following historical order of availability, include clinical presentation, tumor stage, histopathology, qualitative imaging, tumor genomics, patient genomics, quantitative

Please cite this article in press as: Kann et al., Artificial intelligence for clinical oncology, Cancer Cell (2021), https://doi.org/10.1016/j.ccell.2021.04.002





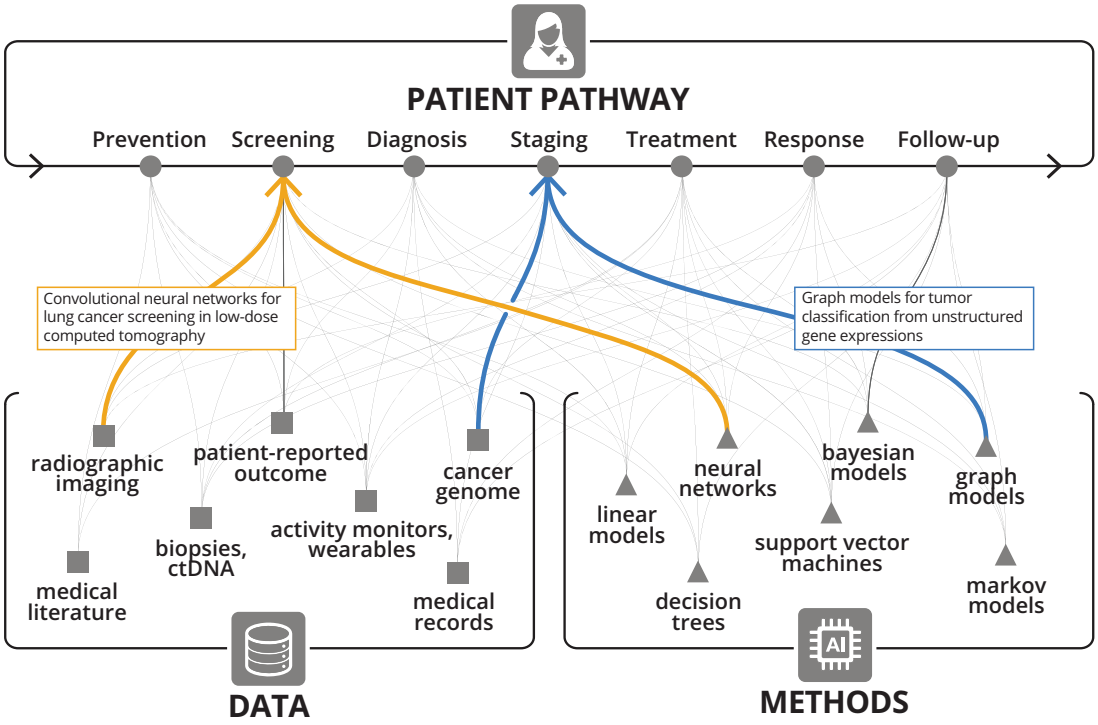


**Figure 1. Narrow-task-specific AI applications addressing a specific cancer care touchpoint along the patient pathway, and utilizing a specific data type and AI method**

imaging, liquid biopsies, electronic medical record mining, wearable devices, and digital behavior (Figure 1). Furthermore, as a patient moves along the cancer care pathway, the number of influxing, intra-patient data streams grows. With each step through the pathway, new data are generated out of the pathway with the potential to be reincorporated at a later time back into the pathway (Figure 2).

As our biological knowledge base and data streams grow in clinical oncology, machine-learning algorithms can be deployed to learn patterns that apply to more and more precise patient groups and generate predictions to guide treatment for the next, "unseen" patient. As we assimilate more data, optimal cancer care, i.e. the care that results in the best survival and quality of life for a patient, inevitably becomes precision care, assuming we have the necessary tools to fully utilize the data. Here, at this intersection of data complexity and precision care in clinical oncology, is where the promise of AI has been so tantalizing, though as of yet unfulfilled.

## AI APPLICATIONS AND TOUCHPOINTS ALONG THE CLINICAL ONCOLOGY CARE PATH

We propose that AI development for clinical oncology should be approached from patient and clinician perspectives across the following cancer care touchpoints: Risk Prediction, Screening, Diagnosis, Prognosis, Initial Treatment, Response Assessment, Subsequent Treatment, and Follow-up (Figure 2). The clinical touchpoint pathway shares features with the "cancer continuum" (Chambers et al., 2018), although it consists of more granular patient and clinician decision-oriented points of contact for AI to add clinical benefit. Each of these touchpoints involves a critical series of decisions for oncologists and patients to make and yields a use case for AI to provide an incremental benefit. Furthermore, touchpoint details will vary by cancer subtype. Within these touchpoints, ideal AI use cases are ones with significant unmet need and large available datasets. In the context of supervised machine learning, these datasets require robust and accurate annotation to form a reliable "ground truth" on which the AI system can train.

## NARROW TASKS WITH HIGH RELIABILITY

As clinical oncology data streams increase in complexity, the tools needed to discern patterns from these data are necessarily more complex. Amid this flood of heterogeneous intra-patient data there is a relative dearth of inter-patient data, which is needed to train large-scale models. Therefore, to accumulate the training data required for generalizable models, it will likely be more fruitful to target and evaluate individual AI models toward specific data streams at a particular touchpoint along the care pathway.

It is tempting to think that, given the increasing data streams that encompass multiple patient characteristics and outcomes, one could develop a unifying, dynamic model to synthesize and drive precision oncology, developing a "virtual guide" of sorts for the oncologist and patient (Topol, 2019). Analogies are often made to transformative technologies, such as self-driving cars and social media recommendations that leverage powerful neural networks on top of streams composed of billions of incoming data points, to predict real-time outcomes and

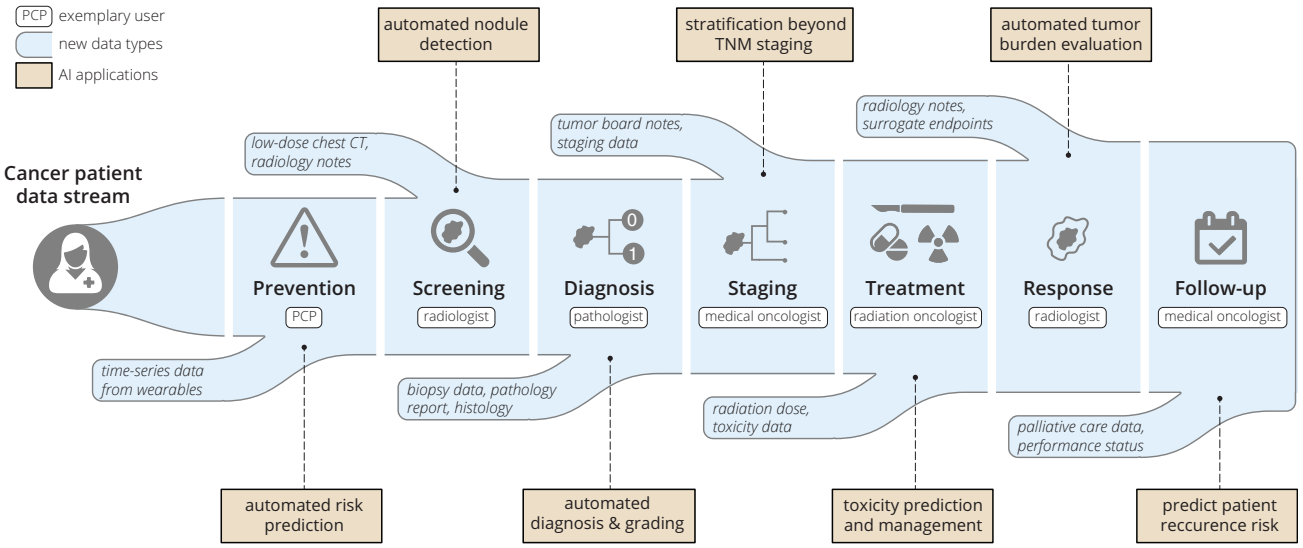


**Figure 2. An example cancer patient pathway converges with an ever-increasing data stream**
Potential AI applications and exemplary clinical users at each touchpoint are also illustrated. PCP, primary care physician.

continually improve performance. While in theory this strategy could one day be deployed in a clinical setting, there are vast differences between these domains that question whether or not we should or even could pursue this strategy currently. One of the most glaring differences between the healthcare and technology domains, in terms of AI application, is the striking difference in data quality and quantity. While there has been a sea change in the collection of data within the healthcare field over the past decade, driven by the adoption of the EHR, datasets still remain virtually siloed, intensely regulated, and, particularly in cancer care, much too small to leverage the most powerful AI algorithms available (Bi et al., 2019; Kelly et al., 2019). One of the most high-profile of these endeavors, IBM's Watson Oncology project, has attempted to develop a broad prediction machine to guide cancer care, but has been limited by suboptimal concordance with human oncologists' recommendations and subsequent distrust (Gyawali, 2018; Lee et al., 2018; Somashekhar et al., 2017).

As our biological perspective has evolved, we now know that cancer is made up of thousands of distinct entities that will follow different trajectories, each with different treatment strategies (Dagogo-Jack and Shaw, 2018; Polyak, 2011). In computational model development, there is thought to be a bare minimum number of data samples required for each model input feature (Mitsa, 2019). As we seek to make recommendations increasingly more bespoke, it becomes more challenging to accrue the quantity of training data necessary to leverage complex algorithms. Fortunately, this data gap in healthcare is well recognized, and a number of initiatives have been proposed to streamline and unify data collection (Wilkinson et al., 2016). However, given the innately heterogeneous, fragmented, and private nature of healthcare data, we in the oncology field may never achieve a level of data robustness enjoyed by other technology sectors. Therefore, strategies are necessary to mitigate the data problem, such as proper algorithm selection, model architecture improvements, data preprocessing, and data-augmentation techniques. Above all, thoughtful selection of narrow use cases across cancer care touchpoints is paramount in order to yield clinical impact.

Once rigorously tested, these narrow AI applications could then be aggregated over the course of a patient's care to provide a measurable, clinical benefit. This sort of AI-driven dimensionality reduction of a patient's feature space allows for optimizing the development process of quality AI applications in the present environment of siloed data, expertise, and infrastructure. As of writing, there are approximately 20 Food and Drug Administration (FDA)-approved AI applications targeted specifically for clinical oncology, and each of these performs a narrow task, utilizing a single data stream at a specific cancer care touchpoint (Benjamens et al., 2020; Hamamoto et al., 2020; Topol, 2019) (Table 1). We hypothesize that the future of AI in oncology will continue to consist of an aggregation of rigorously evaluated, narrow-task models, each one providing small, incremental benefits for patient quantity and quality of life. In the next sections, we review select AI applications that have excelled with this narrow-task approach.

## NARROW-TASK AI EXAMPLES ACROSS THE CLINICAL ONCOLOGY TOUCHPOINTS

### T1. Risk prediction and prevention

Given the burden to people and healthcare systems of cancer diagnosis and management, there is a significant opportunity for AI to help predict an individual's risk of developing cancer, and thereby target screening and early interventions effectively and efficiently. In a mathematical sense, the patient's entire personal history up until diagnosis makes up a vast and extremely heterogeneous data stream to be evaluated, positioning deep learning to have an impact. This is evidenced by the steady development of tools that leverage computational modeling to refine cancer risk. In the past few years, several deep-learning algorithms have been investigated to further tailor risk prediction beyond traditional models. Some of these algorithms utilize

Please cite this article in press as: Kann et al., Artificial intelligence for clinical oncology, Cancer Cell (2021), https://doi.org/10.1016/j.ccell.2021.04.002



**Table 1. FDA approvals to date for deep-learning applications in clinical oncology**

| | Name | Data type | Task | FDA summary | Year |
|---|---|---|---|---|---|
| Thoracic/liver | | | | | |
| 1 | Arterys Oncology DL | CT, MRI | segmentation of lung nodules and liver lesions, automated reporting | https://www.accessdata.fda.gov/cdrh_docs/pdf17/K173542.pdf | 2017 |
| 2 | Siemens AI-Rad Companion (Pulmonary) | CT | segmentation of lesions of the lung, liver, and lymph nodes | https://www.accessdata.fda.gov/cdrh_docs/pdf18/K183271.pdf | 2019 |
| 3 | Riverain ClearRead CT | CT | detection of pulmonary nodules in asymptomatic population | https://www.accessdata.fda.gov/cdrh_docs/pdf16/k161201.pdf | 2016 |
| 4 | Siemens syngo.CT Lung CAD | CT | detection of solid pulmonary nodules, alerts to overlooked regions | https://www.accessdata.fda.gov/cdrh_docs/pdf19/K193216.pdf | 2020 |
| 5 | GE Hepatic VCAR | CT | liver lesion segmentation and measurement | https://www.accessdata.fda.gov/cdrh_docs/pdf19/K193281.pdf | 2020 |
| 6 | Coreline AView LCS | CT | characterization of nodule type, location, measurements, and Lung-RADS category | https://www.accessdata.fda.gov/cdrh_docs/pdf20/K201710.pdf | 2020 |
| 7 | MeVis Veolity | CT | detection of solid pulmonary nodules, alerts to overlooked regions | https://www.accessdata.fda.gov/cdrh_docs/pdf20/K201501.pdf | 2021 |
| 8 | Philips Lung Nodule Assessment and Comparison Option (LNA) | CT | characterization of nodule type, location, and measurements | https://www.accessdata.fda.gov/cdrh_docs/pdf16/K162484.pdf | 2017 |
| 9 | NinesMeasure | CT | characterization of nodule type, location, and measurements | https://www.accessdata.fda.gov/cdrh_docs/pdf20/K202990.pdf | 2021 |
| Breast | | | | | |
| 10 | iCAD ProFound AI | 3D DBT mammography | detection of soft tissue densities and calcifications | https://www.accessdata.fda.gov/cdrh_docs/pdf19/K191994.pdf | 2019 |
| 11 | cmTriage | 2D FFDM | triage and passive notification | https://www.accessdata.fda.gov/cdrh_docs/pdf18/K183285.pdf | 2019 |
| 12 | Screenpoint Transpara | FFDM | detection of suspicious soft tissue lesions and calcifications | https://www.accessdata.fda.gov/cdrh_docs/pdf19/K192287.pdf | 2019 |
| 13 | Zebra Medical Vision HealthMammo | 2D FFDM | triage and passive notification | https://www.accessdata.fda.gov/cdrh_docs/pdf20/K200905.pdf | 2020 |
| 14 | Koios DS for Breast | ultrasonography | classification of lesion shape, orientation, and BI-RADS category | https://www.accessdata.fda.gov/cdrh_docs/pdf19/K190442.pdf | 2019 |
| 15 | Hologic Genius AI Detection | DBT mammography | detection of suspicious soft tissue lesions and calcifications | https://www.accessdata.fda.gov/cdrh_docs/pdf20/K201019.pdf | 2020 |
| 16 | Therapixel MammoScreen | FFDM | detection of suspicious findings and level of suspicion | https://www.accessdata.fda.gov/cdrh_docs/pdf19/K192854.pdf | 2020 |
| 17 | QuantX | MRI | image registration, automated segmentation, and analysis of user-selected regions of interest | https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN170022.pdf | 2020 |
| 18 | ClearView cCAD | ultrasonography | classification of shape and orientation of user-defined regions, and BI-RADS category | https://www.accessdata.fda.gov/cdrh_docs/pdf16/K161959.pdf | 2016 |
| Prostate | | | | | |
| 19 | Quantib Prostate | MRI | semi-automatic segmentation of anatomic structures, volume computations, automated PI-RADS category | https://www.accessdata.fda.gov/cdrh_docs/pdf20/K202501.pdf | 2020 |
| 20 | GE PROView | MRI | prediction of PI-RADS category | https://www.accessdata.fda.gov/cdrh_docs/pdf19/K193306.pdf | 2020 |

(*Continued on next page*)

Please cite this article in press as: Kann et al., Artificial intelligence for clinical oncology, Cancer Cell (2021), https://doi.org/10.1016/j.ccell.2021.04.002



**Table 1. *Continued***

| Name | Data type | Task | FDA summary | Year |
|---|---|---|---|---|
| Central nervous system | | | | |
| 21 Cortechs NeuroQuant | MRI | automated segmentation and volumetric quantification of brain lesions | https://www.accessdata.fda.gov/cdrh_docs/pdf17/K170981.pdf | 2017 |

DBT, digital breast tomosynthesis; FFDM, full-field digital mammography; BI-RADS, Breast Imaging Reporting and Data System; PI-RADS, Prostate Imaging Reporting and Data System.

novel data streams that were not available until recently: satellite imagery (Bibault et al., 2020), internet search history (White and Horvitz, 2017), and wearable devices (Beg et al., 2017). Others maximize the utility of pre-existing data streams, including patient genomics, routine imaging, unstructured health record data, and deeper family history to improve predictions (Ming et al., 2020).

### T2. Screening

Cancer screening involves the input and evaluation of data at a distinct time point to determine whether or not additional diagnostic testing and procedures are warranted. Data streams can be in the form of serum markers, medical imaging, or visual or endoscopic examination. Each of these modalities provides opportunities for the integration of AI to improve the prediction of cancer. For serum markers, such as prostate-specific antigen (PSA), early research suggests that machine-learning algorithms modeling PSA at different time points, in conjunction with other serum markers, may be able to better predict the presence of prostate cancer than PSA alone (Nitta et al., 2019). Perhaps more than in any other application, AI has found high-impact use in medical imaging screening. Narrow-task models have been developed to localize lesions and predict the risk of malignancy on lung cancer computed tomography (CT) (Ardila et al., 2019) and breast cancer mammography (McKinney et al., 2020), with applications that have been shown to perform on par, or sometimes better than expert diagnosticians (Salim et al., 2020). In these applications, raw pixel data of the image is utilized as input into a deep-learning convolutional neural network that is trained on the basis of radiologist-labeled ground-truth outputs. Importantly, while the algorithms demonstrate impressive results in terms of area under the curve, sensitivity, and specificity, they do not evaluate direct clinical endpoints, such as cancer mortality, healthcare costs, or quality of life. Outside of medical imaging, AI has found utility in screening endoscopy for colorectal carcinoma, with an application that guides biopsy-site selection (Guo et al., 2020; Zhou et al., 2020). Furthermore, there are opportunities to improve diagnostic yield for other malignancies for which screening has been traditionally difficult and unproven. This could be accomplished by AI improving the analysis of pre-existing data streams, such as abdominal CT or magnetic resonance imaging (MRI), or via its ability to integrate multi-modal data streams, such as EHR and genomic data. While currently the United States Preventive Services Task Force (USPSTF, 2021) recommends against screening for many cancers, there are a number of ongoing investigations to determine whether incorporation of AI into screening criteria and technology may allow screening to be utilized in a wider array of disease sites, such as pancreatic cancer.

### T3. Diagnosis

Diagnosing involves the exclusion of other benign disease processes and the characterization of cancer by primary site, histopathology, and, increasingly, genomic classification. Diagnosis represents an AI touchpoint for these three domains by analyzing their respective data streams: including clinical examination and medical imaging (i.e., radiomics), digital pathology, and genomic sequencing. A key study that revealed the promise of deep learning for cancer diagnosis showed that convolutional neural networks could achieve dermatologist-level accuracy in the classification of skin cancers utilizing digital photographs (Esteva et al., 2017). Other promising areas of investigation in this realm include non-invasive brain tumor diagnosis (Chang et al., 2018) and prostate cancer Gleason grading (Schelb et al., 2019) via MRI, automated histopathologic diagnosis for breast cancer (Ehteshami Bejnordi et al., 2017) and prostate cancer (Nagpal et al., 2020), and utilization of radiographic and histopathologic data to predict underlying genomic classification (Lu et al., 2018). Thus far, the Screening and Diagnosis touchpoints account for nearly all FDA-approved deep learning applications for clinical oncology, with three algorithms focusing on mammography and three focusing on CT-based lesion diagnosis (Benjamens et al., 2020).

### T4. Risk stratification and prognosis

Historically, risk stratification consisted of TNM staging, although increasingly additional data streams such as genomics, advanced imaging, and serum markers have allowed for more precise risk stratification. Given the vast heterogeneity in cancer risk, risk stratification presents a highly attractive use case for AI. Over the past two decades, genomic classifiers, developed with machine learning, have been integrated into risk stratification for a number of malignancies. Classifiers such as OncotypeDx for breast cancer, a logistic regression-based classifier, and the Decipher score, a random forest-based classifier, have demonstrated the ability to improve prognostication (Spratt et al., 2017) and guide treatment (Sparano et al., 2018). The Decipher score genomic classifier is based on 22 genomic expression markers input into a random forest model that was trained to predict metastasis after prostatectomy for patients with prostate cancer at a single institution (Erho et al., 2013). This classifier has been subsequently validated in several external settings, and is now undergoing investigation in randomized controlled trials (NCT04513717,NCT02783950). Deep-learning strategies have been explored to integrate multi-omic data sources into risk-stratification models utilizing combinations of diagnostic imaging (Kann et al., 2020b), EHR data (Beg et al., 2017; Manz et al., 2020), and genomic information (Qiu et al., 2020). Furthermore, there is the potential for deep learning to better risk-stratify

Please cite this article in press as: Kann et al., Artificial intelligence for clinical oncology, Cancer Cell (2021), https://doi.org/10.1016/j.ccell.2021.04.002



patients based on large population databases, such as the Surveillance, Epidemiology, and End Results program, by learning non-linear relationships between database variables, although preliminary efforts require validation (She et al., 2020).

### T5. Initial treatment strategy

The formulation of initial treatment strategy is arguably the most pivotal touchpoint for AI in the cancer pathway, as it directly influences patient management. The last two decades have seen exponential growth in the number and complexity of initial treatment options for common cancers (Kann et al., 2020a). A common predicament for initial treatment is which combination of systemic therapy, radiotherapy, and surgery is optimal for a given patient. Machine-learning methods utilizing genomic (Scott et al., 2017) and radiomic data (Lou et al., 2019) have been investigated to predict radiation sensitivity. While immunotherapy has been adopted in an increasing number of disease settings, it remains difficult to predict response based on currently available biomarkers, and machine-learning algorithms with radiomic input have demonstrated the ability to improve response prediction (Sun et al., 2018). Furthermore, deep learning has demonstrated the ability to analyze multi-modal data streams within the genomic realm: a recent analysis demonstrated that integration of tumor mutational burden, copy-number alteration, and microsatellite instability code can help predict response to immunotherapy (Xie et al., 2020). AI could also enable more accurate "evidence-based treatment." Natural language processing and powerful language models can help analyze published scientific works and utilize existing oncology literature, for example by extracting medical oncology concepts from EHR and linking these to a literature corpus (Simon et al., 2019).

### T6. Response assessment

Assessment of response to treatment generally includes radiographic and clinical assessments. Quantitative response assessment criteria such as Response Evaluation Criteria in Solid Tumors (RECIST) and Response Assessment in Neuro-Oncology (RANO) have long been established as reproducible ways to assess response to therapy, although in the age of targeted immunotherapies their validity has been questioned (Villaruz and Socinski, 2013). As targeted therapeutics and immunotherapies have entered the clinic, however, it has become clear that response assessment via RECIST is inadequate, due to phenomena such as pseudoprogression (Gerwing et al., 2019). Detailed response assessment is often a time-intensive process that requires a high degree of human expertise and experience, not to mention high intra- and inter-reader variability. Additionally, despite periodic review and revision of these criteria, they remain inapt at capturing edge cases, such as variable lesion response, in the case of patients receiving immunotherapy. Deep learning has demonstrated potential for automated response assessment, including automated RANO assessment (Kickingereder et al., 2019) and RECIST response in patients undergoing immunotherapy (Arbour et al., 2020).

### T7. Subsequent treatment strategy

When approaching AI algorithm development for subsequent treatment strategy, there are a number of specific considerations that generate complexity as compared with initial treatment strategy. To begin with there are additional data streams to consider, such as prior treatments, treatment-related toxicity, restaging imaging, and often multiple tissue specimens. Given the heterogeneity in data streams and the shrinking patient populations from which to build these models, subsequent treatment strategy is a challenging space for evidence-based decision making and, in turn, for reliable AI applications. Algorithms that utilize longitudinal follow-up information may help here. In one example, AI has demonstrated the ability to synthesize serial CT follow-up imaging for lung cancer patients post chemoradiation and to predict later recurrence (Xu et al., 2019). An intervention such as this could guide selection for patients to undergo consolidative treatments such as surgery or immunotherapy.

### T8. Follow-up

Another underexplored area for AI oncologic applications is the development of tools to guide precision follow-up. Diagnostic and screening algorithms may often be transferable to the follow-up setting, but will require retraining and validation for the task of interest. Similar to T7, the effect of prior cancer treatment on the data stream will often shift things significantly. For example, radiomic features extracted from the same tumor, pre- and post-treatment, show significant discrepancies (van Dijk et al., 2019). These "delta" features could be used to predict patient recurrence risk and late toxicity, helping to tailor follow-up plans (Chang et al., 2019). Appropriately triaging patients for escalated follow-up and attention can promote decreased morbidity and more efficient healthcare resource utilization; AI leveraging EHR data has demonstrated the ability to accomplish this by selecting patients at high risk for acute-care visit while undergoing cancer therapy and assigning them to an escalated preventive care strategy (Hong et al., 2020). In cases where patients have untreatable relapse, end-of-life care becomes an extremely important and challenging process. AI has shown potential here as well, as a way to triage patients at high risk of mortality and nudge physicians to converse with patients regarding their values, wishes, and quality-of-life options (Ramchandran et al., 2013).

## CHALLENGES FOR CLINICAL TRANSLATION: BEYOND PERFORMANCE VALIDATION

While tremendous strides have been made in the development of oncologic AI, as evidenced by the surge in publications and published datasets in recent years, there remains a large gap between evidence for AI performance and evidence for clinical impact. While there have been thousands of published studies of deep-learning algorithm performance (Kann et al., 2019), a recent systematic review found only nine prospective trials and two published randomized clinical trials of deep learning in medical imaging (Nagendran et al., 2020).

As alluded to above, perhaps the defining barrier to development of clinical AI applications in oncology, and healthcare overall, is data limitation, both in quality and quantity. The problems with data curation, aggregation, transparency, bias, and reliability have been well described (Norgeot et al., 2020; Thompson et al., 2018). Additionally, the lack of AI model interpretability, trust, reproducibility, and generalizability has received ample

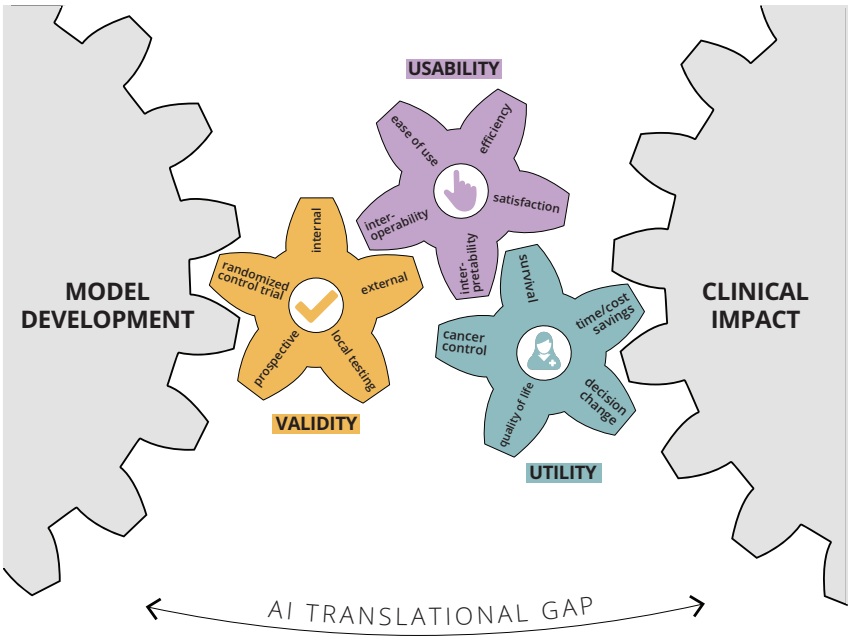


**Figure 3. Bridging the AI translational gap between initial model development and routine clinical cancer care by emphasizing and demonstrating three essential concepts: clinical validity, utility, and usability**

and well-justified attention (Beam et al., 2020). While all of these challenges must be overcome for successful AI development, here we introduce several concepts specific to clinical translation of models that have already succeeded in preliminary stages of development and validation: clinical validity, utility, and usability (Figure 3). Incorporation of these concepts into model design and evaluation is easy to overlook, yet is critical to move clinical AI beyond the research and development stage into real-world cancer care.

To demonstrate *clinical validity*, a model is often evaluated in the following general sequence: internal validation, external validation, prospective testing, and local testing in the real-world population of interest (Park et al., 2020). Recently developed guidelines such as FAIR data, CONSORT/SPIRTAI, and the (in development) TRIPOD-AI checklists should be followed to ensure reproducibility, transparency, and methodologic rigor (Liu et al., 2019). These guidelines are an important step forward in standardizing AI model development pathways and establishing a basis to determine AI study methodological rigor. While the vast majority of AI published reports include an internal blinded test set, far fewer utilize an external validation set, and an even smaller proportion employ prospective testing and benchmark comparisons with human experts (Kim et al., 2019). Given the lack of hypothesis-driven feature selection in most AI models, performance in real-world scenarios can vary dramatically if the test data distribution varies from the training data (Moreno-Torres et al., 2012). For this reason, multiple external validation sets are of utmost importance. Beyond this, it is often difficult to predict how a model will perform on edge cases, i.e., those that were under-represented in training data (Oakden-Rayner et al., 2020). In the practice of oncology, detection of rare findings can be critical to safe cancer care, and thus must be taken into account to demonstrate that a model is clinically valid. One way to mitigate the risk of model failure in real-world use is to conduct trial, run-in periods of "silent" prospective testing in the scenario of interest (Kang et al., 2020). If a model performs well in the run-in period, there is some assurance that it will be safe to use, although its performance on extremely rare cases may be still difficult to presume.

Demonstrating *clinical utility* requires clinical validity as a prerequisite, but goes beyond performance validation to the testing of clinically meaningful endpoints. High performance on commonly used endpoints, such as area under the receiver-operating characteristic curve, sensitivity, or specificity, may suffice for certain diagnostic applications, but real-world impact will require validation of clinical endpoints as appropriate for each touchpoint along the care pathway. In the case of oncology, this includes overall survival, disease control, toxicity reduction, improved quality of life, and decreased healthcare resource utilization. Testing of these endpoints should be ideally performed in the setting of a randomized trial. The gold standard would be randomizing patients to the AI intervention and directly comparing clinical endpoints. A few of these trials have been completed, with one notable example involving testing accuracy for polyp detection rate on colonoscopy (Wang et al., 2019). In this study, the primary outcome was adenoma detection rate. Despite demonstrating the superiority of the AI systems, downstream clinical benefit in terms of quality of life or survival requires yet further investigation. Another approach to AI clinical trials is to apply a validated model to all patients for risk stratification and then to apply randomized interventions. This was pursued successfully in a trial that utilized EHR data to predict patients at high risk for emergency department (ED) visits during radiotherapy (Hong et al., 2020). High-risk patients were then randomized to usual care or extra preventive provider visits. It was found that high-risk patients randomized to extra visits had significantly fewer ED and hospital admissions, while low-risk patients had uniformly low rates of ED and hospital admissions without extra care. While providing a lower level of clinical utility evidence than a true randomized trial, this type of study strategy is attractive and practical for AI-based risk-prediction models, which make up a large proportion of AI models in development. Randomized clinical trials are notoriously difficult and time consuming to execute, and AI interventions have unique characteristics that make such undertakings even more daunting. Notably, AI models are able to adapt to new data and improve over time; how would one take this into account in a traditional randomized trial? While we need AI to embrace randomized trials to truly prove clinical utility, it may be time to recognize that a re-imagining of the traditional randomized clinical trial may be necessary

Please cite this article in press as: Kann et al., Artificial intelligence for clinical oncology, Cancer Cell (2021), https://doi.org/10.1016/j.ccell.2021.04.002



to appropriately study the benefits of AI applications (Haring, 2019).

Beyond validation of clinically meaningful endpoints, demonstrating *clinical usability* involves study of the AI model in a real-world setting, where it interfaces with clinical practitioners and patients. Evaluation of effects of the model on time task, user satisfaction, and acceptance of AI recommendations should be performed (Kumar et al., 2020). A mechanism of feedback should be integrated into the design of the platform to identify weak points and opportunities for improved interface (Cutillo et al., 2020). Additionally, inter-operability between systems at the facility-to-facility, intra-facility, and point-of-care levels are crucial to streamline workflow (He et al., 2019). Usability issues are also specific to the data streams being analyzed. New data streams such as mobile health data and wearable activity monitors each present unique challenges to usability and adoption (Beg et al., 2017). A key component of promoting usability is interpretability of the AI algorithm. As data streams become more inter-related, it is increasingly difficult to discern a biological or clinical rationale supporting an algorithm's predictions. This "black-box" effect may be acceptable in certain consumer electronics industries, but due to the consequential and medicolegal nature of healthcare decision making, lack of interpretability poses a tremendous barrier to clinical use (Doshi-Velez and Kim, 2017; Wang et al., 2020). Fortunately, there is a growing research field dedicated to investigation of interpretability issues, and several techniques, such as saliency maps, hidden-states analysis, variable importance metrics, and feature visualizations can illuminate some aspects of AI prediction rationale (Guo et al., 2019; Olah et al., 2018). Beyond this, an appreciation of advances in Human Factors research and collaboration with appropriate experts can help streamline the adoption of otherwise clinically validated algorithms. Finally, translating algorithms into clinically usable solutions requires robust information technology support services that may require dedicated investment from clinical institutions and departments.

Another key concept related to clinical usability is addressing the challenges that emerge when multiple AI models are deployed sequentially or simultaneously at a given touchpoint or series of touchpoints. Orchestration of these situations, which are expected to become more common, require attention to end-user responsibilities, inter-operability, access, and training. As patients move through the oncology care path, they interact (directly or indirectly) with many different care providers who may be the primary users of a given AI application (Figure 2). These users may have a primarily diagnostic or therapeutic role (or both). From a simplified perspective, the primary diagnosticians of the cancer care path are pathologists and radiologists, while the therapeutic clinicians tend to be medical, radiation, interventional, and surgical oncologists. Multidisciplinary touchpoints along the pathway, e,g., tumor boards, represent opportunities to collate and orchestrate disparate AI applications. In addition to physicians, there are numerous advanced practice providers such as nurses and physician assistants, as well as therapists, social workers, and medical students, who may be users of a specific AI application. If, for example, a patient receives a CT scan with an AI-generated prediction of malignancy, and this prediction is subsequently utilized as input for another algorithm to recommend surgery as treatment, who is the "designated user" primarily responsible for utilizing and disseminating that information? A further issue, which logically follows, is who is legally liable for decisions based on the use of the model. Specific solutions have not yet been developed to address these issues, and are, unfortunately, likely to arise on an ad hoc case-by-case basis. This clinical orchestration of AI models merits further resources, investigation, and guidelines aimed at medical AI developers and cancer care providers to navigate these complex issues.

Despite the vanishingly few FDA-approved AI applications for oncologic indications, with numerous applications in the pipeline, there is substantial interest in streamlining ways to bridge the gap between development and clinical translation. Accordingly, the FDA is in the process of devising AI- and machine-learning-specific guidelines for approved clinical use. The recently released action plan incorporates the above clinical concepts and sets the stage for further defining a framework for safe AI translation to the clinic (FDA, 2021).

## CONCLUSIONS

Increasing data streams and advances in computational algorithms have positioned AI to improve clinical oncology via rigorously evaluated, narrow-task applications interacting at specific touchpoints along the cancer care path. While there are a number of promising AI applications for clinical oncology in development, substantial challenges remain to bridge the gap to clinical translation. The most successful models leverage large-scale, robustly annotated datasets for narrow tasks at specific cancer care touchpoints. Further development of AI applications for cancer care should focus on clinical validity, utility, and usability. Successful incorporation of these concepts will require bringing a patient-provider, clinical decision-centric emphasis to model development and evaluation.

### ACKNOWLEDGMENTS

The authors acknowledge financial support from NIH (H.J.W.L.A.: NIH-USA U24CA194354, NIH-USA U01CA190234, NIH-USA U01CA209414, and NIH-USA R35CA22052; B.H.K.: NIH-K08:DE030216), the European Union – European Research Council (H.J.W.L.A.: 866504), as well as the Radiological Society of North America (B.H.K.: RSCH2017).

### DECLARATIONS OF INTERESTS

H.J.W.L.A. is a shareholder of and receives consulting fees from Onc.Ai and BMS, outside submitted work. A.H. is a shareholder of and receives consulting fees from Altis Labs, outside submitted work.

### REFERENCES

AACR Project GENIE Consortium. (2017). AACR project GENIE: powering precision medicine through an international consortium. Cancer Discov. *7*, 818–831.

Amin, M.B., Greene, F.L., Edge, S.B., Compton, C.C., Gershenwald, J.E., Brookland, R.K., Meyer, L., Gress, D.M., Byrd, D.R., and Winchester, D.P. (2017). The Eighth Edition AJCC Cancer Staging Manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA Cancer J. Clin. *67*, 93–99.

Arbour, K.C., Luu, A.T., Luo, J., Rizvi, H., Plodkowski, A.J., Sakhi, M., Huang, K.B., Digumarthy, S.R., Ginsberg, M.S., Girshman, J., et al. (2020). Deep learning to estimate RECIST in patients with NSCLC treated with PD-1 blockade. Cancer Discov. *11*, 59–67.

Please cite this article in press as: Kann et al., Artificial intelligence for clinical oncology, Cancer Cell (2021), https://doi.org/10.1016/j.ccell.2021.04.002



Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat. Med. *25*, 954–961.

FDA. (2021). Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. https://www.fda.gov/media/145022/download.

Bari, A., Marcheselli, L., Sacchi, S., Marcheselli, R., Pozzi, S., Ferri, P., Balleari, E., Musto, P., Neri, S., Aloe Spiriti, M.A., et al. (2010). Prognostic models for diffuse large B-cell lymphoma in the rituximab era: a never-ending story. Ann. Oncol. *21*, 1486–1491.

Beam, A.L., Manrai, A.K., and Ghassemi, M. (2020). Challenges to the reproducibility of machine learning models in health care. JAMA *323*, 305–306.

Beg, M.S., Gupta, A., Stewart, T., and Rethorst, C.D. (2017). Promise of wearable physical activity monitors in oncology practice. J. Oncol. Pract. *13*, 82–89.

Benjamens, S., Dhunnoo, P., and Meskó, B. (2020). The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med. *3*, 118.

Bhattacharyya, R., Ha, M.J., Liu, Q., Akbani, R., Liang, H., and Baladandayuthapani, V. (2020). Personalized network modeling of the pan-cancer patient and cell line interactome *4*, 399–411.

Bi, W.L., Hosny, A., Schabath, M.B., Giger, M.L., Birkbak, N.J., Mehrtash, A., Allison, T., Arnaout, O., Abbosh, C., Dunn, I.F., et al. (2019). Artificial intelligence in cancer imaging: clinical challenges and applications. CA Cancer J. Clin. *69*, 127–157.

Bibault, J.-E., Bassenne, M., Ren, H., and Xing, L. (2020). Deep learning prediction of cancer prevalence from satellite imagery. Cancers *12*, 3844.

Chambers, D.A., Vinson, C.A., and Norton, W.E. (2018). Advancing the Science of Implementation across the Cancer Continuum (Oxford University Press).

Chang, K., Bai, H.X., Zhou, H., Su, C., Bi, W.L., Agbodza, E., Kavouridis, V.K., Senders, J.T., Boaro, A., Beers, A., et al. (2018). Residual convolutional neural network for the determination of IDH status in low- and high-grade gliomas from MR imaging. Clin. Cancer Res. *24*, 1073–1081.

Chang, Y., Lafata, K., Sun, W., Wang, C., Chang, Z., Kirkpatrick, J.P., and Yin, F.-F. (2019). An investigation of machine learning methods in delta-radiomics feature analysis. PLoS One *14*, e0226348.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al. (2013). The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J. Digit. Imaging *26*, 1045–1057.

Creutzberg, C.L., van Stiphout, R.G.P.M., Nout, R.A., Lutgens, L.C.H.W., Jürgenliemk-Schulz, I.M., Jobsen, J.J., Smit, V.T.H.B.M., and Lambin, P. (2015). Nomograms for prediction of outcome with or without adjuvant radiation therapy for patients with endometrial cancer: a pooled analysis of PORTEC-1 and PORTEC-2 trials. Int. J. Radiat. Oncol. Biol. Phys. *91*, 530–539.

Cutillo, C.M., MI in Healthcare Workshop Working Group, Sharma, K.R., Foschini, L., Kundu, S., Mackintosh, M., and Mandl, K.D. (2020). Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. NPJ Digital Med. *3*, https://doi.org/10.1038/s41746-020-0254-2.

Dagogo-Jack, I., and Shaw, A.T. (2018). Tumour heterogeneity and resistance to cancer therapies. Nat. Rev. Clin. Oncol. *15*, 81–94.

van Dijk, L.V., Langendijk, J.A., Zhai, T.-T., Vedelaar, T.A., Noordzij, W., Steenbakkers, R.J.H.M., and Sijtsema, N.M. (2019). Delta-radiomics features during radiotherapy improve the prediction of late xerostomia. Sci. Rep. *9*, 12483.

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv., 1702.08608.

Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J.A.W.M., the CAMELYON16 Consortium, Hermsen, M., Manson, Q.F., et al. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA *318*, 2199–2210.

Erho, N., Crisan, A., Vergara, I.A., Mitra, A.P., Ghadessi, M., Buerki, C., Bergstralh, E.J., Kollmeyer, T., Fink, S., Haddad, Z., et al. (2013). Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. PLoS One *8*, e66855.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., and Thrun, S. (2017). Corrigendum: dermatologist-level classification of skin cancer with deep neural networks. Nature *546*, 686.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., and Dean, J. (2019). A guide to deep learning in healthcare. Nat. Med. *25*, 24–29.

Gerwing, M., Herrmann, K., Helfen, A., Schliemann, C., Berdel, W.E., Eisenblätter, M., and Wildgruber, M. (2019). The beginning of the end for conventional RECIST — novel therapies require novel imaging approaches. Nat. Rev. Clin. Oncol. *16*, 442–458.

Guo, T., Lin, T., and Antulov-Fantulin, N. (2019). Exploring interpretable LSTM neural networks over multi-variable data. arXiv, 1905.12034.

Guo, L., Xiao, X., Wu, C., Zeng, X., Zhang, Y., Du, J., Bai, S., Xie, J., Zhang, Z., Li, Y., et al. (2020). Real-time automated diagnosis of precancerous lesions and early esophageal squamous cell carcinoma using a deep learning model (with videos). Gastrointest. Endosc. *91*, 41–51.

Gyawali, B. (2018). Does global oncology need artificial intelligence? Lancet Oncol. *19*, 599–600.

Hamamoto, R., Suvarna, K., Yamada, M., Kobayashi, K., Shinkai, N., Miyake, M., Takahashi, M., Jinnai, S., Shimoyama, R., Sakai, A., et al. (2020). Application of artificial intelligence technology in oncology: towards the establishment of precision medicine. Cancers *12*, 3532.

Haring, A. (2019). In the age of machine learning randomized controlled trials are unethical. https://towardsdatascience.com/in-the-age-of-machine-learning-randomized-controlled-trials-are-unethical-74acc05724af.

He, J., Baxter, S.L., Xu, J., Xu, J., Zhou, X., and Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. Nat. Med. *25*, 30–36.

Hong, J.C., Eclov, N.C.W., Dalal, N.H., Thomas, S.M., Stephens, S.J., Malicki, M., Shields, S., Cobb, A., Mowery, Y.M., Niedzwiecki, D., et al. (2020). 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. *38*, 3652–3661.

Kang, J., Morin, O., and Hong, J.C. (2020). Closing the gap between machine learning and clinical cancer care—first steps into a larger world. JAMA Oncol. *6*, 1731–1732.

Kann, B.H., Thompson, R., Thomas, C.R., Jr., Dicker, A., and Aneja, S. (2019). Artificial intelligence in oncology: current applications and future directions. Oncology *33*, 46–53.

Kann, B.H., Johnson, S.B., Aerts, H.J.W.L., Mak, R.H., and Nguyen, P.L. (2020a). Changes in length and complexity of clinical practice guidelines in oncology, 1996-2019. JAMA Netw. Open *3*, e200841.

Kann, B.H., Hicks, D.F., Payabvash, S., Mahajan, A., Du, J., Gupta, V., Park, H.S., Yu, J.B., Yarbrough, W.G., Burtness, B.A., et al. (2020b). Multi-institutional validation of deep learning for pretreatment identification of extranodal extension in head and neck squamous cell carcinoma. J. Clin. Oncol. *38*, 1304–1311.

Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., and King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Med. *17*, 195.

Kickingereder, P., Isensee, F., Tursunova, I., Petersen, J., Neuberger, U., Bonekamp, D., Brugnara, G., Schell, M., Kessler, T., Foltyn, M., et al. (2019). Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. Lancet Oncol. *20*, 728–740.

Kim, D.W., Jang, H.Y., Kim, K.W., Shin, Y., and Park, S.H. (2019). Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J. Radiol. *20*, 405–410.

Kumar, A., Chiang, J., Hom, J., Shieh, L., Aikens, R., Baiocchi, M., Morales, D., Saini, D., Musen, M., Altman, R., et al. (2020). Usability of a machine-learning

Please cite this article in press as: Kann et al., Artificial intelligence for clinical oncology, Cancer Cell (2021), https://doi.org/10.1016/j.ccell.2021.04.002



clinical order recommender system interface for clinical decision support and physician workflow. medRxiv. https://doi.org/10.1101/2020.02.24.20025890.

Kuznar, W. (2015). The push toward value-based payment for oncology. Am. Health Drug Benefits *8*, 34.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444.

Lee, W.-S., Ahn, S.M., Chung, J.-W., Kim, K.O., Kwon, K.A., Kim, Y., Sym, S., Shin, D., Park, I., Lee, U., et al. (2018). Assessing concordance with Watson for oncology, a cognitive computing decision support system for colon cancer treatment in Korea. JCO Clin. Cancer Inform *2*, 1–8.

Liu, X., Faes, L., Calvert, M.J., and Denniston, A.K.; CONSORT/SPIRIT-AI Extension Group (2019). Extension of the CONSORT and SPIRIT statements. Lancet *394*, 1225.

Lou, B., Doken, S., Zhuang, T., Wingerter, D., Gidwani, M., Mistry, N., Ladic, L., Kamen, A., and Abazeed, M.E. (2019). An image-based deep learning framework for individualising radiotherapy dose: a retrospective analysis of outcome prediction. Lancet Digital Health *1*, e136–e147.

Lu, C.-F., Hsu, F.-T., Hsieh, K.L.-C., Kao, Y.-C.J., Cheng, S.-J., Hsu, J.B.-K., Tsai, P.-H., Chen, R.-J., Huang, C.-C., Yen, Y., et al. (2018). Machine learning–based radiomics for molecular subtyping of gliomas. Clin. Cancer Res. *24*, 4429–4436.

Maddox, T.M., Rumsfeld, J.S., and Payne, P.R.O. (2019). Questions for artificial intelligence in health care. JAMA *321*, 31–32.

Manz, C.R., Chen, J., Liu, M., Chivers, C., Regli, S.H., Braun, J., Draugelis, M., Hanson, C.W., Shulman, L.N., Schuchter, L.M., et al. (2020). Validation of a machine learning algorithm to predict 180-day mortality for outpatients with cancer. JAMA Oncol. *6*, 1723–1730.

Marusyk, A., Almendro, V., and Polyak, K. (2012). Intra-tumour heterogeneity: a looking glass for cancer? Nat. Rev. Cancer *12*, 323–334.

McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., et al. (2020). International evaluation of an AI system for breast cancer screening. Nature *577*, 89–94.

Ming, C., Viassolo, V., Probst-Hensch, N., Dinov, I.D., Chappuis, P.O., and Katapodi, M.C. (2020). Machine learning-based lifetime breast cancer risk reclassification compared with the BOADICEA model: impact on screening recommendations. Br. J. Cancer *123*, 860–867.

Mitsa, T. (2019). How do you know you have enough training data?. https://towardsdatascience.com/how-do-you-know-you-have-enough-training-data-ad9b1fd679ee.

Mittendorf, E.A., Hunt, K.K., Boughey, J.C., Bassett, R., Degnim, A.C., Harrell, R., Yi, M., Meric-Bernstam, F., Ross, M.I., Babiera, G.V., et al. (2012). Incorporation of sentinel lymph node metastasis size into a nomogram predicting non-sentinel lymph node involvement in breast cancer patients with a positive sentinel lymph node. Ann. Surg. *255*, 109–115.

Moreno-Torres, J.G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N.V., and Herrera, F. (2012). A unifying view on dataset shift in classification. Pattern Recognit *45*, 521–530.

Nagendran, M., Chen, Y., Lovejoy, C.A., Gordon, A.C., Komorowski, M., Harvey, H., Topol, E.J., Ioannidis, J.P.A., Collins, G.S., and Maruthappu, M. (2020). Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ *368*, m689.

Nagpal, K., Foote, D., Tan, F., Liu, Y., Chen, P.-H.C., Steiner, D.F., Manoj, N., Olson, N., Smith, J.L., Mohtashamian, A., et al. (2020). Development and validation of a deep learning algorithm for Gleason grading of prostate cancer from biopsy specimens. JAMA Oncol. *6*, 1372–1380.

Nitta, S., Tsutsumi, M., Sakka, S., Endo, T., Hashimoto, K., Hasegawa, M., Hayashi, T., Kawai, K., and Nishiyama, H. (2019). Machine learning methods can more efficiently predict prostate cancer compared with prostate-specific antigen density and prostate-specific antigen velocity. Prostate Int. *7*, 114–118.

Norgeot, B., Quer, G., Beaulieu-Jones, B.K., Torkamani, A., Dias, R., Gianfrancesco, M., Arnaout, R., Kohane, I.S., Saria, S., Topol, E., et al. (2020). Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat. Med. *26*, 1320–1324.

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., and Ré, C. (2020). Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. Proc. ACM Conf. Health Inference Learn *2020*, 151–159.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., and Mordvintsev, A. (2018). The building blocks of interpretability. Distill *3*, 1572.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F.L., Walker, M.G., Watson, D., Park, T., et al. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N. Engl. J. Med. *351*, 2817–2826.

Park, Y., Jackson, G.P., Foreman, M.A., Gruen, D., Hu, J., and Das, A.K. (2020). Evaluating artificial intelligence in medicine: phases of clinical research. JAMIA Open *3*, 326–331.

Polyak, K. (2011). Heterogeneity in breast cancer. J. Clin. Invest. *121*, 3786–3788.

Porter, M.E. (2009). A strategy for health care reform—toward a value-based system. N. Engl. J. Med. *361*, 109–112.

Qiu, Y.L., Zheng, H., Devos, A., Selby, H., and Gevaert, O. (2020). A meta-learning approach for genomic survival analysis. Nat. Commun. *11*, 6350.

Ramchandran, K.J., Shega, J.W., Von Roenn, J., Schumacher, M., Szmuilowicz, E., Rademaker, A., Weitner, B.B., Loftus, P.D., Chu, I.M., and Weitzman, S. (2013). A predictive model to identify hospitalized cancer patients at risk for 30-day mortality based on admission criteria via the electronic medical record. Cancer *119*, 2074–2080.

Richens, J.G., Lee, C.M., and Johri, S. (2020). Improving the accuracy of medical diagnosis with causal machine learning. Nat. Commun. *11*, 3923.

Russell, I., and Haller, S. (2003). Introduction: tools and techniques of artificial intelligence. Int. J. Pattern Recognition Artif. Intell. *17*, 685–687.

Salim, M., Wåhlin, E., Dembrower, K., Azavedo, E., Foukakis, T., Liu, Y., Smith, K., Eklund, M., and Strand, F. (2020). External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. JAMA Oncol. *6*, 1581–1588.

Samuel, A.L. (1959). Some studies in machine learning using the game of checkers. IBM J. Res. Dev. *3*, 210–229.

Schelb, P., Kohl, S., Radtke, J.P., Wiesenfarth, M., Kickingereder, P., Bickelhaupt, S., Kuder, T.A., Stenzinger, A., Hohenfellner, M., Schlemmer, H.-P., et al. (2019). Classification of cancer at prostate MRI: deep learning versus clinical PI-RADS assessment. Radiology *293*, 607–617.

Schilsky, R.L. (2010). Personalized medicine in oncology: the future is now. Nat. Rev. Drug Discov. *9*, 363–366.

Scott, J.G., Harrison, L.B., and Torres-Roca, J.F. (2017). Genomic biomarkers for precision radiation medicine—authors' reply. Lancet Oncol. *18*, e239.

She, Y., Jin, Z., Wu, J., Deng, J., Zhang, L., Su, H., Jiang, G., Liu, H., Xie, D., Cao, N., et al. (2020). Development and validation of a deep learning model for non-small cell lung cancer survival. JAMA Netw. Open *3*, e205842.

Simon, G., DiNardo, C.D., Takahashi, K., Cascone, T., Powers, C., Stevens, R., Allen, J., Antonoff, M.B., Gomez, D., Keane, P., et al. (2019). Applying artificial intelligence to address the knowledge gaps in cancer care. Oncologist *24*, 772.

Somashekhar, S.P., Kumarc, R., Rauthan, A., Arun, K.R., Patil, P., and Ramya, Y.E. (2017). Abstract S6-07: Double blinded validation study to assess performance of IBM artificial intelligence platform, Watson for oncology in comparison with Manipal multidisciplinary tumour board—first study of 638 breast cancer cases. Cancer Res. *77*, https://doi.org/10.1158/1538-7445.SABCS16-S6-07.

Sparano, J.A., Gray, R.J., Makower, D.F., Pritchard, K.I., Albain, K.S., Hayes, D.F., Geyer, C.E., Jr., Dees, E.C., Goetz, M.P., Olson, J.A., Jr., et al. (2018). Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. N. Engl. J. Med. *379*, 111–121.

Spratt, D.E., Yousefi, K., Deheshi, S., Ross, A.E., Den, R.B., Schaeffer, E.M., Trock, B.J., Zhang, J., Glass, A.G., Dicker, A.P., et al. (2017). Individual patient-level meta-analysis of the performance of the decipher genomic classifier in high-risk men after prostatectomy to predict development of metastatic disease. J. Clin. Oncol. *35*, 1991–1998.

Please cite this article in press as: Kann et al., Artificial intelligence for clinical oncology, Cancer Cell (2021), https://doi.org/10.1016/j.ccell.2021.04.002



Stephenson, A.J., Scardino, P.T., Kattan, M.W., Pisansky, T.M., Slawin, K.M., Klein, E.A., Anscher, M.S., Michalski, J.M., Sandler, H.M., Lin, D.W., et al. (2007). Predicting the outcome of salvage radiation therapy for recurrent prostate cancer after radical prostatectomy. J. Clin. Oncol. *25*, 2035–2041.

Sun, R., Limkin, E.J., Vakalopoulou, M., Dercle, L., Champiat, S., Han, S.R., Verlingue, L., Brandao, D., Lancia, A., Ammari, S., et al. (2018). A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. Lancet Oncol. *19*, 1180–1191.

Thompson, R.F., Valdes, G., Fuller, C.D., Carpenter, C.M., Morin, O., Aneja, S., Lindsay, W.D., Aerts, H.J.W.L., Agrimson, B., Deville, C., Jr., et al. (2018). Artificial intelligence in radiation oncology: a specialty-wide disruptive transformation? Radiother. Oncol. *129*, 421–426.

Topol, E.J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. *25*, 44–56.

(2021). USPSTF: A and B Recommendations. https://www.uspreventiveservicestaskforce.org/uspstf/recommendation-topics/uspstf-and-b-recommendations.

Villaruz, L.C., and Socinski, M.A. (2013). The clinical viewpoint: definitions, limitations of RECIST, practical considerations of measurement. Clin. Cancer Res. *19*, 2629–2636.

Wang, P., Berzin, T.M., Glissen Brown, J.R., Bharadwaj, S., Becq, A., Xiao, X., Liu, P., Li, L., Song, Y., Zhang, D., et al. (2019). Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut *68*, 1813–1819.

Wang, F., Kaushal, R., and Khullar, D. (2020). Should health care demand interpretable artificial intelligence or accept "black box" medicine? Ann. Intern. Med. *172*, 59–60.

Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., Network, C.G.A.R., et al. (2013). The cancer genome atlas pan-cancer analysis project. Nat. Genet. *45*, 1113.

White, R.W., and Horvitz, E. (2017). Evaluation of the feasibility of screening patients for early signs of lung carcinoma in web search logs. JAMA Oncol. *3*, 398–401.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data *3*, 160018.

Xie, C., Duffy, A.G., Brar, G., Fioravanti, S., Mabry-Hrones, D., Walker, M., Bonilla, C.M., Wood, B.J., Citrin, D.E., Gil Ramirez, E.M., et al. (2020). Immune checkpoint blockade in combination with stereotactic body radiotherapy in patients with metastatic pancreatic ductal adenocarcinoma. Clin. Cancer Res. *26*, 2318–2326.

Xu, Y., Hosny, A., Zeleznik, R., Parmar, C., Coroller, T., Franco, I., Mak, R.H., and Aerts, H.J.W.L. (2019). Deep learning predicts lung cancer treatment response from serial medical imaging. Clin. Cancer Res. *25*, 3266–3275.

Yousuf Zafar, S. (2016). Financial toxicity of cancer care: it's time to intervene. J. Natl. Cancer Inst. *108*, djv370.

Zhong, G., Ling, X., and Wang, L. (2019). From shallow feature learning to deep learning: benefits from the width and depth of deep architectures. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. *9*, e1255.

Zhou, D., Tian, F., Tian, X., Sun, L., Huang, X., Zhao, F., Zhou, N., Chen, Z., Zhang, Q., Yang, M., et al. (2020). Diagnostic evaluation of a deep learning model for optical diagnosis of colorectal cancer. Nat. Commun. *11*, 2961.

Zhu, X., Vondrick, C., Fowlkes, C., and Ramanan, D. (2015). Do we need more training data? Int. J. Comput. Vis. *119*, 76–92.