

Review

From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment

Kyle Swanson,^{1,6} Eric Wu,^{2,6} Angela Zhang,^{3,6} Ash A. Alizadeh,⁴ and James Zou^{1,2,5,*}

¹Department of Computer Science, Stanford University, Stanford, CA, USA

²Department of Electrical Engineering, Stanford University, Stanford, CA, USA

³Department of Genetics, Stanford University, Stanford, CA, USA

⁴Department of Medicine, Stanford University, Stanford, CA, USA

⁵Department of Biomedical Data Science, Stanford University, Stanford, CA, USA

⁶These authors contributed equally

*Correspondence: jamesz@stanford.edu

<https://doi.org/10.1016/j.cell.2023.01.035>

SUMMARY

Machine learning (ML) is increasingly used in clinical oncology to diagnose cancers, predict patient outcomes, and inform treatment planning. Here, we review recent applications of ML across the clinical oncology workflow. We review how these techniques are applied to medical imaging and to molecular data obtained from liquid and solid tumor biopsies for cancer diagnosis, prognosis, and treatment design. We discuss key considerations in developing ML for the distinct challenges posed by imaging and molecular data. Finally, we examine ML models approved for cancer-related patient usage by regulatory agencies and discuss approaches to improve the clinical usefulness of ML.

INTRODUCTION

In the past decade, machine learning (ML) has seen an explosion of applications in medicine, particularly within oncology.¹ As a set of complex, heterogeneous, and prevalent diseases, cancers provide both a challenging set of diagnostic problems and copious data in multiple modalities.² This makes clinical oncology a promising field for ML, which utilizes data to learn patterns and the structure of a dataset (see machine learning primer section for a brief introduction to ML). In particular, rich imaging and molecular data have spurred the application of ML to correlate these data sources with early cancer detection, monitoring of cancer progression, and identification of optimized therapeutic treatment.

Medical imaging has been a powerful tool that has revolutionized cancer diagnostics. In particular, medical imaging enables non-invasive, cheap, and scalable detection, localization, and monitoring of cancer. Radiology images, as well as other image modalities like skin images or colonoscopy videos, are used for screening and diagnosis.³ Pathology images of tissue samples are used to confirm a cancer diagnosis and determine prognostic factors such as cancer subtype.⁴ Both radiology and pathology images can guide treatment by informing the selection of chemotherapy or immunotherapy and aiding radiotherapy planning.⁵ As medical imaging is increasingly fundamental to the clinical oncology workflow, the quantity of imaging data is often growing faster than clinicians can handle.³ This leads to a desire for automated methods of processing medical images to reduce clinician workload, accelerate the analysis of time-sensitive

images, and mitigate clinician errors. Advances in ML for computer vision have been adapted for medical imaging and are already showing great promise for rapidly and accurately analyzing a variety of imaging modalities in clinical oncology.^{6,7}

Although imaging informs many aspects of cancer care, molecular characterization can provide a more fine-grained view of a patient's cancer status.⁸ This is particularly important as cancer therapeutics become increasingly targeted and mechanistic.⁹ Liquid biopsies, which measure molecular biomarkers present in non-invasive physiology samples such as blood or urine, have emerged as a promising approach to profiling tumor states for cancer diagnostics. Liquid and solid tumor biopsies also make it possible to serially profile tumor status and identify characteristics of tumor evolution and heterogeneity that are associated with resistance to therapies, recurrence, and poor survival outcomes.¹⁰ Due to the wealth of information provided by liquid biopsies and solid tumor biopsies, ML has been instrumental in predicting clinical outcomes and cancer status from rich molecular features.

In this review, we explore recent advances in ML applied to clinical oncology. We focus on relatively mature ML technologies already deployed or close to deployment in clinical settings. There is a large body of exciting development of ML for more basic cancer research and drug discovery that we do not cover here. Because imaging and molecular data are two major data modalities in clinical oncology with distinct ML challenges, we structure the review to discuss imaging ML and molecular ML separately. For each modality, we discuss both the major applications of ML and the types of ML models and techniques that

are frequently used. As many of these ML models are moving from lab to clinic, we also review the regulatory process for approving ML methods for cancer diagnostics. We highlight examples of recently approved ML-based devices in this category and discuss the clinical studies necessary to obtain approval. We then discuss how to improve ML model design and evaluation in order to build trust in cancer-related ML and further clinical adoption. Finally, we outline emerging technologies, both in medicine and ML, that are promising directions for future research in clinical oncology.

MACHINE LEARNING PRIMER

ML aims to solve tasks by learning patterns from data rather than using hand-coded rules.⁴ An ML model is trained to perform a task by showing it several examples of input data (e.g., mammograms) and corresponding output labels (e.g., cancer or no cancer) and updating the internal parameters of the model accordingly to make its predictions more accurate. Model evaluation on external test data, which comes from an entirely different source than the training and internal test data (e.g., a different hospital or patient population), is particularly valuable to determine the model's generalizability across diverse settings. While most ML methods for cancer are a form of supervised learning, where each data point has an associated label, unsupervised learning methods such as clustering and dimensionality reduction can produce relevant insights into unlabeled data.⁷

Traditional ML vs. deep learning

Traditional ML algorithms take a wide variety of forms, with most designed to work with tabular data, where each data point has a set of explicit features (e.g., patient age or gene mutation status) that are used to predict the label.³ One common algorithm is called a random forest, which consists of a set of decision trees, each of which is constructed based on the training data to make a series of binary decisions about the input features that culminates in a prediction of the label of the data point. Another algorithm is the support vector machine (SVM), which learns a line (or hyperplane in multiple dimensions) in the coordinate system defined by the input features to separate the data points into two classes. Regression models learn a linear combination of input features that predict either continuous labels (e.g., linear regression) or binary labels (e.g., logistic regression).

With the increasing availability and power of graphics processing units (GPUs), a subfield of machine learning called deep learning (DL) has overtaken traditional ML for many prediction tasks.³ The core component of DL models is a neural network, which consists of one or more layers of units called neurons that compute weighted sums of inputs followed by applying a nonlinear function. These layers of neurons thus compute a representation of the input called an embedding, which is then used by the final layer of neurons to make an output prediction. The DL models are more flexible compared to traditional ML models, and because DL relies less on feature engineering, they are capable of processing a wider variety of unstructured data types including images, text, and speech. However, DL models typically require significantly more training

data, so traditional ML models can still be useful, particularly for data-limited or tabular tasks.²

In order to process non-tabular data, the architecture of a neural network (e.g., number of neurons or layers or connections between neurons) is modified to fit the desired data type.² Convolutional neural networks (CNNs) are primarily designed for processing images. Graph neural networks (GNNs) handle graph data, such as cell-cell interaction graphs. Recurrent neural networks (RNNs) and transformers analyze sequential data, such as genetic sequences or series of images. Each of these classes of models has many specific model architectures, such as ResNet or U-Net for CNNs and LSTM or GRU for RNNs. The models are optimized with stochastic gradient descent. [Figure 1](#) illustrates common traditional ML and DL models.

Both traditional ML and DL models require that the data is cleaned (e.g., modifying data with missing features or extreme values) in order to learn effectively.⁴ Additionally, the input features must be amenable to the type of model. For example, neural networks use vectors of real numbers as input, so categorical features such as cancer type are typically converted to one-hot vectors with all zeros except for a single one in a position that indicates the appropriate category. Many traditional ML methods are available in the scikit-learn package, while deep learning models can be built using packages such as PyTorch and TensorFlow. Because ML models often require tuning hyperparameters to obtain optimal performance, it is important that a validation dataset be used during this step that is distinct from the held-out test dataset, which is only evaluated after the final hyperparameters have been chosen.

Training techniques

One common technique is transfer learning, where a model is first trained on a large dataset that is somewhat related to the task of interest (pre-training) before being trained on a smaller dataset consisting of the actual task of interest (fine-tuning).³ For example, image-based cancer detection models are often pre-trained on large object detection datasets, enabling the model to recognize general shapes, and are then fine-tuned on small cancer detection image datasets. Transfer learning is more useful when the pre-training data are similar to the data of interest. Another common method is data augmentation, where input data are modified (e.g., images are rotated or blurred) to artificially expand the training set and make the model more robust to noise that might appear in real-world data.⁷ Regularization is a technique that controls the size of the parameters of a model to prevent overfitting and encourage sparse feature usage.² Weak supervision involves using data with limited or noisy label information.⁷ A common type of weak supervision is multiple instance learning, in which labeled data points (e.g., images with cancer/no cancer labels) are broken down into smaller pieces (e.g., image tiles) that are easier for an ML model to process. The model makes predictions on each piece of the data separately, and those predictions are then aggregated to form a prediction for the whole data point. Finally, interpretability is a set of methods that aim to explain why a model is making a certain prediction.⁶ For example, an image-based model might highlight regions of an image that led the model to diagnose a patient with cancer.

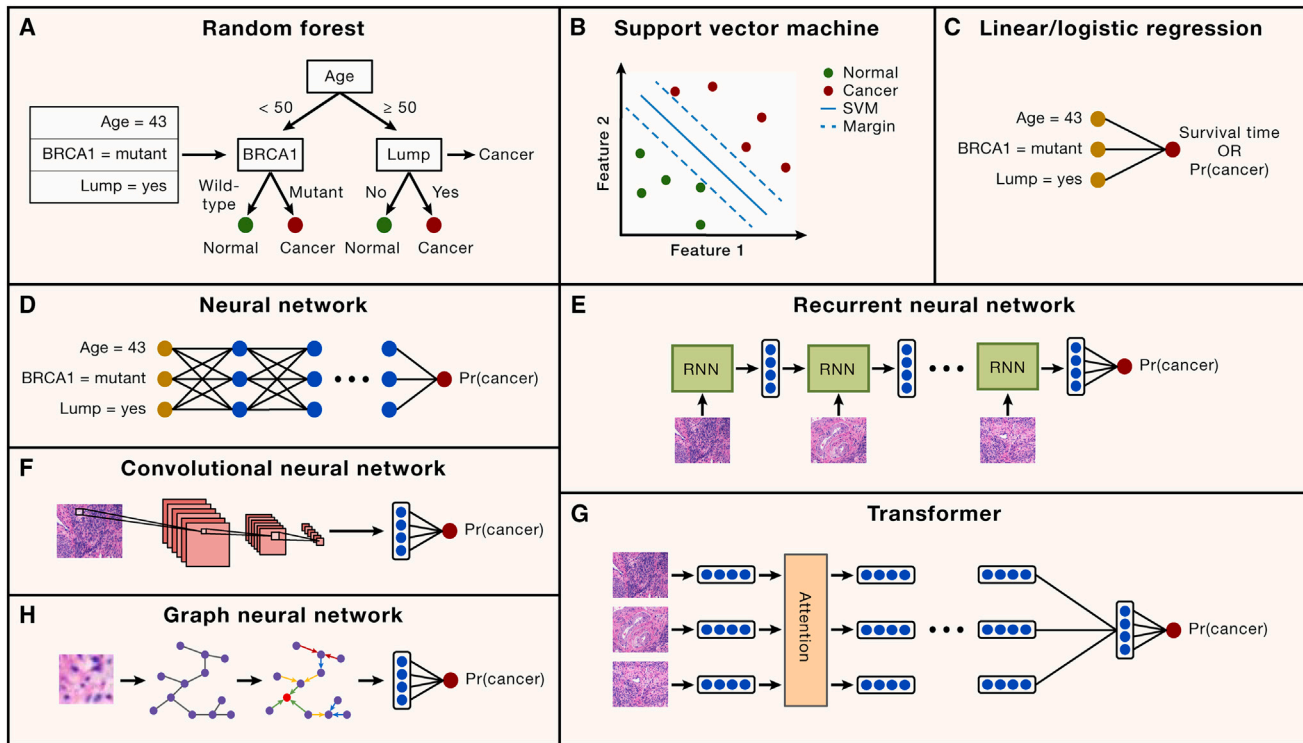


Figure 1. Common machine learning models

(A) A random forest model builds decision trees that make predictions based on a series of binary decisions about the input features.
 (B) A support vector machine (SVM) learns a line (or hyperplane in many dimensions) in feature space that separates two classes of data points with the largest possible margin between the two classes.
 (C) A regression model uses a linear combination of input features to predict either continuous labels (linear regression) or binary labels (logistic regression).
 (D) A neural network consists of multiple layers of neurons that iteratively compute linear combinations of inputs followed by a nonlinear function to predict outcomes such as the probability of cancer.
 (E) An RNN processes sequential data, such as genetic sequences or a series of images, by applying the same neural network layers to each object in the sequence and maintaining a memory of the objects it has seen.
 (F) A CNN applies patches of neurons called filters that scan an image for patterns. Early layers identify low-level features like edges, while later layers identify high-level features such as tumor morphology.
 (G) A transformer analyzes sequential data by repeatedly applying an operation called attention to compare each element in the sequence to all the other elements in order to update its internal representation of the sequence.
 (H) A GNN is designed for graph-structured data such as a graph of neighboring cells. It first encodes basic features of each node and edge in the graph, and then neural network layers pass information across the graph to update the node and edge representations, which are then used to predict the label of the graph. Each of these general classes of models has many specific architectures with different numbers and sizes of layers of neurons.
 Image sources: histology.¹¹

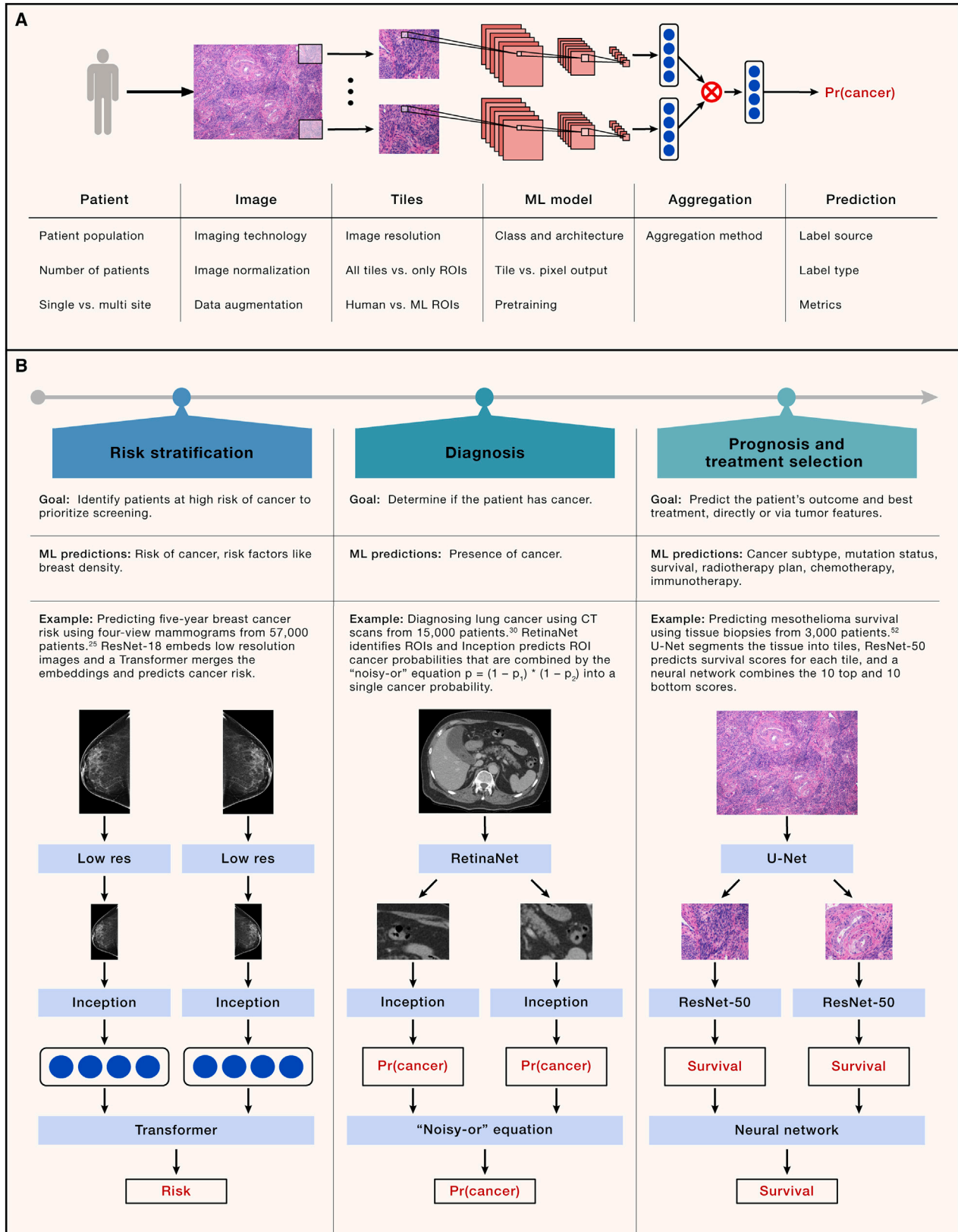
MACHINE LEARNING FOR IMAGE-BASED CANCER DIAGNOSIS, PROGNOSIS, AND TREATMENT

In this section, we highlight applications of image-based ML throughout the clinical workflow for cancer. Early ML approaches used hand-crafted image features such as tumor shape or textural heterogeneity that were computationally extracted from images.⁶ These features were used as inputs to a traditional ML model, such as a support vector machine (SVM) or random forest, to make a clinical prediction. Starting in the early 2010s, a class of ML models called deep learning (DL) models began to take hold as the dominant ML method.¹² DL models automatically learn features from an image to make clinical predictions, thereby simultaneously reducing the need for painstakingly crafting image features while significantly outperforming feature-based ML models.^{3,4} These models can be

applied to virtually any medical imaging modality, including X-ray¹³ and MRI for radiology,¹⁴ H&E stains for pathology,¹⁵ images of skin lesions for dermatology,¹⁶ and videos of colonoscopies for gastroenterology.¹⁷ Here, we discuss examples of ML—primarily DL—applied to three clinical stages: risk stratification, diagnosis, and prognosis and treatment planning. [Figure 2](#) illustrates the general image-based ML model pipeline and each of the three clinical stages. Although we discuss each stage separately, it is worth noting that some ML methods make predictions that cross these boundaries, such as simultaneous diagnosis and prognosis via pathology images.¹⁸

Risk stratification

Understanding a patient's risk of developing cancer is important for early cancer detection and effective treatment. Often, cancer risk is evaluated based on a patient's demographics, family



(legend on next page)

history, and genetics, but imaging can also reveal patient characteristics that might increase cancer risk. Existing work on image-based cancer risk prediction falls into two categories: predicting characteristics associated with cancer risk and directly predicting cancer risk itself.

Risk proxies

A typical example of a characteristic associated with cancer risk is breast density in breast cancer. Breast density is correlated with increased risk of cancer development and missed detection on mammography and therefore indicates who may benefit from additional screening.²¹ To improve breast density assessment with DL, Lehman et al. trained a ResNet-18 CNN model on mammograms to predict breast density categories routinely evaluated in clinical practice.²¹ The model showed a high level of agreement with a panel of five radiologists on a held-out test set of images. Furthermore, the DL model was implemented in clinical practice, and radiologists accepted the binary density predictions of the model 94% of the time. The model was additionally validated at an external site and showed the potential to increase the consistency of breast density evaluations by radiologists at different sites.²²

Risk prediction

More often than quantifying risk proxies, DL is used to directly predict cancer risk. For example, DL models are often trained to use images from a screening mammogram to predict whether a patient will develop cancer at some point.²³ Dombrower et al. highlight the benefit of this direct approach to risk prediction, as they showed that a breast cancer risk score produced by an Inception-ResNet-v2 CNN model was more accurate than using clinical breast density assessments to predict risk.²⁴ Yala et al. developed a DL model on mammograms that could better predict the likelihood that a woman would develop breast cancer within five years than the well-established Tyrer-Cuzick risk model, which is based on clinical features such as patient age.²⁵ Their method consisted of a ResNet-18 model to process each of the four standard mammogram views, followed by a transformer network that aggregated the view embeddings into a single mammogram embedding. This embedding was used to predict known risk factors, a baseline cancer risk score, and a hazard score for additional risk in future years. They also used a conditional-adversarial training scheme to make the model invariant to the mammogram device to ensure consistent risk assessments across devices. The authors later validated their model on test sets from seven hospitals across five countries, demonstrating the generalizability of the model across

diverse patient populations and screening centers.²⁶ Ha et al. designed a CNN model that predicts risk not only at the image level but also at the pixel level, meaning that each risk prediction score comes with a heatmap on the image indicating the regions where cancer is most likely to develop.²⁷ Although most studies have focused on risk stratification for breast cancer, ML has also been used for predicting lung cancer risk from chest X-rays with CNNs¹³ and for predicting prostate cancer risk from MRIs, with support vector machines applied to hand-crafted radiomics features.¹⁴

These methods aim to personalize cancer screening by providing a risk score to a physician, who is then responsible for determining an appropriate screening frequency for the patient. However, since standard, non-ML risk scores are relatively coarse-grained and imprecise, current guidelines place patients in large groups based on high or low risk and suggest the same screening schedule for all patients in a group, rather than adapting the screening frequency uniquely for each patient.²⁸ Yala et al. demonstrated that reinforcement learning, an area of machine learning that involves deciding which actions to take to maximize a reward, can be used in conjunction with DL risk prediction models to automatically design an optimal screening schedule for each patient individually.²⁸ These individual screening schedules significantly improved simulated early detection rates per screening mammogram compared to standard clinical guidelines.

Diagnosis

Diagnosing cancer typically involves two steps. First, either in the course of routine screening or in response to symptoms, patients undergo non-invasive imaging such as radiological scans. Second, if these images reveal suspicious regions of tissue that might indicate cancer, a biopsy is then taken and sent to a pathology lab, which can confirm the diagnosis with the help of histological imaging. ML can improve the diagnostic accuracy of both of these steps by identifying patterns—both known and unknown to clinicians—that indicate the presence or absence of cancer. ML also provides a consistent and detailed image evaluation that can catch cancers missed by time-constrained physicians, which is particularly crucial in radiology for early detection.

Non-invasive imaging

Detecting signs of cancer via ML applied to radiological or other non-invasive imaging has garnered substantial attention and excitement due to the abundance of data and the success of

Figure 2. Machine learning for image-based cancer diagnosis, prognosis, and treatment

(A) An illustration of the general ML model pipeline for image-based cancer prediction tasks, along with key considerations at each step. For each patient in a patient population, an image is captured from radiology, pathology, or another imaging modality. Often, the image is high resolution and is broken down into image tiles—either covering the full image or only ROIs—that are small enough for an ML model to process. An ML model processes each image tile, producing an embedding of the tile or a tile-level or pixel-level prediction. The tile outputs are aggregated into a single output using either a formula or an ML model such as an RNN. A final prediction component, such as a neural network, uses the combined tile output to predict the label, and metrics evaluate the model predictions. Labels may come from different sources (e.g., radiology or biopsy) and can have different types (e.g., binary for classification or real-valued for regression).

(B) The clinical stages of image-based ML predictions for cancer and simplified examples of ML methods for each stage.

Risk Stratification: For certain cancers such as breast cancer, healthy patients regularly undergo radiological screening to assess the patient's risk of developing cancer and prioritize future screening.

Diagnosis: Radiology images are used to identify potentially cancerous lesions during routine screening or in response to symptoms. If cancer is suspected by radiology, then a biopsy is taken, and pathology images are used to confirm the diagnosis.

Prognosis and Treatment Selection: Radiology or pathology images are further used to evaluate prognosis and select treatments.

Image sources: mammography,¹⁹ CT,²⁰ histology.¹¹

ML methods, with several claiming to achieve physician-level performance for cancer detection. These methods hold promise to improve and standardize early detection of cancer, save physicians time, and expand access to high-quality cancer care to patients in low-resource settings. Esteva et al. trained an Inception v3 CNN to classify skin cancer from images of skin lesions, matching the performance of 21 dermatologists on biopsy-proven clinical images.¹⁶ With the prevalence of smartphones, skin lesion classification with DL could potentially be available directly to patients.²⁹ DL also has the potential to aid doctors with diagnostic procedures such as colonoscopies by analyzing live videos and highlighting suspicious regions of tissue in real-time to guide the operation.¹⁷ In radiology, Ardila et al. developed a 3D CNN for lung cancer screening with one component identifying regions of interest (ROIs), another component processing the entire image, and a final classification layer combining the outputs of both components.³⁰ If a prior CT scan is available, the model extracts features from ROIs in both the current and prior CT images. Their model was at least on par with six radiologists and reduced both false positive and false negative rates in some situations. While many such methods were validated on relatively small datasets from a single site, McKinney et al. built a DL model for diagnosing breast cancer from mammograms and evaluated their model on large datasets from the US and the UK.³¹ They found that their model had superior performance compared to six radiologists. They also demonstrated that in many cases, they could replace a second reader, which is standard procedure in the UK, with their model's prediction and save 88% of the time of the second reader without sacrificing performance.

Despite these successes, there has been debate about the transparency, interpretability, reproducibility, and robustness of some of these results.³² Most of these studies are retrospective, single-site, and compare ML performance *post hoc* to human performance rather than evaluating ML models in the way they would be used in the clinic, as a system to assist human decision making. Some recent studies have worked to address these shortcomings to more convincingly demonstrate the benefits of ML in cancer diagnosis. Qian et al. performed a prospective, rather than retrospective, evaluation of a DL model using ultrasound to assess breast cancer.³³ Kim et al. designed a reader study in which radiologists evaluated mammograms either with or without the aid of an ML model trained on mammograms from five institutions in three countries.³⁴ Radiologists from multiple institutions had superior performance when working in conjunction with ML rather than alone. Hekler et al. had dermatologists and an ML model separately evaluate skin images to detect cancer and then combined those predictions using a decision tree-based ML algorithm called XGBoost to achieve performance superior to either method independently.³⁵

Image-based deep learning has also been used in other ways to aid preliminary cancer diagnosis. In Yala et al., a ResNet-18 model was built to triage mammograms by setting a high-sensitivity prediction threshold so that nearly all predicted negative cases were truly negative.³⁶ In a simulation study, these predicted negative cases were skipped by radiologists, allowing radiologists to only read 80.7% of mammograms while maintaining sensitivity and specificity across all cases. Instead of diagnosing

cases, Xu et al. built a CNN model to segment breast ultrasound images into functional tissues to aid clinicians who interpret and diagnose the images.³⁷ Cao et al. designed a model that simultaneously diagnoses and grades prostate cancer at the pixel level from multi-parametric MRI, leveraging the power of DL models to move beyond cancer detection alone.³⁸ A future direction is integrating patient history and pertinent clinical presentation in image-based DL models. Multimodal DL models have become increasingly popular in healthcare applications, given the importance of clinical history in diagnosis. In one instance, Akselrod-Ballin et al. trained a DL model to diagnose breast cancer from mammograms that additionally incorporates information from medical records, finding that it led to improved diagnostic accuracy over models that did not incorporate health records.³⁹

Confirmation by pathology

Pathology samples, typically stained with hematoxylin and eosin (H&E), are assessed by pathologists to confirm a preliminary cancer diagnosis. Due to the large size of digital whole slide images of histopathology, DL models frequently use multiple instance learning (MIL). In MIL, the DL model operates on small image tiles and then aggregates individual tile-level embeddings or predictions into a diagnostic prediction for the whole slide.⁴⁰ Campanella et al. used MIL to train a DL model for prostate, breast, and other cancers. The model could allow pathologists to exclude 65–75% of slides while still identifying cancers with 100% sensitivity.⁴¹ This model has the potential to significantly reduce the workload of pathologists, allowing them to spend more time on difficult cases.

As with preliminary diagnosis via non-invasive imaging, rigorous evaluations of DL-based pathology tools using multi-site, prospective trials with DL-assisted pathologists are needed to evaluate the clinical utility of these models. Several recent works have performed studies with at least some of these criteria, showing improved pathologist performance when assisted by DL models that highlight ROIs of the image and/or provide a diagnostic prediction.^{42,43}

DL models sometimes predict more than a binary cancer versus no cancer label in order to provide clinicians with additional diagnostic information. For example, in cases of cancer of unknown primary origin, determining an appropriate diagnosis and treatment plan requires inferring the origin of cancer. Lu et al. trained a ResNet-50-based model on H&E images to identify a tumor as primary or metastatic and predict its site of origin across 18 different primary origins, with top-3 prediction accuracy on an external set exceeding 90%.¹⁵ The model incorporated attention after the CNN layers, which identified regions in the slide of high diagnostic relevance and provided a form of human interpretability. Coudray et al. built an Inception v3 CNN model for lung cancer to simultaneously diagnose cancer, determine the tumor subtype of positive cases, and predict the presence of six genetic mutations from H&E-stained images.¹⁸

Prognosis and treatment selection

After a cancer diagnosis, physicians and patients are interested in determining the patient's prognosis and selecting the optimal treatment for that patient. Since both prognosis and treatment selection depend on the characteristics of the cancer, many ML methods indirectly aid prognosis and treatment selection

by predicting tumor features such as cancer subtype or mutation status. Other methods directly predict prognosis or guide treatment selection by evaluating or planning potential treatments. Below, we discuss both types of methods.

Tumor features

Prognosis and treatment selection are both informed by a number of tumor features that can be predicted by image-based ML models. For example, ML models have been developed to predict the subtype or grade of a tumor, such as the Gleason grade in prostate cancer,⁴⁴ which gives physicians information about patient survival and which treatments might be most effective. Esteva et al. fuse information from both histology slides and clinical data in a DL model that predicts the likelihood of 5- and 10-year metastasis, which can indicate more aggressive disease that requires additional treatment.⁴⁵ They pre-trained the image portion of their DL model using a self-supervised technique called momentum contrast, in which the model was trained to identify whether two image tiles were augmented versions of the same tile or were different tiles. Besides tumor subtype, another goal is to predict the genetic characteristics of a tumor, such as microsatellite instability,⁴⁶ tumor mutational burden,⁴⁷ or whole-genome duplication.⁴⁸ Some studies use H&E images to predict gene expression and assess survival-related tumor heterogeneity.⁴⁹ Saltz et al. develop a deep learning-based computational stain that identifies tumor-infiltrating lymphocytes whose spatial patterns are correlated with survival.⁵⁰ Wang et al. use a 3D CNN to predict EGFR mutation status in lung adenocarcinoma from ROIs selected manually from CT scans, thus providing a non-invasive method of genotyping cancer and informing potential treatments.⁵¹ When biopsy samples are available, it is still more reliable to measure genotypes using molecular methods that we discuss in the next section.

Prognosis

A number of DL models have been developed to predict patient survival from histology slides. Courtiol et al. provide an example of this type of model and workflow for prognosis in mesothelioma.⁵² First, they trained a U-Net CNN on several hundred manually annotated histology images to perform tissue segmentation. Next, they divided each patient's whole slide histology image into small image tiles and kept all the tiles that were predicted to contain at least 20% tissue according to the U-Net model. Using transfer learning, they took a ResNet-50 CNN pre-trained on an image recognition task called ImageNet and used it to predict a score for each tile. The 10 highest and 10 lowest scores were passed to a neural network that predicts the patient's survival time. The ResNet-50 model and neural network were trained together on 2,300 slides with a loss function based on the Cox proportional hazards model. They demonstrate that their model significantly outperforms simpler survival prediction models that only use histological type or grade without the image. Bychkov et al. instead predict survival for colorectal cancer using all image tiles by applying an RNN to aggregate the embeddings produced by a CNN model for each tile.⁵³ In contrast to methods using histology images, Xu et al. take advantage of the fact that radiology is non-invasive and can easily be repeated over time to develop a combined CNN + RNN model that updates its survival predictions over the course of treatment.⁵⁴

Response to treatment

Predicting response to treatment, either prior to or during the early stages of treatment, can aid physicians in selecting the optimal treatment for a patient. Joo et al. developed a multi-modal DL model to predict whether patients would achieve a pathologic complete response after neoadjuvant chemotherapy (NAC) for breast cancer.⁵⁵ Their model made predictions by fusing information from two different pretreatment MRIs, each processed with a 3D ResNet model, and clinical information, such as age and HER2 status, processed by a neural network. Gu et al. also aimed to predict response to NAC in breast cancer, but they applied DL models to pairs of ultrasonography images, with one image taken before NAC and the other taken after some, but not all, of the NAC treatments.⁵⁶ Through a prospective study, they showed that their model could predict whether a patient would respond to the full course of therapy, indicating that it could be used to alter the course of treatment early for those patients who are predicted not to respond. Tian et al. built a model that extracts features from CT images using a DenseNet CNN and hand-crafted radiomics features, with a neural network classifier processing the concatenation of both sets of features to assess PD-L1 expression in non-small cell lung cancer.⁵⁷ This enables a non-invasive prediction of response to anti-PD-1 antibody immunotherapy. Lu et al. found that deep learning could evaluate tumor morphological change in metastatic colorectal cancer from CT scans, which may allow early adjustments during treatment.⁵⁸ Notably, this study used an RNN to combine image features extracted by CNNs from CT scans at multiple time points during treatment.

Radiotherapy planning

Planning radiotherapy is a time-consuming process that could benefit from the speed of ML models. McIntosh et al. performed a blinded, head-to-head study of human-generated and ML-generated radiotherapy treatment plans for prostate cancer.⁵ ML-generated treatment plans were inferred from the treatment plans of previous patients who were most similar to the current patient according to a learned similarity metric based on features extracted from CT images. In their prospective study of 50 patients, ML-generated plans were selected over human-generated plans 61% of the time while reducing the radiotherapy planning time by 60%, from a median of 118 h to 47 h. Hosny et al. built a U-Net model to segment primary non-small cell lung cancer tumors and involved lymph nodes in CT images, which is a time-consuming step in radiotherapy planning, with validation across eight internal and external clinical sites in multiple countries.⁵⁹ In their study, AI assistance led to a 65% reduction in segmentation time and a 32% reduction in variability between clinicians.

MACHINE LEARNING FOR MOLECULAR CANCER DIAGNOSIS, PROGNOSIS, AND TREATMENT

Recent advances in sample processing, genomic sequencing, and molecular technologies have generated rich datasets from solid tumor biopsies and molecular liquid biopsies, which aim to detect circulating cell-free tumor DNA (cfDNA). ML models have played an instrumental role in mapping these datasets to clinical outputs. We first give an overview of liquid biopsy and solid tumor datasets and discuss how their unique

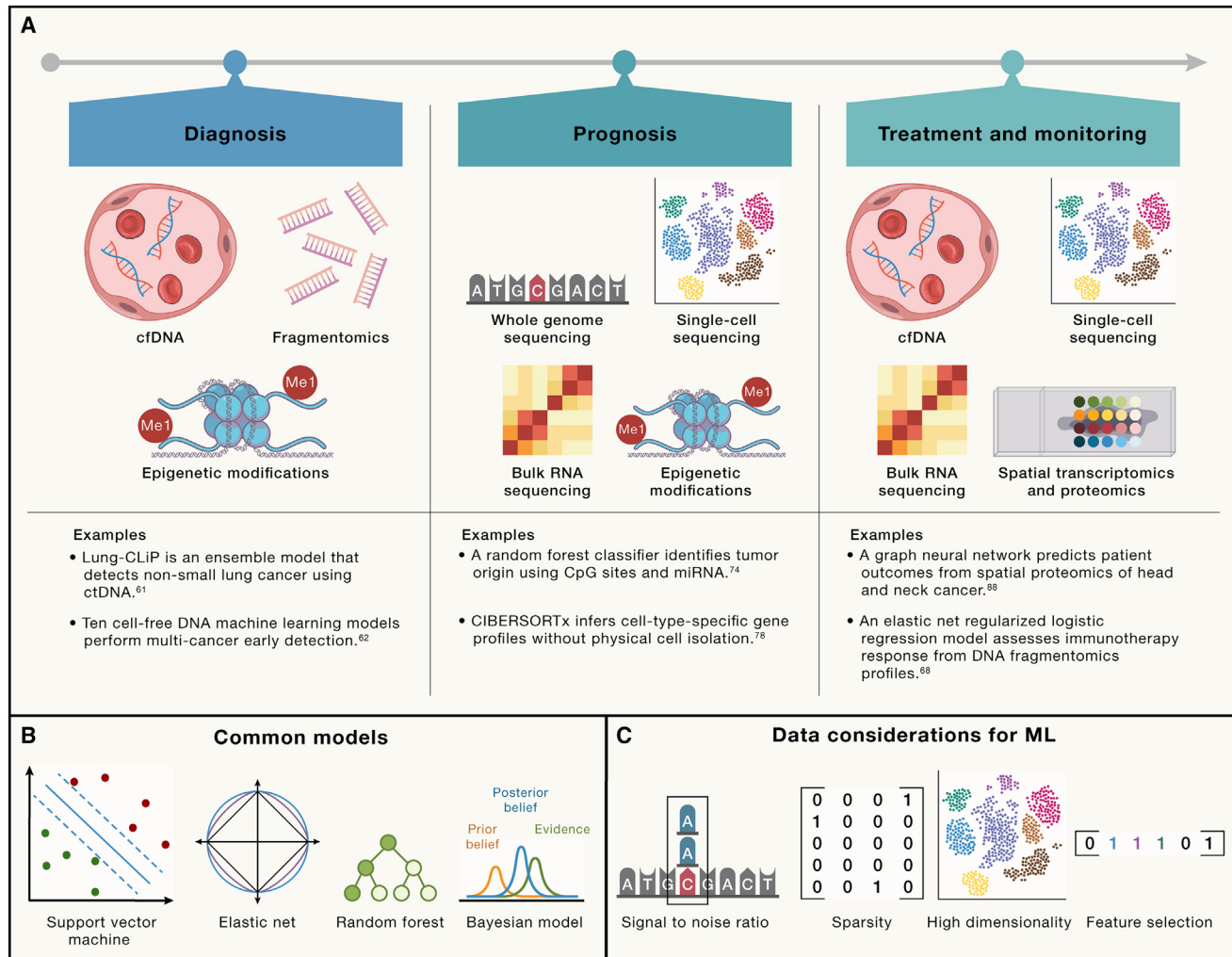


Figure 3. Machine learning for molecular cancer diagnosis, prognosis, and treatment

(A) Common molecular datasets for molecular cancer diagnostics include circulating cell-free DNA (cfDNA), methylation status, and fragmentomics. Many molecular datasets for cancer prognosis have been generated from whole-genome sequencing, single-cell transcriptomics, and bulk RNA sequencing of solid tumor biopsies. Utilizing molecular datasets for cancer treatment prediction and selection is a rapidly developing field incorporating foundational molecular technologies and emerging methods such as spatial omics. Example studies are given.

(B) Designs of common ML models for molecular data.

(C) Considerations of molecular data that inform the choice of ML model.

characteristics influence the ML models utilized. We focus our attention on how ML models have been applied for tertiary analysis of genomic datasets. We then give an overview of how ML models have been applied to facilitate liquid biopsy-based and solid tumor-based diagnosis, prognosis, and treatment selection and tumor monitoring. These advancements, summarized in Figure 3, have spurred a rapidly developing field that has garnered tremendous clinical and commercial interest.

Characteristics and ML models for molecular datasets

Liquid and solid tumor biopsy data sequencing datasets share several characteristics and challenges that guide the design of ML methods. First, dataset size is often limited. Each tumor subtype may only be represented by less than 50 samples.⁶⁰ Given the small number of samples per dataset, ML models tend to be

smaller and leverage careful feature engineering and domain expertise.⁶¹ Ongoing initiatives, such as the Circulating Cell-free Genome Atlas (CCGA), which has recruited 15,000 patients from over 140 sites, will provide valuable new resources that are multi-institutional and balanced in patient and clinical demographics.⁶²

The small sample challenge of liquid and solid tumor biopsy datasets is amplified by the high-dimensional nature of the data. Thus, applying ML to liquid and solid biopsy datasets requires careful consideration of properly selecting features or aggregating existing features for model training. Additionally, high dimensionality warrants vigilance for overfitting to training data.⁶³ Here, regularization, which regularizes or pressures model coefficients toward zero in order to encourage less complex and flexible models that are less susceptible to overfitting,

have been helpful in mitigating problems that arise with high-dimensional datasets. Common regularization methods that have been used with molecular datasets include ridge, LASSO, or elastic net.

Molecular datasets can also suffer from a low signal-to-noise ratio stemming from difficulties in determining the veracity of detected variants.⁶⁴ Of note, circulating tumor DNA (ctDNA) typically comprises only 5%–10% (in late-stage disease) to less than 0.01%–1.0% (in early-stage disease) of total circulating cell-free DNA.⁶⁴ The balance between wide coverage but low sequencing depth versus high sequencing depth of a more limited target is an important factor that affects the signal-to-noise ratio.⁶⁵ This tradeoff is further amplified when creating molecular datasets for ML applications. Targeted sequencing panels can reduce noise; however, emerging work has demonstrated that aggregation variants across the genome can improve ML performance. Careful design of training datasets for ML applications can help to mitigate some of the noisy data limitations. Case-control designs—e.g., cases comprising patients with localized non-small cell lung cancer matched with controls of risk-matched adults undergoing annual radiologic screening for lung cancer—are a common strategy to reduce confounders and improve signal.⁶¹

While DL has become the model of choice for numerous genomic applications, the unique challenges of liquid and solid tumor biopsy data have rendered DL models less directly applicable. Moreover, inductive biases of popular DL architectures (e.g., spatial invariance of CNNs) are less suitable for sequence variants or gene expression. Rather, smaller models such as regularized logistic regression,⁶¹ SVM,⁶⁶ random forest classifiers,⁶⁷ and elastic nets⁶⁸ are commonly used, and they utilize domain expertise to design features.⁶⁶

Applications of ML models to molecular tumor data

In this section, we review how ML is facilitating the use of molecular data for cancer diagnosis, prognosis, and treatment selection and tumor monitoring (Figure 3).

Cancer diagnosis

Early cancer detection is critical for timely interventions that can improve patient outcomes. Liquid biopsy methods utilize detected variants from a targeted sequencing panel to determine the presence of cancer. While detected mutational burden can be predictive, using mutational burden alone can be limited in sensitivity, specificity, and power.⁶¹ Integrating additional variants and genomic features can increase predictive power. ML models have been instrumental in classifying detected variants as pathological, aggregating variants, and identifying variants that are most predictive.

Models such as logistic regression⁶⁹ and elastic net⁶¹ have been used to integrate detected variants. For example, Lung-CLiP (Cancer Likelihood in Plasma) employs an ensemble ML classifier using nearest neighbor classifiers, naive Bayes, logistic regression, and decision trees to determine the likelihood that a plasma sample contains lung cancer ctDNA.⁶¹ While detecting variant burden from cfDNA is promising, ascertaining the tissue of origin of ctDNA is more challenging.

DNA methylation sequences have also been pursued as molecular predictors for early cancer detection. Changes to CpG

DNA methylation are one of the earliest molecular aberrations in cancer initiation and offer enhanced capability to infer tissue origin of ctDNA due to the presence of tissue-specific CpG islands. A systematic evaluation of 10 ML classifiers with various data inputs (whole-genome sequencing of cfDNA, targeted cfDNA panels, and DNA methylation) using CCGA found that classifiers that utilized whole-genome methylation sequences had the highest cancer detection sensitivity and best prediction of cancer signal origin.⁶²

A central challenge in utilizing methylation sequences is determining which methylation features to select, given that there are 30 million CpG sites that can be methylated or unmethylated. This can be tackled through ML methods that facilitate dimensionality reduction or feature selection. Regularized regression, such as elastic net, has been popular in feature selection for methylation datasets.⁷⁰ Maros et al. systematically compared four ML classifiers (random forest, elastic net, SVM, and boosted trees) in combination with post-processing algorithms and found that elastic net delivers the best performance in methylation-based cancer detection and classification.⁷¹ Grail has utilized probability models, such as Bernoulli mixture models, to determine the ranking of positive and negative methylation features likely to distinguish cancer types from one another or non-cancer.⁷²

While previous liquid biopsy technologies have primarily utilized cfDNA sequences or methylation status, the fragmentation patterns of cfDNA, also called fragmentomics, can provide additional features to enhance ML cancer detection models. Several studies have found that incorporating fragmentomics into their classifier improved classifier performance.^{61,67} Similarly, Jamshidi et al. found that fragment length ML classifiers provided similar sensitivity to a classifier based on genomic alterations.⁶² Improved performance could be attributed to additional epigenetic or mechanistic information conveyed by fragmentomic profiles that can increase predictive capability. For example, Esfahani et al. utilized an elastic net model trained on fragmentomics to infer gene expression, classify non-small cell lung cancer, and assess immunotherapy response.⁶⁸

Cancer prognosis

While liquid biopsies hold the potential to revolutionize cancer diagnostics, solid tumor molecular analysis is currently more mature and can provide high-resolution molecular and clinical information that can be leveraged to better characterize cancer prognosis.

Advances in exome and whole-genome sequencing and bulk and single-cell transcriptomic technology offer exciting opportunities to characterize tumor origin, stage, and grade, which influence cancer prognosis. Determining tumor origin, particularly for metastatic tumors, is an important aspect of cancer prognosis that molecular ML models can facilitate. Random forest classifiers have been a popular model of choice for predicting tumor origin. For example, Nguyen et al. utilized an ensemble of binary random forest classifiers trained on 6,756 whole-genome-sequenced primary and metastatic tumors that discriminated between 35 cancer types with an overall recall of 90%.⁷³ Similarly, Tang et al. developed a random forest classifier trained on methylation and miRNA expression data from 17 classes of solid tumors to predict tumor origin.⁷⁴ For metastatic tumors,

researchers developed random forest models that perform feature selection and tissue-of-origin classification using gene expression and mutation data.⁷⁵ Random forest classifiers are popular due to their ease of interpretability, which provides mechanistic justification of predictions and can facilitate novel biomarker discovery. However, random forest classifiers often require hand-selected features that have relied on patterns of somatic mutations and chromatin state for determining tumor origin. Using a fully connected, feedforward neural network, Jiao et al. determined features correlated with tumor origin and found that passenger mutation regional distribution and mutation type strongly predict tumor origin.⁷⁶

Determining cell-type composition in tumors is critical in assessing cancer prognosis, as it gives insight into the differentiation status, tumor origin, and stage. Several methods have been developed to deconvolve bulk RNA-seq data, a common and cost-effective method to profile solid tumors. Methods such as CIBERSORT use SVMs to deconvolve bulk RNA-seq data to estimate cell-type compositions.⁷⁷ CIBERSORTx and CODEFACS have expanded CIBERSORT to deconvolve bulk RNA using nu-support vector regression (ν -SVR) analysis and achieve cell-type-specific gene expression without single-cell data.^{78,79} While most deconvolution efforts have thus far focused on bulk cellular tissue sources such as tumor specimens, ML deconvolution applications to cell-free nucleic acids are emerging. Indeed, inference of cell types of origin within cell-free RNA (cfRNA) transcriptomes has been achieved using adaptations of CIBERSORTx and ν -SVR,⁸⁰ as well as using Bayesian cell proportion reconstruction inferred using statistical marginalization.⁸¹

In addition to DNA mutations and RNA expression, DNA methylation patterns can also differentiate between different cancer types and subtypes. Capper et al. take advantage of this by designing an ML model that can assign central nervous system tumor (CNS) samples to methylation classes that correspond to tumor types based on genome-wide methylation data.⁸² Their model consists of a random forest to compute raw scores for the methylation classes followed by a multinomial logistic regression model to calibrate those scores as probabilities of each class. In two prospective analyses, they showed that the methylation predictions perform comparable to or better than histopathological analysis in subtyping some tumors. As an alternative to genomic and transcriptomic methods, Klein et al. used mass spectrometry to analyze epithelial ovarian cancer, and they developed SVM and 1D CNN models that analyze the mass spectrum and predict the histotype of the tumor.⁸³

Cancer treatment and tumor monitoring

Selecting cancer treatment, predicting response to treatment, and monitoring tumors after treatment are areas of great promise for ML and genomics. Current treatment selection is determined by clinical guidelines and trials that typically use a handful of clinical features. In contrast, molecular profiles of cancers generate a much larger number of features that can be leveraged to inform cancer treatments. For example, Sammut et al. take a multi-omics approach to predict response to chemotherapy by incorporating clinical, genomic, transcriptomic, pathology, and treatment information into an ensemble model that averages

the predictions of logistic regression, SVM, and random forest models.⁸⁴ Bayesian models such as the continuous individualized risk index (CIRI), which are adept at handling small datasets and quantifying uncertainty, have been used to model ctDNA dynamics after treatment in diverse cancers.⁸⁵ Such approaches can model ctDNA responses associated with outcomes after therapy with immune checkpoint inhibitors for non-small cell lung cancer and predict which patients will achieve durable clinical benefit.⁸⁶ New emerging genomic technologies, such as single-cell transcriptomics and spatial transcriptomics, have the potential to revolutionize histopathology characterization of solid tumors. In particular, single-cell transcriptomics can profile the cell composition, which ML models can leverage to predict cancer treatment response and potential resistance.⁸⁷ Graph neural networks trained on spatial proteomics can model the tumor microenvironment and predict patient response to cancer treatments.^{49,88}

REGULATORY APPROVAL OF CANCER ML ALGORITHMS

The ML algorithms reviewed in the previous sections reflect notable advances in the research landscape. However, before ML algorithms can be deployed on patients, they generally require regulatory approval, which entails more rigorous clinical trials and validation testing than what is presented in published academic work. As such, only a small proportion of ML algorithms end up being deployed on patients. Of those that do, they typically perform well in several predefined tasks like detection and triage settings, and they demonstrate reliability and generalizability across different patient populations.

In the US, most ML algorithms are regulated as medical devices by the Food and Drug Administration (FDA). In the past decade, over 300 AI/ML-enabled medical devices have been approved by the FDA, with over 40% approved since 2020.⁸⁹ As an exception to FDA approval, laboratory-developed tests (LDTs) may alternatively receive Clinical Laboratory Improvement Amendments (CLIA) certification by the Centers for Medicare & Medicaid Services (CMS). Certification of such CLIA LDTs generally applies a lower regulatory standard for approval than the FDA.⁹⁰ While FDA-approved medical devices are approved for use by medical practitioners, CLIA-certified LDTs are approved for use only by the laboratory for which the certification is granted. LDTs have become increasingly complex and often use ML. The FDA has called for stricter regulation and oversight particularly over higher-risk LDTs,⁹⁰ though regulatory changes remain to be implemented. [Figure 4](#) summarizes several examples of regulatory-approved ML medical devices for cancer, including clinical study and ML model details, and [Table 1](#) shows additional examples of approved devices.

The European Union's FDA equivalent, the European Medicines Agency (EMA), operates similarly: cancer-diagnostic AI/ML devices are given a CE mark, which grants approval for sale across the EU and other European countries. However, unlike the US FDA, EMA device approval is decentralized, where individual member countries conduct evaluations, and publicly available information on approvals is sparse. In a comparative analysis of ML devices approved by both the US FDA and


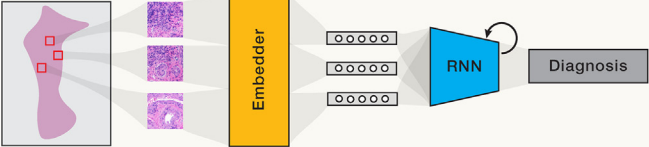
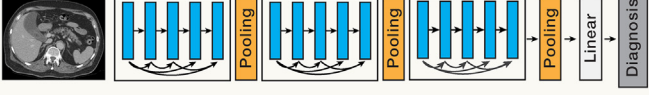
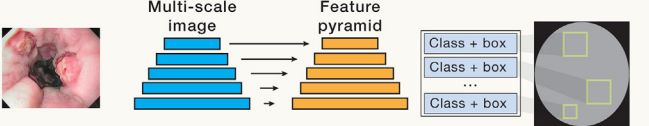
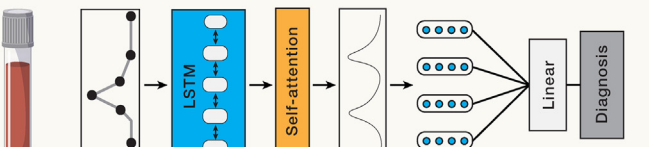
Device description	Clinical study details	Model diagram and description
<p>Transpara</p> <ul style="list-style-type: none"> Breast cancer mammography detection algorithm FDA approved, 2020 	<ul style="list-style-type: none"> AI-assisted and standalone studies <ul style="list-style-type: none"> 18 readers 240 exams 	 <p>Key: ■ Feature processing, ■ Feature aggregation, Linear layers, Prediction</p> <ul style="list-style-type: none"> RetinaNet object detection model Outputs image and lesion scores
<p>Paige Prostate</p> <ul style="list-style-type: none"> Prostate pathology cancer diagnostic algorithm FDA approved, 2019 	<ul style="list-style-type: none"> Standalone analytical testing <ul style="list-style-type: none"> 847 slides AI-assisted study <ul style="list-style-type: none"> 527 slides 16 pathologists 	 <ul style="list-style-type: none"> ResNet-34 CNN feature extractor RNN for score prediction Multiple instance learning
<p>Optellum</p> <ul style="list-style-type: none"> Lung CT cancer nodule detection algorithm FDA approved, 2021 	<ul style="list-style-type: none"> AI-assisted and standalone studies <ul style="list-style-type: none"> 300 subjects 12 readers 	 <ul style="list-style-type: none"> DenseNet CNN classifier
<p>GI Genius</p> <ul style="list-style-type: none"> Lesion detection for endoscopy video FDA approved, 2021 	<ul style="list-style-type: none"> Standalone study <ul style="list-style-type: none"> 150 videos 338 lesions 	 <ul style="list-style-type: none"> RetinaNet object detection model Video frames are individually processed
<p>InterVenn GLORI</p> <ul style="list-style-type: none"> Lab developed test for ovarian cancer diagnosis CLIA certified, 2021 	<ul style="list-style-type: none"> Prospective observational study <ul style="list-style-type: none"> 1,200 participants 	 <ul style="list-style-type: none"> LSTM model for signal processing Regression model for score prediction

Figure 4. Regulatory approval of cancer ML algorithms

Examples of ML medical devices for cancer that have received regulatory approval, including Transpara,⁹¹ Paige Prostate,³⁸ Optellum,⁹² GI Genius,⁹³ and InterVenn GLORI.^{94,95} Clinical study details are based on information available in published works and registered clinical trials. Model details are based on publications by device developers. Image sources: mammography,¹⁹ CT,²⁰ histology,¹¹ endoscopy.⁹⁶

EMA, most devices first received approval in Europe, suggesting a potentially lower regulatory bar compared to the US.¹⁰⁵

Imaging-based algorithms

Imaging-based algorithms comprise over 70% of all FDA-approved AI/ML devices.¹⁰⁶ Of these, radiology applications are the most abundant. Pre-diagnosis algorithms like WRDensity and Densitas use CNN architectures like ResNet¹⁰² to provide breast density category predictions for mammograms. AI-Rad Companion and Quantib Prostate use CNN-based networks¹⁰⁰

like U-Net to perform automated segmentation, density calculation, and volume estimation of the prostate gland. Computer-aided triage devices like Saige-Q¹⁰⁷ and CmTriage use CNN classification algorithms to mark a subset of mammogram cases as suspicious to aid radiologists in workload prioritization. Computer-aided detection/diagnosis devices provide more information by identifying and scoring regions of interest in each image. Examples of breast cancer devices include Lunit Insight, which draws heatmaps (using convolutional layers) with probability percentages over suspicious regions in a mammogram,³⁴ and

Table 1. Additional examples of regulatory-approved cancer diagnostic devices

Approval type (#)	Date approved	Device name	Description	Type of AI/ML	Clinical study details
FDA (P170019)	2017	FoundationOne CDx	Microsatellite instability and tumor mutational burden solid tumor tests	Probit model for level of detection ⁹⁸	Prospective observational studies (1,400 participants)
CLIA certification	2017	Signatera	LDT for ctDNA-based cancer recurrence test	Cox proportional hazards model ⁹⁷	Prospective observational studies (2,000 participants), still recruiting
FDA (K173839)	2017	The Cancer Genetics Tissue of Origin Test	Tissue of origin genetic test	Normalization, classification, and correlation algorithms ⁹⁹	Analytical testing only
FDA (K183271)	2019	AI RAD Companion (Pulmonary)	Lung nodule segmentation	FCOS CNN object detection network ¹⁰⁰	>4,500 cases, standalone study only, reader-annotated ground truth
FDA (K183285)	2019	CmTriage	Breast cancer triage	CNN ¹⁰¹	1,255 exams, standalone study only, biopsy-proven ground truth
FDA (K200595)	2020	CellaVision DC-1	Blood cell counter	CNN	Analytical and clinical testing (598 samples) comparing to predicate device
FDA (K201232)	2020	Limbus Contour	Radiation treatment planning	U-Net CNN ¹⁰³	Benchtop testing only
FDA (K193229)	2020	Transpara	Breast cancer detection	VGG-16 CNN and gradient boosting trees ⁹¹	240 exams, AI-assisted (18 readers) and standalone studies, ground truth unclear
FDA (K202013)	2020	WRDensity	Breast cancer density	Resnet-34 CNN ¹⁰²	871 exams, standalone study only, consensus ground truth
FDA (K211951)	2021	GI Genius	GI lesion detection	CNN object detection network ⁹³	Standalone study only (150 videos with 338 lesions)
CLIA certification	2021	Grail Galleri	Multi-cancer early detection test	Various ML models (logistic/lasso regression, Markov chains, random forest) ¹⁰⁴	Prospective observational and interventional studies (>130K participants)
CLIA certification	2021	InterVenn GLORI	LDT for ovarian cancer diagnosis	Regression models and RNNs ^{94,95}	Prospective observational study (1,200 participants), ground truth by imaging
FDA (DEN200080)	2021	Paige Prostate	Prostate pathology cancer detection	ResNet-34 CNN + RNN (multiple-instance/weakly supervised learning) ⁴¹	Standalone analytical testing on 847 whole slide images (WSIs) and AI-assisted study on 527 WSIs with 16 pathologists, consensus ground truth
FDA (K202300)	2021	Optellum Virtual Nodule Clinic	Lung nodule diagnosis	DenseNet CNN ⁹²	300 subjects, AI-assisted (12 readers) and standalone studies, ground truth unclear

MammoScreen, which uses a RetinaNet CNN architecture to draw a bounding polygon over potential lesions along with the predicted lesion type and risk score out of ten.¹⁰⁸ Another example is Optellum Virtual Nodule Clinic, a lung cancer algorithm for CT images that uses a DenseNet architecture to output malignancy prediction scores for user-selected regions of interest.⁹²

Imaging ML has more recently expanded outside of radiology as well. Paige Prostate is an FDA-approved prostate pathology algorithm, based on the work of Campanella et al.,⁴¹ that uses CNNs and RNNs to diagnose prostate cancer from biopsy slides.¹⁰⁹ Other prostate CLIA-certified pathology ML tests include DeepDx Prostate, which uses semantic segmentation CNNs, and Galen Prostate, which uses multiscale CNNs and

gradient boosting classifiers for automated Gleason scoring.¹¹⁰ GI Genius, an FDA-approved device for polyp detection in endoscopy videos, uses a CNN on individual video frames to produce bounding boxes over suspicious lesions.⁹³

Skin cancer is a promising yet challenging domain. Nevisense, currently the only skin cancer AI device on the market, is a device that works by measuring electrical impedance across a potentially abnormal skin lesion. On the horizon, 3Derm has received FDA breakthrough designation for autonomous detection of skin cancers, which is a fast-track process that signals possible future approval. In the EU, several skin AI devices have already received CE mark approval (TeleSkin and SkinVision), but their efficacy has been questioned by independent validation studies.¹¹¹

Several devices have been approved for post-diagnosis decision making; for instance, Limbus Counter and Ethos are both devices that use segmentation CNNs like U-Net to draw contours of organ structures for radiation treatment planning.¹¹²

Molecular-based algorithms

Most molecular-based algorithms are focused on diagnostic applications in blood samples. FDA-approved cell counting devices like CellaVision and Sight OLO use CNNs to characterize and count white blood cells, red blood cells, and platelets in blood samples and are intended for use by lab technicians.¹¹³ CellSearch uses computer vision algorithms to characterize the morphology of circulating tumor cells in metastatic breast, colorectal, or prostate cancer patients. The Cancer Genetics Tissue of Origin Test is an RNA-based diagnostic algorithm for aiding clinicians in determining the tissue of origin for tumors. Exact Science's Cologuard is a colorectal cancer genomics test that relies on mathematical algorithms to produce risk scores.

Liquid biopsy tests are the most common type of ML-enabled diagnostics performed by CLIA-certified laboratories. LungLife AI's LungLB is a liquid biopsy test that uses a signal-binning algorithm to confirm suspicious lung nodules in CT scans. Galleri is a liquid biopsy test that uses various ML regression and classification models⁷² for early detection of multiple cancers and has received FDA breakthrough designation but not approval. InterVenn has CLIA certification for two products: GLORI is a glycoproteomic liquid biopsy test that utilizes neural networks and logistic regression models for ovarian cancer diagnosis, and DAWN IO is a test that uses tree-based methods and ensemble classifiers for assessing melanoma therapy.⁹⁴ Other genomics tests that are not on the market but are in ongoing large clinical trials include Freenome's Multinomics, a cell-free biomarker patterns blood test using SVM,⁶⁰ and Exact Science's multi-cancer early detection blood test.

Clinical studies evaluating cancer ML algorithms

The types of clinical studies vary depending on the regulatory pathway a device is approved by. For FDA approval, devices must demonstrate evidence of clinical safety and effectiveness for use on patients. Clinical evidence is typically produced via AI-assisted studies and/or standalone studies. AI-assisted studies compare clinicians using AI in diagnostic decision making with those not using AI. In these studies, ground truthing is typically determined by the consensus of several specialists' interpretations. Readers are selected across varying degrees of specialty (generalist versus board certified). Standalone studies provide another form of clinical evidence: the performance of the AI alone is assessed with reference to a reader consensus ground truth, and the metric is compared to the average clinical reader's performance or a standard. In both types of studies, evaluation studies are typically enriched with cancer cases relative to the population incidence rate.

As an example, Transpara, a breast cancer detection algorithm that received FDA approval in 2018, reported clinical evidence from an AI-assisted study and a standalone comparison. Transpara draws regions of interest around suspicious lesions in a mammogram and outputs a score indicating the likelihood of cancer in the image. In the reader study, fourteen board-certified

radiologists read mammograms once with the aid of AI and once without, with a one-month washout period in between. The evaluation dataset consisted of 240 total mammogram studies, with 100 cancer exams, 40 false positive recalls from screening, and 100 normal exams. The primary endpoint was the superiority of performance with AI versus without. Secondary analyses included a superior performance with AI on lesion subtypes and average reading time saved by radiologists. The standalone study compared the AI's performance with the average performance of the fourteen radiologists. In the AI-assisted study, the radiologists' performance improved from 0.866 AUC without AI assistance to 0.886 with AI assistance. In the standalone study, the AI achieved an AUC of 0.887 versus the average clinical reader's performance of 0.866 AUC.

For molecular-based ML device approvals, analytical testing is often conducted in addition to clinical testing. For instance, CellaVision DC-1's FDA evaluation provided evidence demonstrating analytical precision via repeatability (measurements under the same conditions are consistent) and reproducibility (measurements under different conditions are consistent). The clinical testing compared measurements on patient samples with the approved predicate device. Other analytical validation characteristics include accuracy and specificity.

CLIA certifications are less transparent in their evaluation standards compared to the FDA (i.e., no publicly available summaries) but are generally limited to ensuring the analytical validity of lab capabilities. In addition to CLIA certification, most commercially available LDTs have undergone clinical trial validations that are registered with [ClinicalTrials.gov](https://www.clinicaltrials.gov). These studies tend to be prospective and larger in scope than FDA-approved device counterparts, which have a median participant size of 300.⁸⁹ For instance, Grail's Galleri has ongoing clinical trials with over 130,000 participants across multiple settings and countries. Intervenn's GLORI test enrolled 1,200 patients in its clinical trial. Primary endpoints are similar to FDA evaluations and include AUC, sensitivity, specificity, positive predictive value, and negative predictive value.

DISCUSSION

ML is increasingly important in cancer detection, prognosis, and treatment planning. However, the reliability and trust of ML algorithms have lagged behind the pace of technical development. In this section, we discuss some key challenges that ML faces on the path to the clinic, including disparate regulatory standards, stringent criteria for meaningful model evaluation, and barriers to adoption by doctors and hospitals. We then discuss how ML methods differ when applied to various cancer data modalities, and we conclude by highlighting some exciting recent developments in both biomedical and ML technologies that illustrate the potential of ML to transform clinical oncology.

Regulatory standards

Disparate regulatory standards in the US and internationally can lead to under-regulation and mistrust of ML algorithms.¹¹⁴ Within the US, the FDA has historically deferred the regulation of LDTs to CMS. Whereas CMS typically focuses only on analytical validity (i.e., precision, sensitivity, and accuracy of measuring

molecular quantities), the FDA places additional emphasis on clinical validity (whether the test accurately identifies the relevant disease in patients). As LDTs today increasingly provide diagnostic predictions and involve ML-based algorithms, demonstrating that cancer diagnostic tests truly achieve the desired clinical outcomes is necessary for ensuring their trustworthiness and reliability to doctors and patients.

Discrepancies in regulation internationally contribute an additional risk to the trustworthiness of medical ML algorithms. A study of medical devices approved in both the US and EU revealed that devices that gained CE mark approval first in the EU were three times more likely to be recalled due to safety concerns than devices that received US FDA approval first.¹¹⁵ A key difference is that in the US, the FDA requires clinical evaluation prior to approval; in the EU, clinical evaluation is only required after approval as a post-market follow-up study.¹¹⁶ In effect, the CE mark system incentivizes faster adoption of ML into the clinic but at the risk of prematurely approving devices that may pose potential harm to patients.

Limitations of ML model evaluations

The lack of high-quality, diverse evaluations hinders the ability to assess true algorithm performance in patient populations. One factor is the lack of gold-standard test datasets—on-site validations are difficult and patient data are hard to obtain, in part due to privacy concerns and restrictive data use agreements.¹¹⁷ A well-documented phenomenon of ML models is that they can learn spurious correlations present in device types and demographics,⁸⁹ resulting in biased performance when evaluated on different patient populations. Additionally, evaluation test sets are often enriched with positive cases, which can yield imbalanced comparisons.

Metrics

Medical AI studies often use proxy metrics for clinical endpoints, which may generate misleading conclusions. For instance, AUC summarizes model performance across all possible operating points, which is not informative of how an algorithm will perform when deployed at a particular threshold. Algorithms that show an AUC improvement or exceed a certain AUC value (e.g., >0.95 in some FDA-approved devices) may perform differently in real-world populations.¹¹⁸ Fixed-threshold metrics like sensitivity and specificity should reflect the relevant clinical task at hand; for instance, a diagnostic algorithm may be optimized for minimizing missed cancers but should also consider the additional burden to patients caused by false positives (i.e., invasive testing and stress).

Clinical trials and monitoring

Prospective trials are also important to measure appropriate clinical outcomes, rather than a simple comparison to stand-alone references. For example, if an ML device is to be used as a clinical diagnostic aid, then it should be evaluated by comparing clinician performance with and without the device rather than evaluating the device's predictions in isolation.¹¹⁹ Randomizing patient cohorts can minimize biases in selecting test populations. Also, prospective trials can capture human-AI interactions that occur after deployment.¹²⁰ Continuous performance monitoring of ML algorithms after approval and post-market surveillance mechanisms are necessary to ensure that

the purported clinical benefits of ML hold up under various distribution shifts.¹²¹ As a case study, earlier-generation computer-aided detection software for mammography was approved by the FDA in 1998 and widely adopted in part because of Medicare and Medicaid reimbursements. However, a large observational study by Lehman et al. on mammograms from 2003 to 2009 found that CAD software had failed to improve the diagnostic accuracy of mammography.¹²² This was due in part to changes in radiologists' behavior, with increased familiarity with the ML over time.¹²³ Moreover, the original evaluation data included older traditional film mammograms, which have since been phased out. As such, reproducibility and transparency are essential for building trust in the outcomes of validation studies.³²

Interpreting ML models

Interpretability is a common challenge for ML. One important reason is that most models do not explicitly identify causal features but instead rely on correlating input features with outcomes. As such, models may accurately identify phenotypes but rely on spurious confounders present in the data and present misleading conclusions.¹²⁴ Nonetheless, interpretability methods can still be useful for explaining how an ML model makes its predictions, which is important for building trust with clinicians and providing additional diagnostic insight beyond the prediction alone.¹²⁵ Interpretability methods can either be applied *post hoc* to extract explanations from trained models, or they can be incorporated into the model design so that the model learns to simultaneously produce explanations and predictions. Examples of *post hoc* interpretability techniques include using the ML model to generate heatmaps over the input³³ and clustering the inputs into interpretable groups based on the ML model's embedding of the input.¹²⁶ As an example of a model with explainability built into its design, Zhang et al. created an ML model that learns to generate explanations in natural language for its predictions during training.¹²⁷ *Post hoc* methods are convenient because they can be applied to most models without requiring specialized training, but models with interpretability built in may provide more reliable explanations for what the model is doing.¹²⁵ Models that output a probability or range of scores (e.g., from 1 to 10) should be carefully designed and calibrated to user expectations.¹²⁸

Challenges to adoption

While most academic research has been focused on improvements in the diagnostic accuracy of ML algorithms, many of the driving factors for real-world clinical ML adoption fall outside of solely technical progress. Interoperability and integration with existing electronic health records and image storage systems is a significant barrier to adoption by hospital systems.¹²⁹ Clinicians may not trust or understand ML algorithm decisions and outputs. Developers must effectively communicate the economic value of their ML algorithms to hospital decision makers and overcome organizational inertia. Finally, patients and clinicians should also understand the benefits and risks of using ML in decision making.¹³⁰

Different data modalities require different ML techniques

Imaging and molecular data are the two most common data modalities in cancer diagnostics. However, in practice, they require

very different ML approaches due to fundamental differences in the problems each data type presents. Imaging-based tasks typically involve a needle-in-the-haystack problem, where small features associated with cancer are present in a large image space. CNNs are highly effective and have become ubiquitous because they are able to efficiently learn from large amounts of available data, and they can extract spatially distinct hierarchies of features present in an image.

Molecular data, on the other hand, tend to be highly structured and have features that correspond to distinct biological measurements (i.e., DNA sequences). A primary hurdle in analysis is the high dimensionality of biological features and the inherent sparsity present in the data. Here, ML regularization techniques like LASSO regression are used, as well as dimensionality reduction techniques like PCA for selecting salient biomarkers. Finally, statistical ML models like logistic regression and decision trees are used to pick optimal thresholds and minimal levels of detection that correspond to a clinically meaningful presence of disease.

Future developments

New biomedical and ML technologies are rapidly emerging that will change the way ML is applied to cancer diagnostics and may significantly improve the predictive power and clinical usefulness of these models.

Biomedical data

Biomedical advances are enabling physicians to obtain increasingly detailed medical data about patients. In pathology, new multiplexed proteomics technologies like CODEX¹³¹ allow staining for 40–100 proteins simultaneously, providing a much more detailed view of the cellular and proteomic composition of tissues than traditional staining techniques like H&E staining and immunofluorescence. Similarly, spatial transcriptomics¹³² provides a view of the spatial distribution of RNA transcripts across a pathology sample, thereby incorporating another form of omics data into images. Sequencing data from the tumor microbiome might serve as a diagnostic tool for oncology as scientists learn more about the role of bacteria in cancer.¹³³ Data from the immune system, such as T cell receptor sequences, can also provide diagnostic clues for cancer based on the body's response to tumors.¹³⁴ ML methods that use these new sources of data may be able to make more accurate and specific predictions.

Integrating imaging and omics

Imaging and molecular data often provide complementary information about a patient's cancer, so integrating these two data sources can improve ML predictions for diagnostics, prognosis, and treatment. One method of combining the two is through biomedical technologies such as CODEX and spatial transcriptomics, which overlay spatially resolved proteomics and transcriptomics data on images, allowing models to process omics data in image form.^{49,88,135} Another promising direction is the development of multimodal models, which fuse multiple ML models to combine information across several data types (images, genomics, clinical records, etc.) to make better predictions.² Multimodal models can have a more holistic view of each patient and can combine multiple weak signals into a strong signal that can better inform the patient's diagnosis or optimal

treatment. For example, Vanguri et al. predict response to PD-(L)1 blockade in patients with non-small cell lung cancer using a multimodal model that combines medical imaging, histopathologic, and genomic features and outperforms unimodal models.¹³⁶ Although there are many challenges to developing multimodal models, such as linking data across modalities and handling patients with incomplete data, these models may prove to be very powerful because they can reason across multiple sources of information, just as physicians do.

ML methodology

New ML models have emerged that improve upon the standard deep learning architectures, such as CNNs, that are commonly used in cancer diagnostics. Several such models have demonstrated clear improvements in predictive accuracy. One of the best examples is the transformer,¹³⁷ which was originally designed for natural language processing. Transformers have since been modified and applied to pathology images.¹³⁸ Another trend is to re-envision image-based data as a graph and apply GNNs. For example, Wu et al. convert images of tissue samples into graphs of cells, where each cell is a node in the graph and neighboring cells have edges connecting them.⁸⁸ GNNs applied to these graphs can make diagnostic and prognostic predictions that may be more robust against visual artifacts and more sensitive to the interconnections between cells than image-based predictions. Instead of using new ML models, another option is to improve the performance of existing ML models by performing data augmentation with generative ML models, which learn to synthesize new data that look similar to the real training data.¹³⁹ Generative models are also useful for translating between data formats such as generating text reports from medical images.¹⁴⁰

The technological advancements discussed in this Review illustrate the exciting potential of ML to leverage the latest biomedical data to transform the field of clinical oncology. As ML methods are further improved and carefully validated with appropriate monitoring and regulatory oversight, they may soon see wide-scale clinical adoption to improve cancer care for patients.

ACKNOWLEDGMENTS

K.S. is supported by a Knight-Hennessy Scholarship. A.Z. is supported by the National Institutes of Health grant F30HL156478. E.W. is supported by a Stanford Bio-X SIGF Fellowship. J.Z. is supported by a Chan-Zuckerberg Investigator Award.

DECLARATION OF INTERESTS

A.A.A. is an advisor to Celgene, Chugai, Genentech, Gilead, Janssen, Pharmacyclics, and Roche. E.W. is a shareholder of RadNet, Inc. J.Z. is an advisor to Adela, Enable Medicine, and InterVenn Biosciences.

REFERENCES

1. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., and Fotiadis, D.I. (2015). Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **13**, 8–17.
2. Boehm, K.M., Khosravi, P., Vanguri, R., Gao, J., and Shah, S.P. (2022). Harnessing multimodal data integration to advance precision oncology. *Nat. Rev. Cancer* **22**, 114–126.

3. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H., and Aerts, H.J.W.L. (2018). Artificial intelligence in radiology. *Nat. Rev. Cancer* 18, 500–510.
4. Bera, K., Schalper, K.A., Rimm, D.L., Velcheti, V., and Madabhushi, A. (2019). Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* 16, 703–715.
5. McIntosh, C., Conroy, L., Tjong, M.C., Craig, T., Bayley, A., Catton, C., Gospodarowicz, M., Helou, J., Isfahanian, N., Kong, V., et al. (2021). Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. *Nat. Med.* 27, 999–1005.
6. Bera, K., Braman, N., Gupta, A., Velcheti, V., and Madabhushi, A. (2022). Predicting cancer outcomes with radiomics and artificial intelligence in radiology. *Nat. Rev. Clin. Oncol.* 19, 132–146.
7. Shmatko, A., Ghaffari Laleh, N., Gerstung, M., and Kather, J.N. (2022). Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nat. Cancer* 3, 1026–1038.
8. Heitzer, E., Haque, I.S., Roberts, C.E.S., and Speicher, M.R. (2019). Current and future perspectives of liquid biopsies in genomics-driven oncology. *Nat. Rev. Genet.* 20, 71–88.
9. Esposito, M., Ganesan, S., and Kang, Y. (2021). Emerging strategies for treating metastasis. *Nat. Cancer* 2, 258–270.
10. Kwong, G.A., Ghosh, S., Gamboa, L., Patriotis, C., Srivastava, S., and Bhatia, S.N. (2021). Synthetic biomarkers: a twenty-first century path to early cancer detection. *Nat. Rev. Cancer* 21, 655–668.
11. Häggström M. Histology of postmenopausal myometrium, low magnification [Internet]. Wikimedia Commons. Available from: https://commons.wikimedia.org/wiki/File:Histology_of_postmenopausal_myometrium_low_magnification.jpg
12. Levine, A.B., Schlosser, C., Grewal, J., Coope, R., Jones, S.J.M., and Yip, S. (2019). Rise of the machines: advances in deep learning for cancer diagnosis. *Trends Cancer* 5, 157–169.
13. Lu, M.T., Raghu, V.K., Mayrhofer, T., Aerts, H.J., and Hoffmann, U. (2020). Deep learning using chest radiographs to identify high-risk smokers for lung cancer screening computed tomography: development and validation of a prediction model. *Ann. Intern. Med.* 173, 704–713.
14. Varghese, B., Chen, F., Hwang, D., Palmer, S.L., De Castro Abreu, A.L., Ukimura, O., Aron, M., Aron, M., Gill, I., Duddalwar, V., and Pandey, G. (2020). Objective risk stratification of prostate cancer using machine learning and radiomics applied to multiparametric magnetic resonance images. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. New York, NY, USA (Association for Computing Machinery), pp. 1–10. (BCB '20).
15. Lu, M.Y., Chen, T.Y., Williamson, D.F.K., Zhao, M., Shady, M., Lipkova, J., and Mahmood, F. (2021). AI-based pathology predicts origins for cancers of unknown primary. *Nature* 594, 106–110.
16. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115–118.
17. Yamada, M., Saito, Y., Imaoka, H., Saiko, M., Yamada, S., Kondo, H., Takamaru, H., Sakamoto, T., Sese, J., Kuchiba, A., et al. (2019). Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy. *Sci. Rep.* 9, 14465.
18. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., and Tsirigos, A. (2018 Oct). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* 24, 1559–1567.
19. Mammogram - Normal [Internet]. National Cancer Institute Visuals Online. Available from: <https://visualsonline.cancer.gov/details.cfm?imageid=9405>
20. Häggström M. CT of cholecystitis [Internet]. Wikimedia Commons. Available from: https://commons.wikimedia.org/wiki/File:CT_of_cholecystitis.jpg
21. Lehman, C.D., Yala, A., Schuster, T., Dontchos, B., Bahl, M., Swanson, K., and Barzilay, R. (2019). Mammographic breast density assessment using deep learning: clinical implementation. *Radiology* 290, 52–58.
22. Dontchos, B.N., Yala, A., Barzilay, R., Xiang, J., and Lehman, C.D. (2021 Apr). External validation of a deep learning model for predicting mammographic breast density in routine clinical practice. *Acad. Radiol.* 28, 475–480.
23. Arefan, D., Mohamed, A.A., Berg, W.A., Zuley, M.L., Sumkin, J.H., and Wu, S. (2020 Jan). Deep learning modeling using normal mammograms for predicting breast cancer risk. *Med. Phys.* 47, 110–118.
24. Dembrower, K., Liu, Y., Azizpour, H., Eklund, M., Smith, K., Lindholm, P., and Strand, F. (2020 Feb). Comparison of a deep learning risk score and standard mammographic density score for breast cancer risk prediction. *Radiology* 294, 265–272.
25. Yala, A., Mikhael, P.G., Strand, F., Lin, G., Smith, K., Wan, Y.L., Lamb, L., Hughes, K., Lehman, C., and Barzilay, R. (2021). Toward robust mammography-based models for breast cancer risk. *Sci. Transl. Med.* 13, eaba4373.
26. Yala, A., Mikhael, P.G., Strand, F., Lin, G., Satuluru, S., Kim, T., Banerjee, I., Gichoya, J., Trivedi, H., Lehman, C.D., et al. (2022). Multi-institutional validation of a mammography-based breast cancer risk model. *J. Clin. Oncol.* 40, 1732–1740.
27. Ha, R., Chang, P., Karcich, J., Mutasa, S., Pascual Van Sant, E., Liu, M.Z., and Jambawalikar, S. (2019). Convolutional neural network based breast cancer risk stratification using a mammographic dataset. *Acad. Radiol.* 26, 544–549.
28. Yala, A., Mikhael, P.G., Lehman, C., Lin, G., Strand, F., Wan, Y.L., Hughes, K., Satuluru, S., Kim, T., Banerjee, I., et al. (2022). Optimizing risk-based breast cancer screening policies with reinforcement learning. *Nat. Med.* 28, 136–143.
29. Dai, X., Spasić, I., Meyer, B., Chapman, S., and Andres, F. (2019). Machine learning on mobile: an on-device inference app for skin cancer detection. In *2019 Fourth International Conference on Fog and Mobile Edge Computing (FMEC)*, pp. 301–305.
30. 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.
31. McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafiyan, 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.
32. Haibe-Kains, B., Adam, G.A., Hosny, A., Khodakarami, F., Massive Analysis Quality Control MAQC Society Board of Directors, Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., et al. (2020). Transparency and reproducibility in artificial intelligence. *Nature* 586, E14–E16.
33. Qian, X., Pei, J., Zheng, H., Xie, X., Yan, L., Zhang, H., Han, C., Gao, X., Zhang, H., Zheng, W., et al. (2021). Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. *Nat. Biomed. Eng.* 5, 522–532.
34. Kim, H.E., Kim, H.H., Han, B.K., Kim, K.H., Han, K., Nam, H., Lee, E.H., and Kim, E.K. (2020). Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multi-reader study. *Lancet. Digit. Health* 2, e138–e148.
35. Hekler, A., Utikal, J.S., Enk, A.H., Hauschild, A., Weichenthal, M., Maron, R.C., Berking, C., Haferkamp, S., Klode, J., Schadendorf, D., et al. (2019). Superior skin cancer classification by the combination of human and artificial intelligence. *Eur. J. Cancer* 120, 114–121.
36. Yala, A., Schuster, T., Miles, R., Barzilay, R., and Lehman, C. (2019). A deep learning model to triage screening mammograms: a simulation study. *Radiology* 293, 38–46.

37. Xu, Y., Wang, Y., Yuan, J., Cheng, Q., Wang, X., and Carson, P.L. (2019). Medical breast ultrasound image segmentation by machine learning. *Ultrasonics* *97*, 1–9.
38. Cao, R., Mohammadian Bajgiran, A., Afshari Mirak, S., Shakeri, S., Zhong, X., Enzmann, D., Raman, S., and Sung, K. (2019). Joint prostate cancer detection and gleason score prediction in mp-MRI via FocalNet. *IEEE Trans. Med. Imaging* *38*, 2496–2506.
39. Akselrod-Ballin, A., Chorev, M., Shoshan, Y., Spiro, A., Hazan, A., Melamed, R., Barkan, E., Herzel, E., Naor, S., Karavani, E., et al. (2019). Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology* *292*, 331–342.
40. Wang, D., Khosla, A., Gargeya, R., Irshad, H., and Beck, A.H. (2016). Deep Learning for Identifying Metastatic Breast Cancer. Preprint at arXiv. <http://arxiv.org/abs/1606.05718>.
41. Campanella, G., Hanna, M.G., Geneslaw, L., Mirafior, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., and Fuchs, T.J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* *25*, 1301–1309.
42. Song, Z., Zou, S., Zhou, W., Huang, Y., Shao, L., Yuan, J., Gou, X., Jin, W., Wang, Z., Chen, X., et al. (2020). Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. *Nat. Commun.* *11*, 4294.
43. Steiner, D.F., MacDonald, R., Liu, Y., Truszkowski, P., Hipp, J.D., Gammage, C., Thng, F., Peng, L., and Stumpe, M.C. (2018). Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *Am. J. Surg. Pathol.* *42*, 1636–1646.
44. Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van der Laak, J., Hulsbergen-van de Kaa, C., and Litjens, G. (2020). Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol.* *21*, 233–241.
45. Esteva, A., Feng, J., van der Wal, D., Huang, S.C., Simko, J.P., DeVries, S., Chen, E., Schaeffer, E.M., Morgan, T.M., Sun, Y., et al. (2022). Prostate cancer therapy personalization via multi-modal deep learning on randomized phase III clinical trials. *NPJ Digit. Med.* *5*, 71.
46. Kather, J.N., Pearson, A.T., Halama, N., Jäger, D., Krause, J., Loosen, S.H., Marx, A., Boor, P., Tacke, F., Neumann, U.P., et al. (2019). Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* *25*, 1054–1056.
47. Jain, M.S., and Massoud, T.F. (2020). Predicting tumour mutational burden from histopathological images using multiscale deep learning. *Nat. Mach. Intell.* *2*, 356–362.
48. Fu, Y., Jung, A.W., Torne, R.V., Gonzalez, S., Vöhringer, H., Shmatko, A., Yates, L.R., Jimenez-Linan, M., Moore, L., and Gerstung, M. (2020). Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat. Cancer* *1*, 800–810.
49. He, B., Bergensträhle, L., Stenbeck, L., Abid, A., Andersson, A., Borg, Å., Maaskola, J., Lundeberg, J., and Zou, J. (2020). Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat. Biomed. Eng.* *4*, 827–834.
50. Saltz, J., Gupta, R., Hou, L., Kurc, T., Singh, P., Nguyen, V., Samaras, D., Shroyer, K.R., Zhao, T., Batiste, R., et al. (2018). Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep.* *23*, 181–193.e7.
51. Wang, S., Shi, J., Ye, Z., Dong, D., Yu, D., Zhou, M., Liu, Y., Gevaert, O., Wang, K., Zhu, Y., et al. (2019). Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning. *Eur. Respir. J.* *53*, 1800986.
52. Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., Manceron, P., Toldo, S., Zaslavskiy, M., Le Stang, N., et al. (2019). Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* *25*, 1519–1525.
53. Bychkov, D., Linder, N., Turkki, R., Nordling, S., Kovanen, P.E., Verrill, C., Walliander, M., Lundin, M., Haglund, C., and Lundin, J. (2018). Deep learning based tissue analysis predicts outcome in colorectal cancer. *Sci. Rep.* *8*, 3395.
54. 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.
55. Joo, S., Ko, E.S., Kwon, S., Jeon, E., Jung, H., Kim, J.Y., Chung, M.J., and Im, Y.H. (2021). Multimodal deep learning models for the prediction of pathologic response to neoadjuvant chemotherapy in breast cancer. *Sci. Rep.* *11*, 18800.
56. Gu, J., Tong, T., He, C., Xu, M., Yang, X., Tian, J., Jiang, T., and Wang, K. (2022). Deep learning radiomics of ultrasonography can predict response to neoadjuvant chemotherapy in breast cancer at an early stage of treatment: a prospective study. *Eur. Radiol.* *32*, 2099–2109.
57. Tian, P., He, B., Mu, W., Liu, K., Liu, L., Zeng, H., Liu, Y., Jiang, L., Zhou, P., Huang, Z., et al. (2021). Assessing PD-L1 expression in non-small cell lung cancer and predicting responses to immune checkpoint inhibitors using deep learning on computed tomography images. *Theranostics* *11*, 2098–2107.
58. Lu, L., Dercle, L., Zhao, B., and Schwartz, L.H. (2021). Deep learning for the prediction of early on-treatment response in metastatic colorectal cancer from serial medical imaging. *Nat. Commun.* *12*, 6654.
59. Hosny, A., Bitterman, D.S., Guthrie, C.V., Qian, J.M., Roberts, H., Perni, S., Saraf, A., Peng, L.C., Pashtan, I., Ye, Z., et al. (2022). Clinical validation of deep learning algorithms for radiotherapy targeting of non-small-cell lung cancer: an observational study. *Lancet. Digit. Health* *4*, e657–e666.
60. Wan, N., Weinberg, D., Liu, T.Y., Niehaus, K., Ariazi, E.A., Delubac, D., Kannan, A., White, B., Bailey, M., Bertin, M., et al. (2019). Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA. *BMC Cancer* *19*, 832.
61. Chabon, J.J., Hamilton, E.G., Kurtz, D.M., Esfahani, M.S., Moding, E.J., Stehr, H., Schroers-Martin, J., Nabet, B.Y., Chen, B., Chaudhuri, A.A., et al. (2020). Integrating genomic features for non-invasive early lung cancer detection. *Nature* *580*, 245–251.
62. Jamshidi, A., Liu, M.C., Klein, E.A., Venn, O., Hubbell, E., Beausang, J.F., Gross, S., Melton, C., Fields, A.P., Liu, Q., et al. (2022). Evaluation of cell-free DNA approaches for multi-cancer early detection. *Cancer Cell* *40*, 1537–1549.e12.
63. Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., and Telenti, A. (2019). A primer on deep learning in genomics. *Nat. Genet.* *51*, 12–18.
64. Zviran, A., Schulman, R.C., Shah, M., Hill, S.T.K., Deochand, S., Khamnei, C.C., Maloney, D., Patel, K., Liao, W., Widman, A.J., et al. (2020). Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. *Nat. Med.* *26*, 1114–1124.
65. Xiao, W., Ren, L., Chen, Z., Fang, L.T., Zhao, Y., Lack, J., Guan, M., Zhu, B., Jaeger, E., Kerrigan, L., et al. (2021). Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. *Nat. Biotechnol.* *39*, 1141–1150.
66. Peneder, P., Stütz, A.M., Surdez, D., Krumbholz, M., Semper, S., Chircard, M., Sheffield, N.C., Pierron, G., Lapouble, E., Tötzl, M., et al. (2021). Multimodal analysis of cell-free DNA whole-genome sequencing for pediatric cancers with low mutational burden. *Nat. Commun.* *12*, 3230.
67. Mouliere, F., Chandrananda, D., Piskorz, A.M., Moore, E.K., Morris, J., Ahlborn, L.B., Mair, R., Goranova, T., Marass, F., Heider, K., et al. (2018). Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci. Transl. Med.* *10*, eaat4921.
68. Esfahani, M.S., Hamilton, E.G., Mehrmohamadi, M., Nabet, B.Y., Alig, S.K., King, D.A., Steen, C.B., Macaulay, C.W., Schultz, A., Nesselbush,

- M.C., et al. (2022). Inferring gene expression from cell-free DNA fragmentation profiles. *Nat. Biotechnol.* *40*, 585–597.
69. Cohen, J.D., Li, L., Wang, Y., Thoburn, C., Afsari, B., Danilova, L., Douville, C., Javed, A.A., Wong, F., Mattox, A., et al. (2018). Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* *359*, 926–930.
70. Yousefi, P.D., Suderman, M., Langdon, R., Whitehurst, O., Davey Smith, G., and Relton, C.L. (2022). DNA methylation-based predictors of health: applications and statistical considerations. *Nat. Rev. Genet.* *23*, 369–383.
71. Maros, M.E., Capper, D., Jones, D.T.W., Hovestadt, V., von Deimling, A., Pfister, S.M., Benner, A., Zucknick, M., and Sill, M. (2020). Machine learning workflows to estimate class probabilities for precision cancer diagnostics on DNA methylation microarray data. *Nat. Protoc.* *15*, 479–512.
72. Liu, M.C., Oxnard, G.R., Klein, E.A., Swanton, C., and Seiden, M.V.; CCGA Consortium (2020). Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann. Oncol.* *31*, 745–759.
73. Nguyen, L., Van Hoeck, A., and Cuppen, E. (2022). Machine learning-based tissue of origin classification for cancer of unknown primary diagnostics using genome-wide mutation features. *Nat. Commun.* *13*, 4013.
74. Tang, W., Wan, S., Yang, Z., Teschendorff, A.E., and Zou, Q. (2018). Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics* *34*, 398–406.
75. He, B., Lang, J., Wang, B., Liu, X., Lu, Q., He, J., Gao, W., Bing, P., Tian, G., and Yang, J. (2020). TOOme: a novel computational framework to infer cancer tissue-of-origin by integrating both gene mutation and expression. *Front. Bioeng. Biotechnol.* *8*, 394.
76. Jiao, W., Atwal, G., Polak, P., Karlic, R., Cuppen, E., PCAWG Tumor Subtypes and Clinical Translation Working Group, Danyi, A., De Ridder, J., van Herpen, C., Lolkema, M.P., and Steeghs, N. (2020). A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nat. Commun.* *11*, 728.
77. Chen, B., Khodadoust, M.S., Liu, C.L., Newman, A.M., and Alizadeh, A.A. (2018). Profiling Tumor Infiltrating Immune Cells with CIBERSORT. *Methods Mol. Biol.* *1711*, 243–259.
78. Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* *37*, 773–782.
79. Wang, K., Patkar, S., Lee, J.S., Gertz, E.M., Robinson, W., Schischlik, F., Crawford, D.R., Schäffer, A.A., and Ruppin, E. (2022). Deconvolving clinically relevant cellular immune cross-talk from bulk gene expression using CODEFACS and LIRICS stratifies patients with melanoma to anti-PD-1 therapy. *Cancer Discov.* *12*, 1088–1105.
80. Vorperian, S.K., Moufarrej, M.N., and Tabula Sapiens Consortium, and Quake, S.R. (2022). Cell types of origin of the cell-free transcriptome. *Nat. Biotechnol.* *40*, 855–861.
81. Chu, T., Wang, Z., Pe'er, D., and Danko, C.G. (2022 Apr). Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat. Cancer* *3*, 505–517.
82. Capper, D., Jones, D.T.W., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., Koelsche, C., Sahm, F., Chavez, L., Reuss, D.E., et al. (2018). DNA methylation-based classification of central nervous system tumours. *Nature* *555*, 469–474.
83. Klein, O., Kanter, F., Kulbe, H., Jank, P., Denkert, C., Nebrich, G., Schmitt, W.D., Wu, Z., Kunze, C.A., Sehoul, J., et al. (2019). MALDI-imaging for classification of epithelial ovarian cancer histotypes from a tissue microarray using machine learning methods. *Proteomics. Clin. Appl.* *13*, e1700181.
84. Sammut, S.J., Crispin-Ortuzar, M., Chin, S.F., Provenzano, E., Bardwell, H.A., Ma, W., Cope, W., Dariush, A., Dawson, S.J., Abraham, J.E., et al. (2022). Multi-omic machine learning predictor of breast cancer therapy response. *Nature* *601*, 623–629.
85. Kurtz, D.M., Esfahani, M.S., Scherer, F., Soo, J., Jin, M.C., Liu, C.L., Newman, A.M., Dührsen, U., Hüttmann, A., Casasnovas, O., et al. (2019). Dynamic risk profiling using serial tumor biomarkers for personalized outcome prediction. *Cell* *178*, 699–713.e19.
86. Nabet, B.Y., Esfahani, M.S., Moding, E.J., Hamilton, E.G., Chabon, J.J., Rizvi, H., Steen, C.B., Chaudhuri, A.A., Liu, C.L., Hui, A.B., et al. (2020). Noninvasive early identification of therapeutic benefit from immune checkpoint inhibition. *Cell* *183*, 363–376.e13.
87. Wu, Z., Lawrence, P.J., Ma, A., Zhu, J., Xu, D., and Ma, Q. (2020). Single-cell techniques and deep learning in predicting drug response. *Trends Pharmacol. Sci.* *41*, 1050–1065.
88. Wu, Z., Trevino, A.E., Wu, E., Swanson, K., Kim, H.J., D'Angio, H.B., Preska, R., Charville, G.W., Dalerba, P.D., Egloff, A.M., et al. (2022). Graph deep learning for the characterization of tumour microenvironments from spatial protein profiles in tissue specimens. *Nat. Biomed. Eng.* *6*, 1435–1448.
89. Wu, E., Wu, K., Daneshjou, R., Ouyang, D., Ho, D.E., and Zou, J. (2021). How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat. Med.* *27*, 582–584.
90. Food and Drug Administration (2017). Discussion Paper on Laboratory Developed Tests (LDTs). <https://www.fda.gov/media/102367/download>.
91. Rodríguez-Ruiz, A., Krupinski, E., Mordang, J.J., Schilling, K., Heywang-Köbrunner, S.H., Sechopoulos, I., and Mann, R.M. (2019 Feb). Detection of breast cancer with mammography: effect of an artificial intelligence support system. *Radiology* *290*, 305–314.
92. Baldwin, D.R., Gustafson, J., Pickup, L., Arteta, C., Novotny, P., Declercq, J., Kadir, T., Figueiras, C., Sterba, A., Exell, A., et al. (2020). External validation of a convolutional neural network artificial intelligence tool to predict malignancy in pulmonary nodules. *Thorax* *75*, 306–312.
93. Repici, A., Badalamenti, M., Maselli, R., Correale, L., Radaelli, F., Rondonotti, E., Ferrara, E., Spadaccini, M., Alkandari, A., Fugazza, A., et al. (2020). Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. *Gastroenterology* *159*, 512–520.e7.
94. Lindpaintner, K., Mitchell, A., Pickering, C., Xu, G., Vigal, K., Axenfeld, B., Rice, R., Cong, X., Frederick, D.T., Michaud, W., et al. (2022). Glycoproteomics as a powerful liquid biopsy-based predictor of checkpoint inhibitor treatment benefit in metastatic malignant melanoma. *J. Clin. Orthop.* *40*, 9545.
95. Wu, Z., Serie, D., Xu, G., and Zou, J. (2020). PB-Net: Automatic peak integration by sequential deep learning for multiple reaction monitoring. *J. Proteomics* *223*, 103820.
96. Samir. Esophageal varices - post banding [Internet]. Wikimedia Commons. Available from: https://commons.wikimedia.org/wiki/File:Esophageal_varices_-_post_banding.jpg
97. Henriksen, T.V., Tarazona, N., Frydendahl, A., Reinert, T., Gimeno-Vallente, F., Carbonell-Asins, J.A., Sharma, S., Renner, D., Hafez, D., Roda, D., et al. (2022). Circulating tumor DNA in stage III colorectal cancer, beyond minimal residual disease detection, toward assessment of adjuvant therapy efficacy and clinical behavior of recurrences. *Clin. Cancer Res.* *28*, 507–517.
98. Milbury, C.A., Creeden, J., Yip, W.K., Smith, D.L., Pattani, V., Maxwell, K., Sawchyn, B., Gjoerup, O., Meng, W., Skoletsky, J., et al. (2022). Clinical and analytical validation of FoundationOne®CDx, a comprehensive genomic profiling assay for solid tumors. *PLoS One* *17*, e0264138.
99. Dumur, C.I., Lyons-Weiler, M., Sciulli, C., Garrett, C.T., Schrijver, I., Holley, T.K., Rodriguez-Paris, J., Pollack, J.R., Zehnder, J.L., Price, M., et al. (2008). Interlaboratory performance of a microarray-based gene

- expression test to determine tissue of origin in poorly differentiated and undifferentiated cancers. *J. Mol. Diagn.* *10*, 67–77.
100. Homayounieh, F., Digumarthy, S., Ebrahimi, S., Rueckel, J., Hoppe, B.F., Sabel, B.O., Conjeti, S., Ridder, K., Siermanns, M., Wang, L., et al. (2021). An artificial intelligence-based chest X-ray model on human nodule detection accuracy from a multicenter study. *JAMA Netw. Open* *4*, e2141096.
 101. Retson, T.A., Lim, V., and Watanabe, A.T. (2022). High performance of FDA-cleared platform for mammography triage. https://f.hubspotusercontent20.net/hubfs/5209275/NCBC%202021%20cmTriage%20Poster%203_12_21_Final.pdf.
 102. Matthews, T.P., Singh, S., Mombourquette, B., Su, J., Shah, M.P., Pedemonte, S., Long, A., Maffit, D., Gurney, J., Hoil, R.M., et al. (2021). A multi-site study of a breast density deep learning model for full-field digital mammography and synthetic mammography. *Radiol. Artif. Intell.* *3*, e200015.
 103. Wong, J., Fong, A., McVicar, N., Smith, S., Giambattista, J., Wells, D., Kolbeck, C., Giambattista, J., Gondara, L., and Alexander, A. (2020). Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. *Radiother. Oncol.* *144*, 152–158.
 104. Liu, M.C., Oxnard, G.R., Klein, E.A., Swanton, C., and Seiden, M.V.; CCGA Consortium (2020). Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann. Oncol.* *31*, 745–759.
 105. Muehlethaler, U.J., Daniore, P., and Vokinger, K.N. (2021). Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *Lancet. Digit. Health* *3*, e195–e203.
 106. Center for Devices, Radiological Health (2022). Artificial Intelligence and Machine Learning Program: Research on AI/ML-Based Medical Devices (U.S. Food and Drug Administration. FDA). <https://www.fda.gov/medical-devices/medical-device-regulatory-science-research-programs-conducted-osel/artificial-intelligence-and-machine-learning-program-research-aiml-based-medical-devices>.
 107. Lotter, W., Sorensen, G., and Cox, D. (2017). A multi-scale CNN and curriculum learning strategy for mammogram classification. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Springer International Publishing), pp. 169–177.
 108. Pacilè, S., Lopez, J., Chone, P., Bertinotti, T., Grouin, J.M., and Fillard, P. (2020). Improving breast cancer detection accuracy of mammography with the concurrent use of an artificial intelligence tool. *Radiol. Artif. Intell.* *2*, e190208.
 109. Raciti, P., Sue, J., Ceballos, R., Godrich, R., Kunz, J.D., Kapur, S., Reuter, V., Grady, L., Kanan, C., Klimstra, D.S., and Fuchs, T.J. (2020). Novel artificial intelligence system increases the detection of prostate cancer in whole slide images of core needle biopsies. *Mod. Pathol.* *33*, 2058–2066.
 110. Pantanowitz, L., Quiroga-Garza, G.M., Bien, L., Heled, R., Laifenfeld, D., Linhart, C., Sandbank, J., Albrecht Shach, A., Shalev, V., Vecsler, M., et al. (2020). An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. *Lancet. Digit. Health* *2*, e407–e416.
 111. Freeman, K., Dinnes, J., Chuchu, N., Takwoingi, Y., Bayliss, S.E., Matin, R.N., Jain, A., Walter, F.M., Williams, H.C., and Deeks, J.J. (2020). Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies. *BMJ* *368*, m127.
 112. Archambault, Y., Boylan, C., Bullock, D., Morgas, T., Peltola, J., Ruokokoski, E., Genghi, A., Haas, B., Suhonen, P., and Thompson, S. (2020). Making on-line adaptive radiotherapy possible using artificial intelligence and machine learning for efficient daily re-planning. *Med. Phys. Intl. J.* *8*.
 113. Bachar, N., Benbassat, D., Brailovsky, D., Eshel, Y., Glück, D., Levner, D., Levy, S., Pecker, S., Yurkovsky, E., Zait, A., et al. (2021). An artificial intelligence-assisted diagnostic platform for rapid near-patient hematology. *Am. J. Hematol.* *96*, 1264–1274.
 114. Liu, X., Rivera, S.C., Moher, D., Calvert, M.J., and Denniston, A.K.; SPIRIT-AI and CONSORT-AI Working Group (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ* *370*, m3164.
 115. Hwang, T.J., Sokolov, E., Franklin, J.M., and Kesselheim, A.S. (2016). Comparison of rates of safety issues and reporting of trial outcomes for medical devices approved in the European Union and United States: cohort study. *BMJ* *353*, i3323.
 116. Mishra, S. (2017). CE mark or something else?—Thinking fast and slow. *Indian Heart J. Teach. Ser.* *69*, 1–5.
 117. Pesapane, F., Volonté, C., Codari, M., and Sardanelli, F. (2018). Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. *Insights Imaging* *9*, 745–753.
 118. Oakden-Rayner, L., Gale, W., Bonham, T.A., Lungren, M.P., Carneiro, G., Bradley, A.P., and Palmer, L.J. (2022). Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study. *Lancet. Digit. Health* *4*, e351–e358.
 119. Daneshjou, R., He, B., Ouyang, D., and Zou, J.Y. (2021). How to evaluate deep learning for cancer diagnostics - factors and recommendations. *Biochim. Biophys. Acta. Rev. Cancer* *1875*, 188515.
 120. Vodrahalli, K., Daneshjou, R., Gerstenberg, T., and Zou, J. (2022). Do humans trust advice more if it comes from ai? an analysis of human-ai interactions. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA (Association for Computing Machinery), pp. 763–777. (AIES '22).
 121. Ferryman, K. (2020). Addressing health disparities in the Food and Drug Administration's artificial intelligence and machine learning regulatory framework. *J. Am. Med. Inform. Assoc.* *27*, 2016–2019.
 122. Lehman, C.D., Wellman, R.D., Buist, D.S.M., Kerlikowske, K., Tosteson, A.N.A., and Miglioretti, D.L.; Breast Cancer Surveillance Consortium (2015). Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern. Med.* *175*, 1828–1837.
 123. Fenton, J.J. (2015). Is it time to stop paying for computer-aided mammography? *JAMA Intern. Med.* *175*, 1837–1838.
 124. Duffy, G., Clarke, S.L., Christensen, M., He, B., Yuan, N., Cheng, S., and Ouyang, D. (2022). Confounders mediate AI prediction of demographics in medical imaging. *NPJ Digit. Med.* *5*, 188.
 125. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* *1*, 206–215.
 126. Wulczyn, E., Steiner, D.F., Moran, M., Plass, M., Reihls, R., Tan, F., Flament-Auvigne, I., Brown, T., Regitnig, P., Chen, P.H.C., et al. (2021). Interpretable survival prediction for colorectal cancer using deep learning. *NPJ Digit. Med.* *4*, 71.
 127. Zhang, Z., Chen, P., McGough, M., Xing, F., Wang, C., Bui, M., Xie, Y., Sapkota, M., Cui, L., Dhillon, J., et al. (2019). Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nat. Mach. Intell.* *1*, 236–245.
 128. Castelvechi, D. (2016). Can we open the black box of AI? *Nature* *538*, 20–23.
 129. Varghese, J. (2020). Artificial intelligence in medicine: chances and challenges for wide clinical adoption. *Visc. Med.* *36*, 443–449.
 130. 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.
 131. Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Vazquez, G., Black, S., and Nolan, G.P. (2018). Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell* *174*, 968–981.e15.
 132. Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., et al.

- (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 353, 78–82.
133. Nejman, D., Livyatan, I., Fuks, G., Gavert, N., Zwang, Y., Geller, L.T., Rotter-Maskowitz, A., Weiser, R., Mallel, G., Gigi, E., et al. (2020). The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science* 368, 973–980.
134. Beshnova, D., Ye, J., Onabolu, O., Moon, B., Zheng, W., Fu, Y.X., Brugarolas, J., Lea, J., and Li, B. (2020). De novo prediction of cancer-associated T cell receptors for noninvasive cancer detection. *Sci. Transl. Med.* 12, eaaz3738.
135. Levy-Jurgenson, A., Tekpli, X., Kristensen, V.N., and Yakhini, Z. (2020). Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer. *Sci. Rep.* 10, 18802.
136. Vanguri, R.S., Luo, J., Aukerman, A.T., Egger, J.V., Fong, C.J., Horvat, N., Pagano, A., Araujo-Filho, J.d.A.B., Geneslaw, L., Rizvi, H., et al. (2022). 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. Cancer* 3, 1151–1164.
137. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention Is All You Need. *Adv. Neural Inf. Process. Syst.*, 5998–6008.
138. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., and Mahmood, F. (2022). Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE), pp. 16144–16155.
139. Xiao, Y., Wu, J., and Lin, Z. (2021). Cancer diagnosis using generative adversarial networks based on deep learning from imbalanced data. *Comput. Biol. Med.* 135, 104540.
140. Li, C.Y., Liang, X., Hu, Z., and Xing, E.P. (2019). Knowledge-driven encode, retrieve, paraphrase for medical image report generation. *AAAI* 33, 6666–6673.