HHS Public Access

Author manuscript

Cancer. Author manuscript; available in PMC 2024 June 15.

Published in final edited form as:

Cancer. 2024 June 15; 130(12): 2101–2107. doi:10.1002/cncr.35307.

Uses and limitations of artificial intelligence for oncology

Likhitha Kolla, BS[1], **Ravi B. Parikh, MD, MPP**[1,2]

[1]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[2]Corporal Michael J. Crescenz VA Medical Center, Philadelphia, PA, USA

Abstract

Modern artificial intelligence (AI) tools built on high-dimensional patient data are reshaping oncology care, helping to improve goal concordant care, decrease cancer mortality rates, and increase workflow efficiency and scope of care. However, data related concerns and human biases that seep into algorithms during development and post-deployment phases affect performance in real world settings, limiting the utility and safety of AI technology in oncology clinics. To this end, we review the current potential and limitations of predictive AI for cancer diagnosis and prognostication as well as of generative AI, specifically modern chatbots, that interfaces with patients and clinicians. We conclude the review with a discussion on ongoing challenges and regulatory opportunities in the field.

INTRODUCTION

Innovations in oncology have driven a decline in cancer mortality rates by 33% in the last 32 years[1]. The concurrent rise of precision medicine techniques – some of which based on artificial intelligence (AI) – has enabled oncology clinicians to better identify previously unnoticed patterns in radiological scans, predict disease progression, offer tailored therapies, and suggest clinical trial eligibility. Clinical AI technologies have ushered an era of more effective screening, optimized treatment regimens, and improved patient outcomes, marking a significant leap forward in cancer care delivery.

Artificial intelligence describes the creation of data-driven, self-operating algorithms for problem solving. Two key terms central to AI are machine learning (ML) and deep learning (DL). Machine learning is a subset of AI and refers to algorithms that automatically learn and adapt from data without explicitly being programmed[2]. Deep learning is a specialized group within ML that mimics human brains by using a multi-layered web of algorithms to process information from data.

AI algorithms are grouped into two categories: predictive AI and generative AI. Predictive AI tools learn patterns from training data to forecast outcomes in new scenarios. For example, an image-based classification tool used to diagnose breast cancer from

Correspondence: Ravi B. Parikh, Perelman School of Medicine, Departments of Medicine and Health Policy, 423 Guardian Drive, University of Pennsylvania, Philadelphia 19104, PA, USA. ravi.parikh@pennmedicine.upenn.edu.

AUTHOR CONTRIBUTIONS

Likhitha Kolla and Ravi B. Parikh: Conceptualization, writing, editing

mammogram scans is a predictive tool. Generative AI creates novel outputs that were not explicitly in the training data. AI chatbots that interact with patients in conversation are a form of generative AI.

Limitless variations of AI algorithms for cancer care management have been published, yet only a minority have been clinically implemented. Roadblocks in implementation include limited FDA regulatory guidelines, high upfront costs for the integration of AI into clinical workflows, noninterpretability of the algorithms, and limited monitoring of algorithms post-deployment[3]. Of the 71 AI-associated devices that were approved by the FDA in 2021, the majority were cancer diagnostics (>80%) and spanned the fields of cancer radiology (54.9%), pathology (19.7%), and radiation oncology (8.5%). These devices were applied to solid malignancies, and most frequently to breast cancer (31%) and lung/prostate cancer (8.5%)[4].

AI tools that have been integrated into clinical workflows in oncology clinics can analyze medical records and help doctors make more informed decisions, saving time and optimizing care. However, the use of AI tools is limited by their inconsistent performance after deployment. Biases associated with algorithm development and implementation can lead to inaccurate predictions that burden healthcare systems, care teams, and individual patients.

To this end, our review will provide context to three use cases of AI for cancer care delivery: diagnosis and classification, prognostication, and chatbots that improve patient care and optimize clinical workflows. The first two applications rely on predictive tools while the third application relies on generative tools. We choose primary studies in each category to illustrate the potential and current use of AI for optimizing cancer care. Each section will also highlight relevant biases and limitations as well as strategies to overcome these obstacles. We conclude the review with considerations on the regulation of AI in oncology.

DIAGNOSIS

Early-stage cancers and cancers that have relapsed after treatment are difficult to diagnose on radiology and pathology reports, a concern that is compounded in patients who appear clinically stable. Machine learning algorithms that have been trained on thousands to millions of images (i.e., radiological scans, pathology images, mobile photographs) of normal and cancerous lesions learn to classify between the two groups, benefitting cases where subtle differences are undetectable to the human eye. Commonly used AI algorithms for image classification are convolutional neural networks (CNN), deep learning architectures that extract identifying features for each group and use the resulting schema for a new classification task. The algorithm assigns a probability for each output class, and the image is classified into the group assigned the highest probability. The accuracy of the AI tool is measured by comparing the algorithm classifications with clinician classifications, referred to as “ground truth.”

Diagnostic AI in oncology can offer early detection with high accuracy, increase efficiency of care, and be scalable across health systems. Two cancer types that have benefited from automated diagnosis with AI technology are skin and breast cancers.

Early work on automated diagnosis for dermatological cancers laid the groundwork for the design and use of AI for cancer diagnosis. Skin lesions are common among adults, however some, like melanoma, can be malignant. Melanoma constitutes only 5% of skin cancer diagnoses but has a 32% 5-year survival rate if detected at a metastatic stage, compared to 99% when detected earlier[5]. Differentiating between early-stage melanoma and similarly appearing benign skin lesions is visually difficult, often leading to misdiagnosis or a delay in diagnosis. This clinical setting is a prime example for where AI technology help identify fine-grained variability between pathological and benign states, enabling clinicians to make more accurate diagnoses and offer therapeutic interventions earlier in patients' disease trajectories.

A 2017 proof-of-concept study of a deep learning algorithm used to classify between malignant melanoma, benign nevi, and non-neoplastic lesions was the first major work to establish the competence of AI tools for cancer diagnosis. A CNN trained on 129,450 biopsy-proven photographic images was compared against the performance of 21 board-certified dermatologists. The overall classification accuracy of the algorithm on a testing set was on par with that of two dermatologists (72.1% vs 66.0% and 65.6%, respectively)[6]. The promise of this diagnostic aid lies in its performance and scalability. Melanoma diagnosis is standardly obtained with an in-person visual examination of the skin lesion and a procedure to obtain a biopsy for histopathological confirmation of the disease. AI tools that match the diagnosis accuracy and rate of dermatologists without the requirement for invasive procedures or clinic visits can expand the scope of cancer care to areas with limited medical resources and/or access to medical care. These algorithms can also classify skin lesions from readily available mobile images with equal performance to specialist and novice physicians, as shown in a recent multicenter, prospective trial, further eliminating the need for a clinic visit for initial screening[7].

Breast cancer screening is another clinical context where AI diagnostic aids have been beneficial in identifying disease processes earlier. Screening efforts have evolved with improvements in radiological equipment and public outreach in the last few decades, helping to decrease breast cancer mortality by over 43% since 1989[8]. In most settings, mammogram screening undergoes a two-reader evaluation, where two clinicians independently read the scans and their results are combined. A third reader is solicited only to settle discordance. Despite recent advancements, lack of available screening resources and radiologists can constrain breast cancer screening in scan interpretation. In addition, human error and variability in diagnostic interpretation of mammogram results by radiologists is a large challenge. In a study of 359 radiologists surveying more than 1.6 million mammograms, 41% inconsistently failed to meet standard recall rates[9]. AI algorithms reduce the burden of screening by alleviating the need for a second reader and increasing efficiency of detecting lesions in a scan.

Computer-aided detection (CAD) algorithms built on deep learning frameworks identify a suspicious region of interest on scans for radiologists to review and have assisted mammographic interpretation with mixed performance. An external validation study of three commercially tested AI CAD algorithms that screen for breast cancer in 8805 women found that only one achieved sensitivity and specificity metrics that aligned with the US Breast Cancer Surveillance Consortium benchmarks. The accuracy rate of this algorithm was 95.6%, surpassing the other two which scored on average 92.1%[10]. Cancer detection by the best algorithm was surpassed by 8% when combined with assessments from a first reader. This form of AI and human collaboration exceeded AI alone and a double-reader interpretation of mammograms, where the results from the first and second readers are combined. Additionally, a retrospective study on 275,000 breast cancer cases reported that AI software that matched the performance of human radiologists when acting as the second reader of mammography scans can streamline breast cancer diagnoses by cutting radiologists workloads by at least 30%[11].

Beyond algorithm development and reader studies, validating the performance of an AI diagnostic tool in prospective clinical trials is critical for evaluating its effective integration and safety in real world settings. The Mammography Screening with Artificial Intelligence (MASAI) study in Sweden is the first randomized control trial to assess how AI CAD tools can be integrated safely into clinical workflows[12]. In this prospective, non-inferiority, single-blinded study, an AI CAD was used to triage mammogram interpretations to a single or double reading setting. Scans with higher CAD classification scores were prioritized for a two-reader evaluation. Cancer detection rate increased by 20% and overall workload decreased by half, offering an evidence-based framework for integrating AI tools for breast cancer screening into clinical workflows.

The benefit of AI algorithms for image classification is apparent for cancer diagnosis. In many cases, the technology can detect whether a patient has a cancerous lesion with as good of an accuracy as a clinician, reduce physician workload, provide a non-invasive alternative for diagnosis, and increase access to care. However, there are important cases where the algorithm can break, leading to critical questions in bias, fairness, and robustness.

Underreporting, underrepresentation, and heterogeneity in image acquisition can skew the data used to train an AI algorithm[13]. As a result, the algorithm is not generalizable to patient populations that are not well represented in the training dataset. For example, in the case of skin cancers, AI algorithms run the risk of worse performance for people with darker skin[14]. Many published AI algorithms are trained on publicly available image datasets that are biased. A survey of 21 accessible skin lesion datasets covering more than 100,000 pictures revealed that lesions on darker skin were underrepresented[15]. Of the 2,436 images where skin color was indicated, only 11 were of brown or Black skin; and of the 1,585 images with attached ethnicity information, none were from people with African, Afro-Caribbean, or South Asian backgrounds.

Modifications along the algorithm development pipeline can help mitigate these concerns. Training data can be expanded to include representative images from all demographics (e.g. skin color, ages, body types). Training sets with image data should include samples taken

from different angles, lighting, and equipment; and AI technologies should accommodate changes in image acquisition technology by retraining the model with new images. Furthermore, additional clinical trials and validation studies investigating the integration of developed algorithms into clinical workflows as a potential second or third reader of diagnostic scans are needed.

PROGNOSTICATION

Forecasting patient outcomes helps tailor medical plans and optimize resource allocation in oncology. However, predicting prognosis is a challenge for oncologists, with an estimated 63% overestimating and 17% underestimating survival[16]. One reason is that physicians rely on both clinical precedent and nationally published population statistics (e.g., 5-year median survival) to assess an individual patient, which leads to overgeneralization and inaccurate assessments of risk. The consequences of inaccurate predictions in oncology include increased emotional burden on patients and their caregivers, inappropriate allocation of resources, decreased trust in the patient-physician relationship, and delay in crucial therapeutic or end-of-life interventions[17]. AI-based risk prediction models that generate individualized estimates on prognosis have augmented clinician assessments of risk and aided personalized care decisions in oncology.

While diagnostic models evaluate whether a patient has a disease, prognostic models focus on whether a patient will develop a disease or an adverse outcome (e.g., hospitalization or mortality). Prognostic AI algorithms are built with unstructured data, like clinical notes from electronic health records (EHRs), radiology reports, and pathology findings, as well as structured data, like patient demographics, lab results, and patient reported outcomes (PROs) surveys. Benefits of EHR data include the depth and breadth of information available for each patient, the opportunity for frequent, longitudinal collection of data, and the continual tracking of outcomes. In oncology, the resulting predictions are commonly used to stratify patients along a risk continuum, and “high risk” patients who fall above a threshold of risk qualify for additional interventions.

A clinical scenario where prognostic AI models have helped to optimize goal concordant care and healthcare spending is for the prioritization of end-of-life care for patients with advanced cancers. EHR-based machine learning algorithms that calculated the 180-day risk of mortality identified cases with high accuracy (AUC: 0.95–0.96) and provided an individualized, data-driven alternative to standard prognostic models and decision-making frameworks derived from prior randomized control trials[18]. The algorithm generalized well to real-world settings in a prospective trial where it was paired with a behavioral health intervention to prompt serious illness conversations (SICs) for high-risk patients with advanced cancer. Implementation prompted an 11% increase[19] in patient encounters with documented serious illness conversations and decreased end-of-life spending by $75.33 on average per day[20]. The effects of the algorithm on end-of-life care and SIC rates sustained outside of trial settings at its continued deployment in a large healthcare system.

We use the mortality prediction algorithm above as a case study to illustrate the potential of prognostic AI models to augment risk assessments and clinical decision-making in

oncology. Since the algorithm was trained on data from and applied to patients at a single institution, it becomes hard to determine how it could perform at a different health system. This idea of heterogeneity in model performance across testing scenarios (e.g., patient subpopulations, geographic locations, time) raises necessary inquiries on algorithmic fairness and reliability, and the safety in the long-term use of prognostic AI models in oncology.

Equal model performance across patient subpopulations (e.g., by race or gender) ensures algorithmic fairness and promotes fair allocation of interventions. However, risk prediction algorithms can propagate existing social biases, manifesting as less accurate predictions for protected patient groups. Two main factors leading to biased models include limited training data for medically underrepresented patient populations and the use of improper proxy variables in the model that coarsely represent the true mechanism of risk. A prominent 2019 study showed that a widely used risk prediction algorithm that predicted healthcare costs, which are a function of both medical conditions and social determinants of health, underserved Black patients who had similar health profiles as White patients. Predicting the number of chronic conditions instead increased the percentage of Black patients who qualified for additional health interventions by 28.8%[21].

Misrepresentation of data calls for careful consideration of model inputs. The inclusion of socially defined features, such as race and ethnicity, in risk prediction models has been shown to be clinically relevant but socially contentious[22]. In one study, four risk models that predicted postoperative cancer recurrence among patients with colorectal cancer showed that the inclusion of race and ethnicity variables decreased racial bias metrics and increased algorithmic fairness along several metrics of model performance[23]. While some socially defined variables strongly predict risk in this setting and can promote fair resource allocation, the reliance on these predictors masks the true unattainable or unaccounted for drivers (e.g., socioeconomic status, biomarkers associated with disease risk). Including variables like race can also ossify discriminatory generalizations about protected patient subpopulations, widening the gap in access to high-quality cancer care.

Performance drift, the deterioration of model performance with time, affects the reliability of risk predictions post-deployment. Most deployed models in oncology settings are deterministic in nature, and changes in the data generation process without a paired update of the algorithm can lead to unreliable predictions. Two commonly identified reasons for drift in clinical risk models include changes to EHR software and documentation practices as well as changes in healthcare practice patterns. A recent study of drift in the 6-month mortality prediction model discussed above noted a 7% decrease in true positive rate during the COVID-19 pandemic period, a drop that was associated with decreases in laboratory utilization during quarantine[24]. Continuous monitoring and intermittent updating of the model is crucial to mitigate drift-related negative consequences on care decisions and resource allocation.

Quality of data in the training, pre-deployment phase of AI models also affects performance of prognostic AI tools in real world settings. Some forms of EHR data are incomplete, unstructured, subject to human recording error, and unstandardized across health systems.

These concerns are mirrored in other patient data modalities becoming more prevalent for risk prediction, like PROs and mobile data, with the additional cost of noisy and imprecise measurements in their longitudinal collection. Robust data preprocessing, error correction, and standardization procedures as well as data-sharing standards are needed to enhance AI performance.

CHATBOTS AND GENERATIVE AI

As we write this review, modern conversational chatbots are making a wave across healthcare. Chatbots are computer programs that generate human-like language. The underlying learning architecture of modern chatbots evolved from predictive natural language processing and speech recognition software to generative large language models (LLMs) that process large text-based datasets to translate, predict, and produce content. For patients, LLM chatbots offer support to patient education, patient-clinician communication, and mental health services. For physicians, LLMs have the potential to encode clinical knowledge, automate medical documentation (e.g., informed consent), enhance telemedicine interactions, and assist in clinical trial enrollment[25].

Recent work on the use of LLMs for cancer care management has revealed that the technology is still limited in the quality and accuracy of information provided. One retrospective, cross-sectional study evaluated whether the recommendations for breast, prostate, and lung cancer treatment generated by OpenAI's commercial LLM, ChatGPT, aligned with the standard-of-care set by the National Comprehensive Cancer Network (NCCN)[26]. The team found that approximately one-third of the chatbot's treatment recommendations did not fully agree with NCCN guidelines. Recommendations varied with the phrasing of questions, and discrepancies between the chatbot and guidelines were often attributed to uninterpretable responses, suggesting caution is needed when using LLM chatbots for treatment information.

In a related study, researchers compared the quality of information generated on the top Google search queries of 5 common cancers (i.e., lung, skin, colorectal, breast, prostate) provided by ChatGPT v3.5 with those provided by Perplexity, Chatsonic, and Bing AI. While quality of the responses was good, with a median DISCERN score of 5, they were hard to understand with a college-level readability and were not readily actionable[27].

Chatbots and other generative AI technologies are still nascent in medicine with limitations in accuracy, readability, and reliability. Commercial chatbots like ChatGPT are trained on a wide variety of text data found on the internet, with limited quality checks on the validity of the information. In addition, LLMs are a "black-box" and their minimal explainability, the ability to understand and interpret how algorithms arrive at their predictions, remains a significant challenge[28]. LLMs are not able to identify the sources or the exact training data used to generate their text. As a result, LLMs can propagate misinformation, causing confusion and mistrust among users (i.e. patients and physicians). AI hallucinations, the generation of inaccurate information from a prompt founded on false information, are another source of mistrust[29].

The adoption of chatbots for medicine relies on achieving both understandable language and conveying complex medical topics accurately, which current algorithms cannot do consistently as readability scores vary by the user's verbiage of the prompt. And while medical knowledge expands each day, algorithms are not continuously updated to accommodate this change. As a result, the chatbots that are not trained on updated information can become unreliable and more inaccurate with time.

The field of generative AI technology is evolving rapidly, and we expect a parallel expansion in its application to oncology. Better regulations of chatbots for medical care are needed to prioritize patient safety and privacy[30]. And while the use of chatbots can be better regulated in the clinic, it is more difficult to oversee the private use of chatbots by patients who autonomously seek medical knowledge. Oncology clinicians aware of this concern can guide their patients through chatbot-derived medical information.

DISCUSSION

Alan Turing's 1950 question "Can machines think?" posed in his provocative work *Computer Machinery and Intelligence* laid the conceptual groundwork for a new field, artificial intelligence[31]. Turing's thought experiment led to the creation of early AI robotic systems that imitated human decision making, including the prosthetic "Tentacle Arm" and robots for industrial assembly lines. At the turn of the 21st century, with refinements in AI architectures to make them more amenable for high stake medical environments, the reach of AI spread to augment clinician decision-making and as a result has reshaped the landscape of cancer care management.

AI offers endless potential to push cancer care to new frontiers by enabling early diagnoses, offering more precise estimates of risk, informing effective treatment regimens, and freeing clinician time for patient-focused interactions.

This review only touched the surface of the potential for AI systems in oncology. Use cases for AI in medicine and epidemiology that we did not cover here but are equally important to deliberate on include the application of AI to clinical trial enrollment and for the study of disease development and progression; cancer genomics and genetic mutations; digital health and mobile monitoring of disease status; and population-level risk factors.

AI algorithms are only as good as the data and assumptions they are fed. Biased representation of patient populations and medical scenarios in the training datasets can lead to data overfitting and inaccurate generalizations of AI tools in the real world. Dataset shift, the stray of real-world distributions of data from the training set, can lead to a drift in AI performance over time and decrease the reliability of its output. Ensuring diverse, representative data in the training, evaluation, and post-deployment monitoring phases is critical.

Beyond biases baked into the training data, human biases can affect how AI algorithms are utilized in the clinic. In a 2019 survey of doctors in Korea, 83.4% appreciated the usefulness of AI in medicine especially for medical diagnosis, but only 5.9% were familiar with AI and 29.3% acknowledged that AI cannot help in unexpected situations owing to inadequate

information[32]. Factors like physician expertise, technological literacy, and age influence the adoption of these technologies into practice. Physician hesitation in utilizing AI tools can stem from the lack of algorithm explainability in how predictions are generated, unassigned medical liability and economic cost of inaccurate predictions, and unfamiliarity with AI tools. Current methodological work on decoding explainability involves statistical scores ascribed to input variables to determine each model input's contribution to the generation of a prediction[33]. Decoding variable importance can help gain user confidence in the output and enable better integration of these tools into clinical workflows. On the other hand, automation bias, the overreliance on AI to make clinical decisions, at the possible cost of negating one's own clinical intuition about a patient, can equally hinder the proper use of AI for cancer care delivery[34]. False negative and false positive cases are overlooked with automation bias, which could lead to misinformed medical decisions[35].

AI systems cannot perfectly replicate clinician decision-making, with variables like patient composure, cognitive status, and clinical status that are not resolutely captured in data but nonetheless are critical in assessing patient risk. Expanding prognostic AI models to include patient reported outcomes that continuously capture symptoms and functional status outside of the clinic walls can improve model accuracy and clinical relevance. One study reported a 4% increase in AUC in a mortality prediction model that was trained on PRO and EHR data vs EHR data alone[36]. Interactive AI frameworks like human-in-the-loop models or human-machine collaborative models that incorporate real-time feedback and insights from clinicians can improve prediction accuracy and confidence, ensuring a more comprehensive approach to risk assessment and decision-making in healthcare settings.

As more AI algorithms are developed and implemented into the clinic, readjustments to clinical workflows in oncology to accommodate them hinges on defined regulatory oversight. Measures to protect patient privacy, standardize data collection, and maintain algorithmic reliability will ensure the responsible use of AI. Additional focus on hardware requirements, continuous monitoring, and restricted use cases is needed for generative technologies like LLM-based chatbots. In today's converging era of AI and oncology[30], a balance between innovation and responsibility will raise cancer care delivery to new heights to the benefit of patients, physicians, and healthcare systems.

CONFLICT OF INTERST STATEMENT

RBP has received grants from the National Institutes of Health, Department of Defense,, Prostate Cancer Foundation, National Palliative Care Research Center, NCCN Foundation, Conquer Cancer Foundation, Humana, Emerson Collective, Schmidt Futures, Arnold Ventures, and Veterans Health Administration; personal fees and equity from GNS Healthcare, Thyme Care, and Onc.AI; personal fees from the Cancer Study Group, Biofourmis, Genetic Chemistry Therapeutics, CreditSuisse, G1 Therapeutics, Humana, and Nanology; honoraria from Flatiron and Medscape; has board membership (unpaid) at the Coalition to Transform Advanced Care and American Cancer Society; and serves on a leadership consortium (unpaid) at the National Quality Forum, all outside the submitted work.

REFERENCES

1. Siegel RL, Miller KD, Wagle NS & Jemal A Cancer statistics, 2023. CA Cancer J Clin 73, 17–48 (2023). [PubMed: 36633525]

2. Janiesch C, Zschech P & Heinrich K Machine learning and deep learning. doi:10.1007/s12525-021-00475-2/Published.
3. Elemento O, Leslie C, Lundin J & Tourassi G Artificial intelligence in cancer research, diagnosis and therapy. Nature Reviews Cancer vol. 21 747–752 Preprint at 10.1038/s41568-021-00399-1 (2021).
4. Luchini C, Pea A & Scarpa A Artificial intelligence in oncology: current applications and future perspectives. British Journal of Cancer vol. 126 4–9 Preprint at 10.1038/s41416-021-01633-1 (2022). [PubMed: 34837074]
5. Melanoma: Statistics. http://www.cancer.net/about-us/cancernet-editorial-board.
6. Esteva A et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118 (2017). [PubMed: 28117445]
7. Menzies SW et al. Comparison of humans versus mobile phone-powered artificial intelligence for the diagnosis and management of pigmented skin cancer in secondary care: a multicentre, prospective, diagnostic, clinical trial. Lancet Digit Health 5, e679–e691 (2023). [PubMed: 37775188]
8. Giaquinto AN et al. Breast Cancer Statistics, 2022. CA Cancer J Clin 72, 524–541 (2022). [PubMed: 36190501]
9. Lehman CD et al. National performance benchmarks for modern screening digital mammography: Update from the Breast Cancer Surveillance Consortium. Radiology 283, 49–58 (2017). [PubMed: 27918707]
10. Salim M et al. External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. JAMA Oncol 6, 1581–1588 (2020). [PubMed: 32852536]
11. Sharma N et al. Retrospective large-scale evaluation of an AI system as an independent reader for double reading in breast cancer screening. doi:10.1101/2021.02.26.21252537.
12. Lång K et al. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. Lancet Oncol 24, 936–944 (2023). [PubMed: 37541274]
13. Guo LN, Lee MS, Kassamali B, Mita C & Nambudiri VE Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection—A scoping review. J Am Acad Dermatol 87, 153–157 (2022). [PubMed: 34252469]
14. Adamson AS & Smith A Machine learning and health care disparities in dermatology. JAMA Dermatology vol. 154 1247–1248 Preprint at 10.1001/jamadermatol.2018.2348 (2018). [PubMed: 30073260]
15. Wen D et al. Characteristics of publicly available skin cancer image datasets: a systematic review. The Lancet Digital Health vol. 4 e64–e74 Preprint at 10.1016/S2589-7500(21)00252-1 (2022). [PubMed: 34772649]
16. Christakis NA & Lamont EB Extent and determinants of error in doctors' prognoses in terminally ill patients: Prospective cohort study. Br Med J 320, 469–472 (2000). [PubMed: 10678857]
17. Glare P et al. A systematic review of physicians' survival predictions in terminally ill cancer patients. British Medical Journal vol. 327 195–198 Preprint at 10.1136/bmj.327.7408.195 (2003). [PubMed: 12881260]
18. Parikh RB et al. Machine Learning Approaches to Predict 6-Month Mortality Among Patients With Cancer. JAMA Netw Open 2, e1915997 (2019). [PubMed: 31651973]
19. Manz CR et al. Effect of Integrating Machine Learning Mortality Estimates with Behavioral Nudges to Clinicians on Serious Illness Conversations among Patients with Cancer: A Stepped-Wedge Cluster Randomized Clinical Trial. in JAMA Oncology vol. 6 (American Medical Association, 2020).
20. Parikh RB et al. End-of-life spending analysis of randomized trial of machine learning nudges to prompt serious illness communication among patients with cancer. Journal of Clinical Oncology 41, 6515–6515 (2023).
21. Obermeyer Z, Powers B, Vogeli C & Mullainathan S Dissecting racial bias in an algorithm used to manage the health of populations. http://science.sciencemag.org/.

22. Manski CF, Mullahy J & Venkataramani AS Using measures of race to make clinical predictions: Decision making, patient health, and fairness. Proc Natl Acad Sci U S A 120, (2023).
23. Khor S et al. Racial and Ethnic Bias in Risk Prediction Models for Colorectal Cancer Recurrence When Race and Ethnicity Are Omitted as Predictors. JAMA Netw Open 6, E2318495 (2023). [PubMed: 37318804]
24. Parikh RB et al. Performance drift in a mortality prediction algorithm among patients with cancer during the SARS-CoV-2 pandemic. J Am Med Inform Assoc 30, 348–354 (2023). [PubMed: 36409991]
25. Decker H et al. Large Language Model-Based Chatbot vs Surgeon-Generated Informed Consent Documentation for Common Procedures. JAMA Netw Open 6, e2336997 (2023). [PubMed: 37812419]
26. Chen S et al. Use of Artificial Intelligence Chatbots for Cancer Treatment Information. JAMA Oncol (2023) doi:10.1001/jamaoncol.2023.2789.
27. Pan A, Musheyev D, Bockelman D, Loeb S & Kabarriti AE Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries About Cancer. JAMA Oncol (2023) doi:10.1001/jamaoncol.2023.2947.
28. Zhao H et al. Explainability for Large Language Models: A Survey. (2023).
29. Azamfirei R, Kudchadkar SR & Fackler J Large language models and the perils of their hallucinations. Critical Care vol. 27 Preprint at 10.1186/s13054-023-04393-x (2023).
30. Meskó B & Topol EJ The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 6, (2023).
31. Turing AM M I N D A QUARTERLY REVIEW OF PSYCHOLOGY AND PHILOSOPHY I.-COMPUTING MACHINERY AND INTELLIGENCE. https://academic.oup.com/mind/article/LIX/236/433/986238 (1950).
32. Oh S et al. Physician confidence in artificial intelligence: An online mobile survey. J Med Internet Res 21, (2019).
33. Molnar C, König G, Bischl B & Casalicchio G Model-agnostic feature importance and effects with dependent features: a conditional subgroup approach. Data Min Knowl Discov (2023) doi:10.1007/s10618-022-00901-9.
34. Parikh RB, Teeple S & Navathe AS Addressing Bias in Artificial Intelligence in Health Care. JAMA - Journal of the American Medical Association vol. 322 2377–2378 Preprint at 10.1001/jama.2019.18058 (2019). [PubMed: 31755905]
35. Evans KK, Birdwell RL & Wolfe JM If You Don't Find It Often, You Often Don't Find It: Why Some Cancers Are Missed in Breast Cancer Screening. PLoS One 8, (2013).
36. Parikh RB et al. Development of Machine Learning Algorithms Incorporating Electronic Health Record Data, Patient-Reported Outcomes, or Both to Predict Mortality for Outpatients With Cancer. JCO Clin Cancer Inform 6, 2200073 (2022).