NEJM AI

NEJM AI 2025;2(5)
DOI: 10.1056/AIra2401164

REVIEW ARTICLE

The Use of Artificial Intelligence for Cancer Therapeutic Decision-Making

Olivier Elemento, Ph.D.,[1] Sean Khozin, M.D.,[2] and Cora N. Sternberg, M.D.[1]

Received: November 29, 2024; Revised: January 24, 2025; Accepted: February 13, 2025; Published: April 17, 2025

Abstract

Artificial intelligence (AI) has the potential to transform cancer therapeutic decision-making by improving diagnostics and personalizing treatments. This review explores the current and future impact of AI in oncology, focusing on its applications in radiology and pathology and the potential of large language models in treatment selection. Despite significant advancements, AI integration into clinical workflows is limited due to challenges such as data quality, model accuracy, and lack of validation through clinical trials. We propose key strategies to address these challenges, including developing robust multicenter datasets, promoting practical AI model development, researching workflow integration and human–AI collaboration, leveraging lessons from AI in medical imaging, establishing evaluation guidelines, and incentivizing prospective clinical trials. By implementing these strategies, AI can significantly enhance cancer care and patient outcomes, paving the way for its effective integration into oncology practice.

The Rapid Pace of AI Progress in Medicine

Artificial intelligence (AI) has rapidly advanced over the past decade, combining new algorithms, efficient hardware, and large datasets, leading to advancements in deep learning. Medical research has leveraged deep learning for applications such as automated skin lesion classification.[1,2] The emergence of large language models (LLMs), such as GPT-4, Claude, and Med-PaLM2, has enabled sophisticated natural language processing of unstructured clinical data like electronic health records (EHRs).

State-of-the-art medical LLMs, such as Med-PaLM2, perform well on benchmark datasets, achieving up to 86.5% accuracy on the MedQA dataset,[3] which consist of multiple-choice questions based on the United States Medical Licensing Examination. Physicians preferred Med-PaLM2's answers over the physicians' on eight of nine clinical utility axes.[3] However, detailed performance results on oncology questions are lacking. Rydzewski et al.[4] specifically examined the performance of several LLMs on over 2000 oncology questions; GPT-4 achieved the highest accuracy at 68.7%. While this compares reasonably well with human performers, GPT-4 and other LLMs demonstrated significant error rates, including overconfidence, hallucinations, and inaccuracies.

The author affiliations are listed at the end of the article.

Dr. Elemento can be contacted at ole2001@med.cornell.edu.

The Rise of Foundation Models in Cancer Diagnosis

Accurate diagnosis, which is crucial for effective treatment, is advancing with AI, particularly in cancer detection using radiology and pathology data. Studies have shown the clinical utility of AI in detecting breast cancer from mammograms. In a study conducted in Hungary, AI, as an additional reader in breast cancer screening, achieved more cancer detections than double readings by physicians.[5] The ScreenTrustCAD Swedish study showed that replacing one radiologist with an AI model resulted in a 4% higher cancer detection rate.[6] Another study from Sweden showed that AI-assisted screening with triaging high-risk cases to a radiologist had equivalent performance to double readings, but reduced the workload by 44.3%.[7] Similar performances were observed in lung cancer detection studies conducted in the United States and China, although prospective randomized clinical trials (RCTs) are still needed.[8,9]

A common concern is that AI models perform well at medical centers whose data they were trained on but lose performance elsewhere due to dataset shifts.[10] Such shifts, characterized as differences between development and deployment datasets, can affect performance, but data variability across sites and over time is often unclear. Foundation models, trained on large unlabeled datasets using self-supervised learning, address this by enabling fine-tuning for specific tasks.[11] Examples of biomedical foundation models include Pal et al.'s model for cancer imaging biomarker discovery trained from a dataset of 11,467 radiographic lesions. Fine-tuned models based on that foundation model showed robust performance (area under the curve >0.95) in tasks like predicting anatomical site and lesion malignancy.[12]

After initial detection using imaging, a biopsy is performed for definitive diagnosis and staging. Scanned tissue slides can be analyzed using AI models as part of a digital pathology workflow. Pathology AI models have achieved strong performance with deep learning. For example, the Paige digital pathology software achieved 96.6% sensitivity in prostate biopsy readings.[13] Another automated deep-learning system achieved performance similar to pathologists for Gleason grading.[14] Pathology foundation models like CONtrastive learning from Captions for Histopathology and Virchow have achieved high accuracy across benchmarks, and the fine-tuned models performed well in diagnosing rare diseases.[15,16] However, despite promising performance and recent U.S. Food and Drug Administration (FDA) clearances,[17] prospective RCTs assessing these models in clinical settings are lacking, possibly due to integration challenges in existing workflows.

Large Language Models in Oncology: A Unique Opportunity

After diagnosis and staging, AI can theoretically enhance cancer care across multiple disciplines. In surgical oncology, AI may be able to enhance preoperative planning and intraoperative decision-making. Machine learning algorithms can analyze preoperative imaging to optimize surgical approaches and predict potential complications. During procedures, computer vision algorithms may be able to assist surgeons in identifying tumor margins and critical structures in real time, potentially improving surgical precision and reducing positive margin rates. These applications show particular promise in minimally invasive procedures, although implementation challenges, rigorous reliability assessment, and the need for prospective validation remain.[18] Likewise, in radiation oncology, AI may eventually transform treatment planning and delivery.[19-21] Machine learning algorithms can optimize radiation treatment plans, potentially reducing planning time,[22] while improving target coverage and minimizing exposure to healthy tissue.[23]

AI can potentially assist oncologists' central role in identifying treatment options. Active research explores how molecular alterations can be analyzed using AI to predict therapy responses, such as immunotherapies, though few are ready for clinical use.[24-26] AI can assist by providing treatment guidance following accepted guidelines from the National Comprehensive Cancer Network (NCCN), the European Society for Medical Oncology (ESMO), or the Chinese Society of Clinical Oncology (CSCO), which have increased in complexity over the years. AI can also offer guidance when guidelines are lacking, such as after multiple therapy failures,[27] and help identify clinical trials for eligible patients. Managing potential adverse events is another area where AI could play a role.

These tasks rely on clinical data in the EHRs, which are often unstructured, making oncology a field in which LLMs could excel if they can effectively interpret unstructured data. Sushil et al. constructed a dataset of 40 deidentified cancer progress notes and assessed three recent LLMs in a zero-shot extraction of oncological information. GPT-4 exhibited the best performance, with an average accuracy of 68%.[28] These results indicate that LLMs are not yet ready for direct application to EHR data.

Assuming LLMs receive reliable, curated EHR information, their ability to provide accurate cancer treatment recommendations can be assessed. In a study by Chen et al., GPT-3.5 did not perform well; 34.3% of its treatment recommendations included nonconcordant treatments per NCCN guidelines, with 12.5% considered hallucinated.[29] Another study found that GPT-3.5 and Copilot provided completely correct responses in only 36% of scenarios, with inaccurate or misleading information in 24% of cases.[30] Benary et al. found that LLMs deviated substantially from expert recommendations in treatment options for advanced cancer cases.[31] More promising results were achieved in a study by Marchi et al., where GPT-3.5 achieved accuracies of 85.3% for primary treatment selection.[32]

These studies explore the accuracy of LLMs in adhering to best practices and guidelines. When patients exhaust standard-of-care options, they may be eligible for clinical trials. Only a small fraction of cancer patients enroll in trials,[33] partly due to the absence of a systematic approach to matching patients with trials. Studies have addressed this using LLMs with promising performance[34-36]; for example, Ferber et al. achieved 92.7% accuracy in matching patients to trials using GPT-4.[34] Practical implementation challenges remain, especially regarding the integration of real-time information processing and availability to support accelerated decision-making in settings where rapid care is needed. This is particularly crucial in the context of multidisciplinary team (MDT) meetings where timely access to and integration with existing MDT workflows are essential. Streamlining integration with MDT workflows, potentially through AI-powered clinical note generation from conversations, could address these challenges.

While encouraging, these results lack the robustness needed for broad clinical use, except perhaps in matching patients to clinical trials. Variability may stem from different LLMs, prompting strategies and assessment approaches. Early results suggest that LLMs are not ready for autonomous use but point to potential utility in assisting clinicians, similar to AI models in radiology being assessed in prospective trials.

Discussion

Integrating AI into oncology decision-making presents a complex landscape of progress and challenges. While diagnostic applications in radiology and pathology show promising maturity, broader implementation in treatment selection remains premature. A critical limitation is the inconsistent and often inadequate adherence of these models, particularly LLMs, to established clinical guidelines. This underscores the necessity for AI systems to more accurately capture and interpret multifaceted clinical data within EHRs. Current research needs to address standardized prompt engineering techniques and methodologies for processing diverse medical documents. Moreover, the field lacks prospective RCTs, which are crucial for validating the clinical utility and safety of these AI systems before widespread adoption.

Retrieval-augmented generation (RAG) enhances LLM outputs by incorporating references to authoritative knowledge bases, such as ESMO or NCCN guidelines, and providing the added benefit of explainability by grounding responses in established medical knowledge. In a 2024 study, the Almanac approach demonstrated superior performance using RAG compared with standalone LLMs on the ClinicalQA benchmark in nononcology contexts.[37] Example-based prompting techniques, including one-shot and few-shot learning, aim to guide the LLM's reasoning process by providing context-specific examples. However, the effectiveness of RAG and shot learning approaches in improving LLM performance for oncology applications remains largely untested in real-world patient data.

A critical factor potentially limiting AI model accuracy in oncology applications is the quality and quantity of oncology-specific datasets used in training existing LLMs. For most LLMs, specific datasets, their origins, and curation methods are not fully disclosed. This lack of transparency extends to the weighting of different datasets in training pipelines, presenting challenges even for models optimized for medical applications. For instance, while clinical guidelines from organizations such as NCCN, ESMO, and CSCO are likely incorporated into training data, the specific versions used, and their relative importance compared with less reliable sources remain unclear.

Addressing these training data challenges presents significant hurdles. Fine-tuning existing LLMs with specific, high-quality oncology data sources is one potential approach. However, the computational resources required for such endeavors may be prohibitive for most researchers and institutions. Incorporating real-world data into training sets could potentially enhance LLM accuracy in oncology applications, but this approach faces multiple barriers. These include ensuring data integrity and accuracy, given the diverse sources and potential inconsistencies in real-world data. Representation is another crucial issue, as certain real-world datasets may not adequately reflect diverse patient populations, potentially leading to biases in AI models. Standardization poses a significant challenge due

to variations in data collection methods and formats across different health care systems, making it difficult to integrate and analyze real-world data effectively. Moreover, the use of real-world patient data raises important privacy and ethical concerns, as well as the need to navigate the complex landscape of health care data regulations across different jurisdictions. The scarcity of large, publicly available oncology datasets reflecting real-world cases further impedes progress. Although some databases exist, their utility is often limited by restricted access to curated clinical data. The Genomics Evidence Neoplasia Information Exchange (GENIE) database is a notable exception, but is limited in focus and access for most users.[38]

Improving AI models requires addressing potential biases in training data. Rydzewski et al. evaluated LLMs on oncology questions, revealing bias-related concerns with worse performance in female-predominant malignancies.[4] Zack et al. showed that GPT-4 tended to exaggerate known prevalence differences when generating clinical vignettes,[39] underscoring the need for ongoing evaluation and mitigation strategies to prevent perpetuating health care inequities.

AI models in oncology will need to evolve to interpret increasingly complex molecular data from clinical and commercial laboratories. As molecular biomarkers become more prevalent in clinical guidelines, interpreting these sophisticated tests poses significant challenges for many health care professionals. For instance, determining whether a genetic mutation like BRCA is clearly loss-of-function or a variant of unknown significance requires specialized knowledge that some clinicians may not possess. This challenge intensifies as cancer testing grows more sophisticated, incorporating elements such as mutational signatures, complex structural variants, and other biomarkers.[40] AI models could potentially bridge the gap between rapidly advancing molecular diagnostics and clinical practice by assisting in this complex decision-making process. These models, trained on large datasets, could identify clinically relevant alterations, address tumor heterogeneity, and suggest potential treatment options or clinical trials.

Our review of the literature reveals a significant limitation in AI applications aimed at supporting clinical decision-making in oncology: the lack of high-quality data. Such datasets are critical not only for benchmarking and rigorous testing of AI models, followed ideally by clinical trials, but also for technical validation. This validation ensures that AI models consistently demonstrate accuracy, reliability, and robustness across diverse datasets, including those that reflect different patient demographics and clinical conditions. It confirms that models perform as expected, free from biases, and can generalize effectively beyond their training data. Unfortunately, many studies evaluating LLMs rely on limited or proprietary datasets, which hinders reproducibility and broader research. In addition, RCTs may be necessary to clinically validate these models, assessing whether or not their predictions improve patient outcomes in real-world settings in a statistically significant and clinically meaningful way. This step transcends technical accuracy to evaluate the model's safety, effectiveness, and impact on patient care. However, we found that outside radiology, RCTs evaluating the real-world impact of AI models in oncology are exceedingly rare, creating a critical gap in the validation needed for their widespread clinical use. To the best of our knowledge, there has not been any rigorous analysis of why there are so few RCTs of AI models in oncology (and in medicine in general). There is a view that traditional RCTs are impractical for AI models in medicine due to factors like clinical investigators' discomfort with AI, the need to alter existing validated workflows, and a lack of dedicated funding. In addition, the rapid pace of AI development may render RCT results obsolete by the time they are published.

Key Actions for Effective AI Integration in Oncology Care

Our assessment outlined herein identifies several key actions essential for the impactful integration of AI in oncology care decision-making.

DEVELOPMENT OF ROBUST MULTICENTER DATASETS FOR AI TRAINING AND BENCHMARKING

The creation and curation of high-quality multicenter datasets are fundamental for training AI models and establishing reliable benchmarks. High-quality datasets should include diverse patient demographics, comprehensive longitudinal data, and detailed treatment outcomes, ensuring that they reflect the complexities of real-world oncology care. These datasets must accurately represent the diverse real-world data found in EHRs. This effort could involve expanding existing datasets, like those from the GENIE project, or establishing new, bespoke datasets tailored to oncology AI. For example, the EU-funded EUropean Federation for CAncer IMages (EUCAIM) project — a cornerstone of the European Cancer Imaging Initiative — aims to create a federated "AI-ready" infrastructure encompassing over 100,000 cancer cases with multimodal imaging data from distributed repositories across Europe.[41] Given

the need for transparency regarding data provenance, bias minimization, and broad accessibility, the development of oncology AI datasets should be supported by federal initiatives (e.g., National Institutes of Health [NIH] and European Commission [EC]) or disease-specific nonprofits like the American Association for Cancer Research (AACR) or Project Data Sphere.[42] Federal support is crucial to ensure scalability, credibility, and inclusivity, ultimately fostering trust within the oncology community. These datasets should prioritize diverse patient representation to address health care disparities and deliver equitable outcomes across demographic and socioeconomic groups.

PROMOTION OF PRACTICAL ONCOLOGY AI MODEL DEVELOPMENT

Advancing research that develops practical oncology AI models is essential, but must be approached realistically given the challenges in health care environments, such as data privacy concerns, regulatory hurdles, and the need for robust validation in diverse patient populations. Models should prioritize specific, high-impact applications, such as improving diagnosis accuracy, streamlining treatment planning, and enhancing adverse event monitoring. For example, the phased integration of AI-driven decision support tools, such as the Sepsis Watch deep-learning model implemented at Duke University Health System, has demonstrated improved patient outcomes while minimizing disruptions, highlighting the feasibility of a gradual approach.[43] Rather than attempting to cover the entire oncology care continuum, stepwise implementation can make progress more manageable.

RESEARCH ON WORKFLOW INTEGRATION AND HUMAN–AI COLLABORATION

Effective workflow integration requires optimizing human–AI collaboration frameworks in oncology. Developing structured frameworks will define the evolving roles of AI and clinicians, ensuring safety and optimal outcomes. Research should investigate strategies to balance AI autonomy with clinician oversight, facilitating safe incorporation into clinical workflows. Prioritizing user-friendly human–computer interaction will enhance adoption and enable oncologists to interpret AI-generated recommendations for personalized care.

LEVERAGING LESSONS FROM AI-POWERED MEDICAL IMAGING

To maximize the impact of oncology AI models, leveraging insights from AI-powered medical imaging is crucial. In these domains, AI tools have successfully augmented human experts in these domains and are evolving toward greater autonomy. Given the significant computational and data resources required for model training, government agencies such as the Advanced Research Projects Agency for Health, NIH, and EC, alongside industry partners, should provide funding support. Furthermore, these models must incorporate explainable AI features that allow clinicians to understand the rationale behind AI-generated recommendations, which is fundamental for fostering clinical trust and adoption.

ESTABLISHMENT OF GUIDELINES AND BENCHMARKS FOR AI MODEL EVALUATION

Standardized guidelines and benchmarks are imperative for evaluating oncology AI models. These should cover accuracy metrics, transparency, performance thresholds, and validation requirements. Datasets should reflect diverse demographics, and AI models must adhere to high evidence standards. Professional societies such as AACR, American Society of Clinical Oncology, ESMO, CSCO, and American Cancer Society, in collaboration with regulatory bodies such as the FDA and European Medicines Agency, should spearhead guideline development, emphasizing ongoing model updates, ethical considerations, patient privacy, informed consent, and addressing biases.

INCENTIVIZATION OF PROSPECTIVE RANDOMIZED CLINICAL TRIALS FOR ONCOLOGY AI

Oncology AI decision support tools should be seen as a novel paradigm for treatment selection and care management rather than merely diagnostic tools. Unlike diagnostic tools that primarily detect disease presence, decision support tools provide comprehensive guidance on treatment options and care pathways. Similar to drug development, most AI models should undergo prospective RCTs to validate their efficacy and patient benefits. The FDA mandates prospective trials for most therapeutic agents, and real-world data may be considered in specific contexts. Robust clinical evidence is necessary for regulatory approval, reimbursement, incorporation into clinical guidelines, and trust-building among oncologists. To facilitate RCTs of oncology AI models, collaboration and financial support from industry and cooperative groups, such as the European Organisation for Research and Treatment of Cancer in Europe, and the Alliance for Clinical Trials in Oncology, Eastern Cooperative Oncology Group, the American College of Radiology Imaging Network, and Southwest Oncology Group in the United States, are critical. Moreover, developing reimbursement models through Medicare,

European health insurance systems (including national health services and statutory health insurance funds), and other payers is essential, contingent on the proven clinical utility of AI models. The lack of RCTs in oncology AI may be due to the perception that traditional RCTs are not feasible for rapidly evolving AI models. One potential solution to this challenge is the use of adaptive trial designs, which would enable continuous model updates while preserving statistical rigor. In addition, the application of AI-driven tools for automated analysis of EHR data could streamline data collection and analysis, making pragmatic trials a more viable option for evaluating AI models in clinical oncology. On a different note, trials should also evaluate AI's impact on patient engagement, patient understanding, and shared decision-making.

FOSTERING INTERDISCIPLINARY COLLABORATION AND EDUCATION

Successful AI deployment in oncology demands contributions from oncologists, data scientists, ethicists, and policy makers. Collaborative efforts are vital, as is incorporating AI-related education into medical curricula and continuing education for clinicians. Key topics should include machine learning fundamentals, ethical considerations, data privacy, AI model interpretation, and practical clinical integration. An interdisciplinary approach ensures that AI tools are developed with clinical needs, technical feasibility, and ethical considerations in mind, preparing future oncologists to leverage these technologies effectively.

ADDRESSING LEGAL AND LIABILITY CONSIDERATIONS

As AI becomes more integral to clinical decision-making, legal frameworks must evolve to address responsibilities and liabilities. Legal experts, in collaboration with policy makers, should develop guidelines clarifying accountability in cases of AI-assisted decision errors, particularly as systems grow more autonomous. This includes reviewing informed consent requirements for oncology care and clinical trials, as well as evaluating potential impacts on the standard of care.

ADOPTION OF A GLOBAL PERSPECTIVE AND STANDARDIZATION

Cancer is a global health challenge, and AI solutions must be adaptable across international health care systems, with particular attention to the needs of low-resource settings. AI-assisted cervical cancer screening and diagnosis, which can alleviate the need for skilled cytologists and expand screening capabilities to underserved regions through portable AI devices, exemplifies the potential of AI to significantly impact health care in low-resource settings.[44]

To ensure that AI models are applicable in diverse health care environments, it is important to consider variations in resources and regulatory frameworks. Silcox et al. emphasize the need to build the necessary infrastructure to support AI development in global health care settings, focusing on improving data quality, fostering interoperability, and establishing robust governance frameworks.[45] These strategies are crucial for applying AI models effectively across different health care settings, highlighting the importance of international collaborations to establish global standards for oncology AI — including data sharing, model validation, and deployment — to maximize the benefits of AI in improving care quality worldwide. Less stringent regulatory environments, often found in low-resource settings, may also facilitate swifter implementation and broader impact of these AI technologies.

We believe that addressing these strategic actions is essential for bridging the gap between promising AI research and reliable oncology applications for treatment decision-making. A systematic approach to AI implementation will ensure that its transformative potential is realized, providing meaningful benefits to patients, health care providers, and the global oncology community.

Disclosures

Author disclosures are available at ai.nejm.org.

Author Affiliations

[1] Englander Institute for Precision Medicine, Weill Cornell Medicine, New York

[2] CEO Roundtable on Cancer, Cary, NC

References

1. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115-118. DOI: 10.1038/nature21056.

2. Khosravi P, Kazemi E, Imielinski M, Elemento O, Hajirasouliha I. Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images. EBioMedicine 2018;27:317-328. DOI: 10.1016/j.ebiom.2017.12.026.

3. Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models. May 16, 2023 (https://doi.org/10.48550/arXiv.2305.09617). Preprint.

4. Rydzewski NR, Dinakaran D, Zhao SG, et al. Comparative evaluation of LLMs in clinical oncology. NEJM AI 2024;1(5). DOI: 10.1056/aioa2300151.

5. Ng AY, Oberije CJG, Ambrózay É, et al. Prospective implementation of AI-assisted screen reading to improve early detection of breast cancer. Nat Med 2023;29:3044-3049. DOI: 10.1038/s41591-023-02625-9.

6. Dembrower K, Crippa A, Colón E, Eklund M, Strand F, ScreenTrustCAD Trial Consortium. Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study. Lancet Digit Health 2023;5:e703-e711. DOI: 10.1016/S2589-7500(23)00153-X.

7. Lång K, Josefsson V, Larsson A-M, 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 2023;24:936-944. DOI: 10.1016/S1470-2045(23)00298-X.

8. Mikhael PG, Wohlwend J, Yala A, et al. Sybil: a validated deep learning model to predict future lung cancer risk from a single low-dose chest computed tomography. J Clin Oncol 2023;41:2191-2200. DOI: 10.1200/JCO.22.01345.

9. Wang C, Shao J, He Y, et al. Data-driven risk stratification and precision management of pulmonary nodules detected on chest computed tomography. Nat Med 2024;30:3184-3195. DOI: 10.1038/s41591-024-03211-3.

10. Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021;385:283-286. DOI: 10.1056/NEJMc2104626.

11. Lipkova J, Kather JN. The age of foundation models. Nat Rev Clin Oncol 2024;21:769-770. DOI: 10.1038/s41571-024-00941-8.

12. Pai S, Bontempi D, Hadzic I, et al. Foundation model for cancer imaging biomarkers. Nat Mach Intell 2024;6:354-367. DOI: 10.1038/s42256-024-00807-9.

13. Raciti P, Sue J, Retamero JA, et al. Clinical validation of artificial intelligence-augmented pathology diagnosis demonstrates significant gains in diagnostic accuracy in prostate cancer detection. Arch Pathol Lab Med 2023;147:1178-1185. DOI: 10.5858/arpa.2022-0066-OA.

14. Bulten W, Pinckaers H, van Boven H, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncol 2020;21:233-241. DOI: 10.1016/S1470-2045(19)30739-9.

15. Lu MY, Chen B, Williamson DFK, et al. A visual-language foundation model for computational pathology. Nat Med 2024;30:863-874. DOI: 10.1038/s41591-024-02856-4.

16. Vorontsov E, Bozkurt A, Casson A, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. Nat Med 2024;30:2924-2935. DOI: 10.1038/s41591-024-03141-0.

17. Center for Devices, Radiological Health. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. U.S. Food and Drug Administration. 2024 (https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices).

18. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and perils. Ann Surg 2018;268:70-76. DOI: 10.1097/SLA.0000000000002693.

19. Landry G, Kurz C, Traverso A. The role of artificial intelligence in radiotherapy clinical practice. BJR Open 2023;5:20230030. DOI: 10.1259/bjro.20230030.

20. Cilla S, Barajas JEV. Editorial: Automation and artificial intelligence in radiation oncology. Front Oncol 2022;12:1038834. DOI: 10.3389/fonc.2022.1038834.

21. Huynh E, Hosny A, Guthier C, et al. Artificial intelligence in radiation oncology. Nat Rev Clin Oncol 2020;17:771-781. DOI: 10.1038/s41571-020-0417-8.

22. Hosny A, Bitterman DS, Guthier CV, et al. Clinical validation of deep learning algorithms for radiotherapy targeting of non-small-cell lung cancer: an observational study. Lancet Digit Health 2022;4:e657-e666. DOI: 10.1016/S2589-7500(22)00129-7.

23. Lucido JJ, DeWees TA, Leavitt TR, et al. Validation of clinical acceptability of deep-learning-based automated segmentation of organs-at-risk for head-and-neck radiotherapy treatment planning. Front Oncol 2023;13:1137803. DOI: 10.3389/fonc.2023.1137803.

24. Hoang D-T, Dinstag G, Shulman ED, et al. A deep-learning framework to predict cancer treatment response from histopathology images through imputed transcriptomics. Nat Cancer 2024;5:1305-1317. DOI: 10.1038/s43018-024-00793-2.

25. Sinha S, Vegesna R, Mukherjee S, et al. PERCEPTION predicts patient response and resistance to treatment using single-cell transcriptomics of their tumors. Nat Cancer 2024;5:938-952. DOI: 10.1038/s43018-024-00756-7.

26. Gajic ZZ, Deshpande A, Legut M, Imieliński M, Sanjana NE. Recurrent somatic mutations as predictors of immunotherapy response. Nat Commun 2022;13:3938. DOI: 10.1038/s41467-022-31055-3.

27. Stenzl A, Armstrong AJ, Sboner A, et al. Artificial INtelligence to Support Informed DEcision-making (INSIDE) for improved literature analysis in oncology. Eur Urol Focus 2024;10:1011-1018. DOI: 10.1016/j.euf.2024.05.022.

28. Sushil M, Kennedy VE, Mandair D, Miao BY, Zack T, Butte AJ. CORAL: expert-curated oncology reports to advance language model inference. NEJM AI 2024;1(4). DOI: 10.1056/aidbp2300110.

29. Chen S, Kann BH, Foote MB, et al. Use of artificial intelligence chatbots for cancer treatment information. JAMA Oncol 2023;9:1459-1462. DOI: 10.1001/jamaoncol.2023.2954.

30. Kaiser KN, Hughes AJ, Yang AD, et al. Accuracy and consistency of publicly available Large Language Models as clinical decision support tools for the management of colon cancer. J Surg Oncol 2024;130:1104-1110. DOI: 10.1002/jso.27821.

31. Benary M, Wang XD, Schmidt M, et al. Leveraging large language models for decision support in personalized oncology. JAMA Netw Open 2023;6:e2343689. DOI: 10.1001/jamanetworkopen.2023.43689.

32. Marchi F, Bellini E, Iandelli A, Sampieri C, Peretti G. Exploring the landscape of AI-assisted decision-making in head and neck cancer treatment: a comparative analysis of NCCN guidelines and ChatGPT responses. Eur Arch Otorhinolaryngol 2024;281:2123-2136. DOI: 10.1007/s00405-024-08525-z.

33. Unger JM, Shulman LN, Facktor MA, Nelson H, Fleury ME. National estimates of the participation of patients with cancer in clinical research studies based on commission on cancer accreditation data. J Clin Oncol 2024;42:2139-2148. DOI: 10.1200/JCO.23.01030.

34. Ferber D, Hilgers L, Wiest IC, et al. End-to-end clinical trial matching with large language models. July 18, 2024 (https://doi.org/10.48550/arXiv.2407.13463). Preprint.

35. Nievas M, Basu A, Wang Y, Singh H. Distilling large language models for matching patients to clinical trials. J Am Med Inform Assoc 2024;31:1953-1963. DOI: 10.1093/jamia/ocae073.

36. Wong C, Zhang S, Gu Y, et al. Scaling clinical trial matching using large language models: a case study in oncology. August 18, 2023 (https://doi.org/10.48550/arXiv.2308.02180). Preprint.

37. Zakka C, Shad R, Chaurasia A, et al. Almanac — retrieval-augmented language models for clinical medicine. NEJM AI 2024;1(4). DOI: 10.1056/aioa2300068.

38. Choudhury NJ, Lavery JA, Brown S, et al. The GENIE BPC NSCLC cohort: a real-world repository integrating standardized clinical and genomic data for 1,846 patients with non–small cell lung cancer. Clin Cancer Res 2023;29:3418-3428. DOI: 10.1158/1078-0432.CCR-23-0580.

39. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health 2024;6:e12-e22. DOI: 10.1016/S2589-7500(23)00225-X.

40. Cuppen E, Elemento O, Rosenquist R, et al. Implementation of whole-genome and transcriptome sequencing into clinical cancer care. JCO Precis Oncol 2022;6:e2200245. DOI: 10.1200/PO.22.00245.

41. EUCAIM. Cancer Image Europe. 2025 (https://cancerimage.eu/).

42. Green AK, Reeder-Hayes KE, Corty RW, et al. The project data sphere initiative: accelerating cancer research by sharing data. Oncologist 2015;20:464-e20. DOI: 10.1634/theoncologist.2014-0431.

43. Sendak MP, D'Arcy J, Kashyap S, et al. A path for translation of machine learning products into healthcare delivery. Euro Med J Innov 2020. DOI: 10.33590/emjinnov/19-00172.

44. Hou X, Shen G, Zhou L, Li Y, Wang T, Ma X. Artificial intelligence in cervical cancer screening and diagnosis. Front Oncol 2022;12:851367. DOI: 10.3389/fonc.2022.851367.

45. Silcox C, Zimlichmann E, Huber K, et al. The potential for artificial intelligence to transform healthcare: perspectives from international health leaders. NPJ Digit Med 2024;7:88. DOI: 10.1038/s41746-024-01097-6.