BMJ Oncology

Large language models in oncology: a review

David Chen,[1,2] Rod Parsa,[1,3] Karl Swanson,[4] John-Jose Nunez,[5,6] Andrew Critch,[7] Danielle S Bitterman,[8,9,10] Fei-Fei Liu,[11,12] Srinivas Raman [1,2,13]

To cite: Chen D, Parsa R, Swanson K, *et al* . Large language models in oncology: a review. *BMJ Oncology* 2025;**4**:e000759. doi:10.1136/bmjonc-2025-000759

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjonc-2025-000759).

Received 23 January 2025
Accepted 24 April 2025

For numbered affiliations see end of article.

Correspondence to
Dr Srinivas Raman;
srinivas.raman@bccancer.bc.ca

ABSTRACT

Large language models (LLMs) have demonstrated emergent human-like capabilities in natural language processing, leading to enthusiasm about their integration in healthcare environments. In oncology, where synthesising complex, multimodal data is essential, LLMs offer a promising avenue for supporting clinical decision-making, enhancing patient care, and accelerating research. This narrative review aims to highlight the current state of LLMs in medicine; applications of LLMs in oncology for clinicians, patients, and translational research; and future research directions. Clinician-facing LLMs enable clinical decision support and enable automated data extraction from electronic health records and literature to inform decision-making. Patient-facing LLMs offer the potential for disseminating accessible cancer information and psychosocial support. However, LLMs face limitations that must be addressed before clinical adoption, including risks of hallucinations, poor generalisation, ethical concerns, and scope integration. We propose the incorporation of LLMs within compound artificial intelligence systems to facilitate adoption and efficiency in oncology. This narrative review serves as a non-technical primer for clinicians to understand, evaluate, and participate as active users who can inform the design and iterative improvement of LLM technologies deployed in oncology settings. While LLMs are not intended to replace oncologists, they can serve as powerful tools to augment clinical expertise and patient-centred care, reinforcing their role as a valuable adjunct in the evolving landscape of oncology.

CURRENT STATE OF LARGE LANGUAGE MODELS IN MEDICINE

Introduction

Large language models (LLMs) are artificial intelligence (AI) systems focused on the generation of natural language. The field of oncology is well-positioned to benefit from the incorporation of LLM technologies, especially given its emphasis on the synthesis of diverse data types such as clinical, imaging, laboratory and genomic reports integrated with the psychosocial elements of patient-centred medicine. This narrative clinical primer aims to provide a background for the application of LLMs in cancer care and lay the groundwork for their adoption in clinical oncology. Our narrative review adopts a practical approach by offering step-by-step examples of LLM tool integration into clinical oncology workflows as well as discussion of contemporary trends including compound AI systems with human-in-the-loop designs, multi-modal LLMs, and emergent regulatory frameworks to fill the gap between conceptual overviews and clinical realities. We identified relevant studies and potential applications by conducting a comprehensive but non-systematic search of academic databases (PubMed, MEDLINE, EMBASE, Google Scholar) using variations of the keywords 'large language model', 'generative artificial intelligence', 'oncology' and 'cancer'. We also cross-referenced bibliographies of retrieved articles and drew on expert clinical and AI knowledge within the author team to ensure coverage of emerging and notable studies within the scope of large language model development and applications in oncology.

History of NLP and LLMs

Natural language processing (NLP) describes the computer-aided analysis that enables comprehension and generation of human language. In the early 2000s, the first iteration of language generators employed statistical models that estimated the likelihood of the next word in a sequence, based on frequency of occurrence in the training data.[1] To leverage the scale of large natural language datasets such as unstructured text in electronic health records (EHRs), machine learning-based NLP approaches used mathematical models to extract high-level patterns from data to make inferences. The evolution of natural language processing from rule-based algorithms to contemporary large language models is shown as a timeline in figure 1.

Deep learning, inspired by biological neural networks, refers to a sub-field of machine learning models that learn high-level patterns from input data to mimic human-like data processing. Transformer-based LLMs, such as BERT[2] and RoBERTa[3] in the late 2010s as well

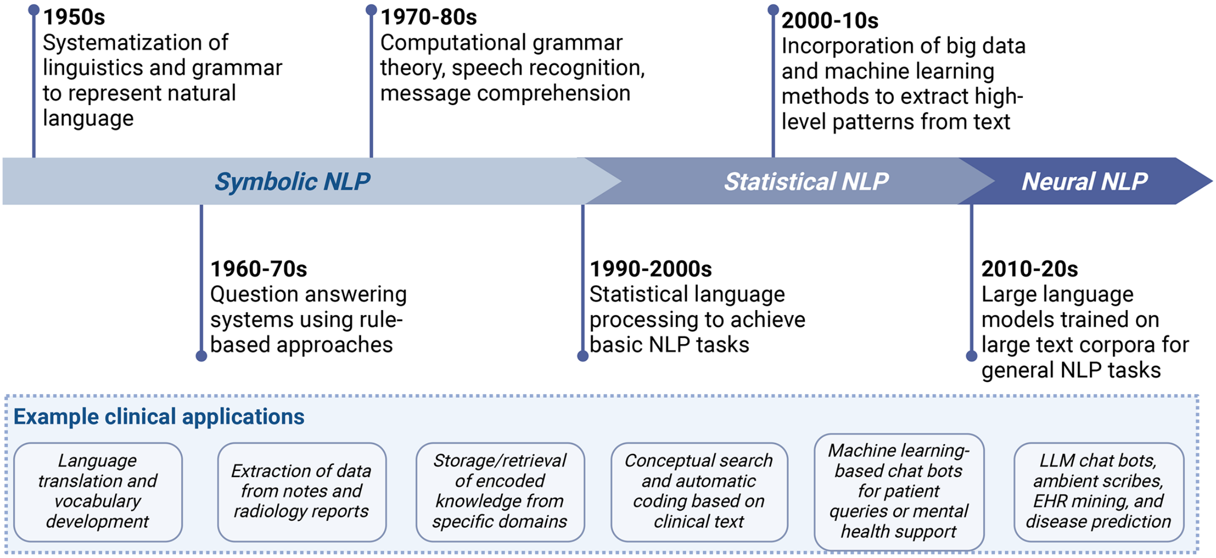
1950s
Systematization of linguistics and grammar to represent natural language
1970-80s
Computational grammar theory, speech recognition, message comprehension
2000-10s
Incorporation of big data and machine learning methods to extract high-level patterns from text
Symbolic NLP
Statistical NLP
Neural NLP
1960-70s
Question answering systems using rule-based approaches
1990-2000s
Statistical language processing to achieve basic NLP tasks
2010-20s
Large language models trained on large text corpora for general NLP tasks
Example clinical applications
Language translation and vocabulary development
Extraction of data from notes and radiology reports
Storage/retrieval of encoded knowledge from specific domains
Conceptual search and automatic coding based on clinical text
Machine learning-based chat bots for patient queries or mental health support
LLM chat bots, ambient scribes, EHR mining, and disease prediction

Figure 1 Timeline of Natural language processing development in medicine.

as GPT in 2020,[4] were critical developments in NLP that exhibited human-like performance in sentiment analysis, question answering, feature extraction, language translation and text summarisation. In clinical contexts, this ability to ingest significant volumes of textual information was first applied to electronic medical records (EMRs); LLMs' ability to process and analyse free text elements enabled clinicians to rapidly create structured datasets, streamline clinical coding, mine academic literature and develop predictive models based on extractable patient features.[5] Oncology-specific applications of NLP models focused initially on case identification (ie, identifying past patients with specific or rare medical characteristics not captured by structured data fields), enhanced cancer staging and codification of staging parameters and the identification of specific clinical outcomes.[6] As conversational capabilities improved, NLP research in oncology turned towards patient-facing applications, including cancer screening campaigns and patient education after diagnosis.[7] The significance of these advances in NLP, and their contribution to human advancement, was most recently underscored by the Nobel Committee's decision to award the 2024 Nobel Prize in Physics to Drs Hopfield and Hinton, pioneers in the development of the artificial neural networks employed in modern LLMs.

LLMs recently entered the public consciousness with OpenAI's 2022 release of ChatGPT 3.5, an AI chatbot based on the GPT LLM that is credited as the fastest-growing consumer software application in history.[8] Today, there are three common categories of clinical LLMs: (1) zero-shot, generalist LLMs that can perform diverse NLP tasks with no pre-training, (2) fine-tuned LLMs that have been trained on custom medical datasets to perform specialised tasks and (3) LLMs equipped with in-context learning or retrieval-augmented generation techniques to enable more accurate recall of medical information from a knowledge base.

Development and validation of clinical LLMs

Standard workflows for the development of clinical LLMs involve model selection, fine-tuning for domain-specific tasks, validation in clinical contexts and deployment in real-world settings as summarised in figure 2. First, clinical LLM applications are typically built using foundational LLMs developed by academic and industry developers, including Google (Gemini), Meta (Llama), OpenAI (GPT) and Anthropic (Claude).[9] Some of the most popular LLMs are open-source (eg, Llama2 by Meta), allowing users to modify their underlying architecture, while others remain proprietary and allow limited interactions through an application programming interface (eg, GPT-4 by OpenAI).

Second, domain-specific LLMs are developed through transfer learning, where foundational LLMs are fine-tuned or trained on specialised datasets, such as clinical notes and EHRs, to function for specialised tasks. Examples of clinical-focused LLMs include Google's Med-PaLM, which focuses on providing high-quality responses to medical questions,[10] and ClinicalBERT, a model which predicts hospital readmission within 30 days by analysing clinical text.[11] Furthermore, a recent study by Wang *et al* demonstrated superior performance of fine-tuned LLMs in the field of radiation oncology, where they outperformed baseline foundational LLMs on tasks related to treatment regimen generation, treatment modality selection and ICD-10 code generation.[12] Likewise, Ferber *et al* demonstrated the superior performance of fine-tuned LLMs compared with baseline foundational LLMs in the field of medical oncology when assessed on management guidelines of pancreatic, colorectal and hepatocellular cancers.[13] LLM outputs can be further optimised by providing model outputs with a few examples through few-shot learning as well as rewarded to steer LLMs towards more truthful and less toxic outputs based on human feedback through reinforcement learning.[14]

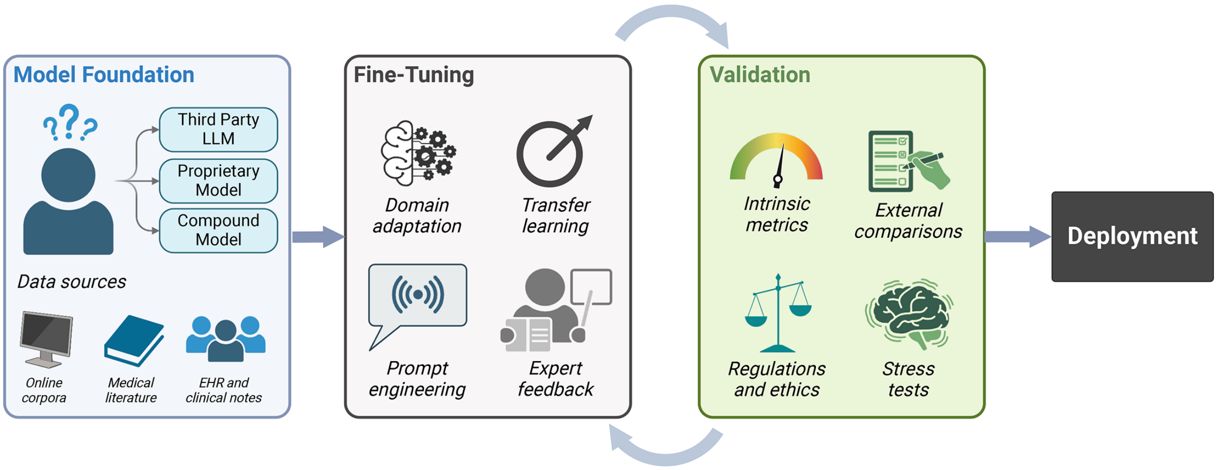

Figure 2 Development and validation process of clinical large language model applications.

Third, LLMs are commonly internally validated through one of three classes of metrics on a human-annotated benchmark dataset, including multiple-classification (classification of text into multiple groups), token-similarity (similarity of generated text with reference text) and question-answering (identifying the answer to a specific question).[15] However, these methods do not capture real-world clinical efficacy, and external validation against expert oncologist decision-making remains essential. Extrinsic evaluations of LLM performance have included comparisons against trained healthcare professionals across test scores or standard of care, measures of clinical efficiency or subjective ratings of performance,[16] with recent recommendations that models be stress-tested via exposure. to diverse clinical scenarios and patient populations to ensure generalisability before real-world deployment.[17] Indeed, a recent cross-sectional study analysing eight LLMs demonstrated an 85% accuracy rate on examination-style multiple choice questions from the American Society of Oncology; however, 81.8% of the incorrect questions were rated as having a medium to high likelihood of moderate to severe harm.[18] Another validation study focused on molecular tumour boards found that LLMs offered equivalent treatment recommendations to clinicians 25% of the time, with a further 37.5% of recommendations as plausible alternative treatments. In generating these recommendations, however, 17% of articles referenced by LLMs were hallucinations, reinforcing the need for clinician supervision.[19] Most recently, researchers have been working towards standardising the human evaluation of LLMs in healthcare. For example, the QUEST framework proposed by Tam was developed through a systematic review of prior evaluation guidelines and addresses gaps in reliability, generalisability and applicability of these guidelines across a variety of medical specialties.[20]

APPLICATIONS OF LARGE LANGUAGE MODELS IN ONCOLOGY

Clinician decision support

Emergent themes of clinician-facing applications of LLMs in oncology include serving as clinical decision support tools for diagnosis, screening and prevention, treatment and management, and automated data extraction and processing for clinician review.[9] Common themes and highlighted examples of potential LLM applications in oncology across a diverse set of application scenarios are shown in figure 3 and online supplemental eTable 1, respectively. In figure 3, we outline three principal themes illustrating how LLMs are being applied in oncology: (1) clinician-facing applications for diagnosis, prognosis, screening, management, data extraction and reference information; (2) patient-facing applications for psychosocial support, communication and reference information; and (3) research facilitation such as screening of trial eligibility, literature evidence synthesis and precision medicine. Notably, this figure highlights the breadth of potential LLM applications grouped by distinct application domains that can help address the multifaceted challenges in modern oncology care.

Diagnosis

LLM-enabled tools have the potential to identify useful patterns from both text-only and multi-modal inputs to recommend clinical diagnoses. In an early evaluation of LLM diagnostic accuracy, Wang *et al* found GPT-4 performed well in generated report structure and clarity but performed worse than physicians in diagnostic accuracy[21] when tested on 109 ultrasound text descriptions of thyroid cancer. However, the Turing test evaluation found that physicians believed that 71% of GPT-generated reports were likely physician-generated, suggesting that GPT and physician-generated reports are largely indistinguishable. Compared with gold-standard clinician annotations, LLMs have shown promising diagnostic performance in exam-style, text-based assessments across several tumour sites including dermatological (85% accuracy),[22] bone (87% accuracy),[23] oropharyngeal (71%

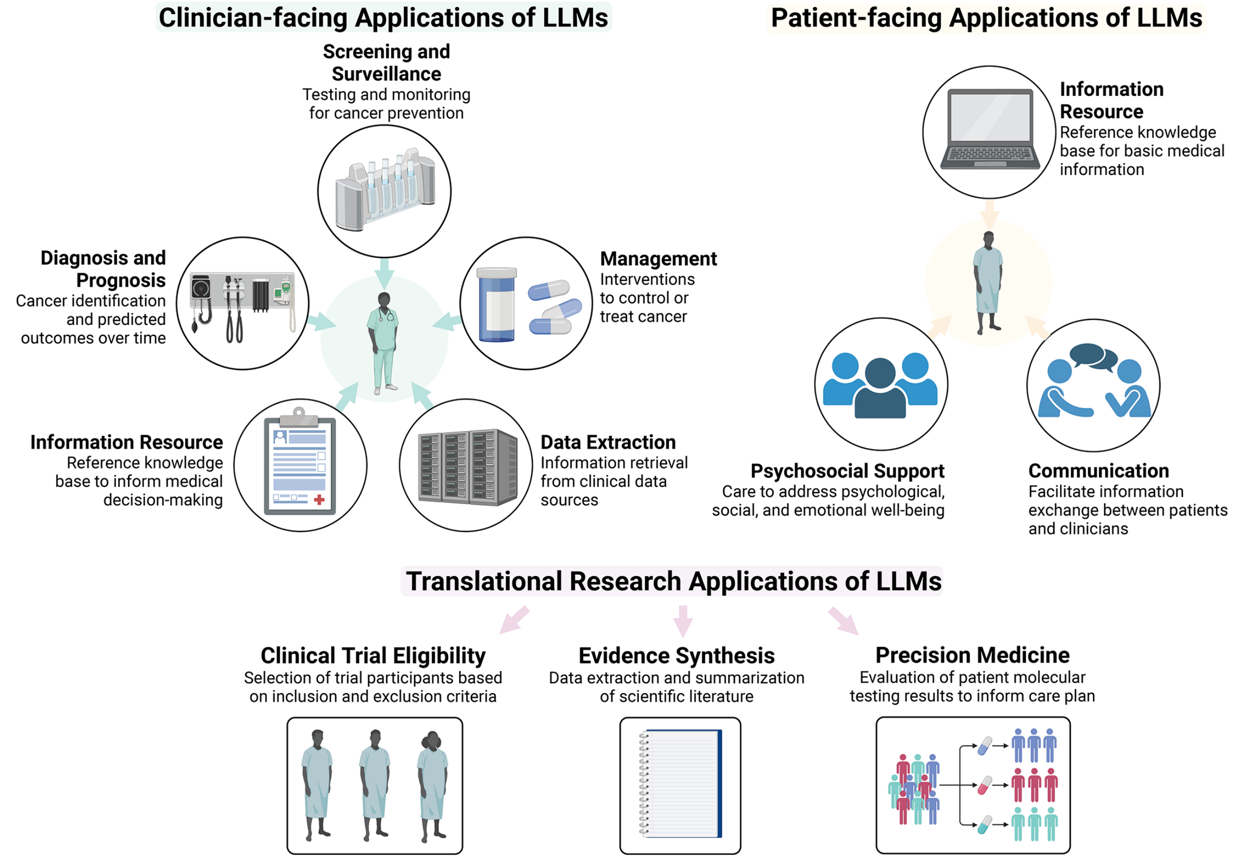
Clinician-facing Applications of LLMs
Screening and Surveillance
Testing and monitoring for cancer prevention
Diagnosis and Prognosis
Cancer identification and predicted outcomes over time
Management
Interventions to control or treat cancer
Information Resource
Reference knowledge base to inform medical decision-making
Data Extraction
Information retrieval from clinical data sources
Patient-facing Applications of LLMs
Information Resource
Reference knowledge base for basic medical information
Psychosocial Support
Care to address psychological, social, and emotional well-being
Communication
Facilitate information exchange between patients and clinicians
Translational Research Applications of LLMs
Clinical Trial Eligibility
Selection of trial participants based on inclusion and exclusion criteria
Evidence Synthesis
Data extraction and summarization of scientific literature
Precision Medicine
Evaluation of patient molecular testing results to inform care plan

Figure 3 Themes of large language model applications in oncology.

accuracy)[24] and neurological cancers (50% accuracy).[25] Notably, the advent of multi-modal LLMs that integrate image processing capabilities is uniquely positioned to analyse both clinical images, including photographs,[22] and multiple radiologic modalities including MRI,[26] CT[27] and ultrasound.[28] While most studies demonstrate human-like performance, there remain concerns about generalisation such as variable diagnostic performance across different skin tones in melanoma,[22] suggesting that the accuracy in minority patient demographics and rare diseases should be evaluated with caution. Similar concerns arise in underrepresented cancer subtypes, where training data scarcity may lead to decreased model performance. Oncologists can use LLM predictions to aid diagnosis but should be cautious about their interpretability and alignment with clinical judgement. However, it is important to note that these findings are derived from early-stage pilot studies in controlled settings and require further validation, especially to assess performance consistency across diverse patient populations, heterogenous practice settings and various cancer diagnoses and stages.

Existing AI-driven approaches in radiology and histopathology have demonstrated high diagnostic accuracy for tasks like tumour localisation and malignancy classification, often through convolutional neural networks or other deep learning architectures that directly analyse imaging data.[29] LLM-based solutions can complement these image-centric models by parsing clinical notes, radiology reports or pathology descriptors, providing a structured synopsis of relevant clinical factors that can enhance diagnostic workflows and communication between radiologists, pathologists and oncologists (60). In this way, LLMs may ultimately function synergistically with established computer vision algorithms to bridge the gap between raw imaging data and patient-level clinical decision-making (61). However, multi-modal LLMs are still in early development, and rigorous testing is needed to evaluate their ability to integrate imaging and text data reliably across diverse patient populations and clinical settings.

Screening and prevention

The integration of LLMs into screening and prevention efforts represents a nascent but rapidly evolving area of research, focusing on text-based knowledge synthesis and guideline-based recommendation.[30] While most AI-based screening applications to date have emphasised image analysis for earlier detection of lesions or nodules, LLMs offer potential in complementary domains like patient risk stratification from EHRs, automated reminder systems for at-risk populations and the generation of patient-specific preventative measures.[31] These text-driven functionalities could be especially valuable for oncologists seeking to optimise large-scale screening programmes or adapt guidelines to individual patient risk profiles.

The utility of LLMs in augmenting decisions related to screening and prevention has been investigated in prostate (85% and 100% accuracy for 30 easy and hard prostate cancer screening questions),[32] colorectal (100% accuracy in 20 colorectal cancer screening questions),[33] breast, ovarian and lung cancer (83% accuracy across 15 select-all-that-apply pan-cancer screening scenarios) contexts.[34] Although these standardised screening evaluations show high accuracy, they are based on predefined question sets rather than real-world clinical scenarios. Chiarelli *et al* tested the reliability of GPT when queried with three prompt variations, showing that there was no difference in accuracy but noted that systematic evaluations of reliability are warranted given the probabilistic nature of LLMs.[32] Despite attempts to ground LLM in evidence-based knowledge such as PubMedBERT and Med-PaLM, oncologists should validate that LLM-based screening recommendations coupled with explanations align with established clinical guidelines and generalise when integrated into patient screening programmes.

Treatment and management

While LLMs can suggest treatment options aligned with established guidelines, studies have found that they may also propose non-concordant treatments, requiring human oversight to align management with patient preferences and evidence-based guidelines.[35] Marchi *et al* found that ChatGPT-3.5 provided accurate suggestions for primary treatment (85.3% accuracy, 100% sensitivity) and adjuvant treatment (96% accuracy, 100% sensitivity) for 68 head and neck cancer cases according to NCCN consensus expert-driven guidelines for cancer management.[36] High sensitivity in treatment recommendations underscores the comprehensive nature of LLM outputs but may lead to over-inclusive lists that require oncologist judgement to refine. For example, in another study, Chen *et al* found that across 104 prompts for 26 pan-cancer diagnoses, GPT-3.5 provided at least 1 NCCN-concordant treatment in 98% of scenarios but also recommended non-concordant treatments in 34.3% of scenarios.[35] Given the occurrence of non-guideline-concordant recommendations, it is critical to underscore that LLM outputs should complement—but not replace—human clinical judgement while future research continues to identify and address the knowledge deficiencies of LLM tools in clinical settings. For the oncologist, this means that LLMs can generate a differential list of treatments for future evaluation of guideline concordance and patient preference, but cannot replace human decision-making. LLMs in specialised oncology tasks have shown mixed performance, with examples in the literature demonstrating that LLMs prescribed chemotherapy protocols with inappropriate dosing (56% accuracy)[37] and were subject to hallucinations when recommending management for immune-related adverse events (44% accuracy).[38] From a medical lens, LLMs may fail to consider important factors such as a patient's comorbid conditions or psychosocio-economic factors that may contraindicate certain regimens in practice. Similarly, while an LLM might produce a seemingly appropriate surgical recommendation, only a trained surgeon or multidisciplinary tumour board can balance tumour resectability, patient preferences, anatomical complexities and the associated risks unique to a specific patient. From a psychosocial lens in palliative scenarios, an LLM's suggestions may overlook family dynamics or cultural values—factors that carry substantial weight in deciding care goals. These examples illustrate how human judgement, guided by clinical expertise and empathy, remains central to comprehensive patient-centred care.

One of the most exciting applications of LLMs in oncology is the recommendation of treatments in complex settings outside of established clinical practice guidelines. These tasks may be ideally suited for LLMs given their ability to process vast amounts of medical literature and patient data to identify patterns that may not be apparent to human experts, but useful for generating novel treatment recommendations in complex and rare cancer cases. In complex breast[39] and colorectal[40] cancer settings, studies have reported 70% and 87% concordance of LLM-generated treatments with tumour board recommendations. Likewise, Chen *et al* found that LLM-generated diagnosis and treatment recommendations of 79 clinical oncology cases with images achieved up to 72% accuracy.[41] However, inaccurate referencing to established guidelines and generation of medically inaccurate outputs with confidence contribute to poor autonomous actionability of LLM-generated recommendations[42]

Data extraction and processing

To enable synthesis of patient data for molecular tumour boards, oncologists may use LLMs to extract key tumour attributes rather than manually extracting this data from the patient EHR.[43] Preston *et al* demonstrated that LLM-based data extraction of tumour attributes, including tumour site and the widely-adopted TNM cancer staging classification, can achieve 94–99% AUROC performance and generalise across multiple health systems and state registries.[43] Notably, LLM-enabled data extractions for well-defined categories, such as TNM stage, can even correct human errors on expert review.[43]

Beyond clinical features, automated extraction of social and behavioural determinants from clinical data[44] can be applied to address several humanistic elements of the cancer patient experience, including identifying at-risk patients who lack advance directives, surrogate decision-makers and decision capacity,[45] and recommending online resources to address psychosocial needs.[46] Instead of prompting LLMs to generate 'black box' predictions from clinical data, oncologists can prompt LLMs to extract important data points from large-scale clinical text, allowing oncologists to prioritise expert synthesis of medical knowledge and patient care over non-patient facing, administrative tasks. While the high performance in structured data extraction is encouraging, variability in EHR systems and documentation practices across institutions may affect extraction performance in external settings, motivating the need for robust, multi-centre evaluations to confirm generalisability in real-world settings.

Patient-facing applications

The familiarity and accessibility of chatbot LLMs with near-human levels of language competency underscore their potential as patient-facing health information resources and supportive management tools to help address patient educational and psychosocial factors of cancer care.

Health information resource

The accurate performance of conversational chatbots on standardised benchmarks of medical competency, such as the USMLE,[47] and common patient queries about cancer[48 49] suggests that LLM applications may serve as a readily accessible, supplementary patient resource for cancer information. Beyond responding to clinician-level queries, cross-sectional studies of ChatGPT-4 reported high accuracy and alignment to oncologists or guidelines if available, when tested on general patient-level questions about genetic counselling,[50] breast,[51] lung,[52] colon[53] and pancreatic[54] cancers. Moreover, pilot evaluations of LLMs for language translation[55] and biomedical text simplification[56] are emergent research directions of clinical LLMs that can facilitate patient education in oncology. LLMs can provide oncology knowledge as an accessible patient resource.

However, we caution that the variable medical accuracy across various cancers and topics,[44] risks of misinterpretation, oversimplification of complex medical information, propagation of outdated or non-personalised advice and decreased readability of chatbot-generated responses compared with physicians may collectively pose serious risks to deploying patient-facing LLMs in real-world settings until effective safeguards for accuracy and misinformation are implemented. Despite the positive results of these pilot studies, oncologists should explain to patients that LLM tools may (1) generate unreliable and outdated information that can lead to harm, (2) fail to personalise recommendations to the individual patient, (3) harbour inherent biases based on their training datasets and (4) provide limited protections to personal health information privacy and security that have yet to be systematically regulated.

Supportive management

Conversational LLMs, known as chatbots, may act as a complementary agent for psychosocial and emotional support in oncology. Chatbots can provide empathetic responses to online patient questions about general medicine[16] and cancer-specific[48] topics in non-inferiority evaluations compared with physicians, supporting their use in generating empathetic template responses under clinician oversight when integrated into patient health portals.[57] LLM tools also pose potential to improve patient communication during post-treatment care, such as improving dialogue rates for patients experiencing oncological aphasia.[58] Combining LLM-enabled language competencies with hardware, such as assistive robots with functional language and physical capabilities, is a promising development towards more human-like levels of emotional connection.[59]

Oncologists should establish clear protocols defining the scope of chatbot use that encompass two major components: development-focused guidelines and patient-facing usage guidelines. From a development perspective, these protocols can inform model creators and industry partners about the clinical and ethical parameters expected in an oncology setting, helping to ensure that chatbot features—such as data handling, language style and management plans—are compatible with current standards of care and privacy regulations. In parallel, usage guidelines aimed at clinicians and patients will clarify the chatbot's intended purposes (eg, providing supplementary educational information, screening for psychosocial support needs or summarising care instructions), limitations (eg, lack of personalised medical advice, potential for erroneous responses) and recommended follow-up actions. This dual approach enables a coordinated effort to shape the chatbot's capabilities during development while also providing straightforward guidelines that support safe, consistent and beneficial interactions between patients, oncology teams and AI-based tools.

Patient perceptions of LLMs

Recent studies show that patients often perceive AI-generated health advice, such as from ChatGPT, as helpful and empathetic, with evidence that users sometimes rate these responses even more favourably than physician-written answers.[60] However, research indicates that patients have only moderately high trust in chatbot responses—Nov *et al* (2023) report average trust scores of around 3.4 out of 5, with confidence falling as question complexity increases.[61] Platt *et al* (2024) similarly found that patients' comfort using ChatGPT for healthcare queries was below mid-range, highlighting accuracy and privacy concerns.[62] A notable risk factor is that lay users cannot always identify when an LLM's advice is inaccurate or outdated, underscoring the potential for harm if chatbots are used without adequate oversight.[60] Nevertheless, public surveys of online users suggest that, despite these reservations, members of the public, including patients, show willingness to adopt AI health tools in the future if privacy, safety and transparency standards are better established.[63 64]

Facilitating and synthesising oncology research

Automated processing of unstructured text is a unique competency of LLMs that can be used to facilitate translational research in oncology. For example, LLMs can structure clinical trial eligibility criteria for cancer patients that achieve moderate performance compared with physician recommendations, with mixed reports of both high false positives[65] and high false negatives.[66] Similar to medical tumour board recommendations, LLMs applied to clinical trial recommendation should be used to generate an

Table 1 How large language models are being used in oncology

Domain	Application
Oncologist	▶ Generate differential diagnoses based on patient clinical notes and data for oncologist review ▶ Prognosticate patient based on risk and survival as a supportive tool ▶ Provide cancer screening information based on established guidelines ▶ Generate treatment recommendations for oncologist review ▶ Extraction of key cancer attributes from clinical text to inform clinical decision-making ▶ Generating summaries of clinical notes, consultations and diagnostic reports
Cancer Patient	▶ Patient health information resource with clinician oversight ▶ Language translation ▶ Biomedical text simplification ▶ Psychosocial and emotional support and counselling
Cancer Research	▶ Clinical trial matching for eligible cancer patients ▶ Extraction of trial outcomes from literature for clinician education ▶ Extraction of mutation and clinical data for precision oncology ▶ Literature mining for drug synergies
Administration	▶ Draft work communications and patient messages ▶ Transcribe and summarise patient and clinician meetings ▶ Generate pre-filled administrative paperwork ▶ Copy-edit and format administrative paperwork

inclusive selection of potentially eligible trials for oncologists to prioritise in their final recommendation.

Translation of oncology clinical trial results into actionable clinical recommendations requires expert synthesis of scientific knowledge prone to time lag between discovery and implementation.[67] To address this problem, LLM systems such as SEETrials have demonstrated proficiency (96% specificity, 94% sensitivity) in automated extraction of intervention outcomes associated with cancer trials reported in conference abstracts,[68] enabling oncologists to glean early insights into the safety and efficacy of novel interventions.

Applied to precision oncology, LLMs have seen success in automating data extraction of driver mutations and clinical data from EHRs to evaluate the prognostic value of these mutations and functional effects.[69] Likewise, literature mining by LLMs may be useful as a research tool for drug synergy predictions applied to complex cancer patient scenarios.[70] Academic oncologists can stay up to date on advancements in LLM applications by engaging with emerging LLM in oncology research, attending interdisciplinary conferences and collaborating with AI experts to safely and effectively integrate these tools into modern oncology practice (table 1).

LIMITATIONS AND FUTURE DIRECTIONS OF LARGE LANGUAGE MODELS

Technical limitations

The implementation of LLMs in medicine is limited both by AI-intrinsic and clinical workflow challenges. Training and testing models on sparse, incompletely labelled datasets risks generating insights that fail to generalise to broad use cases. Furthermore, LLM-based models are prone to generating convincing 'hallucinations', content that is entirely nonsensical or unfaithful to the provided source content,[9] which must be either actively detected or accounted for by healthcare providers. For example, Chen *et al* reported a 12.5% hallucination rate by LLM chatbots asked to generate cancer treatment information.[35] As a result, there has been increasing focus on the development of LLM safeguards that prevent the generation of health disinformation.[71] Other studies have found that LLM chatbots may provide inconsistent responses to the same question asked several times,[72] raising questions about their reliability and reproducibility. Finally, most modern LLM chatbots are trained on fixed time windows—for example, ChatGPT 3.5's initial release was trained on data up to September 2021. This training method may exclude more recent advances and risks generating outdated responses, especially in rapidly evolving fields such as oncology.

Ethical limitations

Collaborations between international institutions, as evidenced by a partnership between the WHO and the European Parliament, have garnered interest in producing ethical guidelines and frameworks for the application of AI in healthcare.[73] These frameworks emphasise the preservation of patient autonomy, technological transparency, accountability and inclusiveness. Until international standards are formalised, significant discussion has focused on adherence to existing national standards, such as the US Health Insurance Portability and Accountability Act (HIPAA). Models that employ identifiable patient information may risk inadvertently storing or disclosing sensitive information in violation of HIPAA regulations or may be vulnerable to cybersecurity breaches.[74] Furthermore, the use of identifiable patient information in the pre-training process may violate principles around informed consent and rights-of-data, especially if previously anonymised data can be re-identified.

Kapsali *et al* have highlighted discrepancies between the aforementioned principles and ChatGPT's features,[75] pointing to its black-box technology and insufficient documentation as causes for concern. Unsurprisingly, it has been shown that patients may prefer human judgement and expertise over AI-generated recommendations,[76] especially when existing legal frameworks around liability and medical malpractice fail to fully address AI-driven outcomes.[77] Indeed, emerging research has revealed the potential for LLMs to perpetuate societal biases, such as race, even without the explicit input of demographic data.[78] A recent systematic review highlighted the prevalence of gender and racial bias in medical LLMs, describing model outputs that leaned on stereotypical gender roles, used gendered language and underrepresented women while overvaluing male competence. Mitigation strategies that limit these biases have relied primarily on prompt engineering methods with varying effectiveness, and there exists a need for standardised metrics that systematically reduce bias in all stages of model development and implementation.[79] Given the historic and ongoing issues with diversity, as evidenced by clinical trial participation for example,[80] there exists an imperative for cancer researchers to interrogate oncology-focused LLMs for data-driven biases.

Resource limitations

Although the economic challenges of LLM deployment in healthcare systems remain underexplored, efforts have been made to estimate the computational, energy and financial costs associated with model development and implementation.[81] The cost burden of LLMs in medicine is based on model training and fine-tuning, integration into existing electronic health systems, input data types and latency requirements. Carbon footprint estimates have been inferred at each lifecycle point of LLMs—pretraining, fine-tuning, and inference—with the inference stage dominating the long-term environmental impact.[82] For example, a single query of a fine-tuned GPT-3 model uses 0.04 kW-h of electricity per 100 pages of generated text,[82] a power consumption that could rapidly balloon when scaled across thousands of patients that each warrant numerous clinician queries to medical records. However, there have been countervailing debates on the economic implications of LLM deployment, with some researchers proposing that the expected cost efficiencies and sustainable practices conferred by automation far outweigh the negative impacts.[83] Cancer care in particular is likely to benefit from these efficiencies, given the longstanding capacity constraints as the number of cancer patients outpaces the number of clinicians available to support them.[84]

Future directions

Emerging directions in LLM implementation will involve advances in technology and model complexity, cohesive regulatory and standardisation frameworks, greater emphasis placed on inclusivity and equity and the incorporation of clinician and patient feedback into development cycles to better align model outputs with desired outcomes (table 2).

Recent research has proposed a paradigm shift from increasing resource usage towards designing specialised component tools that work together as a compound AI system.[85] Roadmaps for the design of compound AI systems in oncology can be informed by previous system designs used for chemical synthesis[86] and geometry theorem proofs.[87] There exists additional potential for multimodal AI models that integrate oncology-focused models in collaboration with other disciplines, such as radiology and pathology, with potential to streamline tasks such as summarisation, patient education, differential diagnosis generation and interdisciplinary collaboration.[88] In the era of precision medicine, the integration of multimodal datasets which combine textual data from medical records, oncology clinic visits, multidisciplinary discussions, genomic pathology reports and imaging findings is likely to enhance patient-specific recommendations. Fine-tuning techniques such as prompt engineering have also shown particular promise; prompts that provide additional clinical context have been shown to generate treatment plans

Table 2 Limitations to large language model adoption in oncology and potential solutions

Limitations	Potential solutions
Technical	► Comprehensive data labelling requirements that employ diverse clinical and patient data ► Continually shifting training window that captures new studies and advancements as they are released ► LLM safeguards that detect and prevent hallucinations and health disinformation ► Development of AI reliability metrics to track output consistency
Ethical	► Development of consensus ethical frameworks around the use of AI in clinical contexts ► Inclusion of both patient and clinician feedback on a continual basis, both into regulatory frameworks and model development ► Adoption of open-source and transparent development, along with clear documentation, to avoid the perception of a black-box technology ► Continual research and benchmarking of societal biases found in LLM inputs and outputs ► Anonymisation of all patient information by LLMs to preserve privacy and security
Economic	► Careful consideration of build vs buy options for institutions considering LLM deployment ► Investment into sustainable energy options that fuel LLM energy consumption while minimising carbon footprint ► Judicious use of LLM model queries, limited to use cases where it improves clinical outcomes ► Comprehensive accounting of the cost efficiencies conferred by LLM deployment (eg, human resources)

AI, artificial intelligence; LLM, Large language model.

in concordance with cancer care guidelines.[89] From a model evaluation perspective, some are proposing more realistic evaluation frameworks using agent-based modelling to create AI structured clinical examinations ('AI-SCE') that test varying degrees of self-governance in dynamic environments.[17]

Deployment of AI in healthcare settings has also engendered ongoing discussion around maximising benefit and minimising risk through standardised regulation. While internationally recognised governance mechanisms for AI in healthcare do not currently exist,[90] there has been increasing consensus in focused areas of interest. For instance, a diverse set of academic, industry, funding agency and publishers has proposed the implementation of Findable, Accessible, Interoperable and Reusable Data Principles to define good data stewardship practices and facilitate data sharing that may be adopted in the precision oncology community.[91] Standardised reporting guidelines for biomedical-focused LLM research, such as TRIPOD+LLM for primary research involving LLMs,[92] QUEST for human evaluation of LLMs[20] or CONSORT-AI for clinical trials involving AI,[93] aim to improve the consistency, reliability and verifiability of future advancements.

Patient-centric regulations for patient privacy, medical malpractice and informed consent lag behind technical innovation. To date, this has only been addressed within the confines of individual partnerships (eg, Google's HIPAA-compliant generative AI at the Mayo Clinic) and not at scale. With time, the adoption of widespread data sharing and ethics frameworks will permit existing models to train on large, open-source and more representative datasets while considering important principles of data privacy and right to use, intellectual property and risk of harm. This will in turn enable the development of accurate, purpose-built LLMs for cancer-specific applications, both via open-source collaborations (eg, RadOnc-GPT, CancerGPT)[70 94] and industry-sponsored offerings (eg, CareIntellect by GE Healthcare, Watson for Oncology by IBM and Intellispace Oncology by Phillips).

The design of human-in-the-loop training cycles, where LLMs are fine-tuned by engineers with clinician feedback, can optimise LLM outputs that are more clinically useful to the oncology care team. Explicit and implicit patient feedback may help LLMs better align outputs with the unique psychosocial experiences of cancer patients. Explicit feedback involves reports from patient users after interactions with the LLM application, such as numerical scores or binary ratings of the text output from the tool. Implicit feedback involves indirect reports from patient users based on user interactions and behaviour patterns with the LLM application, such as monitoring user reactions to LLM outputs through engagement time or characteristics of follow-up queries. The design of oncology LLM applications requires consideration of both emotional and cognitive empathy in order to address the psychosocial demands of the cancer patient experience and prioritise the clinical competencies that impact patient clinical outcomes.

CONCLUSION

LLMs have the potential to impact all aspects of cancer care due to their human-like ability to understand and generate natural language. Clinician and patient-facing applications of LLMs in oncology, ranging from diagnosis, management and emotional support, serve as promising directions of LLM research in oncology. Coupled with emergent multi-modal capabilities and integration into compound AI systems, state-of-the-art LLM applications are well-positioned to move towards addressing clinical and translational research challenges in oncology. However, there remain several limitations of LLM deployment in clinical practice, including medical accuracy, privacy and ethics, which remain to be systematically addressed in order to facilitate their widespread adoption. Validation of LLM applications should demonstrate sufficient benefit in real-world clinical settings necessary to prioritise patient care outcomes in oncology. While the mixed performance of LLMs across oncology-related competencies may suggest that oncologists will not be replaced by AI solutions anytime soon, LLM-based tools may serve as useful clinician decision support and patient-facing management tools under clinician oversight.

Author affiliations
[1]Radiation Medicine Program, Princess Margaret Hospital Cancer Centre, Toronto, Ontario, Canada
[2]Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada
[3]Michael G. DeGroote School of Medicine, McMaster University, Stockholm, Ontario, Canada
[4]Department of Medicine, University of California–San Francisco, San Francisco, California, USA
[5]Department of Psychiatry, University of British Columbia, Vancouver, British Columbia, Canada
[6]BC Cancer Agency, Vancouver, British Columbia, Canada
[7]Center for Human-Compatible Artificial Intelligence, Department of Electrical Engineering and Computer Sciences, UC Berkeley, Berkeley, California, USA
[8]Artificial Intelligence in Medicine (AIM) Program, Mass General Brigham, Harvard Medical School, Boston, Massachusetts, USA
[9]Department of Radiation Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA
[10]Computational Health Informatics Program, Boston Children's Hospital, Boston, Massachusetts, USA
[11]Radiation Medicine Program, Princess Margaret Hospital, Toronto, Ontario, Canada
[12]Department of Radiation Oncology, University of Toronto, Toronto, Ontario, Canada
[13]Radiation Oncology, BC Cancer - Vancouver, Vancouver, British Columbia, Canada

Contributors DC, RP and SR conceptualised, analysed the data, interpreted the data, drafted the manuscript and edited the manuscript. KS, JN, DB and FL edited the manuscript. DC and SR are the guarantors of the manuscript.

Funding This work was partially supported by the CARO CROF studentship and the Robert L. Tundermann and Christine E. Couturier philanthropic funds.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, conduct, reporting or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been

ORCID iD

Srinivas Raman http://orcid.org/0000-0001-5688-9628

REFERENCES

1 Hadi MU, tashi qasem al, Qureshi R, *et al*. A survey on large language models: applications, challenges, limitations, and practical usage. [Preprint] 2023.
2 Devlin J, Chang MW, Lee K, *et al*. BERT: pre-training of deep bidirectional transformers for language understanding. [Preprint] 2019. Available: http://arxiv.org/abs/1810.04805
3 Liu Y, Ott M, Goyal N, *et al*. RoBERTa: a robustly optimized bert pretraining approach. [Preprint] 2019. Available: http://arxiv.org/abs/1907.11692
4 Brown TB, Mann B, Ryder N, *et al*. Language models are few-shot learners. [Preprint] 2020. Available: http://arxiv.org/abs/2005.14165
5 Locke S, Bashall A, Al-Adely S, *et al*. Natural language processing in medicine: A review. *Trends in Anaesthesia and Critical Care* 2021;38:4–9.
6 Yim W-W, Yetisgen M, Harris WP, *et al*. Natural language processing in oncology: a review. *JAMA Oncol* 2016;2:797–804.
7 Raynaud C, Wu D, Levy J, *et al*. Patients facing large language models in oncology: a narrative review. *JCO Clin Cancer Inform* 2024;8:e2400149.
8 Wu T, He S, Liu J, *et al*. A brief overview of chatgpt: the history, status quo and potential future development. *IEEE/CAA J Autom Sinica* 2023;10:1122–36.
9 Thirunavukarasu AJ, Ting DSJ, Elangovan K, *et al*. Large language models in medicine. *Nat Med* 2023;29:1930–40.
10 Singhal K, Azizi S, Tu T, *et al*. Large language models encode clinical knowledge. *Nature* 2023;620:172–80.
11 Huang K, Altosaar J, Ranganath R. Clinicalbert: modeling clinical notes and predicting hospital readmission. [Preprint] 2020.
12 Wang P, Liu Z, Li Y, *et al*. Fine-Tuning Large Language Models for Radiation Oncology, A Specialized Health Care Domain. *Int J Radiat Oncol Biol Phy* 2024;120:e664.
13 Ferber D, Wiest IC, Wölflein G, *et al*. GPT-4 for Information retrieval and comparison of medical oncology guidelines. *NEJM AI* 2024;1.
14 Ouyang L, Wu J, Jiang X, *et al*. Training language models to follow instructions with human feedback. [Preprint] 2022.
15 Hu T, Zhou XH. Unveiling llm evaluation focused on metrics: challenges and solutions. [Preprint] 2024. Available: http://arxiv.org/abs/2404.09135
16 Ayers JW, Poliak A, Dredze M, *et al*. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med* 2023;183:589–96.
17 Mehandru N, Miao BY, Almaraz ER, *et al*. Evaluating large language models as agents in the clinic. *NPJ Digit Med* 2024;7:84.
18 Longwell JB, Hirsch I, Binder F, *et al*. Performance of large language models on medical oncology examination questions. *JAMA Netw Open* 2024;7:e2417641.
19 Berman E, Sundberg Malek H, Bitzer M, *et al*. Retrieval augmented therapy suggestion for molecular tumor boards: algorithmic development and validation study. *J Med Internet Res* 2025;27:e64364.
20 Tam TYC, Sivarajkumar S, Kapoor S, *et al*. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ Digit Med* 2024;7:258.
21 Wang Z, Zhang Z, Traverso A, *et al*. Assessing the role of GPT-4 in thyroid ultrasound diagnosis and treatment recommendations: enhancing interpretability with a chain of thought approach. *Quant Imaging Med Surg* 2024;14:1602–15.
22 Cirone K, Akrout M, Abid L, *et al*. Assessing the utility of multimodal large language models (gpt-4 vision and large language and vision assistant) in identifying melanoma across different skin tones. *JMIR Dermatol* 2024;7:e55508.
23 Yang F, Yan D, Wang Z. Large-Scale assessment of ChatGPT's performance in benign and malignant bone tumors imaging report diagnosis and its potential for clinical applications. *J Bone Oncol* 2024;44:100525.
24 Sievert M, Aubreville M, Mueller SK, *et al*. Diagnosis of malignancy in oropharyngeal confocal laser endomicroscopy using GPT 4.0 with vision. *Eur Arch Otorhinolaryngol* 2024;281:2115–22.
25 Horiuchi D, Tatekawa H, Shimono T, *et al*. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. *Neuroradiology* 2024;66:73–9.
26 Rajendran P, Chen Y, Qiu L, *et al*. Auto-delineation of Treatment Target Volume for Radiation Therapy Using Large Language Model-Aided Multimodal Learning. *Int J Radiat Oncol Biol Phy* 2025;121:230–40.
27 Sun D, Hadjiiski L, Gormley J, *et al*. Outcome Prediction Using Multi-Modal Information: Integrating Large Language Model-Extracted Clinical Information and Image Analysis. *Cancers (Basel)* 2024;16:2402.
28 Guo Y, Wan Z. Performance evaluation of multimodal large language models (LLaVA and GPT-4-based chatGPT) in medical image classification tasks. 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI), IEEE; 541–3. Orlando, FL, USA.
29 Campanella G, Hanna MG, Geneslaw L, *et al*. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med* 2019;25:1301–9.
30 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44–56.
31 Koh D-M, Papanikolaou N, Bick U, *et al*. Artificial intelligence and machine learning in cancer imaging. *Commun Med (Lond)* 2022;2:133.
32 Chiarelli G, Stephens A, Finati M, *et al*. Adequacy of prostate cancer prevention and screening recommendations provided by an artificial intelligence-powered large language model. *Int Urol Nephrol* 2024;56:2589–95.
33 Atarere J, Naqvi H, Haas C, *et al*. Applicability of Online Chat-Based Artificial Intelligence Models to Colorectal Cancer Screening. *Dig Dis Sci* 2024;69:791–7.
34 Nguyen D, Swanson D, Newbury A, *et al*. Evaluation of ChatGPT and Google Bard Using Prompt Engineering in Cancer Screening Algorithms. *Acad Radiol* 2024;31:1799–804.
35 Chen S, Kann BH, Foote MB, *et al*. Use of Artificial Intelligence Chatbots for Cancer Treatment Information. *JAMA Oncol* 2023;9:1459–62.
36 Marchi F, Bellini E, Iandelli A, *et al*. 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;281:2123–36.
37 Erdat EC, Yalciner M, Urun Y. n.d. Accuracy and usability of artificial intelligence chatbot generated chemotherapy protocols. *Future Oncol*:1–6.
38 Burnette H, Pabani A, von Itzstein MS, *et al*. Use of artificial intelligence chatbots in clinical management of immune-related adverse events. *J Immunother Cancer* 2024;12:e008599.
39 Sorin V, Klang E, Sklair-Levy M, *et al*. Large language model (ChatGPT) as a support tool for breast tumor board. *NPJ Breast Cancer* 2023;9:44.
40 Choo JM, Ryu HS, Kim JS, *et al*. Conversational artificial intelligence (chatGPTTM) in the management of complex colorectal cancer patients: early experience. *ANZ J Surg* 2024;94:356–61.
41 Chen D, Huang RS, Jomy J, *et al*. Performance of Multimodal Artificial Intelligence Chatbots Evaluated on Clinical Oncology Cases. *JAMA Netw Open* 2024;7:e2437711.
42 Benary M, Wang XD, Schmidt M, *et al*. Leveraging Large Language Models for Decision Support in Personalized Oncology. *JAMA Netw Open* 2023;6:e2343689.
43 Preston S, Wei M, Rao R, *et al*. Toward structuring real-world data: Deep learning for extracting oncology information from clinical text with patient-level supervision. *Patterns (N Y)* 2023;4:100726.
44 Guevara M, Chen S, Thomas S, *et al*. Large language models to identify social determinants of health in electronic health records. *NPJ Digit Med* 2024;7:6.
45 Song J, Topaz M, Landau AY, *et al*. Using natural language processing to identify acute care patients who lack advance directives, decisional capacity, and surrogate decision makers. *PLoS ONE* 2022;17:e0270220.
46 Leung YW, Park B, Heo R, *et al*. Providing Care Beyond Therapy Sessions With a Natural Language Processing-Based Recommender

System That Identifies Cancer Patients Who Experience Psychosocial Challenges and Provides Self-care Support: Pilot Study. *JMIR Cancer* 2022;8:e35893.
47 Gilson A, Safranek CW, Huang T, *et al*. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med Educ* 2023;9:e45312.
48 Chen D, Parsa R, Hope A, *et al*. Physician and Artificial Intelligence Chatbot Responses to Cancer Questions From Social Media. *JAMA Oncol* 2024;10:956.
49 Pan A, Musheyev D, Bockelman D, *et al*. Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries About Cancer. *JAMA Oncol* 2023;9:1437–40.
50 Patel JM, Hermann CE, Growdon WB, *et al*. ChatGPT accurately performs genetic counseling for gynecologic cancers. *Gynecol Oncol* 2024;183:115–9.
51 Liu HY, Alessandri Bonetti M, De Lorenzi F, *et al*. Consulting the Digital Doctor: Google Versus ChatGPT as Sources of Information on Breast Implant-Associated Anaplastic Large Cell Lymphoma and Breast Implant Illness. *Aesth Plast Surg* 2024;48:590–607.
52 Rahsepar AA, Tavakoli N, Kim GHJ, *et al*. How AI Responds to Common Lung Cancer Questions: ChatGPT vs Google Bard. *Radiology* 2023;307:e230922.
53 Emile SH, Horesh N, Freund M, *et al*. How appropriate are answers of online chat-based artificial intelligence (ChatGPT) to common questions on colon cancer? *Surgery* 2023;174:1273–5.
54 Moazzam Z, Cloyd J, Lima HA, *et al*. Quality of ChatGPT Responses to Questions Related to Pancreatic Cancer and its Surgical Care. *Ann Surg Oncol* 2023;30:6284–6.
55 NLLB Team. Scaling neural machine translation to 200 languages. *Nature* 2024;630:841–6.
56 Swanson K, He S, Calvano J, *et al*. Biomedical text readability after hypernym substitution with fine-tuned large language models. *PLOS Digit Health* 2024;3:e0000489.
57 Garcia P, Ma SP, Shah S, *et al*. Artificial Intelligence–Generated Draft Replies to Patient Inbox Messages. *JAMA Netw Open* 2024;7:e243201.
58 Zeng Y, Tang Q, Chen S, *et al*. Integration of a large language model with augmentative and alternative communication tool for oncological aphasia rehabilitation. *Asia Pac J Oncol Nurs* 2024;11:100344.
59 Kim CY, Lee CP, Mutlu B. Understanding large-language model (llm)-powered human-robot interaction. 2024.
60 Armbruster J, Bussmann F, Rothhaas C, *et al*. “Doctor ChatGPT, Can You Help Me?” The Patient’s Perspective: Cross-Sectional Study. *J Med Internet Res* 2024;26:e58831.
61 Nov O, Singh N, Mann D. Putting chatgpt’s medical advice to the (turing) test: survey study. *JMIR Med Educ* 2023;9:e46939.
62 Platt J, Nong P, Smiddy R, *et al*. Public comfort with the use of ChatGPT and expectations for healthcare. *J Am Med Inform Assoc* 2024;31:1976–82.
63 Chen SY, Kuo HY, Chang SH. Perceptions of ChatGPT in healthcare: usefulness, trust, and risk. *Front Public Health* 2024;12:1457131.
64 Choudhury A, Shamszare H. Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis. *J Med Internet Res* 2023;25:e47184.
65 Hung T, Kuperman G, Sherman EJ, *et al*. Performance of a trained large language model to provide clinical trial recommendation in a head and neck cancer population. *JCO* 2024;42:11081.
66 Wong C, Zhang S, Gu Y, *et al*. Scaling clinical trial matching using large language models: a case study in oncology. [Preprint] 2023. Available: http://arxiv.org/abs/2308.02180
67 Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. *J R Soc Med* 2011;104:510–20.
68 Lee K, Paek H, Huang L-C, *et al*. Seetrials: leveraging large language models for safety and efficacy extraction in oncology clinical trials. *SSRN* [Preprint] 2024.
69 Elsamahy EA, Ahmed AE, Shoala T, *et al*. Deep-GenMut: Automated genetic mutation classification in oncology: A deep learning comparative study. *Heliyon* 2024;10:e32279.
70 Li T, Shetty S, Kamath A, *et al*. CancerGPT for few shot drug pair synergy prediction using large pretrained language models. *npj Digit Med* 2024;7:40.
71 Menz BD, Kuderer NM, Bacchi S, *et al*. Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis. *BMJ* 2024;384:e078538.
72 Funk PF, Hoch CC, Knoedler S, *et al*. ChatGPT’s Response Consistency: A Study on Repeated Queries of Medical Examination Questions. *EJIHPE* 2024;14:657–68.
73 World Health Organization. *Ethics and governance of artificial intelligence for health: WHO guidance*. 1st edn. 2021. Available: https://www.who.int/publications/i/item/9789240029200
74 Ong JCL, Chang S-H, William W, *et al*. Ethical and regulatory challenges of large language models in medicine. *Lancet Digit Health* 2024;6:e428–32.
75 Kapsali MZ, Livanis E, Tsalikidis C, *et al*. Ethical Concerns About ChatGPT in Healthcare: A Useful Tool or the Tombstone of Original and Reflective Thinking? *Cureus* 2024;16:e54759.
76 Rodler S, Kopliku R, Ulrich D, *et al*. Patients’ Trust in Artificial Intelligence-based Decision-making for Localized Prostate Cancer: Results from a Prospective Trial. *Eur Urol Focus* 2024;10:654–61.
77 Shumway DO, Hartman HJ. Medical malpractice liability in large language model artificial intelligence: legal review and policy recommendations. *J Osteopath Med* 2024;124:287–90.
78 Omiye JA, Lester JC, Spichak S, *et al*. Large language models propagate race-based medicine. *NPJ Digit Med* 2023;6:195.
79 Omar M, Sorin V, Agbareia R, *et al*. Evaluating and addressing demographic disparities in medical large language models: a systematic review. *Int J Equity Health* 2025;24:57.
80 Choradia N, Karzai F, Nipp R, *et al*. Increasing diversity in clinical trials: demographic trends at the National Cancer Institute, 2005-2020. *J Natl Cancer Inst* 2024;116:1063–71.
81 Klang E, Apakama D, Abbott EE, *et al*. A strategy for cost-effective large language model use at health system-scale. *NPJ Digit Med* 2024;7:320.
82 Kleinig O, Sinhal S, Khurram R, *et al*. Environmental impact of large language models in medicine. *Intern Med J* 2024;54:2083–6.
83 Ren S, Tomlinson B, Black RW, *et al*. Reconciling the contrasting narratives on the environmental impact of large language models. *Sci Rep* 2024;14:26310.
84 Shulman LN, Sheldon LK, Benz EJ. The Future of Cancer Care in the United States—Overcoming Workforce Capacity Limitations. *JAMA Oncol* 2020;6:327.
85 Zaharia M, Om K, Chen L, *et al*. The shift from models to compound AI systems. n.d. Available: https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/
86 Boiko DA, MacKnight R, Kline B, *et al*. Autonomous chemical research with large language models. *Nature New Biol* 2023;624:570–8.
87 Trinh TH, Wu Y, Le QV, *et al*. Solving olympiad geometry without human demonstrations. *Nature* 2024;625:476–82.
88 Shen Y, Xu Y, Ma J, *et al*. Multi-modal large language models in radiology: principles, applications, and potential. *Abdom Radiol* 2024;50:2745–57.
89 Schulte B. Capacity of ChatGPT to Identify Guideline-Based Treatments for Advanced Solid Tumors. *Cureus* 2023;15:e37938.
90 Morley J, Murphy L, Mishra A, *et al*. Governing Data and Artificial Intelligence for Health Care: Developing an International Understanding. *JMIR Form Res* 2022;6:e31623.
91 Vesteghem C, Brøndum RF, Sønderkær M, *et al*. Implementing the FAIR Data Principles in precision oncology: review of supporting initiatives. *Brief Bioinform* 2020;21:936–45.
92 Gallifant J, Afshar M, Ameen S, *et al*. The tripod-llm statement: a targeted guideline for reporting large language models use. *medRxiv* 2024.
93 Liu X, Cruz Rivera S, Moher D, *et al*. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* 2020;26:1364–74.
94 Liu Z, Wang P, Li Y, *et al*. RadOnc-gpt: a large language model for radiation oncology. [Preprint] 2023.