Advances in Medical Education and Practice

Dovepress
Taylor & Francis Group

Open Access Full Text Article

REVIEW

Artificial Intelligence and Multidisciplinary Oncology Decision-Making: A Systematic Review of LLM Concordance with MDT Recommendations and Implications for Clinical Reasoning Education

Yuan Ren, Wei Su, Youtu Wu, Sheng Dong, Shikai Liang, Xuejun Yang

Department of Neurosurgery, Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua University, Beijing, People's Republic of China

Correspondence: Xuejun Yang, Department of Neurosurgery, Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua University, 168 Litang Road, Changping District, Beijing, 10084, People's Republic of China, Email yangxuejunneuro@163.com

Introduction: Multidisciplinary team (MDT) decision-making is a cornerstone of oncology practice and an important context for developing clinical reasoning skills. Artificial intelligence (AI), particularly large language models (LLMs), has recently been explored as a potential tool to support clinical decision-making in oncology settings.

Methods: This systematic review, conducted according to PRISMA 2020 guidelines, synthesizes evidence on the concordance between LLM-generated recommendations and MDT decisions in oncology. Twenty-two studies were included, primarily comprising retrospective analyses comparing AI-generated outputs with expert MDT consensus across diagnostic, therapeutic, and workflow-related tasks.

Results: LLMs demonstrated generally moderate to high concordance with MDT decisions in guideline-constrained clinical scenarios. However, performance decreased in complex or less-structured cases. Variability in outcomes was associated with model architecture, prompting strategies, and retrieval-augmented approaches. Across studies, AI performance was most consistent when aligned with established clinical guidelines. Substantial heterogeneity existed among models and study designs, limiting direct comparability.

Conclusion: From a cognitive perspective, LLMs may be conceptualized as external information-processing tools that partially mirror structured aspects of MDT reasoning in specific contexts. However, evidence remains limited to retrospective concordance analyses and does not demonstrate educational impact. Any implications for clinical reasoning education should be interpreted as theoretical and require prospective validation using established educational frameworks such as cognitive apprenticeship or shared mental models.

Keywords: multidisciplinary team, artificial intelligence, large language models, oncology education, clinical decision support

Introduction

Cancer care is inherently complex, requiring coordinated expertise across multiple disciplines to address the biological, psychological, and social dimensions of oncology.[1] Previous studies have demonstrated that fragmentation in oncology services, suboptimal interprofessional communication, and limited integration of psychosocial care may adversely affect care quality and patient experience.[2] In response, multidisciplinary team (MDT) models have become a cornerstone of oncology decision-making, contributing to improved diagnostic accuracy, treatment selection, and overall clinical outcomes.[3,4] Traditional MDT education has relied largely on in-person meetings, case discussions, and simulation-based training. However, the COVID-19 pandemic and advances in digital technologies have accelerated the adoption of e-learning, virtual MDT meetings, and hybrid educational formats, enabling remote participation and innovative learning

https://doi.org/10.2147/AMEP.S621485
Received: 1 May 2026
Accepted: 3 July 2026
Published: 10 July 2026

experiences.[5–7] Despite these developments, structured approaches to developing interdisciplinary decision-making competencies remain heterogeneous, and gaps persist in optimizing collaborative clinical reasoning in oncology practice.

Concurrently, artificial intelligence (AI) and clinical decision support systems (CDSS), such as IBM Watson for Oncology, OncoAssist, and NAVIFY Tumor Board, are increasingly integrated into MDT workflows to support clinical decision-making and information synthesis.[8–10] These systems are primarily used as decision-support tools rather than formal educational interventions. At the same time, AI has been explored in medical education for its potential to support knowledge acquisition and reasoning processes; however, its role within MDT contexts has not been clearly defined in empirical educational studies.

Despite the growing adoption of AI in clinical practice, no comprehensive synthesis has systematically evaluated the concordance between AI-generated recommendations and MDT decisions in oncology. Existing evidence is largely derived from retrospective studies comparing outputs from AI systems with expert MDT decisions across diagnostic and therapeutic scenarios. The implications of these findings for clinical reasoning development remain theoretical rather than empirically demonstrated. Therefore, this systematic review aims to synthesize evidence on the concordance between artificial intelligence–generated recommendations and MDT decisions in oncology. This study focuses on patterns of AI–MDT agreement and provides a descriptive synthesis of their reported characteristics, with a brief discussion of their potential relevance to future research on clinical decision-support systems in multidisciplinary oncology care.

Methods

This systematic review was conducted in accordance with PRISMA 2020 guidelines. The aim was to synthesize evidence on the concordance between artificial intelligence–generated recommendations and MDT decisions in oncology.

Eligibility Criteria

Studies were eligible if they evaluated AI-based interventions designed to support MDT decision, or educational applications in oncology. Eligible interventions included, but were not limited to, large language models, virtual patients, simulation-based platforms, and machine learning–enabled educational systems integrated into MDT learning environments. We included randomized controlled trials, observational studies, simulation-based studies, and descriptive educational studies. Outcomes of interest included knowledge acquisition, clinical reasoning, decision-making performance, teamwork competencies, and related educational or patient-relevant outcomes. Studies focusing on rule-based systems, diagnostic-only tools, or non-MDT clinical decision-support applications were excluded.

A systematic search was conducted in PubMed, Embase, Web of Science, and Scopus from January 2020 to March 2026. The search strategy combined terms related to oncology, multidisciplinary team or tumor board, and artificial intelligence or machine learning or large language models or virtual simulation. Boolean operators (AND/OR) were applied to refine the search strategy. The detailed search strategy is provided in Supplementary Appendix S1. A PubMed search was conducted using combinations of terms related to oncology, multidisciplinary team care, and artificial intelligence, including controlled vocabulary and free-text terms. Additional studies were identified through manual screening of the reference lists of included articles.

Study Selection

All retrieved records were imported into a reference management system and deduplicated. Two independent reviewers screened titles and abstracts, followed by full-text assessment against predefined eligibility criteria. Disagreements were resolved through discussion or consultation with a third reviewer. The final review included 22 studies. Details in PRISMA Flow Diagram (Figure 1).

Data Extraction

Data were independently extracted by two reviewers using a standardized extraction form. Extracted variables included study characteristics (author, year, country, study design), participant characteristics, type of AI intervention, MDT educational context, outcome measures, and key findings.

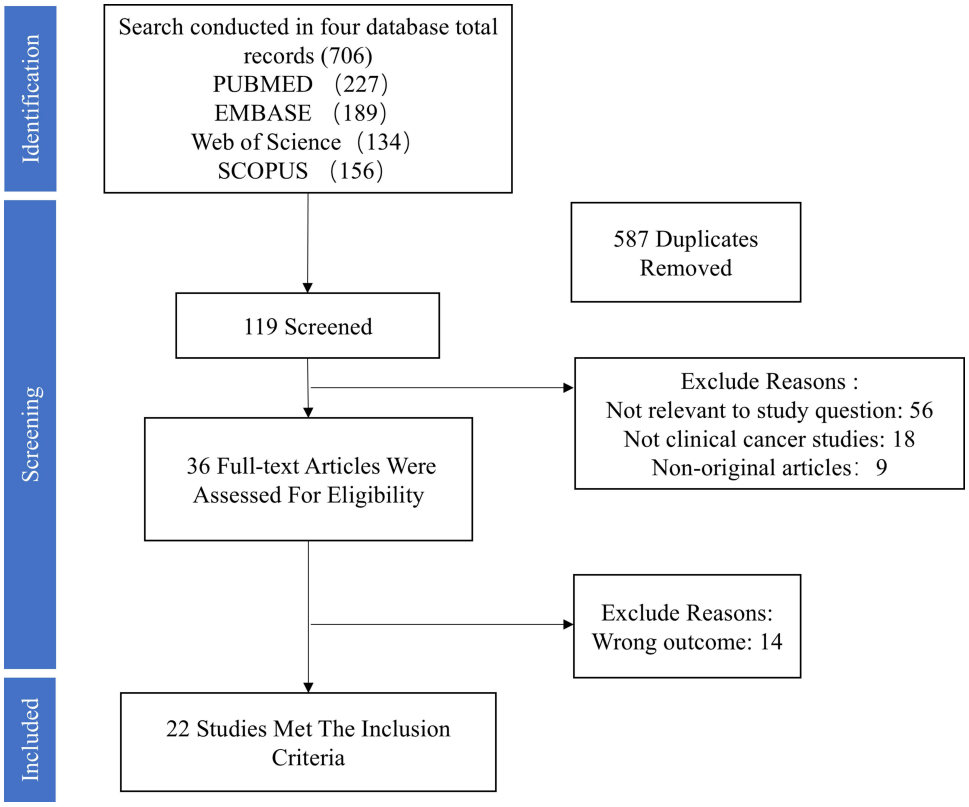

Figure 1 PRISMA flow diagram of study selection.

Data Synthesis

Given the heterogeneity of study designs, interventions, and outcome measures, meta-analysis was not feasible. A structured narrative synthesis was conducted. Studies were grouped according to AI modality, MDT educational application, and reported outcomes. Where applicable, findings were summarized descriptively using ranges and frequencies as reported in the original studies, without transformation of primary outcome data. Patterns of AI–MDT concordance and methodological limitations were synthesized, with a descriptive focus on their potential relevance to future decision-support research in oncology.

Risk of Bias and Quality Assessment

Risk of bias and study quality were assessed using a design-specific approach (Supplementary Tables 1–3). Diagnostic accuracy and concordance studies comparing AI outputs with MDT decisions were evaluated using the QUADAS-2 tool, while observational studies assessing the impact of AI on clinical decision-making were appraised using the ROBINS-I framework. Studies focusing on implementation, clinician perception, or AI-based decision support system evaluation were assessed using the Medical Education Research Study Quality Instrument (MERSQI).

Results

Across nine studies spanning multiple oncology domains, large language models (LLMs) demonstrated overall moderate-to-high concordance with MDT decision-making. Reported agreement varied considerably depending on clinical context, tumor type, case complexity, and the degree of structure in the provided clinical input (Table 1).

Most studies reported moderate-to-high agreement between LLM-generated recommendations and MDT decisions, with concordance ranging approximately from 66% to 93% across clinical scenarios. Performance was generally higher in guideline-driven and well-structured cases. In colorectal liver metastases, ChatGPT demonstrated moderate baseline agreement with MDT decisions (66.7%, $\kappa = 0.606$), which increased to 93.3% ($\kappa = 0.924$) when resectability status was

https://doi.org/10.2147/AMEP.S621485

https://doi.org/10.2147/AMEP.S621485

Table 1 Summary of Included Studies on Artificial Intelligence in Multidisciplinary Oncology Education and MDT Simulation

#	Author (Year)	Country	Study Design	Participants (n)	Oncology Specialty	AI Intervention Type	MDT Context/Simulation	Outcomes (Educational/ Clinical)	MDT Comparison & Key Findings
1	Tokoçin et al, 2026[11]	Türkiye	Retrospective study	100 patients	Breast cancer	Large language model (ChatGPT-35 and GPT-4)	Real-world MDT decision-making	GPT-4: 83% agreement; GPT-35: 75%	GPT-4 showed higher MDT concordance; reduced performance in complex multimorbidity cases
2	Yılmaz et al, 2026[12]	Turkey	Single-center retrospective study	30 CRCLM cases	Hepatobiliary/ Colorectal liver metastases	ChatGPT (LLM, chat-based)	MDT case review of CRCLM management decisions; AI prompted with anonymized clinical summaries	667% agreement (κ=0.606); 93.3% with resectability info (κ=0.924)	High dependence on prompt specificity; conservative bias observed
3	Pinninti et al 2026[13]	-	Prospective blinded concordance study	106 MDT cases	Multi-tumor oncology (breast, ovarian, esophageal, etc)	ChatGPT (standardized prompt-based LLM decision support)	Real MDT cases prospectively evaluated in blinded fashion against AI-generated recommendations	~71% concordance; 78.3% acceptability	Moderate-high MDT agreement; lower performance in systemic/diagnostic decisions
4	Birsin et al, 2026[14]	Turkey	Retrospective cohort study	179 patients	Stage II colorectal cancer	ChatGPT-5 (LLM-based decision support)	MDT: observation/ fluoropyrimidine/oxaliplatin vs AI: clinical-variable–based recommendations (age, ECOG PS, TNM, risk, MMR)	Moderate–substantial agreement; κ not specified	Tendency toward treatment escalation in elderly/frail patients
5	Goh et al, 2025[15]	Singapore	Retrospective comparative accuracy study	50 patients	Postoperative breast cancer (adjuvant therapy decision-making)	Guideline-augmented LLM system (Claude-2-based + NCCN 2023 integration), compared with Claude-2 and GPT-4	MDT reference standard vs AI-generated recommendations across 7 treatment domains	Highest concordance among models	Guideline augmentation improves MDT alignment; domain variability persists
6	Pergialiotis et al, 2026[16]	Greece	Retrospective MDT concordance study	599 patients	Gynecologic oncology	ChatGPT 5.0	AI vs MDT: staging and treatment decisions across tumor types	77% FIGO staging concordance	Lower agreement in molecularly complex and recurrent cases
7	Umihanic et al, 2025[17]	Bosnia and Herzegovina	Retrospective comparative concordance study	91 patients	Breast cancer	ChatGPT-40	MDT vs AI recommendations in resource-limited settings	High concordance; κ/ Fleiss not fully specified	Better performance in guideline-driven cases vs individualized decisions
8	Chatziisaak et al, 2025[18]	Switzerland	Retrospective single-centre cohort study	100 consecutive patients	Colorectal cancer	ChatGPT-4	MDT (pre-/post-therapy) vs ChatGPT treatment plans	Substantial concordance	Higher agreement in post-therapeutic decisions; reduced in frail patients
9	Pamuk et al, 2025[19]	Turkey	Retrospective comparative decision concordance study	25 patients with untreated laryngeal cancer	Head & neck oncology	ChatGPT-4	MDT decisions as reference standard vs AI-generated recommendations	72% full coherence	Stable performance; no incoherent outputs; consistent across subgroups

explicitly provided. Similarly, in a multi-tumor blinded evaluation, overall concordance reached approximately 71%, with higher agreement observed in treatment intent and surgical management compared with diagnostic or systemic therapy recommendations.

Performance differed across oncology subspecialties. Higher concordance was consistently observed in breast cancer and colorectal cancer cohorts, particularly in standardized treatment pathways and guideline-concordant decision settings. In laryngeal cancer, ChatGPT-4 achieved 72% full coherence with MDT decisions without generating incoherent outputs. In breast cancer studies, higher agreement was reported in cases with clear hormone receptor status and chemotherapy indications. In contrast, gynecologic oncology studies showed more heterogeneous performance. While overall agreement in FIGO staging reached 77%, reduced accuracy was observed in complex scenarios such as recurrent ovarian cancer and early-stage endometrial cancer, where individualized and molecularly driven treatment decisions were required.

Across studies, LLM performance declined in cases characterized by higher clinical complexity, multimorbidity, or frailty. Reduced concordance was particularly noted in elderly patients and those with significant comorbidity burden (eg., ASA $\geq$ III), indicating limitations in integrating contextual clinical information beyond structured oncologic variables. In some scenarios, models demonstrated a tendency toward guideline-concordant or conservative recommendations, which in specific populations (eg., elderly patients with stage II colorectal cancer) may not fully align with individualized MDT decision-making.

A consistent finding across studies was the strong dependence of LLM performance on prompt structure and input specificity. The explicit inclusion of key clinical variables, such as tumor resectability and staging details, significantly improved concordance with MDT decisions.

Furthermore, systems incorporating structured clinical guidelines demonstrated improved alignment with MDT reasoning compared with base LLM configurations, suggesting that external knowledge integration enhances decision consistency.

Overall, the evidence indicates that LLs can approximate MDT decision patterns under structured and guideline-driven conditions, with performance diminishing in complex, heterogeneous, or context-dependent clinical scenarios. These findings suggest that LLMs are more suitable as structured educational tools for exposing decision pathways than as autonomous systems for replicating expert multidisciplinary reasoning.

AI-Assisted Clinical Reasoning Training

Across the included studies (Table 2), four investigations examined the role of LLMs in MDT- or tumor board–related clinical reasoning tasks. Despite heterogeneity in tumor types, model architectures, and evaluation designs, the evidence demonstrates a consistent stratified pattern of performance across distinct MDT abstraction levels, ranging from direct benchmarking against expert consensus to model-to-model variability within standardized clinical scenarios.

Two studies (Qu et al and Jiang et al) evaluated LLM performance by directly comparing model outputs with MDT panels or clinician-generated decisions. Qu et al reported that LLMs achieved performance approaching MDT-level decision-making in structured clinical reasoning tasks across a large dataset of 1,423 cases. Similarly, Jiang et al demonstrated that certain models approximated clinician-level performance in selected task categories, although marked variability was observed across model types and clinical domains. Overall, these benchmarking studies indicate that LLMs can approximate MDT decision-making under structured and well-defined conditions, but performance remains sensitive to model selection and task specification.

Huang et al provided intra-MDT level evidence by evaluating multiple LLMs using identical oncology case scenarios. Substantial variability was observed in guideline adherence, treatment prioritization, and internal reasoning consistency across models. Higher-capacity models demonstrated improved alignment with clinical guidelines and more coherent decision pathways, whereas lower-performing models exhibited inconsistent or fragmented reasoning outputs. These findings indicate that LLM-generated recommendations are not uniform within MDT contexts and are strongly model-dependent, even when clinical inputs are standardized.

Across studies, MDT consensus itself demonstrated variability across institutions and clinical settings, introducing an additional layer of uncertainty in benchmarking LLM performance. Differences in institutional protocols, clinician

Table 2 Characteristics of Included Studies Evaluating AI Applications in MDT-Based Oncology Education and Decision Support

#	Author (Year)	Country	Study Design	Participants (n)	Oncology Specialty	AI Intervention Type	MDT Context/ Simulation	Outcomes (Educational/ Clinical)	MDT Comparison & Key Findings
1	Qu et al, 2026[20]	China	Retrospective large-scale MDT simulation study	1423 colorectal cancer cases Colorectal	Colorectal oncology	Multi-LLM comparison (OpenAI o3-mini, DeepSeek-R1, Qwen QwQ-plus)	AI replication of MDT decisions (stability via repeated runs)	Agreement: 62.5–65.4%; κ=0.794 (best model)	Moderate MDT concordance; highest stability in o3-mini; reduced performance in complex cases
2	Jiang et al, 2026[21]		Comparative cross-sectional evaluation study	50 questions + 9 LLMs + physicians (residents, fellows, attendings)	Breast cancer/ Radiology-supported MDT decision-making	Multiple LLMs (ChatGPT-4, ChatGPT-4o, ChatGPT-4o mini, Claude 3 Opus, Claude 3.5 Sonnet, Gemini 1.5 Pro, Tongyi Qianwen 2.5, ChatGLM, Ernie Bot 3.5)	Simulated MDT decision-support using NCCN/BI-RADS breast cancer cases vs physician groups	Comparable accuracy to physicians; model-dependent variability	Top LLMs performed at level of residents/ attendings; no consistent superiority over clinicians
3	Huang et al, 2026[22]	China (single-center MDT dataset)	Comparative evaluation study	95 advanced gastric cancer MDT cases + 14 rare PubMed cases + 10 clinical questions	Gastric cancer/ Medical oncology and surgical MDT	DeepSeek-R1 vs ChatGPT-4o (LLMs)	MDT cases with guideline and rare-case simulations	W=063–0.70; higher accuracy in DeepSeek-R1	DeepSeek-R1 outperformed ChatGPT-4o; good guideline concordance but no clinical outcome validation

https://doi.org/10.2147/AMEP.S621485

experience, and case interpretation contributed to heterogeneity in MDT decisions, limiting the stability of "ground truth" references used for model evaluation. This cross-MDT variability further complicates the assessment of LLM performance, as model outputs are compared against inherently heterogeneous expert decision frameworks.

Collectively, the evidence supports a hierarchical framework of LLM performance in MDT-related clinical reasoning tasks. At the benchmarking level, LLMs can approximate MDT decisions in structured scenarios. At the intra-model level, substantial variability exists across LLM architectures in reasoning consistency and guideline adherence. At the cross-MDT level, heterogeneity in human consensus introduces additional uncertainty in defining reference standards.

AI-Generated MDT Reasoning as a Mirror for Learners

As summarized in Table 3, the included studies collectively position large language models (LLMs) as simulated mirrors of multidisciplinary team (MDT) reasoning, demonstrating a graded spectrum of performance across concordance, decision-support, and simulation-based evaluation contexts. Rather than functioning as purely diagnostic tools, LLMs increasingly operate as educational proxies of MDT cognitive workflows, enabling structured comparison between human expert reasoning and machine-generated clinical deliberation.

Two studies evaluated LLM performance through direct concordance with MDT or molecular tumor board (MTB) decisions. Gabriel et al reported near-perfect alignment between a guideline-structured GPT-4 system and MDT recommendations in prostate cancer cases, suggesting that LLMs can closely replicate deterministic MDT decision pathways when case structures are highly standardized and input information is constrained. This level of performance reflects procedural reproduction of MDT logic, rather than autonomous clinical reasoning. In contrast, Rinderknecht et al conducted a prospective multicenter non-inferiority evaluation and demonstrated that neither GPT-4 nor Claude 3.5 Sonnet achieved non-inferiority compared with real MTB decisions in genitourinary oncology. Although Claude 3.5 Sonnet approached MDT-level performance in selected domains, both models showed reduced accuracy in complex metastatic and multi-algorithm decision scenarios. These findings highlight a key educational distinction: LLMs may reproduce MDT outputs in structured retrospective settings but fail to fully capture context-dependent expert reasoning in prospective clinical environments.

Celsa et al evaluated a multi-model retrieval-augmented framework (PHENO-RAG) in hepatocellular carcinoma and demonstrated improved performance in treatment allocation and clinical complexity stratification. However, performance declined in MDT referral prediction tasks, suggesting that upstream triage decisions require higher-order contextual integration that remains challenging for current LLM systems. Importantly, this study demonstrates that MDT performance is not solely determined by model architecture, but is strongly shaped by retrieval design, structured prompting, and information curation pipelines. From an educational perspective, this highlights that LLMs may function as scaffolded reasoning environments, where system design directly modulates the quality of "learned" MDT-like decision pathways.

Hack et al assessed LLM-generated recommendations in simulated head and neck cancer MDT scenarios using blinded expert scoring frameworks. Results showed high concordance between LLM outputs and expert MDT reasoning, with retrieval-augmented models achieving the highest performance across most evaluation domains. Notably, LLM-generated responses demonstrated strong performance in clarity, structured reasoning presentation, and guideline consistency, occasionally outperforming expert comparators in communication-oriented dimensions. Inter-rater reliability between expert evaluators was high, supporting the internal consistency of simulation-based assessment. From an educational standpoint, these findings suggest that LLMs may function not only as decision-support systems but also as interactive MDT simulation environments, enabling learners to observe structured reasoning patterns that approximate expert deliberation under controlled scenarios.

Overall, Table 3 reveals a structured gradient of LLM capability across three MDT reasoning layers: (1) concordance-level replication of MDT outputs, (2) augmentation-level decision support influenced by system architecture, and (3) simulation-level modeling of expert reasoning processes. Importantly, these layers correspond to distinct educational functions rather than purely performance metrics. At the concordance level, LLMs provide exposure to guideline-aligned MDT decision templates. At the augmentation level, they illustrate how information structuring shapes clinical reasoning outputs. At the simulation level, they function as interactive cognitive mirrors, allowing learners to observe, compare,

Table 3 Comparative Studies of AI and MDT Decision-Making Performance in Oncology

#	Study	Country	Design	Participants (n)	Cancer	AI Type	MDT Context	Outcomes	Key Finding	MDT Reasoning Layer
1	Gabriel et al, 2026[23]	UK	Retrospective concordance study	40 cases	Prostate cancer	RAG-based GPT-4	Real MDT decision replication	Concordance with MDT	Near-perfect agreement (100%) with MDT decisions under guideline-structured conditions	Structured reasoning reproduction (guideline-based MDT logic)
2	Rinderknecht et al, 2025[24]	Germany	Prospective multicenter non-inferiority trial	110 clinical case scenarios	Genitourinary oncology	GPT-4, Claude 3.5	Real MDT as gold standard	Decision quality (mSCS), non-inferiority	LLMs failed to reach non-inferiority vs MDT, especially in complex metastatic cases	Context-dependent reasoning gap (limits of LLM clinical judgment)
3	Celsa et al, 2025[25]	Italy	Retrospective multi-model evaluation	489 clinical reports from 424 patients	Hepatocellular carcinoma	Multi-LLM (PHENO-RAG)	MDT triage + treatment allocation	Concordance, complexity grading	High accuracy in treatment allocation; lower performance in MDT referral decisions; strongly dependent on RAG design	Scaffolded reasoning environment (system-shaped MDT logic)
4	Hack et al, 2026[26]	U.S.-led multicenter	Blinded simulation-based evaluation	14 cases	Head & neck cancer	GPT-4/GPT-5 ± RAG	Simulated MDT scenarios	Expert scoring (appropriateness, clarity, feasibility)	High alignment with expert MDT; RAG improved performance; strong inter-rater reliability	Interactive reasoning simulation (educational MDT environment)

https://doi.org/10.2147/AMEP.S621485

and interrogate MDT-like reasoning pathways. Together, these findings support a conceptual shift in interpreting LLMs within oncology MDT contexts—from static predictive tools to dynamic educational systems that externalize and operationalize clinical reasoning processes.

AI-Supported MDT Workflow and Diagnostic Augmentation

As summarized in Table 4, the included studies extend the role of artificial intelligence beyond decision replication toward process-level augmentation of MDT function. Rather than generating independent treatment recommendations, these systems focus on workflow optimization and diagnostic support, providing educational value by structuring key components of MDT practice.

Wijethilake et al evaluated a deep learning–based automated segmentation and reporting system within simulated MDT settings for vestibular schwannoma management. The introduction of AI-assisted reporting significantly reduced MDT discussion time and improved communication efficiency, despite a modest increase in preparation time associated with system integration. Importantly, the system achieved high segmentation accuracy (Dice score > 0.93), enabling consistent and standardized tumor quantification.

From an educational perspective, this approach shifts MDT interpretation from implicit, experience-dependent reasoning toward explicit and standardized data representation. By structuring imaging outputs and clinical summaries, AI systems may reduce variability in how cases are presented and discussed within MDT environments. This may improve MDT training by exposing learners to more transparent and reproducible information pathways, thereby supporting the development of structured clinical reasoning rather than reliance on tacit expert interpretation.

Liu et al examined the interaction between an AI-based chest CT diagnostic model and MDT consensus in pulmonary nodule evaluation. While MDT assessment alone outperformed AI in diagnostic accuracy, the combined AI–MDT approach achieved the highest agreement with histopathological outcomes.

This finding highlights a complementary cognitive relationship, in which AI contributes pattern recognition and quantitative consistency, while MDT integrates contextual, clinical, and multidisciplinary judgment. From an educational standpoint, this interaction provides a model for augmented clinical reasoning, where learners can observe how algorithmic outputs are incorporated, validated, or overridden within MDT discussions. Such integration may facilitate the development of metacognitive skills, enabling trainees to critically evaluate AI-generated evidence within broader clinical contexts rather than passively accepting automated outputs.

Collectively, these studies suggest that AI contributes to MDT environments not only through decision support but also by reshaping how clinical reasoning is generated, communicated, and validated.

Two complementary educational functions can therefore be identified. First, AI enables workflow transparency and standardization, making implicit MDT processes, such as imaging interpretation and case presentation more explicit and reproducible. Second, AI introduces a paradigm of human–AI co-reasoning, where diagnostic outputs are iteratively refined through multidisciplinary evaluation. Importantly, across both domains, MDT consensus remains the central reference framework, with AI functioning as an augmentation layer rather than an autonomous decision-maker. This positioning is particularly relevant for education-focused applications. It reinforces AI as a tool for supporting, structuring, and interrogating clinical reasoning, rather than replacing the experiential and collaborative learning processes inherent to MDT practice.

LLMs as MDT Simulation-Based Learning Environments

As summarized in Table 5, four studies investigated the use of LLMs and machine learning–based systems to simulate or reproduce MDT decision-making, with an emphasis on their potential educational applications.

Li et al evaluated multiple LLMs in simulating MDT decision-making for complex soft tissue sarcoma cases across 21 expert centers. The models showed low-to-moderate concordance with expert MDT consensus (20–60%) and limited intra-model consistency across repeated runs. Performance varied substantially across cases, particularly in complex scenarios requiring multidisciplinary integration, indicating unstable replication of expert-level MDT reasoning.

Two studies (Thavanesan et al and Pasello et al) examined machine learning models trained on historical MDT decisions. In oesophageal cancer, ML models demonstrated good discrimination between curative and palliative

Table 4 AI-Supported MDT Workflow and Diagnostic Augmentation

#	Study	Country	Design	Participants (n)	Cancer	AI Type	MDT Context	Outcomes	Key Finding	MDT Evidence Level	MDT Function Layer
1	Wijethilake et al, 2026[27]	UK	Pilot comparative simulation study	50 patients	Vestibular schwannoma	Deep learning-based automated tumor segmentation + structured reporting tool	Simulated MDT meetings comparing standard workflow vs AI-assisted reporting	Dice >0.93; reduced MDT prep/discussion time	AI improved MDT workflow efficiency and communication	Workflow augmentation/ efficiency enhancement	Workflow & communication optimization
2	Liu et al, 2025[28]	China	Retrospective diagnostic accuracy study	87 patients	Pulmonary nodules	CT-based AI diagnostic model + MDT consensus integration	AI interpretation compared with MDT consensus; pathology gold standard	AUC, sensitivity, specificity, κ agreement	AI+MDT achieved highest diagnostic accuracy vs AI alone	Hybrid diagnostic support	Diagnostic augmentation (AI–MDT synergy)

https://doi.org/10.2147/AMEP.S621485

Table 5 LLMs as MDT Simulation-Based Learning Environments

	Author (Year)	Country	Study Design	Participants (n)	Oncology Specialty	AI Intervention Type	MDT Context/ Simulation	Outcomes (Educational/Clinical)	MDT Role/ Comparison	Key Findings
1	Li et al, 2025[29]	Germany	Multicenter simulation study	5 patient cases	Soft tissue sarcoma	LLMs (Llama 3.2 Vision 90B; Claude 3.5 Sonnet; DeepSeek-R1; OpenAI o1)	Simulated MDT case-based decisions vs expert MTB consensus	20–60% alignment with expert MDT reasoning; inconsistent outputs across runs	AI-generated MDT reasoning compared with expert consensus	LLMs produce plausible MDT reasoning but lack stable consensus reproduction, showing high variability in decision patterns
2	Webb et al, 2026[30]	UK	Comparative cognitive analysis study	87 participants	Oesophageal cancer	ML feature importance model (decision imitation system)	Clinician vs AI reasoning in MDT decision-making	Partial alignment between clinician reasoning and ML feature weighting; mismatch in psychosocial factors	AI used to model and compare MDT decision priorities	AI reveals latent structure of MDT reasoning but underrepresents holistic and psychosocial factors
3	Thavanesan et al, 2025[31]	UK	Retrospective ML modeling study	1931 patients data	Oesophageal cancer	ML models (XGBoost, RF, MLR)	MDT pathway classification (curative vs palliative decision routes)	High accuracy in curative pathway prediction; reduced performance in heterogeneous palliative cases	AI trained on MDT decisions to replicate clinical pathways	MDT decisions can be partially encoded into predictive models but show limited generalizability to complex cases
4	Pasello et al, 2025[32]	Italy	Multicohort modeling study	Italy	LA-NSCLC	Radiomics + clinicopathological ML model	MDT treatment allocation modeling for learning support	Good replication of MDT treatment allocation patterns across cohorts	AI models trained on MDT decision patterns for teaching stratification logic	MDT decision-making can be decomposed into feature-based rules but reflects imitation rather than true reasoning

https://doi.org/10.2147/AMEP.S621485

Figure 2 Multi-layer framework of AI-supported clinical reasoning in MDT education.
Notes: This figure shows a three-level framework of AI-supported clinical reasoning in MDT education, including model capabilities, workflow integration, and educational development. The layers interact through bidirectional feedback to support clinical decision-making, reasoning, and learning, while enabling factors and key challenges are also indicated.

treatment pathways with strong external validation performance. In locally advanced non-small cell lung cancer, radiomics-integrated models similarly replicated MDT treatment allocation patterns across cohorts. Across both studies, ML models showed consistent ability to reproduce structured MDT decision patterns, although performance decreased in more heterogeneous clinical scenarios.

Webb et al compared clinician decision-making priorities with feature importance derived from ML models in oesophageal cancer MDT settings. The analysis demonstrated partial alignment between human and model-derived decision factors; however, discrepancies were observed in the weighting of demographic and psychosocial variables, which were underrepresented in the ML-derived decision structure.

Overall, the included studies indicate that LLMs and ML-based systems can partially simulate or reproduce MDT decision-making patterns under structured conditions. However, their performance is variable across tumor types and decision complexity, with reduced stability in complex or heterogeneous clinical scenarios. Differences between AI-

derived and clinician-derived decision features further highlight inconsistencies in how MDT reasoning is represented computationally.

Discussion

This review synthesizes emerging evidence on AI, particularly LLMs, in MDT oncology. It repositions their primary value from a clinical decision-support tool to an educational infrastructure for developing clinical reasoning skills.[33] Across included studies, AI systems demonstrate the capacity to approximate MDT decisions, simulate tumor board discussions, and enhance diagnostic workflows; however, their most consistent contribution lies in externalizing and structuring the cognitive processes underlying multidisciplinary decision-making.

Evidence across studies suggests that AI performance in MDT settings is not uniform but varies along a spectrum of clinical reasoning complexity. In structured and guideline-constrained scenarios, LLMs have demonstrated high concordance with MDT decisions, particularly when clinical reasoning is based on standardized knowledge frameworks or augmented with retrieval-based architectures.[34,35] However, as case complexity increases, involving multi-step reasoning, multimorbidity, or less well-defined clinical contexts, model performance becomes increasingly heterogeneous, with reduced reliability and inter-model variability.[36,37] This pattern reflects the inherently non-deterministic nature of clinical reasoning in complex oncology decision-making, as well as limitations in current LLM architectures for sustained long-context reasoning. Furthermore, accumulated evidence indicates that model outputs are highly sensitive to system design factors, including retrieval augmentation, prompting strategy, and knowledge representation format. These factors can substantially influence decision consistency and accuracyy.[34,38]

Beyond model-level differences, system design emerges as a critical determinant of performance. Evidence from clinical AI deployment studies suggests that the effectiveness of artificial intelligence is strongly influenced by its integration into clinical workflows and decision environments, rather than model capability alone.[39] Empirical studies further show that system-level design variations, including data processing pipelines and decision logic, can substantially alter model behavior and downstream clinical outcomes, even when using identical algorithms.[40–42] These findings indicate that MDT-related performance is shaped not only by intrinsic model capacity but also by the broader clinical and informational ecosystem in which reasoning is embedded. This shifts the analytical focus from isolated model evaluation toward system-level AI design and deployment frameworks.

At a higher level of abstraction, LLMs function as interactive simulations of MDT reasoning. They enable dynamic engagement with virtual clinical cases through iterative diagnostic reasoning and structured feedback-based learning.[43] Recent studies have demonstrated that LLM-based systems can simulate patient encounters and support multi-step clinical decision-making through interactive dialogue frameworks (X et al, 2024; Y et al, 2025). These systems also incorporate feedback-driven mechanisms that enhance reflective reasoning and clinical decision-making performance.[44,45]

Importantly, LLMs are increasingly conceptualized as facilitative tools for clinical reasoning development rather than autonomous decision-making systems This perspective emphasizes their role in supporting process-oriented learning and iterative refinement of diagnostic thinking. Through repeated interaction, learners can compare alternative reasoning pathways and identify discrepancies between AI-generated and expert-aligned reasoning trajectories, thereby enhancing reflective practice and clinical reasoning skills.

In addition to reasoning simulation, AI contributes to MDT workflows through diagnostic and organizational augmentation. Imaging automation and structured reporting improve data standardization and communication efficiency. AI–MDT integration also demonstrates complementary diagnostic benefits. AI systems provide quantitative pattern recognition, while MDTs contribute contextual interpretation.[46,47] These findings suggest that AI exerts its greatest impact at the level of process optimization rather than autonomous decision replacement.

Despite these advantages, several limitations remain evident. Performance degradation in complex, multimorbid, or context-rich scenarios underscores current limitations in integrating heterogeneous clinical environments.[39,48,49] In addition, variability across models and tasks, along with occasional systematic biases toward guideline-constrained reasoning, highlights the incomplete nature of algorithmic clinical judgment.[40,49,50] Importantly, these limitations also carry educational value, as they expose learners to uncertainty, model fragility, and the boundaries of computational

reasoning. Collectively, the evidence supports a reconceptualization of AI in MDT education as a layered cognitive framework rather than a singular decision-support tool.[51]

This framework spans multiple layers of clinical reasoning support. These include guideline-level reasoning replication, variability-driven comparative reasoning, system-dependent augmentation, and simulation-based experiential learning. Each layer provides distinct educational value, ranging from foundational knowledge acquisition to advanced reflective reasoning development. As summarized in Figure 2, we propose a layered cognitive framework that conceptualizes AI in MDT oncology across model, system, and cognitive-educational levels. Evidence from included studies supports this stratification (Supplementary Table 4). LLM-based studies show moderate-to-high concordance with MDT decisions at the replication level, particularly in guideline-driven settings. System-level studies highlight improvements in workflow efficiency and diagnostic performance. In contrast, simulation studies reveal variability and instability in AI reasoning, supporting its role as a cognitive training tool rather than an autonomous decision-maker.

This systematic review has several limitations. First, the included studies were highly heterogeneous in design, including observational analyses, simulation-based studies, and educational evaluations, which limited the ability to perform quantitative synthesis. Second, formal risk-of-bias assessment tools were not uniformly applicable across all study types, particularly for descriptive and simulation-based studies. Third, most included studies reported surrogate outcomes rather than direct measures of educational or clinical impact, limiting conclusions regarding real-world effectiveness. Finally, the evidence base was primarily derived from selected oncology subspecialties, which may limit the generalizability of findings to broader clinical contexts.

Future research should move beyond traditional accuracy or concordance metrics and focus on educational outcomes, including clinical reasoning development, decision-making quality, and long-term learning transfer. In parallel, the development of purpose-built AI systems for education optimized for transparency, explainability, and pedagogical adaptability will be essential for realizing the full potential of AI-supported MDT training.

Conclusion

AI systems in oncology MDT contexts should not be viewed solely as decision-making tools, but as representations of clinical reasoning processes that can be leveraged for education. Across included studies, evidence generally demonstrated high levels of AI–MDT concordance in simulation and retrospective settings, emerging but heterogeneous applications in clinical decision support and education, and limited real-world implementation data. By externalizing, structuring, and simulating MDT cognition, AI offers a novel pathway for training the next generation of clinicians. Practically, AI–MDT concordance exercises could be incorporated into oncology training programs by having trainees compare large language model–generated recommendations with actual tumor board decisions, followed by structured faculty-led discussion to identify discrepancies and reasoning gaps. In addition, simulation-based MDT scenarios integrating AI-generated treatment suggestions may be embedded into fellowship training to support the development of clinical reasoning and decision-making skills under expert supervision.

Data Sharing Statement

All data supporting this review are available within the article and its supplementary materials. No new data were created or analyzed in this study.

Author Contributions

Yuan Ren and Xuejun Yang conceived and designed the study, performed statistical analyses, and drafted and critically revised the manuscript. Wei Su and Sheng Dong contributed to data collection, analysis, and interpretation. Youtu Wu was responsible for data collection. Shikai Liang contributed to data collection. All authors contributed to data analysis, drafting or revising the article, have agreed on the journal to which the article will be submitted, gave final approval of the version to be published, and agree to be accountable for all aspects of the work.

Funding

BTCH 1st Young Talent Enlightenment Program (2024-2026).

Disclosure

The authors report no competing interests in this work.

References

1. Horlait M, Baes S, De Regge M, Leys M. Understanding the complexity, underlying processes, and influencing factors for optimal multidisciplinary teamwork in hospital-based cancer teams: a systematic integrative review. *Cancer Nurs*. 2021;44:E476–e492.
2. Nekhlyudov L, Levit LA, Ganz PA. Delivering high-quality cancer care: charting a new course for a system in crisis: one decade later. *J clin oncol*. 2024;42:4342–4351. doi:10.1200/JCO-24-01243
3. de Castro G Jr, Souza FH, Lima J, Bernardi LP, Teixeira CHA, Prado GF. Does Multidisciplinary Team Management Improve Clinical Outcomes in NSCLC? A Systematic Review With Meta-Analysis. *JTO Clin Res Rep*. 2023;4:100580. doi:10.1016/j.jtocrr.2023.100580
4. Williams GJ, Thompson JF. Management changes and survival outcomes for cancer patients after multidisciplinary team discussion; a systematic review and meta-analysis. *Cancer Treat Rev*. 2025;139:102997. doi:10.1016/j.ctrv.2025.102997
5. Ali SR, Dobbs TD, Mohamedbhai H, Whitaker S, Hutchings HA, Whitaker IS. Evaluating remote skin cancer multidisciplinary team meetings in the United Kingdom post-COVID-19. *J Plast Reconstr Aesthet Surg*. 2023;84:250–257. doi:10.1016/j.bjps.2023.04.052
6. Groothuizen JE, Aroyewun E, Zasada M, Harris J, Hewish M, Taylor C. Virtually the same? Examining the impact of the COVID-19 related shift to virtual lung cancer multidisciplinary team meetings in the UK National Health Service: a mixed methods study. *BMJ Open*. 2023;13:e065494. doi:10.1136/bmjopen-2022-065494
7. Kulaksız T, Steinbacher J, Kalz M. Technology-enhanced learning in the education of oncology medical professionals: a systematic literature review. *J Cancer Educ*. 2023;38:1743–1751. doi:10.1007/s13187-023-02329-1
8. Chang LC, Kuo HC, Wang HM, et al. The use of an integrated digital tool to improve the efficiency of multidisciplinary tumor boards-a prospective trial in Taiwan. *Cancers*. 2025;18:17. doi:10.3390/cancers18010017
9. Lin CH, Wang BY, Lin SH, et al. Process re-engineering and data integration using fast healthcare interoperability resources for the multidisciplinary treatment of lung cancer. *JMIR Cancer*. 2025;11:e53887. doi:10.2196/53887
10. Kim MS, Park HY, Kho BG, et al. Artificial intelligence and lung cancer treatment decision: agreement with recommendation of multidisciplinary tumor board. *Transl Lung Cancer Res*. 2020;9:507–514. doi:10.21037/tlcr.2020.04.11
11. Tokoçin M, Pehlivan T, Cin S, et al. Evaluating the role of artificial intelligence in enhancing multidisciplinary team decisions for breast cancer management. *Eur J Breast Health*. 2026;22:184–189. doi:10.4274/ejbh.galenos.2026.2025-10-4
12. Yılmaz M, Abbaslı N, Tuna S, et al. Comparison of artificial intelligence and multidisciplinary team recommendations in the management of colorectal cancer liver metastases. *Sci Rep*. 2026;16. doi:10.1038/s41598-026-38449-z
13. Pinninti R, Gullapalli R, Mallavarapu KM, et al. Comparing artificial intelligence and multidisciplinary tumor board decision making in real-world cancer care: a prospective blinded concordance study. *ESMO Real World Data Digit Oncol*. 2026;11:100688. doi:10.1016/j.esmorw.2026.100688
14. Birsin Z, Jeral S, Cebeci S, et al. Stage II colon cancer: does ChatGPT recommend more intensive adjuvant therapy? A comparison with MDT decisions. *Future Oncol*. 2026;22:427–434. doi:10.1080/14796694.2025.2610463
15. Goh SSN, Mariappan R, Soo Woon Tan G, et al. Augmenting large language models with national comprehensive cancer network guidelines for improved and standardized adjuvant therapy recommendations in postoperative breast cancer cases. *JCO Clin Cancer Inform*. 2025;9:e2400243. doi:10.1200/CCI-24-00243
16. Pergialiotis V, Thomakos N, Lygizos V, et al. Discrepancies Between MDT Recommendations and AI-Generated Decisions in Gynecologic Oncology: a Retrospective Comparative Cohort Study. *Cancers*. 2026;18:452. doi:10.3390/cancers18030452
17. Umihanic S, Osmanovic H, Selak N, et al. Evaluating the Concordance Between ChatGPT and Multidisciplinary Teams in Breast Cancer Treatment Planning: a Study from Bosnia and Herzegovina. *J Clin Med*. 2025;15:14. doi:10.3390/jcm15010014
18. Chatziisaak D, Burri P, Sparn M, Hahnloser D, Steffen T, Bischofberger S. Concordance of ChatGPT artificial intelligence decision-making in colorectal cancer multidisciplinary meetings: retrospective study. *BJS Open*. 2025;9. doi:10.1093/bjsopen/zraf040
19. Pamuk E, Bilen YE, Külekçi Ç, Kuşcu O. ChatGPT-4 vs. multi-disciplinary tumor board decisions for the therapeutic management of primary laryngeal cancer. *Acta Otolaryngol*. 2025;145:714–719. doi:10.1080/00016489.2025.2502563
20. Qu B, Cao L, Wu C, et al. Comparison of large language models and expert multidisciplinary team decisions in colorectal cancer. *BMJ Health Care Inform*. 2026;33:e101780. doi:10.1136/bmjhci-2025-101780
21. Jiang H, Yang C, Zhou W, et al. Evaluation of large language models for radiologists' support in multidisciplinary breast cancer teams: comparative study. *JMIR Med Inform*. 2026;14:e68182. doi:10.2196/68182
22. Huang JB, Li HZ, Zong Z, et al. A comparative study on the application of DeepSeek-R1 and ChatGPT in multidisciplinary treatment decision-making for advanced gastric cancer. *Zhonghua Wei Chang Wai Ke Za Zhi*. 2026;29:114–119. doi:10.3760/cma.j.cn441530-20250409-00149
23. Gabriel A, Antonoglou G, Langley S, Gabriel J. Harnessing artificial intelligence to code a decision-making aid for the prostate cancer brachytherapy multidisciplinary meeting using retrieval-augmented generation. *Cureus*. 2026;18:e103457. doi:10.7759/cureus.103457
24. Rinderknecht E, Haas M, Schnabel MJ, et al. Benchmarking Large Language Models Against Multidisciplinary Tumor Boards in Urological Oncology: results from the Blinded, Prospective CONCORDIA Study. *Eur Urol Oncol*. 2025;9:589–597. doi:10.1016/j.euo.2025.11.016
25. Celsa C, Giuffrè M, Di Maria G, et al. PHENO-RAG: an artificial intelligence tool for guideline-informed management decisions in hepatocellular carcinoma. *JHEP Rep*. 2025;8:101715. doi:10.1016/j.jhepr.2025.101715
26. Hack S, Karni RJ, Maniaci A, et al. Evaluation of large language models as decision support tools for head and neck cancer management: a blinded multidisciplinary simulation study. *Oral Oncol*. 2026;174:107877. doi:10.1016/j.oraloncology.2026.107877

27. Wijethilake N, Connor S, Oviedova A, et al. Artificial intelligence for personalized management of vestibular schwannoma: a multidisciplinary clinical implementation study. *JAMIA Open*. 2026;9:ooaf163. doi:10.1093/jamiaopen/ooaf163
28. Liu XY, Shan FC, Li H, Zhu JB. Analysis on artificial intelligence-based chest computed tomography in multidisciplinary treatment models for discriminating benign and malignant pulmonary nodules. *Clinics*. 2025;80:100734. doi:10.1016/j.clinsp.2025.100734
29. Li CP, Kalisa AT, Roohani S, et al. The imitation game: large language models versus multidisciplinary tumor boards: benchmarking AI against 21 sarcoma centers from the ring trial. *J Cancer Res Clin Oncol*. 2025;151:248. doi:10.1007/s00432-025-06304-9
30. Webb C, Thavanesan N, Naiseh M, Dewar-Haggart R, Underwood T, Vigneswaran G. Health care professionals' perceptions of machine learning based clinical decision support systems for oesophageal cancer management. *Comput Biol Med*. 2026;200:111373. doi:10.1016/j.compbiomed.2025.111373
31. Thavanesan N, Naiseh M, Terol M, et al. Oesophageal cancer multi-disciplinary tool: a co-designed, externally validated, machine learning tool for oesophageal cancer decision making. *EClinicalMedicine*. 2025;89:103527. doi:10.1016/j.eclinm.2025.103527
32. Pasello G, Kotler H, Ferro A, et al. Radiomic approach to support multidisciplinary tumor board decision-making in locally advanced non-small cell lung cancer. *Front Oncol*. 2025;15:1713847. doi:10.3389/fonc.2025.1713847
33. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29:1930–1940. doi:10.1038/s41591-023-02448-8
34. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620:172–180. doi:10.1038/s41586-023-06291-2
35. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2:e0000198. doi:10.1371/journal.pdig.0000198
36. Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of Artificial General Intelligence: early experiments with GPT-4. 2023.
37. Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. *ACM Comput Surv*. 2023;55:248. doi:10.1145/3571730
38. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. *Proceedings of the 36th International Conference on Neural Information Processing Systems*. Curran Associates Inc. New Orleans, LA, USA, 2022.
39. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nature Med*. 2019;25:44–56. doi:10.1038/s41591-018-0300-7
40. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366:447–453. doi:10.1126/science.aax2342
41. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*. 2020;117:12592–12594.
42. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *Npj Digital Med*. 2018;1:18. doi:10.1038/s41746-018-0029-1
43. Chen G, Lin C, Zhang L, Luo Z, Shin YS, Li X. Virtual case reasoning and AI-assisted diagnostic instruction: an empirical study based on body interact and large language models. *BMC Med. Educ*. 2025;25:1493. doi:10.1186/s12909-025-07872-7
44. Brügge E, Ricchizzi S, Arenbeck M, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. *BMC Med Educ*. 2024;24:1391. doi:10.1186/s12909-024-06399-7
45. Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A review of large language models in medical education, clinical decision support, and healthcare administration. *Healthcare*. 2025;13:603. doi:10.3390/healthcare13060603
46. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577:89–94. doi:10.1038/s41586-019-1799-6
47. 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
48. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380:1347–1358. doi:10.1056/NEJMra1814259
49. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med*. 2018;15:e1002683.
50. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature Med*. 2019;25:1337–1340. doi:10.1038/s41591-019-0548-6
51. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc. Neurol*. 2017;2.