Review

Artificial Intelligence in Relation to Accurate Information and Tasks in Gynecologic Oncology and Clinical Medicine – Dunning-Kruger Effects and Ultracrepidarianism

Edward J. Pavlik * , Jamie Land Woodward , Frank Lawton , Allison L. Swiecki-Sikora , Dharani D. Ramaiah , Taylor A. Rives

Posted Date: 7 February 2025

doi: 10.20944/preprints202502.0521.v1

Keywords: artificial intelligence; large language models; generative AI; accuracy; chatbot; gynecologic/oncology

Review

Artificial Intelligence in Relation to Accurate Information and Tasks in Gynecologic Oncology and Clinical Medicine – Dunning-Kruger Effects and Ultracrepidarianism

Edward J. Pavlik [1,*], Jamie Land Woodward [2], Frank Lawton [3], Allison L. Swiecki-Sikora [1], Dharani D. Ramaiah [2] and Taylor A. Rives [1]

[1] Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, Chandler Medical Center-Markey Cancer Center, Lexington, KY 40536-0293, USA

[2] University of Kentucky College of Medicine, Lexington, KY 40536-0293, USA

[3] SE London Gynecological Cancer Centre

* Correspondence: edward.pavlik@uky.edu (E.J.P); Tel.: +1-(859)-323-3830

Abstract: We have published on the accuracy of the Google Virtual Assistant, Alexa, Siri, Cortana [1], Gemini and Copilot [2]. Emerging from this published work was a focus on the accuracy of AI that could be determined through validations. In our work published in 2023, the accuracy of responses to a panel of 24 queries related to gynecologic oncology was low, with Google Virtual Assistant (VA) providing the most correct audible replies (18.1%), followed by Alexa (6.5%), Siri (5.5%), and Cortana (2.3%). In the months following our publication, there was explosive excitement about several generative AIs that continue to transform the landscape of information accessibility by presenting the search results in impressively engaging narratives. This type of presentation has been enabled by combining machine learning algorithms with Natural Language Processing (NLP). In 2024, we published our exploration of the generative AIs Gemini and Copilot as well as the Google Assistant in relation to how accurately they responded to the panel of 24 queries that we used in the 2023 publication. Google Gemini achieved an 87.5% accuracy rate, while the accuracy of Microsoft Copilot was 83.3%. In contrast, the Google VA's accuracy in audible responses improved from 18% in the 2023 report to 63% in 2024. Because of our investigation in this area, we have examined the accuracy of results obtained through different AI models in this review. The landscape of the findings reviewed here surveyed 252 papers published in 2024, topically reporting on AI in medicine of which 83 articles are considered in the present review because they contain evidenced-based findings. In particular, the types of cases considered deal with AI accuracy in initial differential diagnoses, cancer treatment recommendations, board-style exams and performance in various clinical tasks. Importantly, summaries of the validation techniques used to evaluate AI findings are presented. This review focuses on those AIs that have a clinical relevancy evidenced by application and evaluation in clinical publications. This relevancy speaks to both what has been promised and what has been delivered by various AI systems. Readers will be able to understand when generative AI may be expressing views without having the necessary information (ultracrepidarianism) or is responding as if the generative AI had expert knowledge when it does not. Without an awareness that AIs may deliver inadequate or confabulated information, incorrect medical decisions and inappropriate clinical applications can result (Dunning-Kruger effect). As a result, in certain cases a generative AI might underperform and provide results which greatly overestimate any medical or clinical validity.

Keywords: artificial intelligence; large language models; generative AI; accuracy; chatbot; gynecologic/oncology

1. Introduction

Over the past two years, artificial intelligence (AI), particularly generative AI (also known as narrative AI), has garnered immense attention. This surge is largely due to the success of combining machine learning algorithms with Natural Language Processing (NLP) and Large Language Models (LLMs), demonstrated through the release of OpenAI's ChatGPT-3.5 (November 30, 2022), ChatGPT-4 (March 14, 2023) and ChatGPT-4 Turbo (November 2023). [3]. This led to the opportunity to enter commands or queries in natural language and to receive results in natural language when appropriate. It became possible to enter natural language requests for information, image generation, and even programming. Because of the advances in speech-to-text and text-to-speech, it was a short jump from text requests and texts outputs to spoken NLP inputs by humans as well as NLP outputs in responses that could be both read and listened to. Importantly, the ChatGPT family was developed to sound highly coherent. At present, it is difficult to determine the exact number of generative AI systems that exist because of a rapid ongoing emergence of new models. However, including large scale models developed by major companies and smaller-scale specialized models used by startups and researchers, there are probably thousands [4] with a curated list current to October 2024 estimating that over 5000 different generative AI model systems exist [5,6]. This large number alone underscores the reality of evolution in generative AI systems so that some will differentiate into specific niche utilizations, others will pursue a singularity dominating all other forms of intelligence [7].

Google Bard, a generative AI chatbot that was introduced in March of 2023, utilizes its own large language AI model called Gemini [8]. Gemini is able to harvest current information from the Internet and is available in over 40 languages. Microsoft launched Copilot within its Edge browser in February 2023. It uses its own large language model Prometheus, which was built on the Open AI GPT-4 foundation [9].

Highly regarded, comprehensive global reviews of AI and machine learning current to 2023 exist in the literature in relation to conventional and molecular medicine [10,11]. The present review will focus on AI systems that have a clinical relevancy evidenced by application and evaluation in clinical publications. This data-driven relevancy will speak to both what has been promised and what has been delivered by various AI systems.

2. Materials and Methods

Many factors are related to the evaluation of AI based on machine learning (ML), a subfield of AI where computers learn and recognize patterns from sets of examples, rather than being programmed with specific rules. Deep learning is a current widely used ML technique that is based on neural network technology with an architecture that tries to parallel the human brain to learn to recognize complex patterns. With this in mind, research efforts need to develop insights as to why any AI system behaves a certain way or produces a particular result. Some insights into AI performance can be gained by considering that the data in a training model may be inadequate or imperfect. It is important to note that AI capable of addressing varied and complex problems on par with human intelligence is still a long way off. For these reasons, ongoing evaluations of results produced by AI systems are necessary.

One effective test of AI performance is to specifically check the AI's ability to provide exact and real answers to queries that can be verified in the published literature. An example that is pertinent to medical practitioners using AI for practice guidance: in November 2023 Bard was asked for citations to the information that it provided for use in a continuing medical education course. In checking for accuracy, the first citation provided on the journal's web site was not able to be located or by using the DOI (digital object identifier), which returned the message "DOI not found" [12]. Repeated querying returned the original response. Finally, the chatbot was asked "Does this paper actually exist?" and then replied "I apologize for the mistake. I have double-checked, and the paper does not exist. I am still under development, and I am always learning. I will try my best to avoid making similar mistakes in the future." In another documented case, it was reported that 70% of the cited references produced by a chatbot were inaccurate [13]. In these examples, the responses may have been overly governed more by creative style rather than by extracting exact information, resulting in what is widely recognized as "hallucination." Limitations in the degree to which AI tools are trained on or can access medical datasets can contribute to the extent that the AI produces false human-like content and hallucination results [14]. Despite this being a well-known complication of narrative AI, users may not be aware of when it occurs. This example highlights a reality that the reliability of the answers provided by this rapidly emerging technology can be questionable. A generative AI model (GPT-4) performed similarly to other modern differential diagnosis generators achieving a correct most likely diagnosis in only 39% of cases [15] with the authors feeling that it performed as a "black box." Advocates for utilizing large language models in academic surgery have been open to exploring utilizations but recognize the need for validating content [16].

To ensure that use of AI in American healthcare is fair, appropriate, valid, effective, and safe, some have proposed establishing public-private partnerships to create nationwide health AI assurance labs in which best practices concepts could be applied for testing AI models. Reports on performance could be widely shared for managing AI models over time, populations and sites where various AI models are deployed [17].

3. Results

3.1. Gynecologic Oncology

There are many potential utilizations of AI in Gynecologic Oncology as the specialty spans cancer diagnostics and treatment. Gynecologic cancer diagnosis often requires expert physical exams, radiologic studies and pathology review, each of which has been considered for AI utilizations. Narrative AI in gynecologic oncology has been explored to evaluate clinical plans, as well as radiologic and pathologic results. In a structured examination in Obstetrics and Gynecology, spanning 7 sections, ChatGPT was reported to score 3.5% higher than an average historic human score, while in the gynecologic oncology section ChatGPT scored 92% against passing at 71% and a human score of 76.9% [18]. While AI may score higher on knowledge-based exams, physicians provided higher quality responses to gynecologic oncology clinical questions compared to the chatbot. In providing responses to ten questions about the nuanced knowledge and management of gynecologic cancers, physicians were more accurate (86.7%) than ChatGPT (60 %) and Bard (43 %; p < 0.001 for both) [19]. Importantly, physician responses were judged to be best in 76.7 % of evaluations versus ChatGPT (10.0 %) and Bard (13.3 %; p < 0.001). Preliminary work has been reported that describes a machine-learning model based on clinical characteristics and qualitative radiological sonographic indicators operating as a waterfall classification model (benign tumor, invasive cancer, or ovarian metastasis) for characterizing solid adnexal structures, operating with an accuracy of 86.4%, sensitivity of 93.8% and specificity of 86.96%, and distinguishing benign from malignant 90.91% of the time and nonmetastatic from metastatic 91.4% of the time [20]. Importantly, this system includes a tool to identify the contribution of each feature to the classification outcome generated by the machine learning model, thereby providing the clinician with a means for understanding how the automated system arrived at a given decision. An automatic approach of sonogram analysis based on a deep convolutional neural network (ConvNeXt-Tiny), showed robust and stable performance in classifying ultrasound images in the O-RADS v2022 (categories 2–5) radiology reporting systems with an accuracy of 85.9% [21]. A recent systematic review explored the AI role in ultrasound imaging in gynecologic oncology [22], and identified the aggregate shortcomings of published studies as having high risks of bias for subject selection (i.e., sample size, source, or unspecified scanner model, as well as data not derived from open-source datasets, lack of imaging preprocessing) and AI models that were not externally validated. However, it was felt that the current evidence supports the role of AI as a complementary clinical and research tool in diagnosis, patient stratification, and prediction of histopathological correlation in gynecological malignancies. These efforts concluded that AI models will have the ability to discriminate between benign and malignant ovarian masses. With these issues in mind, a recent systematic review and meta-analysis whittled down 3726 potential papers for inclusion to 14 papers, 8 of which utilized machine learning while 6 employed deep learning. They found that there were wide ranges in sensitivity (0.4 to 0.99) and specificity (0.82-0.99) in ovarian cancer diagnosis [23]. This report concluded that overall sensitivity was 81% (95% CI, 0.80–0.82), and specificity was 92% (95% CI, 0.92–0.93), indicating from this global assessment that AI "demonstrates good performance in ultrasound diagnoses of ovarian cancer." However, it did state that "Further prospective work is required to further validate AI for its use in clinical practice." Another preliminary investigation used machine learning to integrate an automated network for MRI image segmentation in combination with multiple conventional test indicators. This approach achieved an overall sensitivity of 91.9% and specificity of 86.9% in the detection of ovarian cancer that appeared independent of stage, with sensitivities for the detection of early and advanced stage ovarian cancer being similar [24]. In addition to radiological uses, machine learning has also been studied with regard to pathology and visual screening. Utilization of a deep-learning model, designed to automatically detect serous tubal intraepithelial carcinoma (STIC), the precursor lesion found in the fallopian tube to high-grade serous ovarian carcinoma, was reported to demonstrate increased accuracy and sensitivity with a significant reduction in slide review time when evaluated by a diverse group of pathologists from 11 countries [25]. Another deep-learning model has been reported to predict homologous recombination deficiency (HRD) from H&E slides so that breast and ovarian cancers can be treated with poly(ADP-ribose) polymerase inhibitors without the need for expensive next generation sequencing [26]. A recent study reported enhancing the diagnostic accuracy of transvaginal ultrasound for distinguishing endometrial cancer and endometrial atypical hyperplasia through the integration of artificial intelligence in women with postmenopausal bleeding, achieving high sensitivity (0.87) and specificity (0.86) using an automated segmentation approach [27]. In cervical cancer screening there have been a plethora of reports using cervical cancer images to predict cervical cancer with outstanding accuracy coupled with a disappointing lack of confirmation subsequently [28–32].

Finally, efforts integrating ChatGPT treatment recommendations to those from an institutional molecular tumor board, resulted in a 45.5% agreement with the molecular tumor board on 114 cases involving endometrial/uterine, ovarian, vulvar, cervical and undefined gynecologic cancers [33].

3.2. Clinical Medicine in General

In addition to gynecologic oncology, AI has also been adapted in various clinical applications throughout other fields of medicine. Narrative AI was heralded initially and within a year various capabilities and improvements were promoted, however, there have been published reports questioning chatbot accuracy. ChatGPT achieved an accuracy of only 60.3% in forming accurate initial differential diagnoses and its performance was described as inferior [34]. In another report, Chatbot GPT4 performed similarly to physicians and outperformed residents with regard to clinical reasoning outcomes [35]. ChatGPT-3.5 turbo-0301 significantly underperformed at providing accurate cancer treatment recommendations, and generated narratives that were discordant with NCCN recommendations a third of the time [36]. Similarly, investigation on the quality of information and misinformation about skin, lung, breast, colorectal, and prostate cancers provided by 4 AI chatbots (ChatGPT version 3.5, Perplexity (Perplexity.AI), Chatsonic (Writesonic), and Bing AI (Microsoft)) concluded that the limitations observed suggest that AI chatbots should be used as supplementary and not as a primary source for medical information [37]. Faced with answering "Questions to Ask About Your Cancer," recommended by the American Cancer Society, both ChatGPT-3.5 and Bing responded correctly in less than 80% of the cases [38]. ChatGPT-4 correctly diagnosed only 57% of complex clinical cases [39].

Machine learning also had mixed performance on board-style examinations compared to humans. When presented with a board-style neurology examination, ChatGPT4 correctly answered 85% correctly on over 1900 questions in behavioral, perceptive, and psychological–related areas, using confident language for both correct as well as incorrect answers [40]. A study on responses to subspecialty questions in nephrology by several LLM AIs ((Llama2-70B, Koala 7B, Falcon 7B, Stable-Vicuna 13B, and Orca-Mini 13B utilizing GPT-4 and Claude 2) concluded that overall 30.6% of 858 multiple choice questions in the Nephrology Self-Assessment Program were correctly answered, although Claude 2 (54.4%) and GPT-4 (73.3%) performed better [41]. However, an Israeli study compared ChatGPT-3.5 and ChatGPT-4 in head-to-head competition with 849 practicing physicians on the 2022 Israeli board residency exams and found that ChatGPT4 passed board residency exams in 4 out of 5 specialties with a score higher than the official passing score of 65%, while outperforming a considerable fraction of practicing physicians [42]. In contrast GPT-3.5 did not pass the board exam in any discipline and was inferior to the majority of physicians in the five disciplines. Using AI guidance in clinical practice can increase risk of errors. In a trial involving 457 clinicians who were randomized to study 6 vignettes using AI model input with or without AI model explanations (among these 6 vignettes, 3 vignettes included standard-model predictions, and 3 vignettes included systematically biased model predictions), physician accuracy was 73% alone and increased just 4.4% with AI explanations [43]. However, the systematically biased AI model information decreased clinician accuracy by 11.3 percentage points. Thus, there is an inherent risk that errant information provided by AI can negatively affect medical decision making by clinicians. In another report on responses by GPT-4V, Gemini Pro, and 4 language-only models: GPT-4, GPT-3.5, and 2 open-source models, Llama 2 (Meta) and Med42 to clinical vignette questions that included both case descriptions and medical imaging, all LLM models were less accurate than human responders and displayed a range of accuracies (44.1%-88.7%) with GPT-4 having the best performance [44]. Comparative evaluations of LLMs (GPT-3.5, GPT-4, PaLM 2, Claude-v1, and LLaMA 1) on 2044 clinical oncology questions also yielded variable accuracies (25.6%-68.7%), with random guess correctness estimated at 25.2% leading to the conclusion that there are models that appear to perform no better than random chance, whereas others may achieve a level of accuracy competitive with resident physicians [45]. Using additional LLMs (ChatGPT-3.5, ChatGPT-4, Mistral-7B-Instruct-v0.2, Mixtral-8x7B-v0.1, Llama-2-13b-chat, Nous-Hermes-Llama2-70b, openchat-3.5-1210, BioMistral-7B DARE) on medical oncology examination questions, and extending evaluations to 3 multimodal and 7 unimodal text only chatbots (ChatGPT-4 Vision, Claude-3 Sonnet Vision, Gemini Vision, ChatGPT-3.5, ChatGPT-4, Claude-2.1, Claude-3 Sonnet, Gemini, Llama2, and Mistral Large), similar results were reported [46,47]. In addition, ranges of performance of these same LLM models have been reported for clinical knowledge and reasoning in ophthalmology [48]. The use of AI has also been explored in day-to-day clinical tasks, such as documentation and patient communication. However, physicians were not more efficient when using voice-enabled AI to automatically document conversations between physicians, patients and their families, intended to make it possible for physicians to give their full attention to the patient while AI technology created complete, accurate clinical notes directly in the electronic hospital record for the clinician to review and sign (ambient clinical documentation via DAX Copilot) [49]. A report that considered generative AI-drafted replies for patient messages found increased physician time spent on answering messages. There was a significant increase in read time, no change in reply time, and significantly longer replies. Increased read time seemed to be attributable to the need to read both the patient's original message and the draft reply in order to avoid AI hallucinations [50]. Similarly, a significant subset of clinicians using an AI-powered clinical documentation tool did not find time-saving benefits or improved electronic health record experiences [51]. Answers by ChatGPT have been reported previously to be ~4X longer in length, but in the context that chatbot responses were rated to behave with significantly higher quality (3.6X) and ~10X better empathy than physicians [52,53]. In a study of 1600 emergency medicine patient medical records, LLM-generated emergency medicine-to-inpatient handoff notes were assessed as superior to physician-written summaries, but marginally inferior in usefulness and safety [54]. AI has also been investigated in extracting information from the medical record. A utilization of large language model AI (LLM AI) with wide appeal is the extraction of unstructured data from clinical notes in electronic medical records. In a recent report, high detail prompts from the LLM AI

(ChatGPT-4) were needed for agreement with text string searches; however, the retest reliability observed that the LLM was consistent in misclassification-type hallucinations and did not perfectly replicate its findings across new session queries on 5 different days [55]. Thus, the application of LLMs to data extraction from electronic clinical notes can accomplish considerable efficiency that is accompanied by trade-offs in accuracy, problems in interpretation by AI ("unhealed," "w/o healing," "no healing") and ambiguous terms ("S/P healing"). It is apparent at present that LLM performance requires meticulous engineering of the prompt syntax methods for understanding nuanced clinical language [56]. Other investigators reported that there was no change in reply action time, write time or read time in the application of AI-generated draft replies to patient portal messaging [57]. Of note, the evaluation of GPT-3.5, GPT-4 and Claude AI found that Claude AI provided the best overall measures of quality, empathy and readability that were comparable to responses from physicians with respect to drafts of responses to cancer patient questions [58]. Similar findings were reported for ChatGPT 3.5 in terms of correctness, conciseness, completeness, and potential for harm with the chatbot scoring 30% higher than human experts on readability of AI-generated responses to questions in radiation oncology [59]. In the Radiology medical specialty, attempts have been made to implement AI to recognize specific findings in images. A report on lower performing radiologists challenged the notion that AI assistance would increase their performance, finding that AI errors adversely influenced radiologist performance and treatment outcomes [60]. An encouraging proof-of-concept has been reported on for Med-PaLM Multimodal as a generalist AI system that interprets clinical terms, imaging, genomics using zero-shot clinical reasoning, to the extent that 4 clinicians preferred reports generated by this AI system over those produced by radiologists in 40.05% of 246 radiology case evaluations with error rates similar to human radiologists [61]. The reader of this report will notice that end results are quite nuanced across model sizes and presented more from the standpoint of computer scientists than clinicians. Similarly, it has been reported that the availability of ChatGPT Plus [GPT-4] to physicians as a diagnostic aid did not significantly improve clinical reasoning over conventional resources (UpToDate and/or Google). Importantly, the LLM alone demonstrated higher performance than the physician group using GPT-4 or the physician group using conventional resources, revealing a need for physician development to achieve elevated performance through physician-artificial intelligence collaboration in clinical practice [62]. Recently the FDA-authorized Sepsis ImmunoScore, an AI-based software designed to identify patients at risk of sepsis with the potential to identify both sepsis disease risk and sepsis mortality [63]. When queried on the energy content of 222 food items, both ChatGPT-3.5 and ChatGPT-4 provided accurate answers less than half of the time [64].

Histology and pathology have also been recognized as having potential for AI use in clinical medicine. Artificial intelligence of the nucleus based on a deep learning method has been reported to identify specific nuclear signatures at nanoscale resolution based on the spatial arrangement of core histone H3, RNA polymerase II or DNA from super-resolution microscopy images and correctly identified human somatic cells, human-induced pluripotent stem cells, very early stage infected cells transduced with DNA herpes simplex virus type 1 and even cancer cells [65].

Existing AI methods for histopathology image analyses have been limited to optimizing specialized models for each diagnostic task [66,67]. Although such methods have achieved some success, they often have limited generalizability to images obtained through different digitization protocols or samples collected from different populations [68]. When Bard and GPT-4 were tasked with interpreting 1134 pathology reports for patients, both generated reports that were significantly easier to read; however, GPT-4 interpreted reports correctly 10% better than Bard and had hallucinations that were 10% of those made by Bard [69]. As such, these types of AI interpretations need to be subject to review by clinicians prior to being made available to patients. However, a recent report described a model that was found effective for 19 anatomical sites in samples from diverse populations and processed by different slide preparation methods [70]. This system provides hope when taken in the context of a reported commercial AI algorithm developed for breast cancer detection (INSIGHT MMG, version 1.1.7.2) claiming to identify women 4-6 years prior to eventual detection in retrospective mammograms, and thereby offering a pathway that can lead to earlier breast cancer diagnosis [71]. Similarly, the identification of breast cancer relapses in the text of unstructured computed tomography (CT) reports using natural language processing (BlueBERT), achieved an accuracy of 93.6%, a sensitivity of 70.1%, and a specificity of 95.3% for regional relapses and an accuracy of 88.1%, a sensitivity of 91.8%), and a specificity of 83.7% for distant relapses [72]. When ChatGPT-4 was used to interpret clinical ophthalmic images, it accurately answered two-thirds of multiple-choice questions that required interpretations of ophthalmic images [73]. However, performance was better on questions unrelated to images than on image-based questions (82% vs 65%).

Attempts to utilize LLM AI for decisions that impact medical systems have been investigated. ChatGPT-4 Turbo was used for evaluating surgical risk stratification and postoperative outcomes. This AI was able to predict that physical status, hospital admission, intensive care unit (ICU) admission, unplanned admission, hospital mortality, and post anesthesia care fairly well; however prediction of the durations of postoperative issues were universally poor, especially by the large language model for PACU phase 1 duration, for hospital duration, and for ICU duration prediction [74]. AI using policy learning trees tested an approach, which can be linked to an electronic health record data platform to improve real-time allocation of scarce treatments [75]. By learning from multiple molecular tumor boards about biomarkers with low evidence levels, an AI system was reported to have a post-training concordance of 88% with molecular tumor boards [76]. ChatGPT-4 has been explored in evaluations of hospital emergency department triage, which involves prioritizing patients based on both the severity of their condition and their expected resource needs, retrospectively using 10,000 patient pairs finding that GPT-4 correctly identified the individual with the higher acuity with an accuracy of 89% which was similar to the accuracy of physician reviewers [77]. Finally, insurance billing and coding has been an additional way AI can be used. A machine learning AI algorithm, utilizing diagnostic item categories and diagnostic cost group methods has been reported on that can reliably price even rare diseases, avoiding serious underpayments even for the 3% of people who have at least one diagnosis as rare as 1 in 1 million [78]. Importantly this AI approach can circumvent diagnostic vagueness and attempts to game the payer system. However, GPT-3.5, GPT-4, Gemini Pro, and Llama2-70b have been reported to be poor medical coders in using ICD-9-CM, ICD-10-CM, and CPT code descriptions to generate appropriately specific billings with GPT-4 having the highest exact match rate (45.9% on ICD-9-CM, 33.9% on ICD-10-CM, 49.8% on CPT), but often generating codes conveying imprecise or fabricated information [79].

4. Conclusions

This review has considered the range of results from various AIs, which in some cases have been poor and underperformed results from trained clinicians. While AI may be successful at answering board questions, responses to patients questions are often incorrect and AI writing clinical notes does not seem to improve efficiency. When AI was combined with physician diagnostics in radiology, the AI negatively influenced physician decision making. While AI usage may be hopeful for radiology, treatment matching and pathology, expensive and diverse learning is required for accuracy. It has been the position of physicians that patients should consent to AI use, and that patients should have opportunities of choosing between physician- and AI-recommended treatment plans [80]. These physician-based perspectives highlight the need for rigorous assessments of how AI impacts oncology care, along with clearly defined accountability for decision-making when issues arise from AI utilization. Key concerns among these physicians include ethical considerations, such as explainability, patient consent, and responsibility, all of which are essential for optimal adoption of AI into cancer care. Similarly, 62.7% of people surveyed in the US stated it was very true that they want to be notified if AI would be involved in their care, while less than 5% answered that they did not find notification important [81].

Issues of AI accuracy have been outlined here in reviews of evidence-based reports. A recent search in Copilot ends with the statement, “AI-generated content may be incorrect.” In essence, this is to say that this AI exhibits the Dunning-Kruger effect by demonstrating unconscious incompetence (i.e. not knowing what it doesn't know). It is important that users and adopters of AI are also sensitive to the ultracrepidarian identity of AI, which tempts users to believe that an AI application is expert in all fields or medical specialties (an "everythingologist") even though it's expertise is limited to only certain areas. The old Russian phrase ("*Doveryai, No Proveryai*") meaning "*trust but verify*" quoted by Ronald Reagan [82,83] certainly has application to AI, but it raises the question of how to verify. Will there be efforts in place to identify AI hallucination or confabulation? Indeed, how is accountability to be positioned? To what extent is the AI creator/vendor, hospital, clinic, or performing physician responsible for inaccurate or faulty results from an AI that is utilized? It should not be ignored that artificial intelligence can generate or be intentionally used for prevarication, misinformation or targeted disinformation. The ability of generative AI to rapidly generate diverse and large amounts of convincing disinformation about vaccination and vaping has been reported to be profound, especially when operation is allowed with few or no guiderails in place [84]. Such results when targeted to blogs or social media postings can amount to intentional Weapons of Mass Disinformation replete with scientific-looking reference citations. In these instances, there must be tools in place for fact checking the language content of information originating from generative AI. Lastly, in an AI dominated clinical world will human clinical skills be subject to atrophy? Will this possibility have a negative effect on both clinical training and clinical research? The ideas considered here are current to the early part of 2025. It is entirely possible that developments in the near future will address the issues raised here. The AI genie has indeed left the bottle. The degree to which our wishes are granted by the AI genie must be examined.

Author Contributions: Conceptualization, E.J.P. and J.L.W.; methodology, J.L.W.; software, E.J.P. and J.L.W.; validation, E.J.P.; formal analysis, E.J.P.; investigation, J.L.W., D.D.R., T.A.R. and A.L.S.-S.; resources, E.J.P.; data curation, E.J.P.; writing—original draft preparation, E.J.P.; writing/review/editing, J.L.W., F.L., D.D.R., T.A.R., A.L.S.-S. and E.J.P.; visualization, J.L.W.; supervision, E.J.P.; project administration, E.J.P.; funding acquisition, E.J.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data will be made available upon request.

Conflicts of Interest: The authors declare no conflict of interest

References

1. Land JM, Pavlik EJ, Ueland E, et al. Evaluation of replies to voice queries in gynecologic oncology by virtual assistants Siri, Alexa, Google, and Cortana. BioMed Informatics. 2023;3(3):553-562. https://doi.org/10.3390/biomedinformatics3030038
2. Pavlik EJ, Ramaiah DD, Rives TA, Swiecki-Sikora AL, Land JM. Replies to Queries in Gynecologic Oncology by Bard, Bing, and the Google Assistant. *BioMed Informatics*. 2024;4(3):1773-1782. https://doi.org/10.3390/biomedinformatics4030097.
3. Brandl R, Ellis C. ChatGPT statistics 2024: All the latest statistics about OpenAI's chatbot. Tooltester. Published 2024. Available at: https://www.tooltester.com/en/blog/chatgpt-statistics/. Accessed January 27, 2025.
4. Google. (2024). Gemini [Large language model]. https://ludwig.guru/s/if+available
5. Vogel M. A curated list of resources on generative AI. Medium. Updated October 9, 2024. Available at: https://medium.com/@maximilian.vogel/5000x-generative-ai-intro-overview-models-prompts-technology-tools-comparisons-the-best-a4af95874e94#id_token=eyJhbGciOiJSUzI1NiIsImtpZCI6IjFkYzBmMTcyZThkNmVmMzgyZDZkM2EyMzFmNmMxOTdkZDY4Y2U1ZWYiLCJ0eXAiOiJKV1QifQ. Accessed January 27, 2025.

6. Yang J, Jin H, Tang J, et al. The practical guides for large language models. Available at: https://github/com/Mooler0410/LLMsPracticalGuide. Accessed January 27, 2025
7. Kurzweil R. The singularity is nearer: When we merge with AI. Penguin Books; 2024. ISBN 9780399562761.
8. Ortiz S. What is Google Bard? Here's everything you need to know. ZDNET. February 7, 2024. Available at: https://www.zdnet.com/article/what-is-google-bard-heres-everything-you-need-to-know/. Accessed February 13, 2024.
9. Microsoft Copilot. Wikipedia. Available at: https://en.wikipedia.org/wiki/Microsoft_Copilot. Accessed February 13, 2024.
10. Haug C.J., & Drazen, J.M. (2023). Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. New England Journal of Medicine, 388(13), 1201-1208. https://doi.org/10.1056/NEJMra2302038
11. Gomes B, Ashley EA. Artificial Intelligence in Molecular Medicine. N Engl J Med. 20224 Jun 29;388(26):2456-2465. https://doi.org/10.1056/NEJMra2204787. PMID: 37379136.
12. Colasacco CJ, Born HL. A Case of Artificial Intelligence Chatbot Hallucination. JAMA Otolaryngol Head Neck Surg. 2024 Jun 1;150(6):457-458. https://doi.org/10.1001/jamaoto.2024.0428. PMID: 38635259.
13. Kacena MA, Plotkin LI, Fehrenbacher JC. The Use of Artificial Intelligence in Writing Scientific Review Articles. Curr Osteoporos Rep. 2024 Feb;22(1):115-121. https://doi.org/10.1007/s11914-023-00852-0. Epub 2024 Jan 16. PMID: 38227177; PMCID: PMC10912250.
14. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large Language Models in Medicine: The Potentials and Pitfalls : A Narrative Review. Ann Intern Med. 2024 Feb;177(2):210-220. https://doi.org/10.7326/M23-2772. Epub 2024 Jan 30. PMID: 38285984.
15. Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA. 2023;330(1):78–80. https://doi.org/10.1001/jama.2023.8288
16. Rengers TA, Thiels CA, Salehinejad H. Academic Surgery in the Era of Large Language Models: A Review. JAMA Surg. 2024;159(4):445–450. https://doi.org/10.1001/jamasurg.2023.6496
17. Shah NH, Halamka JD, Saria S, et al. A Nationwide Network of Health AI Assurance Laboratories. JAMA. 2024;331(3):245–249. https://doi.org/10.1001/jama.2023.26930
18. Li SW, Kemp MW, Logan SJS, et al. ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. *Am J Obstet Gynecol*. 2023;229(2):172.e1-172.e12. https://doi.org/10.1016/j.ajog.2023.04.020
19. Anastasio MK, Peters P, Foote J, Melamed A, Modesitt SC, Musa F, Rossi E, Albright BB, Havrilesky LJ, Moss HA. The doc versus the bot: A pilot study to assess the quality and accuracy of physician and chatbot responses to clinical questions in gynecologic oncology. *Gynecol Oncol Rep*. 2024;55:101477. https://doi.org/10.1016/j.gore.2024.101477. PMID:39224817; PMCID:PMC11367046.
20. Fanizzi A, Arezzo F, Cormio G, Comes MC, Cazzato G, Boldrini L, Bove S, Bollino M, Kardhashi A, Silvestris E, Quarto P, Mongelli M, Naglieri E, Signorile R, Loizzi V, Massafra R. An explainable machine learning model to solid adnexal masses diagnosis based on clinical data and qualitative ultrasound indicators. Cancer Med. 2024 Jun;13(12):e7425. https://doi.org/10.1002/cam4.7425. PMID: 38923847; PMCID: PMC11196372.
21. Liu T, Miao K, Tan G, et al. A study on automatic O-RADS classification of sonograms of ovarian adnexal lesions based on deep convolutional neural networks. *Ultrasound Med Biol*. 2025;51(2):387-395. https://doi.org/10.1016/j.ultrasmedbio.2024.11.009. Epub 2024 Nov 26. PMID: 39603844.
22. Moro F, Ciancia M, Zace D, Vagni M, Tran HE, Giudice MT, Zoccoli SG, Mascilini F, Ciccarone F, Boldrini L, D'Antonio F, Scambia G, Testa AC. Role of artificial intelligence applied to ultrasound in gynecology oncology: A systematic review. Int J Cancer. 2024 Nov 15;155(10):1832-1845. https://doi.org/10.1002/ijc.35092. Epub 2024 Jul 11. PMID: 38989809.
23. Mitchell S, Nikolopoulos M, El-Zarka A, Al-Karawi D, Al-Zaidi S, Ghai A, Gaughran JE, Sayasneh A. Artificial intelligence in ultrasound diagnoses of ovarian cancer: A systematic review and meta-analysis. *Cancers (Basel)*. 2024;16(2):422. https://doi.org/10.3390/cancers16020422. PMID:38275863; PMCID:PMC10813993.
24. Feng, Y. An integrated machine learning-based model for joint diagnosis of ovarian cancer with multiple test indicators. J Ovarian Res 17, 45 (2024). https://doi.org/10.1186/s13048-024-01365-9

25. Bogaerts JM, Steenbeek MP, Bokhorst JM, et al. Assessing the impact of deep-learning assistance on the histopathological diagnosis of serous tubal intraepithelial carcinoma (STIC) in fallopian tubes. J Pathol Clin Res. 2024 Nov;10(6):e70006. https://doi.org/10.1002/2056-4538.70006. PMID: 39439213; PMCID: PMC11496567.
26. Bergstrom EN, et al. Deep learning artificial intelligence predicts homologous recombination deficiency and platinum response from histologic slides. *J Clin Oncol*. 2024;42:3550-3560. https://doi.org/10.1200/JCO.23.02641.
27. Capasso I, Cucinella G, Wright DE, et al. Artificial intelligence model for enhancing the accuracy of transvaginal ultrasound in detecting endometrial cancer and endometrial atypical hyperplasia. Int J Gynecol Cancer. 2024 Oct 7;34(10):1547-1555. https://doi.org/10.1136/ijgc-2024-005652. PMID: 39089731.
28. Hu L, Bell D, Antani S, et al. An observational study of deep learning and automated evaluation of cervical images for cancer screening. *JNCI: Journal of the National Cancer Institute*. 2019;111(9):923-932. https://doi.org/10.1093/jnci/djy225.
29. Xue Z, Novetsky AP, Einstein MH, et al. A demonstration of automated visual evaluation of cervical images taken with a smartphone camera. Int. J. Cancer. 2020; 147: 2416–2423. https://doi.org/10.1002/ijc.33029
30. Desai KT, Befano B, Xue Z, et al. The development of "automated visual evaluation" for cervical cancer screening: The promise and challenges in adapting deep-learning for clinical testing. Int. J. Cancer. 2022; 150(5): 741-752. https://doi.org/10.1002/ijc.33879
31. Parham, G.P., Egemen, D., Befano, B. et al. Validation in Zambia of a cervical screening strategy including HPV genotyping and artificial intelligence (AI)-based automated visual evaluation. Infect Agents Cancer 18, 61 (2023). https://doi.org/10.1186/s13027-023-00536-5
32. Egemen D, Perkins RB, Cheung LC, et al. Artificial intelligence–based image analysis in clinical testing: lessons from cervical cancer screening. JNCI: J Natl Cancer Inst. 2024 Jan;116(1):26-33. https://doi.org/10.1093/jnci/djad202.
33. Rios-Doria E, Wang J, Rodriguez I, et al. Artificial intelligence powered insights: Assessment of ChatGPT's treatment recommendations in gynecologic oncology. *Gynecol Oncol*. 2024;190(Suppl 1):S45. https://doi.org/10.1016/j.ygyno.2024.07.071.
34. Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. J Med Internet Res. 2023;25:e48659. https://doi.org/10.2196/48659.
35. Cabral S, Restrepo D, Kanjee Z, et al. Clinical reasoning of a generative artificial intelligence model compared with physicians. *JAMA Intern Med*. 2024;184(5):581-583. https://doi.org/10.1001/jamainternmed.2024.0295.
36. Chen S, Kann BH, Foote MB, Aerts HJWL, Savova GK, Mak RH, Bitterman DS. Use of Artificial Intelligence Chatbots for Cancer Treatment Information. JAMA Oncol. 2023 Oct 1;9(10):1459-1462. https://doi.org/10.1001/jamaoncol.2023.2954
37. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries About Cancer. JAMA Oncol. 2023;9(10):1437–1440. https://doi.org/10.1001/jamaoncol.2023.2947
38. Janopaul-Naylor JR, Koo A, Qian DC, et al. Physician assessment of ChatGPT and Bing answers to American Cancer Society's questions to ask about your cancer. Am J Clin Oncol. 2024;47(1):17-21. https://doi.org/10.1097/COC.0000000000001050.
39. Shea YF, Lee CMY, Ip WCT, et al. Use of GPT-4 to analyze medical records of patients with extensive investigations and delayed diagnosis. *JAMA Netw Open*. 2023;6(8):e2325000. https://doi.org/10.1001/jamanetworkopen.2023.25000.
40. Schubert MC, Wick W, Venkataramani V. Performance of large language models on a neurology board-style examination. *JAMA Netw Open*. 2023;6(12):e2346721. https://doi.org/10.1001/jamanetworkopen.2023.46721.
41. Wu S, Koo M, Blum L, et al. Benchmarking open-source large language models, GPT-4 and Claude 2 on multiple-choice questions in nephrology. *NEJM AI.* 2024;1(2):doi:10.1056/AIdbp2300092.
42. Katz U, Cohen E, Shachar E, et al. GPT versus resident physicians — a benchmark based on official board scores. *NEJM AI*. 2024;1(5). https://doi.org/10.1056/AIdbp2300192.

43. Jabbour S, Fouhey D, Shepard S, et al. Measuring the Impact of AI in the Diagnosis of Hospitalized Patients: A Randomized Clinical Vignette Survey Study. JAMA. 2023;330(23):2275–2284. https://doi.org/10.1001/jama.2023.22295
44. Han T, Adams LC, Bressem KK, et al. Comparative analysis of multimodal large language model performance on clinical vignette questions. *JAMA*. 2024;331(15):1320-1321. https://doi.org/10.1001/jama.2023.27861.
45. Rydzewski NR, Dinakaran D, Zhao SG, et al. Comparative evaluation of LLMs in clinical oncology. *NEJM AI*. 2024;1(5). https://doi.org/10.1056/aioa2300151. Epub 2024 Apr 16..
46. Longwell JB, Hirsch I, Binder F, et al. Performance of large language models on medical oncology examination questions. *JAMA Netw Open*. 2024;7(6):e2417641. https://doi.org/10.1001/jamanetworkopen.2024.17641.
47. Chen D, Huang RS, Jomy J, et al. Performance of multimodal artificial intelligence chatbots evaluated on clinical oncology cases. *JAMA Netw Open*. 2024;7(10):e2437711. https://doi.org/10.1001/jamanetworkopen.2024.37711.
48. Thirunavukarasu AJ, Mahmood S, Malem A, et al. Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: A head-to-head cross-sectional study. *PLOS Digit Health*. 2024;3(4):e0000341. https://doi.org/10.1371/journal.pdig.0000341.
49. Liu T-L, Hetherington TC, Dharod A, et al. Does AI-powered clinical documentation enhance clinician efficiency? A longitudinal study. *NEJM AI*. 2024;1(12). https://doi.org/10.1056/AIoa2400659.
50. Tai-Seale M, Baxter SL, Vaida F, et al. AI-Generated Draft Replies Integrated Into Health Records and Physicians' Electronic Communication. JAMA Netw Open. 2024;7(4):e246565. https://doi.org/10.1001/jamanetworkopen.2024.6565
51. Liu TL, Hetherington TC, Stephens C, McWilliams A, Dharod A, Carroll T, Cleveland JA. AI-Powered Clinical Documentation and Clinicians' Electronic Health Record Experience: A Nonrandomized Clinical Trial. JAMA Netw Open. 2024 Sep 3;7(9):e2432460. https://doi.org/10.1001/jamanetworkopen.2024.32460. PMID: 39240568; PMCID: PMC11380097.
52. Small WR, Wiesenfeld B, Brandfield-Harvey B, et al. Large language model–based responses to patients' in-basket messages. *JAMA Netw Open*. 2024;7(7):e2422399. https://doi.org/10.1001/jamanetworkopen.2024.22399.
53. 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(6):589-596. https://doi.org/10.1001/jamainternmed.2023.1838. PMID:37115527; PMCID:PMC10148230.
54. Hartman V, Zhang X, Poddar R, et al. Developing and evaluating large language model-generated emergency medicine handoff notes. *JAMA Netw Open*. 2024;7(12):e2448723. https://doi.org/10.1001/jamanetworkopen.2024.48723. PMID:39625719; PMCID:PMC11615705.
55. . Burford KG, Itzkowitz NG, Ortega AG, Teitler JO, Rundle AG. Use of Generative AI to Identify Helmet Status Among Patients With Micromobility-Related Injuries From Unstructured Clinical Notes. JAMA Netw Open. 2024 Aug 1;7(8):e2425981. https://doi.org/10.1001/jamanetworkopen.2024.25981. PMID: 39136946; PMCID: PMC11322845.
56. Shah SV. Accuracy, Consistency, and Hallucination of Large Language Models When Analyzing Unstructured Clinical Notes in Electronic Medical Records. JAMA Netw Open. 2024 Aug 1;7(8):e2425953. https://doi.org/10.1001/jamanetworkopen.2024.25953. PMID: 39136951.
57. Garcia P, Ma SP, Shah S, et al. Artificial Intelligence–Generated Draft Replies to Patient Inbox Messages. JAMA Netw Open. 2024;7(3):e243201. https://doi.org/10.1001/jamanetworkopen.2024.3201
58. Chen D, Parsa R, Hope A, et al. Physician and artificial intelligence chatbot responses to cancer questions from social media. *JAMA Oncol*. 2024;10(7):956-960. https://doi.org/10.1001/jamaoncol.2024.0836.
59. Yalamanchili A, Sengupta B, Song J, et al. Quality of large language model responses to radiation oncology patient care questions. *JAMA Netw Open*. 2024;7(4):e244630. https://doi.org/10.1001/jamanetworkopen.2024.4630.
60. Yu F, Moehring A, Banerjee O, et al. Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nat Med*. 2024;30:837-849. https://doi.org/10.1038/s41591-024-02850-w.

61. Tu T, Azizi S, Driess D, et al. Towards generalist biomedical AI. *NEJM AI*. 2024;1(3). https://doi.org/10.1056/AIoa2300138.
62. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Netw Open*. 2024;7(10):e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969.
63. Bhargava A, López-Espina C, Schmalz L, et al. FDA-authorized AI/ML tool for sepsis prediction: development and validation. *NEJM AI*. 2024;1(12). https://doi.org/10.1056/AIoa2400867.
64. Hoang YN, Chen YL, Ho DK, et al. Consistency and accuracy of artificial intelligence for providing nutritional information. *JAMA Netw Open*. 2023;6(12):e2350367. https://doi.org/10.1001/jamanetworkopen.2023.50367.
65. Carnevali, D., Zhong, L., González-Almela, E. et al. A deep learning method that identifies cellular heterogeneity using nanoscale nuclear features. Nat Mach Intell 6, 1021–1033 (2024). https://doi.org/10.1038/s42256-024-00883-x
66. van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. Nat Med. 2021 May;27(5):775-784. https://doi.org/10.1038/s41591-021-01343-4. Epub 2021 May 14. PMID: 33990804.
67. Shmatko A, Ghaffari Laleh N, Gerstung M, Kather JN. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. Nat Cancer. 2022 Sep;3(9):1026-1038. https://doi.org/10.1038/s43018-022-00436-4. Epub 2022 Sep 22. PMID: 36138135.
68. Reis-Filho JS, Kather JN. Overcoming the challenges to implementation of artificial intelligence in pathology. *JNCI J Natl Cancer Inst*. 2023;115(6):608-612. https://doi.org/10.1093/jnci/djad048.
69. Steimetz E, Minkowitz J, Gabutan EC, et al. Use of artificial intelligence chatbots in interpretation of pathology reports. *JAMA Netw Open*. 2024;7(5):e2412767. https://doi.org/10.1001/jamanetworkopen.2024.12767.
70. Wang, X., Zhao, J., Marostica, E. et al. A pathology foundation model for cancer diagnosis and prognosis prediction. Nature 634, 970–978 (2024). https://doi.org/10.1038/s41586-024-07894-z
71. Gjesvik J, Moshina N, Lee CI, Miglioretti DL, Hofvind S. Artificial Intelligence Algorithm for Subclinical Breast Cancer Detection. JAMA Netw Open. 2024 Oct 1;7(10):e2437402. https://doi.org/10.1001/jamanetworkopen.2024.37402. PMID: 39361281; PMCID: PMC11450515.
72. Lee JJ, Zepeda A, Arbour G, et al. Automated identification of breast cancer relapse in computed tomography reports using natural language processing. *JCO Clin Cancer Inform*. 2024;8:e2400107. https://doi.org/10.1200/CCI.24.00107.
73. Mihalache A, Huang RS, Popovic MM, Patil NS, Pandya BU, Shor R, Pereira A, Kwok JM, Yan P, Wong DT, Kertes PJ, Muni RH. Accuracy of an Artificial Intelligence Chatbot's Interpretation of Clinical Ophthalmic Images. JAMA Ophthalmol. 2024 Apr 1;142(4):321-326. https://doi.org/10.1001/jamaophthalmol.2024.0017. PMID: 38421670; PMCID: PMC10905373.
74. Chung P, Fong CT, Walters AM, et al. Large language model capabilities in perioperative risk prediction and prognostication. *JAMA Surg*. 2024;159(8):928-937. https://doi.org/10.1001/jamasurg.2024.1621.
75. Xiao M, Molina KC, Aggarwal NR, et al. A Machine Learning Method for Allocating Scarce COVID-19 Monoclonal Antibodies. JAMA Health Forum. 2024;5(9):e242884. https://doi.org/10.1001/jamahealthforum.2024.2884
76. Sunami K, Naito Y, Saigusa Y, et al. A Learning Program for Treatment Recommendations by Molecular Tumor Boards and Artificial Intelligence. JAMA Oncol. 2024;10(1):95–102. https://doi.org/10.1001/jamaoncol.2023.5120
77. Williams CYK, Zack T, Miao BY, et al. Use of a large language model to assess clinical acuity of adults in the emergency department. *JAMA Netw Open*. 2024;7(5):e248895. https://doi.org/10.1001/jamanetworkopen.2024.8895.
78. Andriola C, Ellis RP, Siracuse JJ, et al. A Novel Machine Learning Algorithm for Creating Risk-Adjusted Payment Formulas. JAMA Health Forum. 2024;5(4):e240625. https://doi.org/10.1001/jamahealthforum.2024.0625
79. Soroush A, Glicksberg BS, Zimlichman E, et al. Large language models are poor medical coders — benchmarking of medical code querying. *NEJM AI*. 2024;1(5). https://doi.org/10.1056/AIdbp2300040.

80. Hantel A, Walsh TP, Marron JM, et al. Perspectives of oncologists on the ethical implications of using artificial intelligence for cancer care. *JAMA Netw Open*. 2024;7(3):e244077. https://doi.org/10.1001/jamanetworkopen.2024.4077. PMID:38546644; PMCID:PMC10979310.
81. Platt J, Nong P, Carmona G, Kardia S. Attitudes toward notification of use of artificial intelligence in health care. *JAMA Netw Open*. 2024;7(12):e2450102. https://doi.org/10.1001/jamanetworkopen.2024.50102.
82. Centripetal: Evolving Trust But Verify. https://www.centripetal.ai/blog/trust-but-verify-threat-intelligence/
83. User clip: trust but verify December 8, 1987. https://www.c-span.org/clip/white-house-event/user-clip-trust-but-verify/4757483 Accessed January 27, 2025
84. Menz BD, Modi ND, Sorich MJ, Hopkins AM. Health disinformation use case highlighting the urgent need for artificial intelligence vigilance: Weapons of mass disinformation. *JAMA Intern Med*. 2024;184(1):92-96. https://doi.org/10.1001/jamainternmed.2023.5947. PMID:37955873.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.