Towards Clinical Generative Artificial Intelligence: Conceptual Framework

Nicola Bragazzi, Sergio Garbarino

Submitted to: JMIR AI
on: December 30, 2023

Table of Contents

Towards Clinical Generative Artificial Intelligence: Conceptual Framework

Nicola Bragazzi[1]; Sergio Garbarino[2]

[1]University of Parma Parma IT
[2]University of Genoa Genoa IT

Corresponding Author:
Nicola Bragazzi
University of Parma
Via Volturno 39
Parma
IT

Abstract

Clinical decision-making is a crucial aspect of healthcare, involving the balanced integration of scientific evidence, clinical judgment, ethical considerations, and patient involvement. This process is dynamic and multifaceted, relying on clinicians' knowledge, experience, and intuitive understanding to achieve optimal patient outcomes through informed, evidence-based choices. The advent of generative Artificial Intelligence (AI) presents a revolutionary opportunity in clinical decision-making. AI's advanced data analysis and pattern recognition capabilities can significantly enhance the diagnosis and treatment of diseases, processing vast medical data to identify patterns, tailor treatments, predict disease progression, and aid in proactive patient management. However, the incorporation of AI into clinical decision-making raises concerns regarding the reliability and accuracy of AI-generated insights. To address these concerns, eleven “verification paradigms” are here proposed, with each paradigm offering unique methods to verify the evidence-based nature of AI in clinical decision-making. The paper also frames the concept of “clinically explainable, fair, and responsible, clinician-, expert-, and patient-in-the-loop AI”. This model focuses on ensuring AI's comprehensibility, collaborative nature, and ethical grounding, advocating for AI to serve as an augmentative tool, with its decision-making processes being transparent and understandable to clinicians and patients. The integration of AI should enhance, not replace, the clinician’s judgment and should involve continuous learning and adaptation based on real-world outcomes and ethical and legal compliance. In conclusion, while generative AI holds immense promise in enhancing clinical decision-making, it is essential to ensure that it produces evidence-based, reliable, and impactful knowledge. Employing the outlined paradigms and approaches can help the medical and patient communities harness AI's potential while maintaining high patient care standards.

(JMIR Preprints 30/12/2023:55957)
DOI: https://doi.org/10.2196/preprints.55957

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Towards Clinical Generative Artificial Intelligence: Conceptual Framework

Nicola Luigi Bragazzi[1,2,3,*], Sergio Garbarino[3,4]

[1]Laboratory for Industrial and Applied Mathematics (LIAM), Department of Mathematics and Statistics, York University, Toronto, ON, Canada.

[2]Human Nutrition Unit (HNU), Department of Food and Drugs, University of Parma, Parma, Italy.

[3]Department of Neuroscience, Rehabilitation, Ophthalmology, Genetics and Maternal/Child Sciences (DINOGMI), University of Genoa, Genoa, Italy.

[4]Post-Graduate School of Occupational Health, Università Cattolica del Sacro Cuore, Rome, Italy.

Corresponding Author Email: nicolaluigi.bragazzi@unipr.it

Corresponding Author address: Human Nutrition Unit (HNU), Department of Food and Drugs, Medical School, Building C, Via Volturno, 39, 43125 Parma, Italy

Abstract

Clinical decision-making is a crucial aspect of healthcare, involving the balanced integration of scientific evidence, clinical judgment, ethical considerations, and patient involvement. This process is dynamic and multifaceted, relying on clinicians' knowledge, experience, and intuitive understanding to achieve optimal patient outcomes through informed, evidence-based choices. The advent of generative Artificial Intelligence (AI) presents a revolutionary opportunity in clinical decision-making. AI's advanced data analysis and pattern recognition capabilities can significantly enhance the diagnosis and treatment of diseases, processing vast medical data to identify patterns, tailor treatments, predict disease progression, and aid in proactive patient management. However, the incorporation of AI into clinical decision-making raises concerns regarding the reliability and accuracy of AI-generated insights. To address these concerns, eleven “verification paradigms” are here proposed, with each paradigm being a unique method to verify the evidence-based nature of AI in clinical decision-making. The paper also frames the concept of “clinically explainable, fair, and responsible, clinician-, expert-, and patient-in-the-loop AI”. This model focuses on ensuring AI's comprehensibility, collaborative nature, and ethical grounding, advocating for AI to serve as an augmentative tool, with its decision-making processes being transparent and understandable to clinicians and patients. The integration of AI should enhance, not replace, the clinician’s judgment and should involve continuous learning and adaptation based on real-world outcomes and ethical and legal compliance. In conclusion, while generative AI holds immense promise in enhancing clinical decision-making, it is essential to ensure that it produces evidence-based, reliable, and impactful knowledge. Employing the outlined paradigms and approaches can help the medical and patient communities harness AI's potential while maintaining high patient care standards.

Keywords: clinical intelligence; artificial intelligence; iterative process; abduction; benchmarking; verification paradigms

Clinical decision-making and clinical intelligence

Clinical decision-making can be defined as a fundamental aspect of healthcare practice, encompassing a wide set of skills, competencies, processes, and outcomes through which clinicians gather and analyze relevant patient data, differentiate between various conditions, diagnose, treat, and manage patient care, balancing the effectiveness, risks, and benefits of each treatment, patient preferences and other related values within broader societal and cultural contexts, and guidelines or standards of care [1-3].

Clinical decision-making involves a complex interplay of research and biomedical knowledge, experience, and intuitive understanding, developed through years of practice, contextual analytical reasoning, patient-centeredness, and compliance with ethical standards and legal requirements, with the goal of arriving at optimal health outcomes for patients by making informed, evidence-based, and shared choices, while ensuring patient autonomy and confidentiality [4,5].

The major four pillars of clinical decision-making are scientific evidence, clinical judgment (in some complex cases, not isolated to one clinician but involving a team of healthcare professionals, each contributing their expertise), ethical considerations, and patient involvement, which are pivotal to the delivery of high-quality healthcare [6,7].

Clinical decision-making is not a static but rather dynamic, multifaceted, iterative process, based on reflective practice, which implies reviewing and auditing clinical decisions and outcomes to continuously learn and improve decision-making skills, in the face of uncertainty and epistemic risks [5,8].

The advent of generative artificial intelligence and its role in supporting clinical decision-making

Artificial Intelligence (AI) [9] and, in particular, generative AI [10] have the potential to revolutionize the field of clinical decision-making with their advanced capabilities in data analysis and pattern recognition. However, together with their rise, there is a growing necessity to ensure that the knowledge utilized and produced is evidence-based and reliable. This necessity stems from the potential risks and biases associated with AI-generated insights that may not align with established medical knowledge or practices.

Generative AI can process vast amounts of medical data, including patient records, imaging data, laboratory results, other diagnostic inputs, and clinical studies, as well as research papers, to identify patterns and correlations that might be missed by clinicians. By analyzing patient data, generative AI can help in tailoring treatments to individual patients, improving the efficacy of therapies and reducing side effects, predicting disease progression and potential complications, aiding clinicians in proactive patient management, and assisting in diagnosing diseases, potentially identifying conditions earlier and more accurately than traditional methods [11].

On the other hand, generative AI can produce “hallucinations” or even “fabrications” and “falsifications”, generating inaccurate or misleading information that does not accurately reflect the data it was trained on or reality [12,13], which is of particular concern in the medical realm.

Addressing these challenges requires a multifaceted approach, including improving dataset quality and diversity, refining model architectures, and incorporating mechanisms for fact-checking and validation. Moreover, developing methodologies for the model to express uncertainty or request clarification when generating outputs on topics where it has less confidence could enhance reliability. In real-world clinical applications where accuracy and truthfulness are paramount, it is crucial to implement safeguards such as human oversight, rigorous testing across diverse scenarios, and continuous monitoring and updating of AI-based models to mitigate the risks associated with these inaccuracies.

In the present conceptual paper, to address these concerns, we will introduce eleven “verification paradigms”, with each paradigm being a unique method to verify the evidence-based nature of AI in clinical decision-making.

Comparing and contrasting clinical *versus* Artificial Intelligence reasoning

Interesting parallelisms between clinical decision-making and AI reasoning can be drawn (Figure 1), especially in the context of frequentist/Bayesian thinking and large language models (LLMs) like GPT-4, which use conditional probability, revealing an interesting interplay of similarities and contrasts [5].

In clinical decision-making, the reliance on scientific evidence mirrors AI's dependence on extensive datasets for training. Clinicians, through years of practice, develop an intuitive sense of diagnosis and treatment. Clinical reasoning often involves abductive reasoning, which is a form of logical inference that starts with an observation or set of observations and, then, seeks to find the simplest and most likely explanation. In clinical practice, this means forming hypotheses based on symptoms and available data to diagnose a patient's condition. AI, particularly in fields like machine learning and diagnostic algorithms, also frequently uses abductive reasoning: AI-based systems are, indeed, designed to analyze data, identify patterns, and make predictions or decisions based on that analysis. In many ways, this mirrors the process of abductive reasoning where the most likely conclusion is drawn from the available information. For example, in medical diagnostics, AI-based systems might analyze patients' symptoms, medical history, and test results to suggest possible diagnoses. The aspect of human expertise underlying clinical reasoning somewhat parallels how AI-enhanced models develop a form of "intuition" from their vast training data [14,15].

When faced with complex cases, clinical decision-making often involves a collaborative approach among healthcare professionals, akin to the multi-faceted approach of AI that integrates diverse data sources and algorithms. Ethical considerations and patient involvement are central to clinical decisions, much like how AI-based models need to be ethically aligned and user-centric. Furthermore, both fields are dynamic and iterative: clinicians continually adapt their methods based on new research and patient feedback, similar to how AI-enhanced models evolve with new data and interactions.

On the AI side, traditional models often align with frequentist statistics, where the frequency of past events informs future predictions, somewhat like clinicians using epidemiological data. Modern AI, particularly in machine learning, employs Bayesian methods, updating the likelihood of outcomes with new data, reflecting how clinicians revise their hypotheses about diagnoses or treatments as new patient information comes to light. LLMs like GPT-4 can predict outcomes based on conditional probability, which can be compared to clinicians using symptoms to predict diagnoses [16].

AI's proficiency in pattern recognition and predictive analysis also finds a parallel in clinical practice, where patterns in patient symptoms and test results are crucial for effective decision-making. However, despite these parallelisms, significant differences remain, with AI lacking the empathetic and deeply intuitive component inherent in human decision-making and with clinicians interpreting data within a broader human context, an ability AI has yet to fully replicate.

In essence, while there are notable similarities in the use of statistical methods and data analysis between clinical decision-making and AI reasoning, the human aspects of intuition, empathy, and ethical considerations underscore the unique characteristics of each field. The future of healthcare may lie in the harmonious integration of these two domains, leveraging the strengths of each to enhance medical care and improve patient outcomes (Figure 1).

Towards Clinical Large Language Models: Necessity of Verifying Evidence-Based Knowledge

However, the integration of generative AI into clinical decision-making necessitates a rigorous verification process to ensure the reliability and accuracy of the AI-generated insights. This verification is crucial because, as previously mentioned, AI-based models can sometimes generate conclusions based on flawed or biased data, leading to inaccurate or even harmful recommendations. It is essential that AI-generated advice aligns with current medical standards and best practices, besides adhering to ethical standards, respecting patient autonomy, and ensuring equitable treatment

[17,18].

Clinically oriented LLMs [19-25], like ClinicalBERT, BlueBERT, CAML, DRG-LLaMA, GatorTronGPT, or PaLM, have shown impressive capabilities, yet their application in clinical settings faces stringent requirements. Traditional methods of assessing these models' clinical knowledge often depend on automated evaluations using narrow benchmarks. To overcome these shortcomings, Singhal et al. [25] recently introduced MultiMedQA, a comprehensive benchmark that merges six medical question-answering datasets covering a range of areas from professional medicine to consumer queries and includes HealthSearchQA, a new dataset of medically-related online search questions. This novel approach includes a human evaluation framework that examines model answers across various dimensions: namely, i) accuracy, ii) understanding, iii) reasoning, iv) potential harm, and v) bias. The authors tested both PaLM and its instruction-tuned version, Flan-PaLM, on MultiMedQA. Flan-PaLM, using diverse prompting techniques, sets a new standard in accuracy across all MultiMedQA multiple-choice datasets, including MedQA, MedMCQA, PubMedQA, and MMLU clinical topics, achieving a remarkable 67.6% accuracy in MedQA (USMLE-style questions), which is over 17% higher than the previous best. However, human assessments uncovered significant shortcomings. To address these, the authors introduced “instruction prompt tuning”, an efficient method for adapting LLMs to new domains with just a few examples. The resultant model, Med-PaLM, shows promise, yet it still does not match clinician performance, even if the authors could observe that model scale and instruction prompt tuning significantly enhance comprehension, knowledge recall, and reasoning.

A further risk is that LLMs might reinforce existing biases and provide inaccurate medical diagnoses, potentially leading to detrimental effects on healthcare. Zack and colleagues [26] aimed to evaluate whether GPT-4 harbors biases that could influence its application in healthcare settings. Employing the Azure OpenAI interface, the authors scrutinized GPT-4 for racial and gender biases and assessed the impact of such biases on four clinical applications of LLMs: namely, i) medical education, ii) diagnostic reasoning, iii) development and implementation of clinical plans, and iv) subjective patient evaluations, involving experiments using prompts mimicking typical GPT-4 usage in clinical and medical educational settings, and drawing from NEJM Healer clinical vignettes and research on implicit bias in healthcare. The study compared GPT-4's estimates of demographic distributions of medical conditions against actual U.S. prevalence data. For differential diagnosis and treatment planning, the research analyzed variations across demographic groups using standard statistical methods to identify significant differences. The study revealed that GPT-4 inadequately represents demographic diversity in medical conditions, often resorting to stereotypical demographic portrayals in clinical vignettes. The differential diagnoses generated by GPT-4 for standardized clinical vignettes tended to reflect biases associated with race, ethnicity, and gender. Furthermore, the model's assessments and plans demonstrated a notable correlation between demographic characteristics and recommendations for costlier procedures, as well as varied perceptions of patients.

All this, taken together, suggests the potential role of LLMs in medicine, but human evaluations also highlight the current models' limitations, underscoring the importance of comprehensive evaluation frameworks and continued methodological advancements to develop safe, effective LLMs for clinical use.

Implementing “Verification Paradigms”: a Comprehensive Evaluation Framework

Several “simulation and scenario testing” or “verification” paradigms can be particularly effective in verifying the evidence-based nature of generative AI in clinical decision-making. The eleven paradigms here proposed have been devised following thorough familiarization with existing literature and extensive consultation with experts in the field, to ensure that the methodologies are not only grounded in the latest academic research and theoretical frameworks but also shaped by practical insights and recommendations from medical professionals and AI technology specialists

(Table 2).

- ***The quiz/vignette/knowledge survey paradigm***

This approach involves assessing the AI's proficiency in various domains such as medical knowledge, diagnostic reasoning, and its understanding of therapeutic interventions by using quizzes/vignettes/validated knowledge surveys designed to mimic real-world clinical scenarios [27]. This would require the AI to not only have a vast knowledge base of medical information but also, and especially, the ability to apply this knowledge contextually, thus demonstrating an understanding of the nuances of patient presentations and how they correlate with various medical conditions and treatments. Additionally, this format could incorporate elements of both frequentist and Bayesian thinking, reflecting the probabilistic nature of clinical reasoning: in other words, as previously mentioned, the AI would have to weigh the likelihood of different diagnoses based on the presented symptoms and history, similar to how clinicians use Bayesian reasoning to update their probability assessments as new information becomes available.
This approach has a number of strengths, including comprehensive evaluation, real-world relevance, contextual understanding, probabilistic reasoning assessment, and adaptability to new information. On the other hand, it suffers from some weaknesses, such as design complexity and resource intensiveness, potential bias in test creation, and lack of interdisciplinary evaluation.
Currently, this approach is the most leveraged. An extensive body of literature has found that LLMs, such as ChatGPT, can successfully pass medical examinations [28], even though with varying degrees of heterogeneity and variability [29], exhibiting strong abilities in explanation, reasoning, memory, and accuracy. On the other hand, LLMs struggle with image-based questions [30] and, in some circumstances, lack insight and critical thinking skills [31].
Some of the studies that exploit quizzes/vignettes/validated knowledge surveys [32,33] have quantified the fluency and accuracy of AI-based tools using validated and reliable instruments, like the “Artificial Intelligence Performance Instrument” (AIPI) [32]. This tool includes nine items related to medical and surgical history: namely, symptoms; physical examination; diagnosis; additional examinations; management plan, and treatments. The AIPI score ranges from 0 (“inadequate clinical case management by the AI”) to 20 (“excellent clinical case management by the AI”). This score can be further subdivided into four sub-scores: i) patient feature, ii) diagnosis, iii) additional examination, and iv) treatment score.

- ***The Historical Data Comparison Paradigm***

This approach involves comparing AI-generated recommendations with outcomes from historical data: by analyzing cases where the clinical outcomes are well-known, one can assess how well the AI's suggestions would have aligned with actual scenarios. This would help in the comprehension of the AI's accuracy in real-world healthcare settings, providing insights into its potential benefits and limitations. This is a crucial step in understanding AI's performance and guiding its integration into clinical practice, ensuring that AI-supported decisions are in line with evidence-based medical standards and ultimately enhance patient care outcomes.
Strengths of this approach include real-world applicability, evidence-based evaluation, and objective benchmarking, by offering a clear, objective, data-driven, and evidence-based way to benchmark AI performance against known outcomes, facilitating a straightforward and comprehensive assessment of its accuracy. Furthermore, this method enables the identification of potential gaps and improvement areas: through direct comparison with historical outcomes, specific areas where AI recommendations may fall short can be identified, guiding further refinements. Demonstrating AI's ability to match or surpass historical outcomes can build trust among clinicians and patients regarding AI's utility in healthcare. However, this method has some weaknesses too, including dependence on data quality, in that the approach is heavily reliant on the availability and quality of

historical data, with poor data quality skewing results, and misleading about AI's true performance. Also, historical data may contain biases (e.g., diagnostic, treatment, or outcome biases), which can inadvertently be reinforced by AI, affecting the fairness and accuracy of its recommendations. This shortcoming is known as “historical bias”, which arises when the data or *corpora* utilized to train AI-based tools no longer accurately reflect the current reality. The potential lack of novel insights is another limitation, since this method benchmarks against known outcomes, and may not fully capture AI's potential to provide novel insights or diagnose conditions that were previously undetected or misdiagnosed. Further, this paradigm evaluates AI against past standards of care, which may not account for advancements in medical knowledge or changes in clinical guidelines over time (“static evaluation”) and its performance on complex, multifactorial cases might not be accurately assessed if historical data are limited or if such cases were managed differently due to evolving standards of care.
Currently, to the best of our knowledge, no published studies have leveraged this approach in the biomedical arena.

- ***The Expert Consensus Paradigm***

In this paradigm, AI-generated diagnoses or treatment plans are evaluated by a panel of medical experts, with the consensus among these experts on the validity of the AI's recommendations serving as a measure of its reliability. This paradigm is particularly useful in assessing the AI's performance in complex cases where human expertise is invaluable, ranging from the psychiatric field, in dealing with issues such as suicide risk assessment [34], to occupational medicine [35], oncology, with the management of malignancies [36], and complex surgical procedures, like bariatric surgery [37].
Strengths include high-quality validation of AI’s performance, ensuring that AI-generated recommendations are thoroughly vetted by experts, and bringing a high level of scrutiny and quality control that is particularly important in complex medical fields. Incorporation of human expertise and adaptability to complex cases are other strengths, by relying on medical experts to evaluate AI advice, and integrating nuanced human judgment and clinical experience that AI might lack or in those instances for which AI algorithms might not have sufficient training data or might lack the capability to understand context deeply. Further, expert feedback provides continuous learning opportunities, offering a platform for AI-based systems to be continuously updated and improved, enhancing their accuracy and reliability over time. This leads to heightened acceptance of AI tools, as having a consensus from medical experts can increase trust among healthcare providers and patients in AI-generated diagnoses or treatment plans.
On the other hand, expert feedback is time- and resource-intensive: gathering a panel of experts and reaching a consensus can be time-consuming and expensive, which may not be feasible for every clinical decision or in settings with limited resources. Also, despite being experts, humans are subject to biases that might affect their judgment, potentially leading to the validation of inaccurate AI recommendations. Scalability issues represent a further shortcoming: the approach may not scale well to everyday clinical practice, where quick decision-making is often required, and the luxury of convening an expert panel for each AI recommendation is not practical. Further, variability in expert opinion could lead to inconsistent validation of AI-generated recommendations and uncertainty in their reliability. Finally, there is a risk that this paradigm could discourage direct validation of AI algorithms through objective measures or independent verification, potentially overlooking errors or biases in the AI-based systems themselves.

- ***The Cross-Discipline Validation Paradigm***

This paradigm is rooted in the understanding that healthcare delivery increasingly relies on the expertise and coordination of diverse professionals to address complex health issues effectively. This approach recognizes that no single professional has all the knowledge and skills necessary to provide

comprehensive care, especially in cases that involve multifaceted medical, psychological, social, and ethical considerations. As clinical decision-making is seen as a multidisciplinary teamwork process, this verification paradigm involves cross-verifying AI-generated insights with experts from various medical disciplines. For example, a diagnosis made by an AI based on radiology images could be evaluated by experts in radiology, oncology, and pathology. This multidisciplinary approach ensures comprehensive evaluation and mitigates the risk of siloed decision-making, which is known to result in incomplete information, lack of coordination, and duplication of efforts, leading to inefficient care, higher costs, increased risk of medical errors, and decreased patient satisfaction, ultimately impacting the quality of patient care and health outcomes.

Currently, little is known about the multidisciplinary nature of LLMs: Li et al. [38] evaluated the proficiency of AI-based tools in addressing interdisciplinary queries in cardio-oncology, leveraging a questionnaire consisting of 25 questions compiled based on the 2022 European Society of Cardiology guideline on cardio-oncology. ChatGPT-4.0 showed the highest percentage of good responses at 68%, followed by Bard, Claude 2, and ChatGPT-3.5 at 52%, and Llama 2 at 48%. A specific area of concern was in treatment and prevention, where all LLMs scored poorly or borderline, particularly when their advice deviated from current guidelines, such as the recommendation to interrupt cancer treatment for patients with acute coronary syndrome. Other studies have assessed LLMs as support tools for multidisciplinary tumor boards in the planning of therapeutical programs for patients with cancer [39,40].

- ***The Rare or Complex Simulation and Scenario Testing Paradigm***

In this method, the AI-based tool is tested against a variety of simulated clinical scenarios, including rare and complex cases such as frail patients with multiple comorbidities, unusual presentations of diseases, or cases where symptoms are ambiguous or misleading. This comprehensive testing can identify areas for innovation and reveal the strengths and limitations of the AI-based tool in diverse clinical situations, like AI's capabilities in handling diversity. Conversely, this paradigm can be resource-intensive and potentially limited by available data.

A recent study [41] explored ChatGPT’s potential contributions to the diagnosis and management of rare and complex diseases, such as idiopathic pulmonary arterial hypertension, Klippel-Trenaunay syndrome, early-onset Parkinson’s disease, and Rett syndrome. LLMs can early detect the disease through AI-driven analysis of patient symptoms and medical imaging data, rapidly analyze an extensive body of biomedical literature for a better understanding of the mechanisms underlying the disease, and offer access to the latest research findings and personalized treatment plans.

Another study [42] examined the efficacy of three popular LLMs in medical education, particularly for diagnosing rare and complex diseases, and explored the impact of prompt engineering on their performance. Experiments were conducted on 30 cases from a diagnostic case challenge collection, utilizing various prompt strategies and a majority voting approach to compare the LLMs' performance against human consensus and MedAlpaca, an LLM designed for medical tasks. The findings revealed that all tested LLMs surpassed the average human consensus and MedAlpaca's performance by margins of at least 5% and 13%, respectively. In categories of frequently misdiagnosed cases, Google Bard equaled MedAlpaca but exceeded human consensus by 14%. GPT-4 and ChatGPT-3.5 showed superior performance over MedAlpaca and human respondents in moderately often misdiagnosed cases, with minimum accuracy improvements of 28% and 11%, respectively. Using a majority voting strategy, particularly with GPT-4, yielded the highest overall accuracy across the diagnostic complex case collection. On the Medical Information Mart for Intensive Care-III datasets, Google Bard and GPT-4 reached the highest diagnostic accuracy scores of 93% with multiple-choice prompts, while ChatGPT-3.5 and MedAlpaca scored 73% and 47%, respectively.

- ***The False Myth Paradigm***

This paradigm involves deliberately introducing known medical myths or outdated concepts into the AI's training data. The AI's ability to identify and reject these myths serves as a test of its understanding of current medical knowledge and its ability to discern evidence-based information. On the other hand, this approach requires a careful selection of myths and, if used in an inappropriate way, can reinforce incorrect information.

A few studies have harnessed this approach [43,44]. These studies have evaluated the accuracy of two artificial intelligence tools, ChatGPT-4 and Google Bard, in debunking 20 sleep-related myths, using a 5-point Likert scale for falseness and public health significance, and compared their performance with expert opinions. ChatGPT labeled 85% of the statements as either “false” (45%) or “generally false” (40%), showing high reliability in identifying inaccuracies, especially regarding sleep myths surrounding timing, duration, and behaviors during sleep. The tool demonstrated varying success in other categories like pre-sleep behaviors and brain function related to sleep. On a 5-point Likert scale, ChatGPT scored an average of 3.45 (SD=0.87) in identifying the falseness of statements and 3.15 (SD=0.99) in understanding their public health significance, indicating a good level of accuracy and understanding. Similarly, Google Bard identified 19 out of 20 statements as false, which was not significantly different from ChatGPT-4's accuracy. Google Bard's average falseness rating was 4.25 (SD=0.70), with skewness of -0.42 and kurtosis of -0.83, indicating a distribution with fewer extreme values compared to ChatGPT-4. For public health significance, Google Bard scored an average of 2.4 (SD=0.80), with skewness and kurtosis of 0.36 and -0.07, respectively, suggesting a more normal distribution than ChatGPT-4. The intra-class correlation coefficient between Google Bard and sleep experts was 0.58 for falseness and 0.69 for public health significance, showing moderate agreement. Text-mining showed Google Bard focused on practical advice, whereas ChatGPT-4 emphasized theoretical aspects. A readability analysis found Google Bard's responses matched an 8th-grade reading level, making them more accessible than ChatGPT-4's, which aligned with a 12th-grade level.

- ***The Challenging (or Controversial) Question Paradigm***

Here, the AI-based tool is presented with controversial or complex medical questions that do not have straightforward answers. The way AI navigates these questions, balancing different viewpoints and evidence, can reveal its depth of understanding and its ability to handle nuanced medical issues. In the realm of medicine, evidence is hierarchical, with systematic reviews and meta-analyses at the top. An analytical evaluation would consider how the AI prioritizes and evaluates/appraises different levels of evidence, and whether it can differentiate between high-quality and lower-quality studies. Also, AI should detect and minimize biases present in medical literature and data sources. Analytically, this involves evaluating the algorithms for their ability to identify potential biases in studies (e.g., publication bias, selection bias) and adjust their conclusions accordingly. Shortcomings of this paradigm include subjective evaluation criteria and dependence on the quality of input questions.

A few studies [45,46] have assessed the skills of AI-based tools in understanding or generating complex and nuanced clinical documents, such as guidelines.

- ***The Real-Time Monitoring Paradigm***

Here, the AI's recommendations are implemented in a controlled clinical environment, and patient outcomes are closely monitored, simulating a randomized controlled trial (RCT). This real-world testing provides valuable feedback on the AI's efficacy and safety in actual clinical settings.

While this paradigm can provide direct insights into practical impact and simulate real-world testing, it requires a controlled clinical environment and may be limited by ethical concerns related to the

experimental use of AI.

So far, only a few RCTs have been implemented. A recent blinded RCT [47] explored the efficacy of ChatGPT, alongside traditional typing and dictation methods, in assisting healthcare providers with clinical documentation, specifically in writing a history of present illness (HPI) based on standardized patient histories. Eleven participants, including medical students, orthopedic surgery residents, and attending surgeons, were tasked with documenting HPI using one of the three methods for each of the three standardized patient histories. The methods were assessed for speed, length, and quality of documentation. Results indicated that while dictation was the fastest method and resulted in longer and higher-quality patient histories according to the Physician Documentation Quality Instrument score, ChatGPT ranked intermediate in terms of speed. However, ChatGPT-generated documents were more comprehensive and organized than those produced by typing or dictation. A significant drawback noted was the inclusion of erroneous information in slightly more than one-third of ChatGPT-generated documents, raising concerns about accuracy. Additionally, there was a lack of consensus among reviewers regarding the quality of patient histories.

In another controlled trial [48], ChatGPT's utility in providing empathetic responses to people with multiple sclerosis was assessed. The study recruited a sample of 1,133 participants (aged 45.26 (SD=11.50) years; 68.49% females), who were surveyed through an online form distributed *via* digital communication platforms. Participants, blinded to the authors of the responses, evaluated alternate responses to four questions on a 1-5 Likert scale for overall satisfaction and used the Consultation and Relational Empathy scale for assessing perceived empathy. Results showed that ChatGPT's responses were perceived as significantly more empathetic than those from neurologists. However, there was no significant association between ChatGPT's responses and mean satisfaction. College graduates were significantly less likely to prefer ChatGPT's responses compared to those with a high school education.

- ***The Algorithm Transparency and Audit Paradigm***

This paradigm focuses on the transparency of the AI algorithms and the ability to audit their decision-making processes. By understanding how the AI-based tool arrives at its conclusions, clinicians can better assess the validity of its recommendations, which is crucial for building trust in AI-based systems among healthcare professionals.

Strengths include improved decision-making and enhanced trust and confidence by demystifying how decisions are made, thus building trust among clinicians and patients, crucial for the acceptance and integration of AI in healthcare. Clinicians can make more informed decisions by understanding the reasoning behind AI recommendations, potentially leading to better patient outcomes. AI-based tools can also facilitate regulatory compliance: transparency is key to meeting regulatory standards for medical devices and software, including AI-based systems used in healthcare. AI enables continuous improvement, as a transparent decision-making process allows for easier identification of errors or biases in the AI system, facilitating ongoing refinement and improvement. Further, exposing the decision-making process has educational benefits for healthcare professionals, helping them to understand complex AI methodologies and enhancing their ability to work alongside AI tools. On the other hand, this approach has some weaknesses that should be acknowledged, including complexity for end-users: AI decision-making processes, especially in deep learning, can be incredibly complex and difficult for end-users to understand, potentially limiting the effectiveness of transparency. Understanding and/or trusting the AI process might lead some clinicians to over-rely on AI recommendations without applying their judgment, especially in ambiguous or complex cases. Complete transparency might expose proprietary algorithms to potential theft or misuse, challenging companies to balance transparency with protecting their intellectual property. Moreover, there is potential room for misinterpretation: there is a risk that transparency could lead to misinterpretation of how AI algorithms work, especially without a strong foundation in data science or AI

methodologies among healthcare professionals. Finally, developing transparent AI systems that are also understandable to clinicians requires significant resources, including time and expertise, potentially slowing innovation.

- ***The Feedback Loop Paradigm***

This approach involves the continuous updating of the AI system based on feedback from its practical applications, with clinicians providing feedback on the AI's performance, which is then used to refine and improve the AI models. This iterative, ongoing process ensures that the AI-based system properly evolves and adapts to changing medical knowledge and practices. Conversely, it requires as well ongoing efforts and resources, besides depending on the quality of feedback.
A few studies have investigated reproducibility and repeatability [49,50]. In a study [49] involving emergency physicians, six unique prompts were used in conjunction with 61 patient vignettes to assess ChatGPT's ability to assign Canadian Triage and Acuity Scale (CTAS) scores through 10,980 simulated triages. ChatGPT returned a CTAS score in 99.6% of the queries. In terms of temporal reproducibility and repeatability, the study found considerable variation in the results: 21.0% due to repeatability (using the same prompt multiple times) and 4.0% due to reproducibility (using different prompts). ChatGPT's overall accuracy in triaging patients was 47.5%, with an under-triage rate of 13.7% and an over-triage rate of 38.7%. Of note, providing more detailed prompts resulted in slightly greater reproducibility but did not significantly improve accuracy.
In another study [50] assessing ChatGPT's proficiency in answering frequently asked questions (FAQs) about endometriosis, detailed internet searches were used to compile questions, which were then aligned with the European Society of Human Reproduction and Embryology (ESHRE) guidelines. An experienced gynecologist rated ChatGPT's responses on a scale of 1-4. To test repeatability, each question was asked twice, with reproducibility determined by the consistency of ChatGPT's scoring within the same category for repeated questions. Out of the FAQs, 91.4% (n=71) were answered completely, accurately, and sufficiently by ChatGPT. The model showed the highest accuracy in addressing symptoms and diagnosis (94.1%, 16/17 questions) and the lowest in treatment-related questions (81.3%, 13/16 questions). Among the 40 questions related to the ESHRE guidelines, 27 (67.5%) were rated grade 1, seven (17.5%) grade 2, and six (15.0%) grade 3. The reproducibility rate was highest (100%) for questions in the categories of prevention, symptoms and diagnosis, and complications. However, it was lowest for questions aligned with the ESHRE guidelines, at 70.0%.
These contrasting findings warrant further investigation.

- ***The Ethical and Legal Review Paradigm***

The “Ethical and Legal Review Paradigm” emphasizes the importance of ensuring that AI recommendations in healthcare settings adhere to established ethical guidelines and legal standards, which involves regular review rounds of the AI's recommendations by an ethics committee and/or legal team. This is particularly important in sensitive areas like critical care, emergency management, end-of-life care, or genetic testing, where the stakes of decisions are particularly high, and the moral and legal implications are significant. This approach aims to safeguard patients' rights, maintain trust in AI-assisted healthcare, and ensure that the implementation of AI technologies in medicine is both ethically sound and legally compliant [51,52].
The deployment of AI-based tools like ChatGPT in sensitive fields raises, indeed, several ethical and legal concerns. One significant issue is the potential for bias in AI algorithms, which can lead to unfair or incorrect outcomes. Moreover, the use of AI in these fields touches on privacy concerns, especially with the processing of personal data. Furthermore, issues regarding accountability and liability for malpractices and bad outcomes associated with LLM AI-influenced medical decision-making represent an emerging topic in the arena of legal medicine and, more broadly, forensic

science.
These concerns underscore the need for strict ethical guidelines and robust legal frameworks governing AI use in biomedical and clinical practices, with the final goal of leveraging AI's strengths while mitigating its limitations, ensuring it serves as a tool for progress rather than a source of bias and error [52,53].

Integrating the "Verification Paradigms"

These various paradigms for assessing AI in healthcare contexts underscore the multifaceted and complex nature of integrating AI technologies like ChatGPT into medical practices. These paradigms reflect a concerted effort to evaluate AI systems' proficiency, ethical alignment, and practical utility in clinical settings comprehensively. Each of these paradigms offers a unique perspective and method for verifying the reliability and accuracy of generative AI in clinical decision-making, and they can be used in combination to provide a robust validation framework (Tables 3 and 4, Figure 2).
It is of paramount importance to note that all these paradigms do not necessarily have the same weight or importance; their relevance can vary depending on the context, the specific healthcare domain, and the goals of the AI system being assessed. Integrating and combining these paradigms can provide a comprehensive, robust evaluation framework that leverages the strengths of each approach while mitigating their individual limitations.
Contextual and/or clinical relevance can be used to prioritize these approaches: in clinical settings where decision-making is complex and highly nuanced (e.g., oncology or psychiatry), paradigms that emphasize expert consensus and cross-discipline validation may be more critical, whereas for emerging treatments or rare diseases, paradigms focusing on simulation and scenario testing, and challenging questions can be invaluable to explore AI's capacity to contribute novel insights or support rare condition management. In contexts where AI is being directly implemented into clinical workflows and related follow-up, real-time monitoring, and feedback loop paradigms become essential to ensure patient safety and system efficacy.
Combining paradigms for comprehensive evaluation requires a "layered, sequential, strategic integrative approach", starting with broad assessments like the quiz/vignette/survey paradigm to gauge general knowledge and reasoning abilities, followed by more specific tests, such as historical data comparison for accuracy in real-world scenarios, and expert consensus for nuanced judgment calls. The cross-discipline validation paradigm can be harnessed to assess AI's recommendations from multiple professional perspectives, ensuring a holistic evaluation of AI's clinical recommendations. Throughout all stages of evaluation, the ethical and legal review paradigm is continuously applied to ensure adherence to ethical standards and legal requirements, safeguarding patient rights and data privacy.
This "layered, sequential, strategic integrative approach" enables continuous improvement of the entire process. An initial assessment utilizes paradigms like the quiz/vignette/survey and historical data comparison to initially evaluate AI's knowledge base and practical accuracy, which are iteratively refined and optimized by applying the feedback loop paradigm, using insights from real-time monitoring and expert consensus, followed by algorithm transparency and audits to ensure the system's decisions are understandable and justifiable.
For AI-based systems targeting specific or novel medical fields, the rare or complex simulation and scenario testing should be integrated, alongside challenging question paradigms, to push the boundaries of AI's capabilities and uncover areas for innovation. The feedback loop paradigm should be implemented where AI systems are regularly updated based on new clinical evidence, shifts in expert consensus, and outcomes from real-time monitoring, to ensure that AI remains aligned with current medical standards and practices, through continuous evolution and adaptive learning.
This evolution is maintained transparently in terms of how feedback and new data influence AI algorithms, fostering trust among healthcare professionals and patients. Regular ethical and legal reviews should accompany these updates to address any emerging concerns.

Throughout the entire process, which is dynamic, adaptive, and iterative, a broad range of stakeholders — including patients, healthcare professionals, ethicists, and legal experts— should be engaged. This ensures diverse perspectives are considered, particularly in applying paradigms like expert consensus, ethical and legal review, and real-time monitoring. As previously said, integrating these paradigms creates an ongoing process for evaluating and improving AI in healthcare, acknowledging the complexity of medical decision-making and the importance of maintaining ethical standards, and ensuring that AI systems are not only accurate and effective but also trusted and ethical components of healthcare delivery.

Towards a model of "clinically explainable, fair, and responsible, clinician-, expert-, and patient-in-the-loop Artificial Intelligence"

Clinical decision-making is a cornerstone of healthcare, demanding a blend of knowledge, intuition, and experience. It is a dynamic process where clinicians sift through patient data, balancing the effectiveness and risks of treatments against patient preferences and ethical standards, with the goal of optimal health outcomes, achieved through informed, evidence-based choices that respect patient autonomy and confidentiality [54-56].

As previously said, clinical decision-making is built on four pillars: scientific evidence, clinical judgment, ethical considerations, and patient involvement. The integration of generative AI into this realm presents exciting possibilities and challenges: on the one hand, AI's capacity to analyze vast amounts of medical data can enhance diagnosis, tailor treatments, and predict disease progression. However, its incorporation demands rigorous verification to align AI-generated insights with medical standards and ethical practices.

In the present conceptual paper, to ensure the reliability of AI in clinical decision-making, various verification paradigms have been proposed. The quiz/vignette/knowledge survey paradigm assesses AI's proficiency in medical domains by using realistic scenarios to test its knowledge and contextual application, incorporating frequentist and Bayesian reasoning in clinical diagnosis, whilst the Historical Data Comparison Paradigm examines AI recommendations against known clinical outcomes, assessing real-world accuracy. The Expert Consensus Paradigm involves a panel of medical experts evaluating AI-generated diagnoses and treatment plans, whereas the Cross-Discipline Validation Paradigm cross-checks AI insights with professionals from different medical fields, ensuring comprehensive evaluation. Additionally, the Rare or Complex Simulation and Scenario Testing Paradigm tests AI against a range of clinical scenarios, revealing its strengths and limitations. The False Myth Paradigm tests the AI's ability to reject outdated concepts or information/ content not substantiated by scientific evidence, whereas the Challenging Question Paradigm assesses how AI handles nuanced medical issues. The Real-Time Monitoring Paradigm involves implementing AI recommendations in controlled environments to monitor patient outcomes. The Algorithm Transparency and Audit Paradigm focuses on understanding how AI reaches its conclusions, essential for clinician trust. The Feedback Loop Paradigm ensures AI's continuous improvement based on practical application feedback. Lastly, the Ethical and Legal Review Paradigm ensures AI recommendations comply with ethical guidelines and legal requirements. Each paradigm offers a unique perspective for verifying AI in clinical decision-making, and, when used in combination, they provide a comprehensive framework for ensuring the accuracy and reliability of AI, crucial for its effective integration into healthcare. This blend of AI and traditional clinical expertise promises a future of enhanced healthcare delivery, marked by precision, efficacy, and patient-centered care.

The convergence of generative AI in clinical decision-making, when rigorously verified and integrated with traditional healthcare practices, paves the way for a model of "clinically explainable, fair, and responsible, clinician-, expert-, and patient-in-the-loop Artificial Intelligence". This model emphasizes not just the technical prowess of AI but also its comprehensibility, collaborative nature, and ethical grounding, ensuring that AI acts as an augmentative tool rather than an opaque,

autonomous decision-maker (“AI as a black box”). Clinically explainable AI demystifies the often complex and opaque decision-making processes of AI systems. In particular, the Algorithm Transparency and Audit Paradigm plays a crucial role here, ensuring that AI’s reasoning is accessible and understandable to clinicians. This transparency is vital for trust and effective collaboration between human experts and AI-based systems: clinicians need to understand the rationale behind AI-generated recommendations to make informed decisions, particularly in complex or critical cases.

This understanding would also facilitate discussions and interactions with patients, who are increasingly seeking active roles in their healthcare decisions. By demystifying AI outputs, healthcare providers can offer clear, comprehensible explanations to patients, fostering trust and informed consent. Incorporating clinicians and experts in the loop is, indeed, fundamental in realizing this model: the Expert Consensus and Cross-Discipline Validation Paradigms highlight the importance of human expertise in evaluating and interpreting AI-generated insights, with clinicians bringing invaluable context, experience, and judgment to the table, which are crucial for nuanced decision-making. AI, in this context, is a tool that augments but does not replace the clinician’s judgment. This collaboration ensures that AI recommendations are not only based on data and algorithms but also tempered by human insight and ethical considerations. Patient involvement is another cornerstone of this model: patient-centric care is increasingly recognized as a key component of quality healthcare.

The integration of AI in clinical decision-making should not diminish the patient's role but rather enhance it. By providing tailored and precise medical insights, AI can empower patients with information that is specific to their condition and treatment options. This approach aligns with the growing trend towards personalized/individualized medicine, where treatments are tailored to individual patient profiles. AI can facilitate this by analyzing patient data in-depth, offering insights that help in crafting personalized treatment plans. Moreover, engaging patients in the decision-making process, aided by AI's insights, respects their autonomy and preferences, leading to better satisfaction and adherence to treatment plans. Implementing a clinically explainable, fair, and responsible, clinician-, expert-, and patient-in-the-loop AI model also necessitates continuous learning and adaptation: the Feedback Loop Paradigm ensures that AI systems evolve based on real-world outcomes and clinician inputs. This ongoing refinement is crucial for the AI-based tool to stay relevant and effective in the ever-changing landscape of medical knowledge and practice.

Finally, the Ethical and Legal Review Paradigm ensures that AI recommendations are continually assessed for ethical and legal compliance, an aspect critical in maintaining public trust and upholding professional standards. Trust in this context extends beyond mere reliability to include ethically relevant and value-laden aspects of AI systems' design and usage. This broadened understanding of trust aims to encompass concerns about fairness, transparency, privacy, and the prevention of harm, among others. While pure epistemic accounts of trust focus solely on rational and performance-based criteria, more broadly speaking trust encompasses the full spectrum of ethical considerations necessary for truly trustworthy AI, fully integrating ethical considerations into the core of what it means for an AI system to be considered trustworthy. AI-based systems not only function effectively and reliably but also and especially operate within ethical boundaries, adhering to ethical standards and principles that respect human autonomy, prevent harm, and promote fairness and transparency [57].

In summary, the envisioned model of AI in healthcare is one where AI acts as an intelligent, transparent, and adaptable assistant in the complex process of clinical decision-making, enhancing, rather than replacing, human expertise, and keeping clinicians, experts, and patients central to the decision-making process. This approach not only leverages the strengths of AI in data processing and pattern recognition but also upholds the irreplaceable value of human judgment, experience, and ethical reasoning, all crucial for delivering high-quality, patient-centered healthcare.

Current state-of-the-art and Future Directions

Currently, in a great portion of articles, the authors have limited themselves to querying the AI-based tools on a variety of topics, without fully leveraging their potential. If that was understandable at the beginning of the revolution posed by LLMs, where early fascination and curiosity were prevalent, it is time to go beyond just chatting with ChatGPT and shift toward a deeper, comprehensive, and robust assessment of the capabilities of smart chatbots in real-world clinical settings. Researchers should make responsible use of AI, utilize standardized reporting guidelines [58], systematically compare different types of AI-based tools, evaluate the accuracy, repeatability, and reproducibility of the tools, and incorporate ethical and legal considerations. Validated and reliable reporting checklists are essential for ensuring that research findings and advancements are communicated clearly and consistently, facilitating comparative analyses across different AI-enhanced tools. This will not only help in identifying the most effective solutions but also in uncovering potential biases, limitations, and areas for improvement. By systematically comparing different AI-based tools and rigorously evaluating their performance, the research community can establish a benchmark for what constitutes successful integration of AI in clinical settings. A composite set of performance and outcome metrics is essential for validating the reliability of AI in clinical applications and for ensuring that tools can be confidently utilized across various settings without loss of performance quality. Currently, only accuracy is being investigated, with only a few studies exploring the repeatability and reproducibility of AI-generated medical responses and recommendations.

Scholars can harness the eleven paradigms here proposed to make AI-enhanced applications more clinically relevant and meaningful, as well as robust and safe.

Conclusions

Generative AI holds immense promise in enhancing clinical decision-making and offering personalized, accurate, and efficient healthcare solutions. However, ensuring that this technology produces evidence-based, reliable, impactful knowledge is paramount. By employing paradigms and approaches like those outlined in the present conceptual paper, the medical and patient communities can better harness the potential of AI while safeguarding against misinformation and maintaining high standards of patient care.

Table 1. Overview of the verification paradigms.

Verification Paradigm	Brief Description
Quiz/Vignette/Knowledge Survey	Uses clinical scenarios to test AI's medical knowledge and reasoning.
Historical Data Comparison	Compares AI recommendations with known clinical outcomes to gauge accuracy.
Expert Consensus	Evaluates AI-generated diagnoses or treatment plans against expert medical opinion.
Cross-Discipline Validation	Verifies AI insights with professionals from various medical disciplines for comprehensive evaluation.
Rare or Complex Simulation and Scenario Testing	Assesses AI's ability to handle rare and complex medical cases through simulated scenarios.
False Myth Paradigm	Tests AI's capability to identify and reject medical myths or outdated concepts.
Challenging (or Controversial) Question	Presents AI with complex medical questions to evaluate its nuanced understanding and reasoning.
Real-Time Monitoring	Monitors AI recommendations in clinical settings to observe real-world efficacy and safety.
Algorithm Transparency and Audit	Focuses on the transparency of AI's decision-making process and its ability to be audited.
Feedback Loop	Involves continuous AI improvement based on feedback from practical applications and outcomes.
Ethical and Legal Review	Regularly reviews AI recommendations to ensure they adhere to ethical guidelines and legal standards.

Table 2. Verification paradigms, with their strengths and weaknesses.

Verification Paradigm	Strengths	Weaknesses
Quiz/Vignette/Knowledge Survey	Comprehensive evaluation Real-world relevance Assessment of contextual understanding and probabilistic reasoning	Complex to design Resource-intensive Potential bias in test creation
Historical Data Comparison	Real-world applicability Evidence-based evaluation Objective benchmarking	Dependent on data quality Historical bias May not capture AI's potential for novel insights
Expert Consensus	Leverages human expertise Valuable in complex cases Incorporates ethical judgment	Subjective Time-consuming Potential for expert bias
Cross-Discipline Validation	Comprehensive evaluation from multiple perspectives Mitigates risk of siloed decision-making	Coordination challenges Requires broad expert availability
Rare or Complex Simulation and Scenario Testing	Reveals AI's capabilities in handling diversity Can identify areas for innovation	Potentially limited by available data Resource-intensive
False Myth Paradigm	Tests AI’s current knowledge base Assesses ability to discern evidence-based information	Requires careful selection of myths Risk of reinforcing incorrect information
Challenging (or Controversial) Question	Evaluates AI's handling of ambiguity and complexity Assesses balance of different viewpoints	Subjective evaluation criteria Depends on quality of input questions
Real-Time Monitoring	Direct insight into practical impact Simulates real-world testing	Requires controlled clinical environment Ethical concerns with experimental use
Algorithm Transparency and Audit	Enhances trust and understanding Facilitates regulatory compliance	Complexity for end-users Risk of exposing proprietary information
Feedback Loop	Ensures continuous improvement Adapts to changing medical knowledge	Requires ongoing effort and resources Dependence on quality of feedback
Ethical and Legal Review	Safeguards patient rights Ensures adherence to ethical guidelines	Time-consuming Needs multidisciplinary expertise

Table 3. Overview of the Layered Integrative Approach for Evaluating AI in Healthcare, delineating the structured, multi-stage framework for the comprehensive assessment and continuous improvement of AI systems.

Stage	Verification paradigm	Objective	Integration
Initial Assessment	Quiz/Vignette/Knowledge Survey	To gauge the AI's foundational medical knowledge and its ability to apply this knowledge in simulated real-world scenarios.	Forms the baseline assessment of the AI's capabilities, setting the stage for more targeted evaluations.
Refinement:	Historical Data Comparison	To refine the AI's understanding and application of medical knowledge by comparing its recommendations or diagnoses against known outcomes from historical data	Uses the insights gained from initial assessments to focus on areas requiring improvement, ensuring that the AI's recommendations are grounded in real-world evidence
Expert Feedback	Expert Consensus	To incorporate nuanced clinical insights and expert judgments into the AI's learning, ensuring it aligns with current clinical practices and expert opinions	Builds on the refined knowledge base by integrating expert clinical insights, further improving the AI's decision-making processes
Comprehensive Evaluation	Cross-Discipline Validation	To evaluate the AI's recommendations and diagnostics across various medical disciplines, ensuring a comprehensive and holistic assessment	Leverages the foundational knowledge, refined understanding, and expert insights to test the AI's capabilities in a multidisciplinary context, identifying any gaps or biases
Complexity Handling	Rare or Complex Simulation and Scenario Testing	To test the AI's ability to handle complex, rare, or novel medical scenarios, ensuring it can adapt to a wide range of clinical challenges	Utilizes the comprehensive evaluations as a foundation to challenge the AI with scenarios that require sophisticated reasoning, further refining its decision-making abilities
Knowledge	False Myth Paradigm	To ensure the AI's	Builds on the

Accuracy		current knowledge base is accurate and up-to-date, identifying and correcting any misconceptions or outdated information	previous layers by specifically targeting and rectifying inaccuracies in the AI's knowledge, ensuring reliability
Complexity and Nuance Handling	Challenging (or Controversial) Question	To evaluate the AI's ability to navigate complex medical questions that may not have straightforward answers, assessing its reasoning in ambiguous situations	Further refines the AI's decision-making process by exposing it to nuanced clinical scenarios, enhancing its ability to provide balanced and informed recommendations
Real-World Efficacy	Real-Time Monitoring	To monitor the AI's recommendations and diagnoses in real-world clinical settings, assessing its practical efficacy and safety.	Applies all previous layers of assessment in a live clinical environment, providing direct feedback on the AI's performance and areas for improvement.
Transparency and Trust	Algorithm Transparency and Audit	To ensure the decision-making processes of the AI are transparent and understandable, building trust among healthcare providers and patients.	Uses insights from real-world applications and previous evaluations to demystify the AI's logic, ensuring it is both effective and comprehensible
Continuous Improvement:	Feedback Loop	To continuously refine and improve the AI system based on real-world data, feedback, and evolving medical knowledge.	Represents the culmination of the integrative approach, where feedback from all previous stages is used to iteratively enhance the AI system, ensuring it remains effective, safe, and ethically compliant over time
Ethical and Legal Compliance	Ethical and Legal Review	To ensure all AI recommendations and processes adhere to established ethical guidelines and legal	Runs parallel to all stages, providing a constant check on the AI's compliance with ethical norms and

		standards	legal requirements, safeguarding against potential malpractices and ensuring patient rights are protected.

Table 4. Engagement and impact of key healthcare stakeholders — physicians, patients, nurses, administrators, AI developers, ethicists, and regulators — across various AI evaluation paradigms, highlighting their roles and interactions in the process of assessing and integrating AI technologies in healthcare.

Verification Paradigm	Stakeholders						
	Physicians	**Patients**	**Nurses**	**Healthcare Administrators**	**AI Developers**	**Ethicists**	**Regulators**
Quiz/Vignette/ Knowledge Survey	Participate in creating/testing	May be subjects in scenarios	Assist in scenario design	Oversee implementation	Design relevant quizzes/surveys	Evaluate scenario ethics	Establish standards for testing
Historical Data Comparison	Use outcomes to validate AI	Benefit from improved outcomes	Observe AI's real-world accuracy	Use data for strategic decisions	Analyze comparison outcomes for improvement	Assess the ethical use of historical data	Monitor data use and outcomes
Expert Consensus	Contribute expertise	Trust in consensus-driven AI	Support expert consensus	Involve in consensus-building	Incorporate expert feedback	Participate in consensus discussions	Ensure expert consensus meets guidelines
Cross-Discipline Validation	Collaborate across specialties	Benefit from holistic care approaches	Facilitate multidisciplinary care	Ensure interdisciplinary cooperation	Work with diverse healthcare teams	Ensure ethical cross-discipline validation	Regulate multidisciplinary validation processes
Rare or Complex Simulation and Scenario Testing	Engage in scenario creation/testing	Receive personalized care for rare conditions	Involved in patient care scenarios	Plan for innovative care solutions	Design simulations for complex conditions	Scrutinize simulations for ethical considerations	Oversee testing for safety and efficacy
False Myth Paradigm	Input on relevant myths	Protected from misinformation	Educate patients on myths vs. facts	Promote accurate patient education	Correct and update AI knowledge	Highlight the ethical handling of myths	Regulate misinformation management
Challenging (or Controversial) Question	Address complex questions	Empowered by nuanced AI assistance	Assist in managing complex cases	Address policy implications	Develop algorithms for nuanced questions	Engage in ethical debates	Set standards for addressing controversial topics
Real-Time Monitoring	Monitor patient outcomes	Directly affected by AI recommendations	Monitor and report on patient responses	Supervise operational integration	Refine AI through real-time data	Monitor ethical implications of real-time use	Ensure patient safety in real-time monitoring
Algorithm Transparency and Audit	Require understanding of AI decisions	Seek transparency for trust	Advocate for clear AI explanations	Demand system transparency	Ensure algorithmic transparency	Advocate for transparent decision-making	Enforce transparency and auditability
Feedback Loop	Provide clinical feedback	Benefit from ongoing improvements	Offer practical feedback	Implement system feedback	Use feedback for technical refinement	Provide ethical oversight in feedback	Facilitate regulatory feedback loops
Ethical and Legal Review	Ensure AI aligns with ethical/legal standards	Protected by ethical/legal safeguards	Uphold ethical standards in AI use	Ensure compliance with regulations	Adhere to ethical and legal standards	Lead ethical and legal reviews	Conduct legal reviews and compliance checks

Figure 1. Integrating Clinical Expertise with Artificial Intelligence (AI) for Enhanced Healthcare Outcomes: A schematic representation of the flow and interplay between traditional clinical reasoning, data acquisition, AI-driven predictive analytics, and the continuous learning cycle leading to improved patient care and diagnostics.

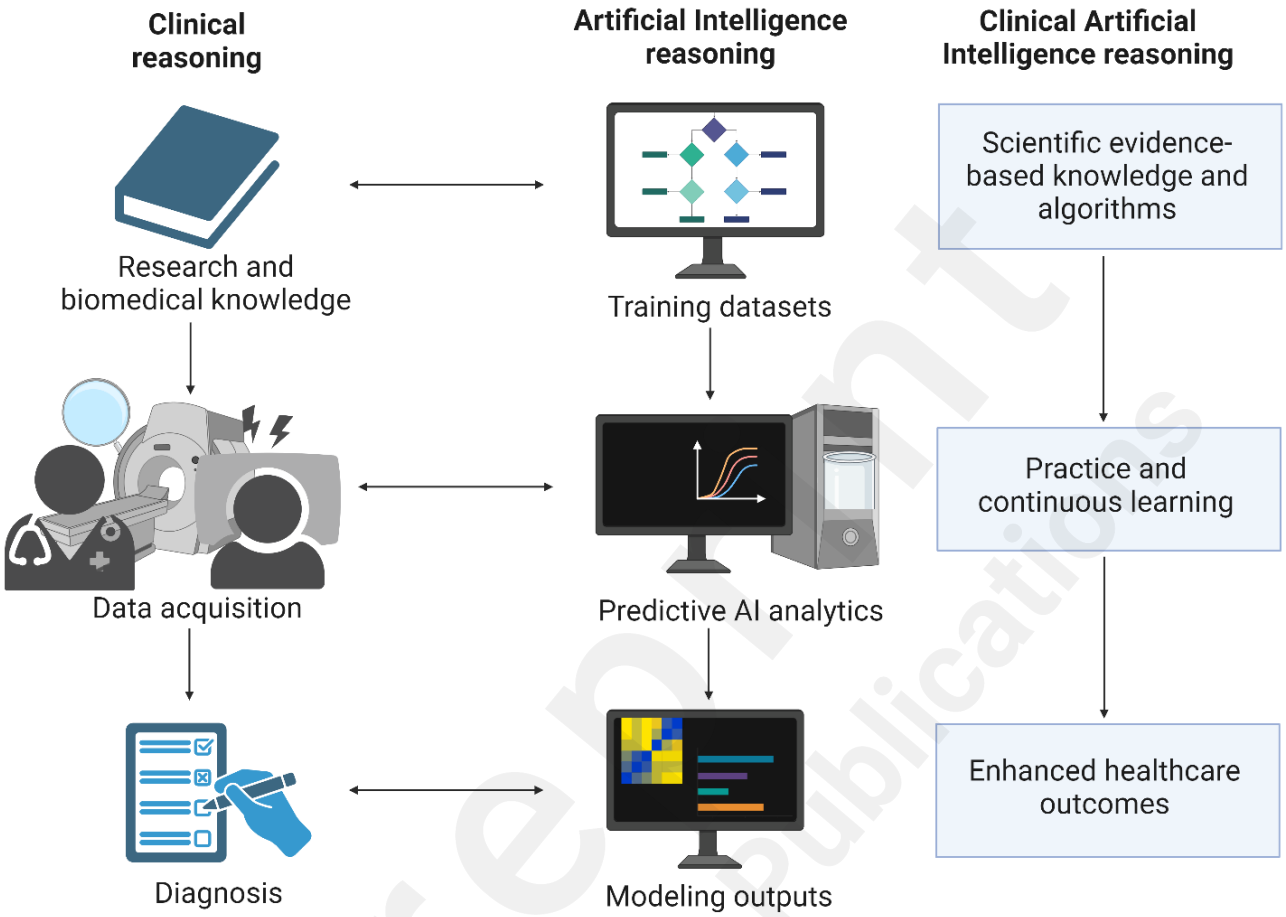

Figure 2. Integrating Verification Paradigms for AI in Healthcare.

References

1. Young M, Thomas A, Lubarsky S, Ballard T, Gordon D, Gruppen LD, Holmboe E, Ratcliffe T, Rencic J, Schuwirth L, Durning SJ. Drawing Boundaries: The Difficulty in Defining Clinical Reasoning. Acad Med. 2018 Jul;93(7):990-995. doi: 10.1097/ACM.0000000000002142. PMID: 29369086.
2. Young ME, Thomas A, Lubarsky S, Gordon D, Gruppen LD, Rencic J, Ballard T, Holmboe E, Da Silva A, Ratcliffe T, Schuwirth L, Dory V, Durning SJ. Mapping clinical reasoning literature across the health professions: a scoping review. BMC Med Educ. 2020 Apr 7;20(1):107. doi: 10.1186/s12909-020-02012-9. PMID: 32264895; PMCID: PMC7140328.
3. Benner P, Hughes RG, Sutphen M. Clinical Reasoning, Decisionmaking, and Action: Thinking Critically and Clinically. In: Hughes RG, editor. Patient Safety and Quality: An Evidence-Based Handbook for Nurses. Rockville (MD): Agency for Healthcare Research and Quality (US); 2008 Apr. Chapter 6. Available from: https://www.ncbi.nlm.nih.gov/books/NBK2643/
4. Andreoletti M, Berchialla P, Boniolo G, Chiffi D. Introduction: Foundations of Clinical Reasoning—An Epistemological Stance. Topoi. 2019;38:389–394. doi: 10.1007/s11245-018-9619-4.
5. Chiffi D. Clinical Reasoning: Knowledge, Uncertainty, and Values in Health Care. 2 October 2020: Springer Cham. doi: 10.1007/978-3-030-59094-9.
6. Worrall J. Evidence: philosophy of science meets medicine. J Eval Clin Pract. 2010;16(2):356–362.
7. Larson EB. How can clinicians incorporate research advances into practice? J Gen Intern Med. 1997 Apr;12 Suppl 2(Suppl 2):S20-4. doi: 10.1046/j.1525-1497.12.s2.3.x. PMID: 9127240; PMCID: PMC1497224.

8. Parascandola M. Epistemic risk: empirical science and the fear of being wrong. Law Probab Risk. 2010;9(3–4):201–214.
9. Müller VC. Philosophy and Theory of Artificial Intelligence. 24 August 2012: Springer Berlin, Heidelberg. doi: 10.1007/978-3-642-31674-6.
10. Schopow N, Osterhoff G, Baur D. Applications of the Natural Language Processing Tool ChatGPT in Clinical Practice: Comparative Study and Augmented Systematic Review. JMIR Med Inform. 2023 Nov 28;11:e48933. doi: 10.2196/48933. PMID: 38015610; PMCID: PMC10716749.
11. Bagde H, Dhopte A, Alam MK, Basri R. A systematic review and meta-analysis on ChatGPT and its utilization in medical and dental research. Heliyon. 2023 Nov 29;9(12):e23050. doi: 10.1016/j.heliyon.2023.e23050. PMID: 38144348; PMCID: PMC10746423.
12. Shorey S, Mattar C, Pereira TL, Choolani M. A scoping review of ChatGPT's role in healthcare education and research. Nurse Educ Today. 2024 Apr;135:106121. doi: 10.1016/j.nedt.2024.106121. Epub 2024 Feb 6. PMID: 38340639.
13. Emsley R. ChatGPT: these are not hallucinations - they're fabrications and falsifications. Schizophrenia (Heidelb). 2023 Aug 19;9(1):52. doi: 10.1038/s41537-023-00379-4. PMID: 37598184; PMCID: PMC10439949.
14. Chiffi D, Zanotti R. Fear of knowledge: clinical hypotheses in diagnostic and prognostic reasoning. J Eval Clin Pract. 2017;23(5):928–934
15. Christakis NA, Sachs GA. The role of prognosis in clinical decision-making. J Gen Intern Med. 1996;11(7):422–425.
16. Savcisens G, Eliassi-Rad T, Hansen LK, Mortensen LH, Lilleholt L, Rogers A, Zettler I, Lehmann S. Using sequences of life-events to predict human lives. Nat Comput Sci. 2023. doi: 10.1038/s43588-023-00573-5.
17. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021 Dec;27(12):2176-2182. doi: 10.1038/s41591-021-01595-0. Epub 2021 Dec 10. PMID: 34893776; PMCID: PMC8674135.
18. The Lancet Digital Health. Large language models: a new chapter in digital health. Lancet Digit Health. 2024 Jan;6(1):e1. doi: 10.1016/S2589-7500(23)00254-6. PMID: 38123249.
19. Wang H, Gao C, Dantona C, Hull B, Sun J. DRG-LLaMA : tuning LLaMA model to predict diagnosis-related group for hospitalized patients. NPJ Digit Med. 2024 Jan 22;7(1):16. doi: 10.1038/s41746-023-00989-3. PMID: 38253711; PMCID: PMC10803802.
20. Hwang S, Reddy S, Wainwright K, Schriver E, Cappola A, Mowery D. Using Natural Language Processing to Extract and Classify Symptoms Among Patients with Thyroid Dysfunction. Stud Health Technol Inform. 2024 Jan 25;310:614-618. doi: 10.3233/SHTI231038. PMID: 38269882.
21. Chen F, Bokhari SMA, Cato K, Gürsoy G, Rossetti SC. Examining the Generalizability of Pretrained De-identification Transformer Models on Narrative Nursing Notes. Appl Clin Inform. 2024 Mar 6. doi: 10.1055/a-2282-4340. Epub ahead of print. PMID: 38447965.
22. Talebi S, Tong E, Li A, Yamin G, Zaharchuk G, Mofrad MRK. Exploring the performance and explainability of fine-tuned BERT models for neuroradiology protocol assignment. BMC Med Inform Decis Mak. 2024 Feb 7;24(1):40. doi: 10.1186/s12911-024-02444-z. PMID: 38326769; PMCID: PMC10848624.
23. Bernstein IA, Koornwinder A, Hwang HH, Wang SY. Automated Recognition of Visual Acuity Measurements in Ophthalmology Clinical Notes Using Deep Learning. Ophthalmol Sci. 2023 Jul 19;4(2):100371. doi: 10.1016/j.xops.2023.100371. PMID: 37868799; PMCID: PMC10587603.
24. Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, Martin C, Flores MG, Zhang Y, Magoc T, Lipori G, Mitchell DA, Ospina NS, Ahmed MM, Hogan WR, Shenkman EA,

Guo Y, Bian J, Wu Y. A study of generative large language model for medical research and healthcare. NPJ Digit Med. 2023 Nov 16;6(1):210. doi: 10.1038/s41746-023-00958-w. PMID: 37973919; PMCID: PMC10654385.

25. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P, Seneviratne M, Gamble P, Kelly C, Babiker A, Schärli N, Chowdhery A, Mansfield P, Demner-Fushman D, Agüera Y Arcas B, Webster D, Corrado GS, Matias Y, Chou K, Gottweis J, Tomasev N, Liu Y, Rajkomar A, Barral J, Semturs C, Karthikesalingam A, Natarajan V. Large language models encode clinical knowledge. Nature. 2023 Aug;620(7972):172-180. doi: 10.1038/s41586-023-06291-2. Epub 2023 Jul 12. Erratum in: Nature. 2023 Jul 27;: PMID: 37438534; PMCID: PMC10396962.
26. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, Jurafsky D, Szolovits P, Bates DW, Abdulnour RE, Butte AJ, Alsentzer E. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health. 2024 Jan;6(1):e12-e22. doi: 10.1016/S2589-7500(23)00225-X. PMID: 38123252.
27. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, Ratcliffe T, Gordon D, Heist B, Lubarsky S, Estrada CA, Ballard T, Artino AR Jr, Sergio Da Silva A, Cleary T, Stojan J, Gruppen LD. Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance. Acad Med. 2019 Jun;94(6):902-912. doi: 10.1097/ACM.0000000000002618. PMID: 30720527.
28. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: A systematic review and a meta-analysis. BJOG. 2024 Feb;131(3):378-380. doi: 10.1111/1471-0528.17641. Epub 2023 Aug 21. PMID: 37604703.
29. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: A systematic review and meta-analysis. J Biomed Inform. 2024 Mar;151:104620. doi: 10.1016/j.jbi.2024.104620. Epub 2024 Mar 8. PMID: 38462064.
30. Haver HL, Bahl M, Doo FX, Kamel PI, Parekh VS, Jeudy J, Yi PH. Evaluation of Multimodal ChatGPT (GPT-4V) in Describing Mammography Image Features. Can Assoc Radiol J. 2024 Apr 6:8465371241247043. doi: 10.1177/08465371241247043. Epub ahead of print. PMID: 38581353.
31. Sumbal A, Sumbal R, Amir A. Can ChatGPT-3.5 Pass a Medical Exam? A Systematic Review of ChatGPT's Performance in Academic Testing. J Med Educ Curric Dev. 2024 Mar 13;11:23821205241238641. doi: 10.1177/23821205241238641. PMID: 38487300; PMCID: PMC10938614.
32. Lechien JR, Maniaci A, Gengler I, Hans S, Chiesa-Estomba CM, Vaira LA. Validity and reliability of an instrument evaluating the performance of intelligent chatbot: the Artificial Intelligence Performance Instrument (AIPI). Eur Arch Otorhinolaryngol. 2024 Apr;281(4):2063-2079. doi: 10.1007/s00405-023-08219-y. Epub 2023 Sep 12. PMID: 37698703.
33. Dronkers EAC, Geneid A, Al Yaghchi C, Lechien JR. Evaluating the Potential of AI Chatbots in Treatment Decision-making for Acquired Bilateral Vocal Fold Paralysis in Adults. J Voice. 2024 Apr 6:S0892-1997(24)00059-6. doi: 10.1016/j.jvoice.2024.02.020. Epub ahead of print. PMID: 38584026.
34. Elyoseph Z, Levkovich I. Beyond human expertise: the promise and limitations of ChatGPT in suicide risk assessment. Front Psychiatry. 2023 Aug 1;14:1213141. doi: 10.3389/fpsyt.2023.1213141. PMID: 37593450; PMCID: PMC10427505.
35. Padovan M, Cosci B, Petillo A, Nerli G, Porciatti F, Scarinci S, Carlucci F, Dell'Amico L, Meliani N, Necciari G, Lucisano VC, Marino R, Foddis R, Palla A. ChatGPT in Occupational Medicine: A Comparative Study with Human Experts. Bioengineering (Basel). 2024 Jan 6;11(1):57. doi: 10.3390/bioengineering11010057. PMID: 38247934; PMCID: PMC10813435.

36. Peng W, Feng Y, Yao C, Zhang S, Zhuo H, Qiu T, Zhang Y, Tang J, Gu Y, Sun Y. Evaluating AI in medicine: a comparative analysis of expert and ChatGPT responses to colorectal cancer questions. Sci Rep. 2024 Feb 3;14(1):2840. doi: 10.1038/s41598-024-52853-3. PMID: 38310152; PMCID: PMC10838275.
37. Jazi AHD, Mahjoubi M, Shahabi S, Alqahtani AR, Haddad A, Pazouki A, Prasad A, Safadi BY, Chiappetta S, Taskin HE, Billy HT, Kasama K, Mahawar K, Gawdat K, Rheinwalt KP, Miller KA, Kow L, Neto MG, Yang W, Palermo M, Ghanem OM, Lainas P, Peterli R, Kassir R, Puy RV, Da Silva Ribeiro RJ, Verboonen S, Pintar T, Shabbir A, Musella M, Kermansaravi M. Bariatric Evaluation Through AI: a Survey of Expert Opinions Versus ChatGPT-4 (BETA-SEOV). Obes Surg. 2023 Dec;33(12):3971-3980. doi: 10.1007/s11695-023-06903-w. Epub 2023 Oct 27. PMID: 37889368.
38. Li P, Zhang X, Zhu E, Yu S, Sheng B, Tham YC, Wong TY, Ji H. Potential Multidisciplinary Use of Large Language Models for Addressing Queries in Cardio-Oncology. J Am Heart Assoc. 2024 Mar 19;13(6):e033584. doi: 10.1161/JAHA.123.033584. Epub 2024 Mar 18. PMID: 38497458.
39. Lukac S, Dayan D, Fink V, Leinert E, Hartkopf A, Veselinovic K, Janni W, Rack B, Pfister K, Heitmeir B, Ebner F. Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. Arch Gynecol Obstet. 2023 Dec;308(6):1831-1844. doi: 10.1007/s00404-023-07130-5. Epub 2023 Jul 17. PMID: 37458761; PMCID: PMC10579162.
40. Vela Ulloa J, King Valenzuela S, Riquoir Altamirano C, Urrejola Schmied G. Artificial intelligence-based decision-making: can ChatGPT replace a multidisciplinary tumour board? Br J Surg. 2023 Oct 10;110(11):1543-1544. doi: 10.1093/bjs/znad264. PMID: 37595064.
41. Zheng Y, Sun X, Feng B, Kang K, Yang Y, Zhao A, Wu Y. Rare and complex diseases in focus: ChatGPT's role in improving diagnosis and treatment. Front Artif Intell. 2024 Jan 11;7:1338433. doi: 10.3389/frai.2024.1338433. PMID: 38283995; PMCID: PMC10808657.
42. Abdullahi T, Singh R, Eickhoff C. Learning to Make Rare and Complex Diagnoses With Generative AI Assistance: Qualitative Study of Popular Large Language Models. JMIR Med Educ. 2024 Feb 13;10:e51391. doi: 10.2196/51391. PMID: 38349725; PMCID: PMC10900078.
43. Bragazzi NL, Garbarino S. Assessing the Accuracy of Generative Conversational Artificial Intelligence in Debunking Sleep Health Myths: Mixed-Methods Comparative Study with Expert Analysis. JMIR Form Res. 2024 Mar 14. doi: 10.2196/55762. Epub ahead of print. PMID: 38501898.
44. Garbarino S, Bragazzi NL. Evaluating the effectiveness of artificial intelligence-based tools in detecting and understanding sleep health misinformation: Comparative analysis using Google Bard and OpenAI ChatGPT-4. J Sleep Res. 2024 Apr 5:e14210. doi: 10.1111/jsr.14210. Epub ahead of print. PMID: 38577714.
45. Saturno MP, Mejia MR, Wang A, Kwon D, Oleru O, Seyidova N, Henderson PW. Generative artificial intelligence fails to provide sufficiently accurate recommendations when compared to established breast reconstruction surgery guidelines. J Plast Reconstr Aesthet Surg. 2023 Nov;86:248-250. doi: 10.1016/j.bjps.2023.09.030. Epub 2023 Sep 15. PMID: 37793197; PMCID: PMC10965244.
46. Zaidat B, Shrestha N, Rosenberg AM, Ahmed W, Rajjoub R, Hoang T, Mejia MR, Duey AH, Tang JE, Kim JS, Cho SK. Performance of a Large Language Model in the Generation of Clinical Guidelines for Antibiotic Prophylaxis in Spine Surgery. Neurospine. 2024 Mar;21(1):128-146. doi: 10.14245/ns.2347310.655. Epub 2024 Mar 31. PMID: 38569639; PMCID: PMC10992653.
47. Baker HP, Dwyer E, Kalidoss S, Hynes K, Wolf J, Strelzow JA. ChatGPT's Ability to Assist with Clinical Documentation: A Randomized Controlled Trial. J Am Acad Orthop Surg. 2024

Feb 1;32(3):123-129. doi: 10.5435/JAAOS-D-23-00474. Epub 2023 Nov 17. PMID: 37976385.

48. Maida E, Moccia M, Palladino R, Borriello G, Affinito G, Clerico M, Repice AM, Di Sapio A, Iodice R, Spiezia AL, Sparaco M, Miele G, Bile F, Scandurra C, Ferraro D, Stromillo ML, Docimo R, De Martino A, Mancinelli L, Abbadessa G, Smolik K, Lorusso L, Leone M, Leveraro E, Lauro F, Trojsi F, Streito LM, Gabriele F, Marinelli F, Ianniello A, De Santis F, Foschi M, De Stefano N, Morra VB, Bisecco A, Coghe G, Cocco E, Romoli M, Corea F, Leocani L, Frau J, Sacco S, Inglese M, Carotenuto A, Lanzillo R, Padovani A, Triassi M, Bonavita S, Lavorgna L; Digital Technologies, Web, Social Media Study Group of the Italian Society of Neurology (SIN). ChatGPT vs. neurologists: a cross-sectional study investigating preference, satisfaction ratings and perceived empathy in responses among people living with multiple sclerosis. J Neurol. 2024 Apr 3. doi: 10.1007/s00415-024-12328-x. Epub ahead of print. PMID: 38568227.
49. Franc JM, Cheng L, Hart A, Hata R, Hertelendy A. Repeatability, reproducibility, and diagnostic accuracy of a commercial large language model (ChatGPT) to perform emergency department triage using the Canadian triage and acuity scale. CJEM. 2024 Jan;26(1):40-46. doi: 10.1007/s43678-023-00616-w. Epub 2024 Jan 11. PMID: 38206515.
50. Ozgor BY, Simavi MA. Accuracy and reproducibility of ChatGPT's free version answers about endometriosis. Int J Gynaecol Obstet. 2023 Dec 18. doi: 10.1002/ijgo.15309. Epub ahead of print. PMID: 38108232.
51. Shumway DO, Hartman HJ. Medical malpractice liability in large language model artificial intelligence: legal review and policy recommendations. J Osteopath Med. 2024 Jan 31. doi: 10.1515/jom-2023-0229. Epub ahead of print. PMID: 38295300.
52. Guleria A, Krishan K, Sharma V, Kanchan T. ChatGPT: Forensic, legal, and ethical issues. Med Sci Law. 2024 Apr;64(2):150-156. doi: 10.1177/00258024231191829. Epub 2023 Aug 1. PMID: 37528607.
53. Amram B, Klempner U, Shturman S, Greenbaum D. Therapists or Replicants? Ethical, Legal, and Social Considerations for Using ChatGPT in Therapy. Am J Bioeth. 2023 May;23(5):40-42. doi: 10.1080/15265161.2023.2191022. PMID: 37130418.
54. Hood L, Auffray C. Participatory medicine: a driving force for revolutionizing healthcare. Genome Med. 2013 Dec 23;5(12):110. doi: 10.1186/gm514. PMID: 24360023; PMCID: PMC3978637.
55. Gorini A, Pravettoni G. P5 medicine: a plus for a personalized approach to oncology. Nat Rev Clin Oncol. 2011 May 31;8(7):444. doi: 10.1038/nrclinonc.2010.227-c1. PMID: 21629214.
56. Bragazzi NL. From P0 to P6 medicine, a model of highly participatory, narrative, interactive, and "augmented" medicine: some considerations on Salvatore Iaconesi's clinical story. Patient Prefer Adherence. 2013 Apr 24;7:353-9. doi: 10.2147/PPA.S38578. PMID: 23650443; PMCID: PMC3640773.
57. Zanotti G, Petrolo M, Chiffi D, Schiaffonati V. Keep trusting! A plea for the notion of Trustworthy AI. AI & Soc. 2023;1-12. doi: 10.1007/s00146-023-01789-9.
58. Cacciamani GE, Collins GS, Gill IS. ChatGPT: standard reporting guidelines for responsible use. Nature. 2023 Jun;618(7964):238. doi: 10.1038/d41586-023-01853-w. PMID: 37280286.

Supplementary Files

Untitled.
URL: http://asset.jmir.pub/assets/961bd9c9d417644f89643a5cc52fe871.docx