Clinical Risks of Artificial Intelligence in Medicine

Maxim Glebov MD[1,2] and Alex Mintz PhD[3,4]

[1]Department of Anesthesiology, Sheba Medical Center, Tel Hashomer, Israel
[2]Department of Information Systems, University of Haifa, Haifa, Israel
[3]Research Lab on Computerized Decision Making, Recanati School of Medicine, Reichman University, Herzliya, Israel
[4]Lauder School of Government, Reichman University, Herzliya, Israel

ABSTRACT Generative artificial intelligence (GenAI), particularly large language models, is increasingly used in medicine to support diagnosis, risk prediction, treatment planning, workflow optimization, documentation, and clinical decision-making. Its ability to process large amounts of data and identify complex patterns may greatly improve patient care and assist clinicians in increasingly data-rich healthcare environments. However, GenAI systems are not inherently neutral, and their use in medicine creates a broad spectrum of clinical risks. These risks include biased or unrepresentative data, poor generalizability, inconsistent recommendations, inaccurate labels, inappropriate outcome proxies, opacity, automation bias, erosion of clinical expertise, hallucinations, fabricated evidence, cybersecurity vulnerabilities, privacy concerns, and unclear accountability. In clinical settings, these risks may lead to misdiagnosis, over- or undertreatment, unequal allocation of resources, propagation of false evidence, inappropriate reliance on automated outputs, and loss of trust. In this review, we discuss the major clinical risks and mechanisms of harm associated with GenAI in medicine, provide clinical examples, and suggest strategies to reduce harm. GenAI should be viewed as a powerful clinical tool, but not as a replacement for human judgment, professional accountability, or compassionate medical care.

IMAJ 2026; 28: 717–723

KEY WORDS: generative artificial intelligence (GenAI), biases, clinical decision-making, patient safety, risks

Artificial intelligence (AI) has become one of the most influential technological developments in modern medicine [1]. Within this broad field, generative AI (GenAI), particularly large language models (LLMs), has rapidly become accessible to clinicians, researchers, healthcare organizations, and patients. AI-based tools are increasingly used to support diagnosis, risk prediction, triage, treatment selection, image interpretation, documentation, workflow optimization, and clinical decision-making [1-3]. These systems process large volumes of biomedical data, identify complex patterns, integrate evidence from multiple sources, and provide decision support at speed and scale [1-3].

The potential benefits of AI in medicine are substantial. AI may improve diagnostic accuracy, detect patterns that are difficult for humans to recognize, support earlier identification of clinical deterioration, personalize risk prediction, reduce administrative burden, improve documentation, and help clinicians work more efficiently in increasingly complex healthcare environments. For patients, AI may also improve access to medical information, help them understand symptoms, prepare for medical appointments, navigate health services, and participate more actively in shared decision-making. In this sense, AI represents not only a major technological innovation but also a significant opportunity to improve the quality, accessibility, and efficiency of care.

However, the clinical adoption of AI, especially GenAI, creates new risks. Some are related to bias. AI may reproduce existing inequities, amplify hidden distortions in medical data, and generate systematically inaccurate outputs for certain patients or patient groups [4,5]. Other risks extend beyond traditional bias, including hallucinated or fabricated medical information, unsafe reliance on automated and inconsistent recommendations, loss of clinician oversight, erosion of clinical expertise, cybersecurity vulnerabilities, privacy risks, and unclear responsibility when AI contributes to harm [6-13]. Together, these risks make AI-related harm a broad patient-safety problem rather than only a technical or ethical concern.

These issues are particularly relevant to healthcare systems with advanced digital infrastructure, where GenAI tools can be connected to electronic health records, institutional documents, clinical communication, and evidence-retrieval systems. The same conditions that enable rapid innovation may also accelerate the spread of unvalidated tools, biased models, or poorly governed

AI systems. AI adoption in digitally advanced healthcare systems should therefore be accompanied by clinical governance, local validation, post-deployment monitoring, and physician education.

This review is not intended to argue against the use of AI in medicine, which we view as extremely promising. Rather, we examine biases in data, models, and Human–AI interaction, as well as risks related to inconsistent recommendations, hallucinations, unclear accountability, privacy breaches, security vulnerabilities, and implementation failures. Our aim is to support safer, fairer, and more responsible clinical use of AI. Within this special issue of the *Israel Medical Association Journal* (*IMAJ*), this review focuses particularly on technical AI risks, automation bias, and hallucination.

DEFINITIONS AND SCOPE

Several related terms require clear definition. *AI-related risk* refers to the possibility that the design, training, validation, deployment, recommendations, or clinical use of an AI system may contribute to patient harm, inequitable care, misinformation, privacy breaches, workflow disruption, or unclear accountability. *Bias in medical AI* refers to a systematic and algorithmic distortion that leads to unequal, inaccurate, or clinically inappropriate outputs for certain patients, groups, or settings [4,5]. Generative AI risk arises from systems that generate text, summaries, explanations, citations, or recommendations, including hallucination, fabricated evidence, and fluent but clinically inaccurate outputs. Throughout this issue of *IMAJ*, the generation of plausible but false information by a language model is called *hallucination*. The term *confabulation* is used synonymously in parts of the literature and denotes the same phenomenon.

GENERATIVE ARTIFICIAL INTELLIGENCE OUTPUTS SHOULD BE INTERPRETED LIKE ANY OTHER CLINICAL INFORMATION SOURCE: USEFUL WHEN APPLIED APPROPRIATELY, DANGEROUS WHEN ACCEPTED WITHOUT CONTEXT, AND ALWAYS SUBORDINATE TO CLINICAL JUDGMENT.

This distinction is important because not all dangers of AI are biases. A fabricated citation, a prompt-injection attack, inconsistent recommendations, or unclear legal accountability may harm patients even if no demographic group is specifically disadvantaged. Conversely, biased data and biased labels can create harm even when the model appears technically accurate.

CLINICAL TAXONOMY OF AI-RELATED RISKS

AI-related harm may arise at several points in the clinical AI lifecycle, including data collection, label definition, model development, validation, deployment, clinician interaction, monitoring, publication, and governance [4,5,7]. Table 1 summarizes the major biases and risks discussed in this review. The following sections present a clinically oriented taxonomy of these risks.

HUMAN–AI INTERACTION RISKS: AUTOMATION BIAS, INCONSISTENT RECOMMENDATIONS, AND EXPERTISE EROSION

Some of the most important dangers arise not from the algorithm alone, but from the interaction between clinicians and AI systems. Automation bias is the tendency to over-rely on AI recommendations because they appear objective, precise, or technologically advanced [6,7]. A clinician may accept an AI-generated diagnosis despite contradictory clinical signs. A radiologist may be influenced by an algorithmic image interpretation. A physician may accept an AI-generated summary without noticing that key clinical details are missing. In each case, the danger is not only that AI may be wrong, but that clinicians may become less likely to challenge it.

The opposite problem is algorithm aversion, in which clinicians reject useful AI recommendations because they distrust the system, do not understand it, or remember previous errors. Both overtrust and undertrust can harm patients.

A further concern is erosion of expertise. If trainees repeatedly rely on AI-generated differentials, image interpretations, or treatment suggestions, they may lose opportunities to develop independent clinical reasoning. AI should assist clinical reasoning, not replace it. The reasoning competencies on which safe AI use depends (the capacity to interrogate a recommendation, weigh the quality of the underlying evidence, and recognize when an output does not fit the patient at hand) are examined in the review by Flugelman and Riskin in this issue of *IMAJ* [14].

Inconsistent results and recommendations represent another clinically important form of AI-related risk. The same clinical question may generate different outputs across AI systems or model versions, or even after repeated prompts to the same system. Recommendations may also change according to prompt wording, missing contextual information, the order in which clinical details are provided, or the way uncertainty is expressed. In medicine, such inconsistency may confuse clinicians

Table 1. Biases and risks of AI, including generative AI, in clinical medicine

Domain	Biases and risks
Human–AI interaction	Automation bias, algorithm aversion, overtrust, undertrust, inconsistent recommendations, selective reliance, erosion of clinical expertise
Data and representation	Representation bias, missing-data bias, measurement bias, unrepresentative datasets, fragmented electronic health records, access-to-care bias
Labels, outcomes, and proxies	Labeling bias, outcome bias, proxy bias, cost-proxy bias, inappropriate use of healthcare cost, reproduction of historical treatment patterns
Demographic and social	Race and ethnicity bias, sex and gender bias, age-related bias, language bias, socioeconomic bias, disability-related bias, geographic bias, proxy-variable bias
Model development and validation	Overfitting, poor calibration, reliance on clinically irrelevant proxies, lack of subgroup validation, whole-cohort performance masking subgroup failure, funneling bias, distribution shift
Deployment and workflow	Implementation bias, workflow bias, alert fatigue, delayed or misplaced alerts, local practice mismatch, performance drift, end-user variability
Generative AI	Inconsistent outputs, inconsistent recommendations, hallucinations, fabricated evidence, fabricated citations, misleading summaries, prompt sensitivity, fluent but clinically inaccurate recommendations, loss of control of AI agents
Research, publication, and funding	Research-priority bias, publication bias, positive-result bias, high-resource-setting bias, commercial incentive bias, AI-generated reference errors, rapid dissemination of unvalidated tools
Transparency and accountability	Black-box opacity, limited explainability, unclear responsibility, weak audit trails, diffusion of accountability
Security and privacy	Privacy breach, adversarial manipulation, corrupted training data, prompt injection, indirect prompt injection, data-silo/generalizability tension
Clinical consequences	Diagnostic error, overtreatment, undertreatment, unsafe triage, inconsistent clinical recommendations, inequitable resource allocation, false reassurance, unnecessary escalation, delayed care, propagation of false evidence, loss of trust

AI = artificial intelligence

and patients, especially when AI is used for symptom interpretation, differential diagnosis, medication advice, triage, or preparation for medical appointments. It may also encourage selective reliance, in which users repeat prompts or compare systems until they receive an answer that appears more convincing or confirms a pre-existing expectation. Inconsistent AI recommendations should therefore be treated as a clinical safety concern. Important outputs should be verified against the patient record, current guidelines, institutional protocols, and clinician judgment rather than accepted as an independent clinical authority [5-7,15].

DATA AND REPRESENTATION RISKS

Data-related problems are the foundation of many AI failures in medicine. AI systems learn from the data on which they are trained; therefore, if the data are incomplete, inaccurate, biased, or unrepresentative, the model may reproduce these limitations [4,5]. Medical data are not neutral. They reflect who had access to care, who was tested, how clinicians documented disease, which populations were included in research, and which patients were absent from the dataset.

A common problem is representation bias. If a model is trained mainly on data from one ethnic group, one sex, one age range, one country, or one type of hospital, its performance may decline when applied to other populations [4,5,15]. Dermatology provides a clinically important example. AI tools trained mainly on lighter-skin images may perform worse in patients with darker skin, potentially delaying melanoma diagnosis [16].

Data-related risk also includes missing-data bias. Patients with fragmented care, limited access to diagnostic testing, lower digital literacy, or lower socioeconomic status may have less complete electronic health records [4,5]. The danger is that missing or distorted data may be interpreted by the model as absence of disease, absence of risk, or absence of need. In this way, AI can convert gaps in healthcare access into apparently objective predictions.

LABEL, OUTCOME, AND PROXY RISKS

Many AI systems are trained using labels such as diagnosis, mortality, length of stay, readmission, complications, or treatment response. These labels are often treated as ground truth, but in medicine they are frequently products of human judgment [4,5] and may therefore introduce la-

beling bias. A diagnosis may reflect clinician experience, implicit bias, incomplete investigation, coding practices, local workflows, or historical standards of care. If a disease was historically underdiagnosed in a specific group, an AI model trained on those diagnostic labels may reproduce the same underdiagnosis.

Outcome and proxy bias are closely related. Some outcomes are convenient to measure but may not represent the true clinical goal. Length of stay may reflect hospital capacity as much as disease severity, and healthcare cost may reflect access to care rather than medical need. A well-known example is the use of healthcare cost as a proxy for clinical need, which may underestimate the needs of groups that historically received less healthcare despite similar or greater disease burden. Such problems may be difficult to detect because the model may appear statistically successful while optimizing the wrong target. A model that accurately predicts historical treatment patterns may not improve care; it may simply reproduce past clinical behavior, including past inequities [17].

FABRICATED CITATIONS, FAKED FINDINGS, PROMPT-INJECTION ATTACKS, AND UNCLEAR LEGAL ACCOUNTABILITY CAN HARM PATIENTS.

DEMOGRAPHIC AND SOCIAL RISKS

Demographic and social risks and biases arise when AI systems perform differently across groups defined by race, ethnicity, sex, gender, age, language, socioeconomic status, disability, geography, or other social factors [4,5]. These risks may enter the model directly, through variables such as race or sex, or indirectly, through proxies such as postal code, insurance status, language, hospital site, previous healthcare use, or frequency of testing.

Race and ethnicity require caution. They are social categories, not simple biological variables. Racial differences in model outputs may reflect social, environmental, or structural determinants rather than genetic causation [5]. Sex and gender bias are also clinically important. Cardiovascular disease, for example, may present differently in males and females, while some prediction models may be trained predominantly on male datasets [4]. Similar concerns apply to older adults, children, pregnant women, patients with disabilities, migrants, and non-native-speaking patients who may be underrepresented or poorly characterized in datasets. The empirical evidence quantifying these disparities in current large language models, including the systematic under- and over-referral of specific patient groups, is presented in the review of sociodemographic biases in this issue of *IMAJ* [18].

MODEL DEVELOPMENT AND VALIDATION RISKS

Even when the data are adequate, risk may be introduced during model development. A model may be optimized to perform well overall but still perform poorly in specific groups [4,5]. For clinicians, the key questions are therefore not only "How accurate is this AI tool?" but also "For which patients, in which settings, and for which clinical decisions does it remain accurate?"

Overreliance on whole-cohort performance metrics may conceal clinically important failures. Overall performance measures, such as accuracy, area under the curve, or calibration statistics, may hide poor performance in underrepresented groups or in specific clinical subgroups [5]. Other algorithmic risks and distortions include overfitting, poor calibration, inappropriate feature selection, reliance on clinically irrelevant proxies, lack of subgroup validation, and failure to account for changes in disease prevalence or clinical practice over time [5,7].

Funneling bias is another potential vulnerability of LLM-assisted clinical decision-making. It occurs when a model is trained within a narrow clinical pathway or selected population and then applied more broadly. For example, an algorithm trained mainly on confirmed cases of one disease may encourage clinicians to focus on that diagnosis, narrowing the differential and reinforcing diagnostic anchoring.

DEPLOYMENT AND WORKFLOW RISKS

A model may perform well during development but fail when introduced into real clinical care. Clinical workflows, documentation habits, disease prevalence, devices, staffing, local protocols, and electronic health record systems may differ among development and deployment environments, resulting in workflow bias [5,7]. An AI alert may be ignored because of alert fatigue, arrive too late to change care, or appear in a part of the electronic record that clinicians do not routinely use. End-user behavior may also introduce risk. Physicians may follow AI recommendations for some patients but override them for others.

Deployment risk highlights an important principle. AI tools require local validation, workflow integration, user training, monitoring for performance drift, and mechanisms for reporting errors and unintended consequences [6,7].

GENERATIVE AI, HALLUCINATIONS, AND FABRICATED EVIDENCE

LLMs introduce a new set of risks. Unlike traditional prediction models, GenAI produces text, summaries, explanations, and recommendations. Its output may sound

fluent and authoritative even when it is wrong. This situation is particularly important in medicine, where clinicians may use GenAI to summarize records, draft letters, explain guidelines, generate recommendations, or support clinical reasoning [5].

The most widely recognized problem is hallucination, the generation of plausible but false information [5]. In clinical practice, hallucination is not a minor inconvenience. A fabricated guideline recommendation, incorrect contraindication, false drug statement, or misleading clinical summary may influence care if not verified by a clinician.

A particularly important form of GenAI-related risk is fabricated evidence and citation failure. This situation includes references to studies that do not exist, citations to real articles that do not support the claim being made, and AI-generated summaries that present fabricated or misrepresented evidence as if it were part of the medical literature. Recent analyses suggest that this weakness is no longer a theoretical problem. A 2026 reference-integrity audit examined 2.5 million biomedical papers and reported thousands of fabricated references [8], while a news analysis reported a steep rise in fabricated citations since 2023 [9].

INCONSISTENCY IN RECOMMENDATIONS IS A CLINICAL SAFETY CONCERN. THE SAME CLINICAL QUESTION MAY YIELD DIFFERENT ANSWERS ACROSS LARGE LANGUAGE MODELS OR AFTER REPEATED PROMPTING OF THE SAME SYSTEM, WHICH INVITES SELECTIVE RELIANCE.

This problem may occur at several levels. Clinicians or researchers using GenAI may receive fabricated citations in response to clinical or academic questions. AI-assisted manuscripts may introduce invalid references into the medical literature. If fabricated or misrepresented references are later indexed, cited, or incorporated into future training datasets or retrieval systems, they may be reproduced and amplified by subsequent AI systems. This situation creates a potential feedback loop in which false evidence becomes increasingly difficult to distinguish from real evidence.

PUBLICATION, RESEARCH, AND FUNDING RISKS

Research-priority bias may occur when what is studied, funded, published, and adopted reflects the priorities of technology-rich institutions rather than the needs of all patients [5,15]. If most published AI tools come from high-resource settings, they may not address the needs of low-resource hospitals, community clinics, rural medicine, rehabilitation, geriatrics, mental health, palliative care, or primary care. Commercial incentives may also favor scalable and profitable applications rather than tools designed to reduce inequity.

Publication bias may create the impression that medical AI is more mature and ready for clinical practice than it truly is. Negative studies, failed implementations, subgroup failures, workflow problems, and safety concerns should be reported because they are essential for safe adoption [5,7].

TRANSPARENCY, TRUST, AND ACCOUNTABILITY RISKS

Many AI systems, especially deep learning models, are difficult to interpret. They may produce output without giving clinicians a clear explanation of how the conclusion was reached. This black-box problem creates difficulties of trust, informed consent, accountability, and clinical reasoning [6,10].

Accountability risk refers to uncertainty over who is responsible when AI contributes to harm. Is responsibility held by the clinician who accepted the recommendation, the hospital that purchased the system, the vendor who developed it, or the regulator who approved it? Without clear governance, AI may diffuse responsibility rather than improve care. Clear lines of responsibility, documentation, audit trails, and escalation pathways are therefore essential.

SECURITY, PRIVACY, AND ADVERSARIAL RISKS

Medical AI systems may be vulnerable to adversarial manipulation. Small changes to input, corrupted training data, hidden triggers, or prompt-injection attacks may alter model outputs [10-13]. In practice, this could involve a manipulated radiology image, altered laboratory input, corrupted dataset, or prompt injection against a clinical language model.

Prompt injection is a cybersecurity attack in which malicious instructions are inserted into user input or external content, such as a document, webpage, or email, causing a language model to ignore its intended instructions or generate manipulated output [11-13]. This issue is particularly relevant to medicine because GenAI systems may be used to summarize clinical records, retrieve guidelines, draft discharge letters, or support clinical recommendations. A hidden instruction in an uploaded referral letter, discharge summary, medical document, or webpage could distort the AI-generated summary or advice. As AI agents gain autonomy to plan and execute tasks, a further concern is the potential loss of human control, in which the actions diverge from intended goals and evade effective oversight or timely interruption [19].

Privacy creates another challenge. Strong privacy protection is essential, but excessive data silos may limit dataset diversity and generalizability. Privacy-preserving approaches such as anonymization and federated learning may allow collaboration without moving patient data between institutions [4,10].

CLINICAL CONSEQUENCES

The clinical consequences of AI-related harm may manifest as diagnostic error, inappropriate treatment, unsafe triage, inconsistent clinical recommendations, inequitable resource allocation, false reassurance, unnecessary escalation, delayed care, or propagation of false evidence [4-9]. This harm may occur even when an AI system appears technically accurate because clinical impact depends not only on model performance but also on how the output is interpreted, trusted, and incorporated into care.

False-positive outputs may trigger unnecessary imaging, prescriptions for antibiotics, monitoring, consultations, or intensive care admission. False-negative outputs may delay diagnosis, reduce monitoring, prevent specialist referral, or falsely reassure clinicians and patients. In GenAI systems, an incorrect summary, fabricated guideline recommendation, or unsupported citation may distort clinical reasoning if not verified against the original record or primary source.

Triage and resource-allocation risks are especially concerning because AI systems may increasingly be used to prioritize patients for beds, clinics, imaging, operating rooms, or intensive care. If the underlying model underestimates risk in disadvantaged or underrepresented groups, it may worsen inequitable access to care [4,5,15]. Feedback-loop risk may then reinforce the problem. If an AI system classifies a group as low risk, clinicians may order fewer tests or provide less monitoring, generating future data that appear to confirm that the group was low risk.

The central question is therefore not whether an AI system is technically impressive, but whether it improves care safely, fairly, transparently, and responsibly in the patients and settings where it will be used.

MITIGATION STRATEGIES

Reducing the dangers of AI in medicine requires action across the entire AI lifecycle. First, datasets should be diverse, representative, well documented, and evaluated for missingness, measurement error, and structural bias [4,5,15]. Second, labels and outcomes should be clinically meaningful rather than merely convenient, because poorly chosen proxies may reproduce historical inequities or optimize the wrong clinical target [4,5,17]. Third, model performance should be reported not only for the overall cohort but also across clinically important subgroups, settings, and deployment environments [4,5,7].

Fourth, AI tools should undergo external validation and, where possible, prospective evaluation before routine clinical deployment [6,7]. Fifth, implementation should include workflow analysis, clinician training, alert-fatigue assessment, monitoring for performance drift, and clear procedures for reporting errors or unintended consequences [6,7]. Sixth, GenAI outputs should be verified against primary sources, especially when they include citations, guideline recommendations, drug information, or patient-specific advice [8,9].

Healthcare-specific AI platforms may help address some of these risks, but they should be viewed as mitigation tools rather than complete solutions. OpenAI for Healthcare, for example, is a suite of products offering healthcare-optimized models, evidence retrieval with transparent citations, role-based access controls, audit logs, and privacy safeguards for regulated clinical environments [20]. In parallel, tools that validate medical data and algorithms, verify references, detect fabricated citations, and monitor AI-assisted publications are becoming increasingly important as the use of AI in medicine expands [8,9,19]. Several LLM developers have released newer model versions that attempt to address some of these concerns. These developments are encouraging, but they do not eliminate the need for local validation, human oversight, source verification, and post-deployment monitoring [6,7,10-13,20].

Last, institutions should establish governance structures that define responsibility, documentation standards, privacy safeguards, cybersecurity protections, and procedures for post-deployment monitoring [6,10-13]. Clinicians should also be trained to use AI critically. GenAI outputs should be interpreted like any other clinical test: useful when applied appropriately, dangerous when accepted without context, and always subordinate to clinical judgment and patient-centered care.

CONCLUSIONS

GenAI has great potential to improve medicine. It supports earlier diagnosis, risk assessment, treatment planning, triage, documentation, workflow efficiency, and improved access to medical information for both clinicians and patients [1-3]. Used well, GenAI may help clinicians manage increasing volumes of data, reduce cognitive and administrative burden, and support more informed conversations with patients.

At the same time, GenAI may reproduce and amplify existing inequities, generate false or misleading information and recommendations, create unsafe reliance on automation, weaken clinical reasoning, threaten privacy and security, threaten privacy and security, and blur accountability [21]. These risks do not mean that GenAI should be rejected. Rather, they mean that it should be implemented with the same caution, validation, monitoring, and professional responsibility expected from any powerful clinical tool.

The most important principle is that GenAI must remain a tool in the hands of clinicians. It should support, but not replace, clinical judgment, human responsibility, and the patient–physician relationship. The consequences of GenAI for that relationship, including patient acceptance and the risk of substituting fluent text for genuine empathy, will be examined in a review in the January 2027 issue of *IMAJ*. Physicians should learn to use these tools critically, just as they use laboratory tests, imaging, and clinical guidelines critically. GenAI may provide information, identify patterns, and suggest options, but the physician must interpret that information in the context of the individual patient.

The future of GenAI in medicine should therefore not be judged by technical performance alone. It should be judged by whether GenAI improves care safely, fairly, transparently, and responsibly. The goal is not to slow innovation, but to ensure that it strengthens medicine while protecting patients, clinicians, and public trust.

Correspondence

Prof. A. Mintz
Lauder School of Government, Reichman University, Herzliya 4610101, Israel
Email: mintz.alex@runi.ac.il

References

1. Angus DC, Khera R, Lieu T, et al. AI, health, and health care today and tomorrow: the JAMA Summit Report on Artificial Intelligence. *JAMA* 2025; 334 (18): 1650-64.
2. D'Adderio L, Bates DW. Transforming diagnosis through artificial intelligence. *NPJ Digit Med* 2025; 8 (1): 54.
3. Fahim YA, Hasani IW, Kabba S, Ragab WM. Artificial intelligence in healthcare and medicine: clinical applications, therapeutic advances, and future perspectives. *Eur J Med Res* 2025; 30 (1): 848.
4. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: a call for open science. *Patterns (N Y)* 2021; 2 (10): 100347.
5. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: implications for clinical decision-making. *PLOS Digit Health* 2024; 3 (11): e0000651.
6. Upadhyay U, Gradisek A, Iqbal U, et al. Call for the responsible artificial intelligence in the healthcare. *BMJ Health Care Inform* 2023; 30 (1): e100920.
7. Abbott EE, Rehman T, Rosania A, et al. Understanding and addressing bias in artificial intelligence systems: a primer for the emergency medicine physician. *J Am Coll Emerg Physicians Open* 2025; 7 (1): 100311.
8. Topaz M, Roguin N, Gupta P, Zhang Z, Peltonen LM. Fabricated citations: an audit across 2.5 million biomedical papers. *Lancet* 2026; 407 (10541): 1779-81.
9. Naddaf M. Surge in fake citations uncovered by audit of 2.5 million biomedical-science papers. *Nature* 2026; 653 (8115): 658-9.
10. Subasi A, Ozaltin O, Mitra A, Subasi ME, Sarirete A. Trustworthy artificial intelligence in healthcare. In: Ordóñez de Pablos P, Zhang X (eds.), Accelerating Strategic Changes for Digital Transformation in the Healthcare Industry. Vol 2. London: Academic Press, 2023: 145-77.
11. Greshake K, Abdelnabi S, Mishra S, et al. Not what you've signed up for: compromising real-world LLM-integrated applications with indirect prompt injection. In: Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security. New York: Association for Computing Machinery, 2023: 79-90.
12. National Cyber Security Centre. Exercise caution when building off LLMs. Published August 30, 2023. [Available from www.ncsc.gov.uk/blog-post/exercise-caution-building-off-llms]. [Accessed 31 May 2026].
13. Ruck D, Sutton M. Indirect prompt injection: generative ai's greatest security flaw. London: Centre for Emerging Technology and Security, Alan Turing Institute, 2024. [Available from cetas.turing.ac.uk/publications/indirect-prompt-injection-generative-ais-greatest-security-flaw]. [Accessed 31 May 2026].
14. Flugelman MY, Riskin A. Clinical reasoning and medical decision-making: science, art, and challenges in the era of artificial intelligence. *IMAJ* 2026; 28 (11): 724-9.
15. Celi LA, Cellini J, Charpignon ML, et al. Sources of bias in artificial intelligence that perpetuate healthcare disparities: a global review. *PLOS Digit Health* 2022; 1 (3): e0000022.
16. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. *JAMA Dermatol* 2018; 154 (11): 1247-8.
17. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; 366 (6464): 447-53.
18. Riskin A, Gover A. Sociodemographic biases in medical decision-making. *IMAJ* 2026; 28 (11): 709-16.
19. Bengio Y, Clare S, Prunkl C, et al. International AI Safety Report 2026. DSIT 2026/001, 2026. [Available from internationalaisafetyreport.org/publication/international-ai-safety-report-2026]. [Accessed 13 September 2026].
20. OpenAI. Introducing OpenAI for Healthcare. Published 8 January 2026. [Available from openai.com/index/introducing-openai-for-healthcare]. [Accessed 13 September 2026].
21. Mintz A, Riskin A, Shoenfeld Y. Biases in medical decision-making. *IMAJ* 2026; 28 (11): 686-9.

Three grand essentials to happiness in this life are something to do, something to love, and something to hope for.
Joseph Addison (1672–1719), English essayist, poet, playwright, and politician