

06-01-2026

Artificial intelligence in healthcare management: Clinical applications, evidence, governance, and implementation challenges

Wiky Fhalyang Razaki

To cite this article: Razaki, W. F. (2026). Artificial intelligence in healthcare management: Clinical applications, evidence, governance, and implementation challenges. *Health Economics Insights Journal*, 1(1), 52–63. <https://journal.privietlab.org/index.php/HEIJ/article/view/1861>

To link to this article: <https://journal.privietlab.org/index.php/HEIJ/article/view/1861>



Follow this and additional works at: <https://journal.privietlab.org/index.php/HEIJ>
Health Economics Insights Journal is licensed under a Creative Commons Attribution 4.0 International License.

This HEIJ: Original Article is brought to you for free and open access by Privietlab. It has been accepted for inclusion in Health Economics Insights Journal by an authorized editor of Privietlab Journals

Full Terms & Conditions of access and use are available at: <https://journal.privietlab.org/index.php/HEIJ/about>

Artificial intelligence in healthcare Management: Clinical applications, evidence, governance, and implementation challenges

Wiky Fhalyang Razaki

Faculty of Public Health, University of Indonesia, Depok, West Java, Indonesia

email: wiky.fhalyang@ui.ac.id

Received 15 April 2026

Revised 21 May 2026

Accepted 01 June 2026

ABSTRACT

Artificial intelligence (AI) has evolved from experimental computer science into a practical component of modern healthcare systems. AI is increasingly applied in medical imaging, oncology, dermatology, ophthalmology, electronic health record analytics, clinical decision support, drug discovery, patient communication, and hospital management. However, the effectiveness of AI depends not only on algorithmic performance but also on safety, equity, explainability, workflow integration, regulatory oversight, and public trust. This paper presents an integrative narrative review of AI in healthcare based on evidence from peer-reviewed literature indexed in major databases, particularly Scopus and Web of Science. The review focuses on five key themes: diagnostic support, predictive analytics, treatment personalization, generative AI, and responsible implementation. Evidence indicates that AI can achieve specialist-level performance in tasks such as skin cancer classification, diabetic retinopathy detection, breast cancer screening, and medical image interpretation. AI systems using electronic health records can also predict deterioration, mortality, readmission, and acute kidney injury. Recent advances in large language models demonstrate potential for medical question answering, documentation assistance, and patient communication. Despite these benefits, many AI systems remain inadequately validated in real-world settings. Major concerns include algorithmic bias, lack of transparency, privacy risks, automation bias, and weak external validity. Safe AI adoption therefore requires rigorous clinical validation, continuous monitoring, transparent governance, and human-centered implementation that supports rather than replaces professional clinical judgment.

Keywords: artificial intelligence; healthcare management; clinical decision support; machine learning; deep learning; digital health; ethics

priviet lab.
RESEARCH & PUBLISHING



1. INTRODUCTION

Artificial intelligence has become one of the most significant technological developments in contemporary healthcare. AI refers to computational systems that perform tasks commonly associated with human intelligence, such as pattern recognition, prediction, classification, language understanding, reasoning, and decision support. In healthcare, these capabilities are especially relevant because clinical practice generates large amounts of heterogeneous data, including imaging, laboratory results, electronic health records, clinician notes, genomic profiles, prescriptions, sensor data, and administrative information. Earlier discussions of big data and machine learning emphasized that data alone do not improve medicine; value is created when data are analyzed, interpreted, and translated into clinical action (Obermeyer & Emanuel, 2016; Beam & Kohane, 2018). Broad reviews have since argued that AI may reshape diagnosis, prognosis, treatment planning, and healthcare delivery if it is developed and implemented responsibly (Jiang et al., 2017; Yu et al., 2018; Topol, 2019). The most persuasive promise of AI is not replacing clinicians but strengthening the capacity of health systems to detect disease earlier, reduce preventable errors, personalize treatment, and allocate scarce resources more effectively.

A major reason for the rapid expansion of AI in health is the success of deep learning in image-rich specialties. In dermatology, Esteva et al. (2017) demonstrated dermatologist-level classification of skin cancer using deep neural networks. In ophthalmology, Gulshan et al. (2016) developed and validated a deep learning algorithm for diabetic retinopathy detection, and Ting et al. (2017) extended this work across multiethnic retinal image datasets. Abràmoff et al. (2018) later reported a pivotal trial of an autonomous AI system for diabetic retinopathy detection in primary care settings. In radiology and oncology, AI has been used for mammography, computed tomography, magnetic resonance imaging, tumor segmentation, and radiomics, with McKinney et al. (2020) reporting international evaluation of an AI system for breast cancer screening and Hosny et al. (2018) outlining the broader role of AI in radiology. These studies show that AI is strongest when a clinical task can be represented as a pattern-recognition problem using well-labeled datasets.

Beyond imaging, AI is increasingly applied to electronic health records and clinical prediction. Rajkomar et al. (2018) showed that deep learning models could use EHR data to predict multiple medical events across two academic medical centers. Tomašev et al. (2019) developed a model for continuous prediction of acute kidney injury, while Komorowski et al. (2018) used reinforcement learning to explore optimal treatment strategies for sepsis in intensive care. These examples illustrate a shift from AI as a diagnostic classifier toward AI as a broader decision-support infrastructure. In principle, AI can support preventive medicine, population health management, risk stratification, care coordination, and hospital operations. For healthcare managers, this makes AI a strategic technology as well as a clinical tool, because its success depends on governance, staff capability, digital infrastructure, budgeting, procurement, regulation, and organizational change.

However, the use of AI in healthcare also creates new risks. Ethical analyses warn that AI systems can reproduce inequities already present in data, clinical practice, and healthcare financing (Char et al., 2018; Challen et al., 2019). Obermeyer et al. (2019) demonstrated racial bias in a widely used population health algorithm because the model used healthcare cost as a proxy for health need, thereby underestimating the needs of Black patients. This finding is especially important because it shows that bias may arise not only from technical flaws but also from problem formulation, proxy variables, and structural inequities. A model can be statistically accurate and still clinically unjust. Therefore, responsible AI requires attention to data representativeness, fairness, interpretability, human oversight, clinical responsibility, and continuous monitoring after deployment.

The evidence-to-practice gap remains large. Kelly et al. (2019) argued that although AI research in healthcare is expanding quickly, comparatively few tools have been successfully integrated into routine practice with clear evidence of improved outcomes. This gap has led to reporting and evaluation guidelines such as CONSORT-AI, SPIRIT-AI, DECIDE-AI, and TRIPOD+AI, which emphasize transparent reporting of AI interventions, model inputs and outputs, human-AI interaction, error analysis, early-stage clinical evaluation, and prediction model reporting (Liu et al., 2020; Rivera et al., 2020; Vasey et al., 2022; Collins et al., 2024). These frameworks are important because clinical AI must be judged not only by area under the curve, sensitivity, or specificity, but also by safety, usability, workflow impact, equity, and patient-centered outcomes.

Recently, large language models and foundation models have expanded the scope of medical AI. Moor et al. (2023) proposed the concept of generalist medical AI, in which flexible models can perform multiple tasks across modalities with limited task-specific training. Singhal et al. (2023, 2025) demonstrated rapid progress in large language models for medical question answering and clinical knowledge tasks. These models may support documentation, triage, education, decision support, patient communication, literature synthesis, and administrative automation. Yet they also raise concerns about hallucination, misinformation, privacy, overreliance, accountability, and the difficulty of validating open-ended outputs. The next phase of AI in health will therefore require integrated evaluation across technical performance, clinical outcomes, user behavior, institutional governance, and societal trust.

The objective of this paper is to review the use of AI in healthcare through an IMRAD structure. Specifically, the paper asks: What are the major applications of AI in healthcare? What evidence supports its clinical and managerial value? What limitations and risks must be addressed? What implementation principles are needed for responsible adoption?

2. METHOD

2.1. Research Design

This paper uses an integrative narrative review design. A narrative review is appropriate because the topic spans diverse clinical domains, methods, technologies, and governance issues. Unlike a narrow systematic review focused on a single intervention or outcome, this paper synthesizes evidence across diagnostic AI, predictive analytics, decision support, generative AI, ethics, and implementation. The IMRAD structure is used to present the paper in a journal-style format.

2.2. Literature Selection

The literature base was selected from peer-reviewed journal articles published in internationally recognized outlets commonly indexed in Scopus and Web of Science. Priority was given to high-impact journals and articles frequently cited in AI-healthcare scholarship, including Nature Medicine, Nature, Nature Biomedical Engineering, Nature Reviews Cancer, JAMA, JAMA Internal Medicine, New England Journal of Medicine, Science, The Lancet Digital Health, BMJ, BMJ Quality & Safety, BMC Medicine, npj Digital Medicine, and Stroke and Vascular Neurology. Selection emphasized studies and reviews that met at least one of the following criteria: direct relevance to AI in healthcare; empirical validation of AI models; systematic review or meta-analysis; ethical or safety analysis; reporting guideline; or recent contribution to generative AI and foundation models.

2.3. Inclusion and Exclusion Criteria

Articles were included if they addressed AI, machine learning, deep learning, foundation models, or large language models in relation to health, medicine, clinical decision support, healthcare delivery, patient safety, or governance. Both empirical studies and high-quality review or guideline

articles were included because the field combines technical validation, clinical translation, and policy concerns. Articles were excluded if they focused only on non-healthcare AI, were not peer reviewed, lacked relevance to clinical or health-system application, or were conference papers without journal publication.

2.4. Data Extraction and Synthesis

Key information was extracted on publication year, clinical domain, AI method, application area, evidence type, main contribution, and implementation implications. The synthesis was organized thematically rather than statistically because included studies varied widely in design and outcome measures. Five themes emerged: diagnostic and imaging AI; EHR-based predictive analytics; treatment personalization and decision support; generative and generalist medical AI; and governance, safety, and implementation. No patient-level data were used; therefore, formal ethical approval was not required for this review draft.

3. RESULTS AND DISCUSSION

3.1. Overview of the Evidence Base

The reviewed literature shows a consistent pattern: AI performs best in bounded tasks with high-quality data, clear labels, and defined outputs. Medical imaging is the strongest example because images can be converted into structured inputs for convolutional neural networks and other deep learning architectures. AI systems have demonstrated high diagnostic performance in dermatology, ophthalmology, radiology, pathology, and oncology. However, the same literature also shows that excellent retrospective performance does not automatically produce clinical benefit. Real-world value depends on external validation, workflow fit, clinician trust, patient acceptability, cost-effectiveness, and institutional readiness.

A second pattern is the expansion of AI from diagnosis to prediction and management. EHR-based models can identify high-risk patients, predict acute deterioration, estimate readmission risk, detect adverse events, and support resource planning. These applications are relevant for hospitals and health systems because they connect clinical care with operational management. A third pattern is the emergence of generative AI, which changes the focus from classification to communication and reasoning. Large language models can answer medical questions, summarize records, draft documentation, and assist with patient communication, but they are difficult to validate because outputs are open-ended and context dependent.

3.2. AI in Diagnostic Imaging and Screening

Diagnostic imaging is the most mature area of AI in healthcare. [Esteva et al. \(2017\)](#) trained a deep neural network to classify skin lesions and reported performance comparable to dermatologists. [Gulshan et al. \(2016\)](#) demonstrated automated detection of diabetic retinopathy from retinal fundus photographs, while [Ting et al. \(2017\)](#) validated a deep learning system for diabetic retinopathy and related eye diseases across multiethnic populations. [Abràmoff et al. \(2018\)](#) extended this evidence through a pivotal trial of an autonomous AI diagnostic system for diabetic retinopathy in primary care offices. In breast cancer screening, [McKinney et al. \(2020\)](#) evaluated an AI system internationally and reported strong performance in mammography interpretation.

These studies suggest several benefits. First, AI may increase screening capacity in health systems with workforce shortages. Second, AI may reduce diagnostic variability by applying consistent criteria across cases. Third, AI may improve access in settings where specialists are scarce. Fourth, AI can support triage by prioritizing high-risk images for expert review. These benefits are especially

relevant in countries with unequal distribution of medical specialists, where AI-enabled screening may help extend specialist-level assessment to primary care and community settings.

Nevertheless, diagnostic AI has limitations. Models trained in one population may perform less well in another due to differences in devices, ethnicity, disease prevalence, image quality, referral patterns, or clinical workflow. Medical imaging models may also be vulnerable to dataset shift, where performance declines after deployment because real-world cases differ from training data. Systematic reviews have warned that many AI imaging studies have methodological weaknesses, including inadequate external validation, retrospective designs, selective datasets, and incomplete reporting (Liu et al., 2019; Aggarwal et al., 2021). Therefore, imaging AI should be implemented with local validation, monitoring, and clear escalation pathways.

3.3. AI in Electronic Health Records and Predictive Analytics

EHR-based AI uses structured and unstructured clinical data to predict outcomes, recommend interventions, and identify risk. Rajkomar et al. (2018) demonstrated scalable deep learning using EHR data across two hospitals, showing that models could predict multiple clinical events without extensive site-specific harmonization. Tomašev et al. (2019) developed a model to predict acute kidney injury up to 48 hours in advance. Early warning systems like these may support timely intervention, reduce avoidable complications, and improve resource allocation.

For healthcare managers, predictive AI is significant because it links clinical risk with operational decision-making. For example, models predicting readmission risk can guide discharge planning; models predicting deterioration can inform staffing and monitoring; and models predicting emergency demand can support bed management. In population health, risk stratification can help organizations identify patients who need preventive outreach, chronic disease management, or care coordination.

However, predictive analytics also illustrates the danger of hidden bias. Obermeyer et al. (2019) showed that a commercial health algorithm produced racial bias because it predicted healthcare cost rather than illness burden. This case demonstrates that the choice of outcome variable is a managerial and ethical decision, not merely a technical one. If a health system optimizes cost, utilization, or billing as a proxy for need, it may reproduce inequities in access and care. Therefore, predictive AI should be evaluated using fairness metrics, subgroup performance, clinical relevance, and patient-centered outcomes.

3.4. AI for Treatment Personalization and Clinical Decision Support

AI can support personalized treatment by integrating patient-specific data with evidence-based recommendations. In intensive care, Komorowski et al. (2018) developed the "AI Clinician," a reinforcement learning model that explored treatment strategies for sepsis. The model suggested individualized decisions for intravenous fluids and vasopressors. Although such work is promising, it also illustrates the challenge of moving from retrospective modeling to prospective clinical use. Treatment decisions involve uncertainty, patient values, clinician experience, and dynamic physiological changes. An algorithm trained on historical decisions may learn from both good and poor practice, and its recommendations may be unsafe if applied without careful testing.

Clinical decision support systems are most useful when they fit clinician workflows and provide actionable, timely, and interpretable recommendations. Poorly designed systems may increase alert fatigue, create confusion, or encourage automation bias. AI decision support should therefore be designed around human-AI collaboration. The clinician should remain responsible for contextual judgment, while AI contributes pattern recognition, risk estimation, and information synthesis. This is consistent with Topol's argument that AI should create "high-performance medicine" by allowing

clinicians to focus more on human care, empathy, and complex decision-making rather than repetitive cognitive tasks.

3.5. Generative AI, Large Language Models, and Generalist Medical AI

The development of large language models has created a new phase of healthcare AI. Unlike traditional diagnostic models, LLMs can generate text, answer questions, summarize information, translate language, and support conversation. [Singhal et al. \(2023\)](#) showed that large language models can encode clinical knowledge and can be evaluated using medical question-answering benchmarks. Later work reported progress toward expert-level medical question answering with large language models ([Singhal et al., 2025](#)). [Moor et al. \(2023\)](#) proposed generalist medical AI, a paradigm in which models can perform many medical tasks across data types, including text, images, signals, and structured records.

LLMs may support healthcare in several ways. They can draft clinical notes, summarize long records, explain medical information to patients, support medical education, assist literature review, and help clinicians retrieve relevant information. [Ayers et al. \(2023\)](#) found that chatbot responses to patient questions on a public forum were rated higher in quality and empathy than physician responses in that study context. However, this finding should not be interpreted as evidence that chatbots can replace clinicians. The study evaluated responses to online questions, not full clinical care with examination, diagnosis, responsibility, and follow-up.

The central risk of generative AI is that fluent language can hide error. LLMs may produce hallucinations, outdated information, fabricated citations, or unsafe recommendations. In clinical settings, this is dangerous because users may confuse confidence with correctness. LLMs also create privacy concerns if patient data are entered into insecure systems. Therefore, generative AI in health should be used with guardrails: verified clinical knowledge sources, human review, audit trails, privacy protection, clear scope of use, and performance evaluation in real-world settings.

3.6. Governance, Reporting, and Implementation

The reviewed literature strongly suggests that healthcare AI is not only a technical innovation but also an organizational transformation. Implementation requires governance structures that define accountability, procurement standards, safety testing, monitoring, clinician training, patient communication, and incident response. [Char et al. \(2018\)](#) emphasized ethical challenges in implementing machine learning in healthcare, while [Challen et al. \(2019\)](#) connected AI bias with clinical safety. [Kelly et al. \(2019\)](#) argued that translation into clinical impact requires robust evaluation beyond technical accuracy.

Reporting guidelines are central to this process. CONSORT-AI extends randomized trial reporting for AI interventions by requiring clear description of the AI system, input and output handling, human-AI interaction, and error analysis ([Liu et al., 2020](#)). SPIRIT-AI provides guidance for clinical trial protocols involving AI interventions ([Rivera et al., 2020](#)). DECIDE-AI addresses early-stage clinical evaluation of AI-based decision support systems ([Vasey et al., 2022](#)). TRIPOD+AI updates prediction model reporting for regression and machine learning methods ([Collins et al., 2024](#)). Together, these frameworks show that AI must be evaluated across the entire lifecycle: development, validation, clinical testing, deployment, monitoring, updating, and retirement.

3.7. Clinical, Managerial, and Ethical Implications

This review finds that AI has already demonstrated strong technical performance in selected healthcare tasks, especially imaging-based diagnosis and risk prediction. AI can classify images, detect disease, identify high-risk patients, support documentation, and assist clinical decision-making. The strongest evidence comes from applications where data are abundant, labels are reliable, and clinical

tasks are clearly bounded. Diabetic retinopathy screening, skin cancer classification, breast cancer screening, and radiological image analysis are prominent examples. EHR-based models show promise for prediction and population health management, while LLMs open new opportunities for communication and knowledge work.

However, the main finding is not simply that AI "works." Rather, the evidence indicates that AI works under specific conditions. Technical performance is necessary but insufficient. A model that performs well in a retrospective dataset may fail when deployed in a new hospital, population, device environment, or workflow. Moreover, a model that improves accuracy may not improve patient outcomes if clinicians ignore it, misunderstand it, overtrust it, or lack resources to act on its recommendations. Therefore, AI should be evaluated as a socio-technical intervention embedded in people, processes, technologies, policies, and institutional incentives.

For clinicians, AI should be viewed as an augmentative decision-support tool. Its value lies in helping clinicians detect patterns, prioritize cases, synthesize information, and reduce cognitive burden. In imaging, AI may serve as a second reader or triage tool. In primary care, autonomous or semi-autonomous AI screening tools may expand access for conditions such as diabetic retinopathy. In hospitals, predictive models may identify deterioration earlier. In documentation, generative AI may reduce administrative workload and allow more time for patient interaction.

Yet clinical autonomy and accountability must be preserved. Clinicians need to understand the intended use, limitations, and uncertainty of AI tools. They do not need to know every mathematical detail, but they need sufficient knowledge to evaluate whether a recommendation is plausible and clinically appropriate. Training in AI literacy should become part of health professional education. This includes understanding model bias, false positives, false negatives, calibration, uncertainty, explainability, and patient communication.

For healthcare managers, AI implementation is a strategic governance issue. Successful adoption requires investment in digital infrastructure, data quality, cybersecurity, staff training, workflow redesign, and evaluation capacity. Managers should avoid purchasing AI tools based only on vendor claims or published accuracy metrics. Instead, they should require evidence of external validation, usability, interoperability, fairness, regulatory status, clinical safety, and post-deployment monitoring.

AI procurement should include multidisciplinary review involving clinicians, data scientists, ethicists, legal experts, patient representatives, and operational managers. Health systems should define who is responsible when AI contributes to a clinical decision, how errors are reported, how model performance is monitored, and when a model should be updated or withdrawn. AI should also be assessed for return on investment, not only in financial terms but also in patient outcomes, staff workload, equity, quality, and safety.

Ethics is central to AI in healthcare because AI systems can affect diagnosis, treatment, access, and resource allocation. The [Obermeyer et al. \(2019\)](#) study is a landmark because it shows how a seemingly neutral algorithm can generate racial bias through a flawed proxy variable. Healthcare organizations should therefore avoid assuming that removing sensitive variables such as race or gender automatically prevents bias. Bias can be embedded in labels, utilization patterns, missing data, historical decisions, and structural inequities.

Fair AI requires representative datasets, subgroup performance analysis, transparent outcome definitions, community engagement, and accountability. Equity should be built into model development from the beginning rather than added after deployment. Patient privacy is also critical. AI systems often require large datasets, and generative AI systems may process sensitive information. Health systems must implement strong data governance, including consent policies, de-identification, access control, audit logs, and compliance with relevant regulations.

Explainability is often presented as a solution to AI mistrust, but explanation alone is not enough. Some explanations may be technically impressive but clinically unhelpful. Clinicians need explanations that answer practical questions: Why is this patient high risk? What data influenced the recommendation? What uncertainty exists? What action is suggested? When should the recommendation be ignored? Explainability should therefore be user-centered and linked to clinical decision-making.

Trust should be earned through evidence, transparency, and monitoring. A trustworthy AI system should have clear intended use, validated performance, known limitations, human oversight, fairness assessment, cybersecurity protection, and mechanisms for feedback. Trust also depends on organizational culture. Clinicians may resist AI if it is imposed without consultation or if it increases workload. Conversely, clinicians may overtrust AI if it is marketed as superior to human judgment. Both extremes are dangerous. Implementation should promote calibrated trust.

3.8. Limitations and Future Research Directions

The literature has several limitations. Many AI studies are retrospective and use curated datasets that may not represent real clinical practice. External validation is often limited. Few studies measure patient outcomes, cost-effectiveness, workflow impact, or long-term safety. Reporting is inconsistent, although newer guidelines are improving standards. Many studies compare AI with clinicians under artificial conditions rather than evaluating human-AI teams in real workflows. Generative AI evidence is especially early, and benchmark performance does not necessarily translate into safe clinical use.

Another limitation is the concentration of AI development in high-income settings. Models trained in well-resourced health systems may not generalize to low- and middle-income countries. Data infrastructure, disease prevalence, device quality, clinical workflows, and regulatory capacity differ across contexts. Global AI health research should therefore prioritize local validation, inclusive datasets, and context-sensitive implementation.

Future research should move from model development to implementation science. The field needs more prospective trials, pragmatic evaluations, health economic analyses, and real-world monitoring studies. Studies should compare usual care, AI-only support, and clinician-AI collaboration. Outcomes should include diagnostic accuracy, patient safety, morbidity, mortality, waiting time, cost, clinician workload, patient experience, and equity. Generative AI research should develop standards for evaluating factuality, reasoning, hallucination, empathy, privacy, and harm.

There is also a need for lifecycle governance. AI models can degrade over time due to changes in population, clinical practice, equipment, coding, or disease patterns. Continuous monitoring should therefore be mandatory. Health systems should establish AI registries, model cards, audit processes, and incident reporting pathways. Regulators should require evidence proportional to risk, with stricter standards for autonomous systems and systems used in high-stakes decisions.

4. CONCLUSION

Artificial intelligence is transforming healthcare by expanding the capacity to detect disease, predict risk, personalize treatment, support communication, and improve operational decision-making. The strongest evidence exists in diagnostic imaging and screening, where deep learning models have demonstrated high performance in dermatology, ophthalmology, mammography, and radiology. EHR-based models show promise for predicting clinical deterioration and supporting population health management. Generative AI and foundation models may further reshape documentation, education, patient communication, and multimodal clinical reasoning.

However, AI is not a simple technological solution to healthcare problems. Its benefits depend on responsible design, representative data, external validation, ethical governance, workflow integration, clinician training, patient trust, and continuous monitoring. Poorly implemented AI may worsen inequity, increase errors, create accountability gaps, and undermine professional judgment. The future of AI in healthcare should therefore be human-centered, evidence-based, and equity-oriented. AI should support clinicians and health systems, not replace the relational, ethical, and contextual dimensions of care.

Ethical Approval

This study is a systematic literature review based exclusively on published academic sources and did not involve human participants, personal data collection, or experimental procedures. Therefore, formal ethical approval was not required.

Informed Consent Statement

Not applicable because this study did not involve human participants.

Authors' Contributions

Not applicable

Disclosure Statement

The author declares no potential conflict of interest.

Data Availability Statement

No primary dataset was generated for this study. All materials analyzed are available in the published sources cited in the reference list.

Funding

This research received no external funding.

Notes on Contributor

Wiky Fhalyang Razaki

Wiky Fhalyang Razaki is associated with Faculty of Public Health, University of Indonesia, Depok, West Java

REFERENCES

- Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Medicine*, 1, 39. <https://doi.org/10.1038/s41746-018-0040-6>
- Aggarwal, R., Sounderajah, V., Martin, G., Ting, D. S. W., Karthikesalingam, A., King, D., Ashrafian, H., & Darzi, A. (2021). Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis. *npj Digital Medicine*, 4, 65. <https://doi.org/10.1038/s41746-021-00438-z>

- Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, 183(6), 589–596. <https://doi.org/10.1001/jamainternmed.2023.1838>
- Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA*, 319(13), 1317–1318. <https://doi.org/10.1001/jama.2017.18391>
- Bajwa, J., Munir, U., Nori, A., & Williams, B. (2021). Artificial intelligence in healthcare: Transforming the practice of medicine. *Future Healthcare Journal*, 8(2), e188–e194. <https://doi.org/10.7861/fhj.2021-0095>
- Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, 28(3), 231–237. <https://doi.org/10.1136/bmjqs-2018-008370>
- Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care: Addressing ethical challenges. *New England Journal of Medicine*, 378(11), 981–983. <https://doi.org/10.1056/NEJMp1714229>
- Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., et al. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, e078378. <https://doi.org/10.1136/bmj-2023-078378>
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118. <https://doi.org/10.1038/nature21056>
- Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22), 2402–2410. <https://doi.org/10.1001/jama.2016.17216>
- Haug, C. J., & Drazen, J. M. (2023). Artificial intelligence and machine learning in clinical medicine, 2023. *New England Journal of Medicine*, 388(13), 1201–1208. <https://doi.org/10.1056/NEJMr2302038>
- Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. W. L. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8), 500–510. <https://doi.org/10.1038/s41568-018-0016-5>
- Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, 2(4), 230–243. <https://doi.org/10.1136/svn-2017-000101>
- Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17, 195. <https://doi.org/10.1186/s12916-019-1426-2>
- Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., & Faisal, A. A. (2018). The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11), 1716–1720. <https://doi.org/10.1038/s41591-018-0213-5>

- Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., Denniston, A. K., & SPIRIT-AI and CONSORT-AI Working Group. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nature Medicine*, 26(9), 1364–1374. <https://doi.org/10.1038/s41591-020-1034-x>
- Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., Ledsam, J. R., Schmid, M. K., Balaskas, K., Topol, E. J., Bachmann, L. M., Keane, P. A., & Denniston, A. K. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *The Lancet Digital Health*, 1(6), e271–e297. [https://doi.org/10.1016/S2589-7500\(19\)30123-2](https://doi.org/10.1016/S2589-7500(19)30123-2)
- McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), 89–94. <https://doi.org/10.1038/s41586-019-1799-6>
- Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956), 259–265. <https://doi.org/10.1038/s41586-023-05881-4>
- Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future: Big data, machine learning, and clinical medicine. *New England Journal of Medicine*, 375(13), 1216–1219. <https://doi.org/10.1056/NEJMp1606181>
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>
- Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., et al. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1, 18. <https://doi.org/10.1038/s41746-018-0029-1>
- Rivera, S. C., Liu, X., Chan, A.-W., Denniston, A. K., Calvert, M. J., & SPIRIT-AI and CONSORT-AI Working Group. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *Nature Medicine*, 26(9), 1351–1363. <https://doi.org/10.1038/s41591-020-1037-7>
- Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972), 172–180. <https://doi.org/10.1038/s41586-023-06291-2>
- Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S. R., Cole-Lewis, H., Neal, D., et al. (2025). Toward expert-level medical question answering with large language models. *Nature Medicine*, 31, 943–950. <https://doi.org/10.1038/s41591-024-03423-7>
- Ting, D. S. W., Cheung, C. Y.-L., Lim, G., Tan, G. S. W., Quang, N. D., Gan, A., Hamzah, H., Garcia-Franco, R., Yeo, I. Y. S., Lee, S. Y., Wong, E. Y. M., Sabanayagam, C., Baskaran, M., Ibrahim, F., Tan, N. C., Finkelstein, E. A., Lamoureux, E. L., Wong, I. Y., Bressler, N. M., et al. (2017). Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*, 318(22), 2211–2223. <https://doi.org/10.1001/jama.2017.18152>

- Tomašev, N., Glorot, X., Rae, J. W., Zielinski, M., Askham, H., Saraiva, A., Mottram, A., Meyer, C., Ravuri, S., Protsyuk, I., et al. (2019). A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767), 116–119. <https://doi.org/10.1038/s41586-019-1390-1>
- Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56. <https://doi.org/10.1038/s41591-018-0300-7>
- Vasey, B., Nagendran, M., Campbell, B., Clifton, D. A., Collins, G. S., Denaxas, S., Denniston, A. K., Faes, L., Geerts, B., Ibrahim, M., Liu, X., Mateen, B. A., Mathur, P., McCradden, M. D., Morgan, L., Ordish, J., Rogers, C., Saria, S., Ting, D. S. W., et al. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine*, 28(5), 924–933. <https://doi.org/10.1038/s41591-022-01772-9>
- Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, 2(10), 719–731. <https://doi.org/10.1038/s41551-018-0305-z>