



Artificial Intelligence in Clinical Decision-Making: Applications, Challenges, and Implications for Patient Safety

Elena Marlowe

Independent Researcher in Digital Health, Bristol, United Kingdom

Abstract

Artificial intelligence is increasingly positioned as a means of supporting diagnosis, predicting deterioration, selecting treatment, and organizing clinical information. Its contribution to patient safety, however, depends on the interaction between algorithms, professionals, patients, and healthcare organizations. This narrative review examines applications of artificial intelligence in clinical decision-making and develops a safety-oriented interpretation of the challenges surrounding implementation. Selected clinical studies, methodological publications, reporting guidelines, and international ethical guidance are discussed without claiming a systematic search or pooled estimate of effectiveness. Applications in diagnostic imaging, physiological monitoring, medication management, and language-based assistance illustrate both the opportunities and the limits of current approaches. Major hazards include unrepresentative data, inappropriate outcome proxies, poor calibration, distribution shift, automation bias, alert fatigue, privacy failures, and fabricated generative outputs. A hypothetical graph demonstrates how an apparently accurate alert can have low positive predictive value when the target condition is uncommon. A comparative table links application areas with plausible harms and practical safeguards. The review argues that predictive performance is an intermediate property rather than proof of patient benefit. Safe implementation requires explicit intended use, independent and local validation, prospective assessment of the clinician–system interaction, transparent communication of uncertainty, and continuous monitoring after deployment. Accountability must be distributed clearly across developers and healthcare institutions while preserving meaningful clinical responsibility. Artificial intelligence can strengthen clinical decision-making when its authority remains proportionate to its evidence and when institutions maintain reliable mechanisms for questioning, correcting, and withdrawing unsafe systems.

Keywords: *artificial intelligence, clinical decision-making, patient safety, algorithmic bias, clinical decision support*

INTRODUCTION

Clinical decision-making requires clinicians to interpret incomplete information, estimate probabilities, weigh competing harms, and incorporate patient preferences. These activities occur under time pressure and within organizations that vary in staffing, resources, and access to specialist knowledge. Artificial intelligence offers computational methods for recognizing patterns across images, laboratory results, physiological signals, and clinical narratives. The appeal is understandable: a well-designed system might identify a deteriorating patient earlier, direct attention to an overlooked abnormality, or reduce the cognitive burden of reviewing extensive records. Nevertheless, an additional source of information can also introduce an additional source of error. Building on the clinical perspective described by Topol (2019), this article treats artificial intelligence as a component of a clinical service rather than as an independent substitute for medical judgment. Machine learning models generate classifications or predictions from relationships

learned in data; language models generate sequences of text and may support information organization or reasoning tasks. Neither capability automatically establishes that a recommendation is correct for a particular patient. Clinical benefit requires an appropriate action, timely delivery, available resources, and a favorable balance between expected benefit and harm. The review therefore asks how applications can be evaluated and governed so that their practical effects improve safety. The discussion is a selective narrative synthesis rather than an original patient study or systematic review. Sources were chosen to represent established applications, documented failures, trial reporting standards, and ethical guidance; no exhaustive database search, formal risk-of-bias scoring, or meta-analysis was performed. The article separates reported evidence from the author's conceptual analysis and from an explicitly hypothetical graph. Its central proposition is that safe adoption requires evaluation of the complete decision pathway, including what clinicians do after receiving an output and what happens when the system is unavailable.

Foundations of Artificial Intelligence in Clinical Decision-Making:

Clinical artificial intelligence encompasses several approaches with different strengths and limitations. Supervised learning maps inputs to labels such as a diagnosis, a subsequent event, or a treatment response. Deep learning can identify complex patterns in images and signals, while natural language processing can structure information in free text. Generative systems can draft summaries or propose explanations, but the fluency of their language should not be confused with verified knowledge. These distinctions matter because a classification error, a poorly calibrated probability, and an invented statement create different safety problems and demand different controls. The intended decision must be specified before selecting a model. A tool that prioritizes radiographs for review serves a different purpose from one that rules out disease or initiates treatment. The relevant patient population, users, input requirements, timing, and consequences of error should therefore be documented. A broad claim that a product supports diagnosis is insufficient for evaluating safety. Institutions need to know which diagnosis, in which setting, using which information, and with what expected response. The permissible degree of autonomy should follow the evidence for that particular use rather than the sophistication of the underlying architecture. Prediction must also be distinguished from causal inference. A model may identify patients likely to experience complications without showing that a particular treatment will prevent those complications. Historical treatment decisions can reflect clinician judgment, resource constraints, and unequal access, all of which complicate learning from observational records. Recommending an intervention on the basis of predicted risk alone can produce inappropriate treatment when competing explanations are ignored. This conceptual distinction is especially important for systems proposing medication changes, intensive monitoring, or invasive procedures. Estimating risk is only one part of deciding whether an action is justified. A useful safety model follows information from collection through interpretation to action and outcome. Errors may arise when the wrong patient's data are supplied, when an input is delayed, when the model is used outside its supported population, or when a correct output is misunderstood. Each stage deserves explicit checking. Clinicians should be able to see whether an output is current and whether required inputs are missing. Organizations should preserve an ordinary clinical pathway for situations in which the system cannot produce a dependable answer. This pathway-based analysis is a conceptual framework proposed here, not an empirical finding about any named product.

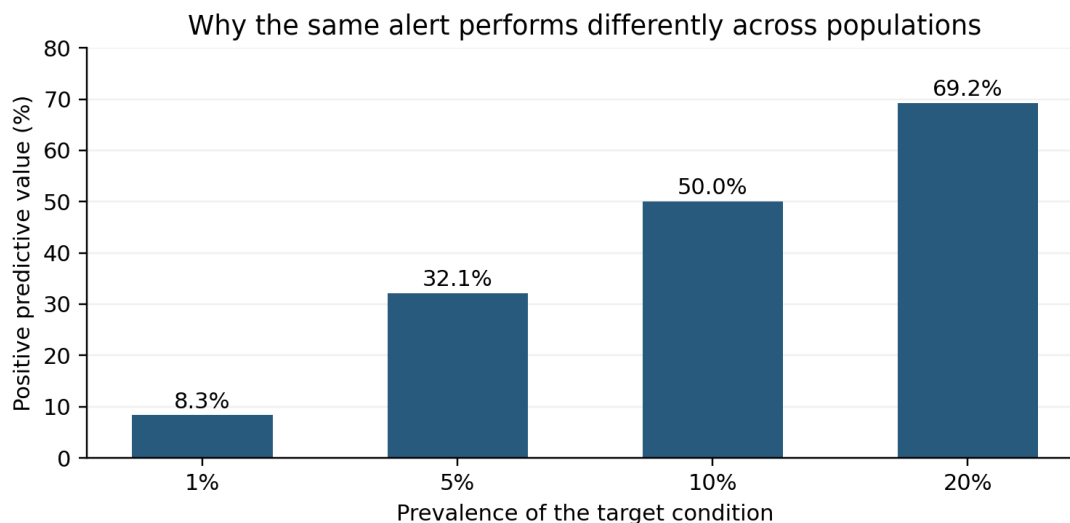
Diagnostic Applications and Clinical Interpretation:

Diagnostic applications are particularly attractive when routine care produces large volumes of images or physiological recordings. In radiology and pathology, algorithms may identify suspicious regions, categorize findings, or prioritize examinations for professional review. In ophthalmology, retinal imaging supports bounded screening tasks with comparatively explicit reference standards. These applications can potentially improve access when specialist availability is limited. Their safety depends on matching the deployed task to the validated task: detecting a specified abnormality is different from interpreting an entire examination or excluding every relevant disease. Abramoff et al. (2018) evaluated an autonomous diabetic retinopathy diagnostic system in a prospective pivotal trial involving primary care settings. The study illustrates the value of defining a target population and a reference standard before claiming clinical usefulness. Its findings should not be generalized automatically to other retinal conditions, imaging devices, or patient groups. A screening result also creates a service obligation. Patients with positive or ungradable

examinations need an accessible referral process; otherwise, technical detection can fail to produce appropriate follow-up. Evaluation of access should therefore include completed referrals rather than screening volume alone. The interpretation of diagnostic evidence requires attention to the comparison group. A retrospective experiment using selected images and expert labels may not reproduce the uncertainty of ordinary practice. Clinical histories, image quality, incidental findings, and competing workload can substantially alter how a tool is used. Nagendran et al. (2020) identified weaknesses in the design and reporting of studies comparing deep learning with clinicians. That review supports caution about translating performance comparisons into claims of clinical superiority. It does not establish that every imaging application is ineffective; rather, it demonstrates why the quality and relevance of the evaluation matter. An institution assessing diagnostic assistance should ask whether it reduces missed actionable disease without creating excessive investigation. Increased detection may identify important pathology, but it can also expose patients to unnecessary tests and procedures. The balance depends on the condition, the consequences of delay, and the downstream pathway. Workflow evaluations should examine disagreement between users and the system, changes in reporting time, and whether attention is diverted from examinations without alerts. These are proposed evaluation questions derived from safety reasoning. They help move assessment from isolated image classification toward the effects of diagnostic assistance on real patients.

Predictive Monitoring and Early Warning Systems

Predictive monitoring uses information accumulated over time to estimate the probability of deterioration, sepsis, readmission, or other adverse events. Electronic records provide repeated observations, but those observations reflect care processes as well as physiology. Frequent blood testing may indicate clinical concern, while missing measurements may reflect limited access, rapid discharge, or documentation practices. A model can learn these patterns without learning a stable representation of disease. Hospitals should therefore examine whether an apparent signal would remain meaningful after changes in staffing, testing policy, or recording systems. Wong et al. (2021) reported an external evaluation of a widely implemented proprietary sepsis prediction model and found substantially poorer discrimination than the developer's previously reported internal performance. This finding concerns the evaluated model and setting, not every sepsis model or subsequent software version. Its broader methodological lesson is the need for independent assessment before relying on a product's development results. An alerting system may appear attractive in procurement materials while producing an unacceptable combination of missed cases and unnecessary alerts in a new environment. Local testing should examine the actual operating threshold and clinical workload. Positive predictive value is especially sensitive to the prevalence of the target condition. In the hypothetical graph included here, sensitivity and specificity are each held at ninety percent while prevalence varies. At one percent prevalence, most alerts are false positives despite these seemingly favorable performance characteristics. The calculation follows Bayes' rule and is offered solely as an educational illustration; it is not an estimate for a clinical product. The example explains why alert burden must be measured in the population where a tool will operate. It also shows why quoting accuracy without the underlying event rate can mislead decision-makers. An early warning alert is useful only when someone can respond appropriately. The response must identify responsibility, available escalation routes, and how quickly reassessment should occur. Excessive low-value alerts can interrupt work and reduce attention to other alarms. Conversely, suppressing alerts aggressively can miss deterioration. Institutions should evaluate the balance using patient-centered outcomes alongside operational measures such as alerts per shift, time to clinical review, and unnecessary investigations. These proposed measures should be interpreted together, because a faster response is not beneficial if it initiates unnecessary or harmful treatment.



Hypothetical calculation • Sensitivity 90% • Specificity 90% • No patient data

Figure 1 Effect of prevalence on positive predictive value.

Treatment Support Medication Safety and Personalized Care:

Treatment support extends beyond identifying disease to helping clinicians choose what to do next. Potential applications include checking medication histories, identifying possible interactions, estimating treatment-associated risks, and organizing options for shared decision-making. Conventional rule-based decision support remains relevant, and a machine learning component should demonstrate added value over simpler alternatives. A more complex model is not inherently safer. Procurement and evaluation should begin with the clinical problem, such as an avoidable prescribing error or delayed recognition of contraindications, rather than with a desire to introduce a particular technology. Medication decisions illustrate why completeness and context matter. A recommendation can be inappropriate when renal function is outdated, an allergy is incorrectly recorded, or a medicine obtained elsewhere is absent from the record. Safe interfaces should distinguish confirmed information from assumptions and should make the basis of a recommendation inspectable. Where critical inputs are unavailable, withholding a specific recommendation may be preferable to generating a confident answer. This is a proposed design principle rather than evidence that a particular medication model improves outcomes. Its purpose is to connect the system's behavior to the consequences of incomplete information. Personalized treatment models also face a causal challenge. Patients receiving different treatments in historical data may differ in severity, prognosis, or access to care before treatment begins. Associations between treatment and outcome can therefore be misleading. An algorithm that reproduces past decisions may preserve both professional expertise and historical mistakes. Prospective comparative evaluation is needed when recommendations could materially alter management. The comparator should be the relevant clinical service, including existing decision support, not an artificially weak alternative. Important outcomes include adverse drug events, symptom burden, functional recovery, and treatment choices consistent with patient priorities. Shared decision-making should remain visible within personalization. Some patients value avoiding hospitalization, while others prioritize maximum diagnostic certainty or the possibility of additional treatment. A probability estimate cannot determine these preferences. Clinicians need to explain reasonable options and uncertainties in language patients understand, especially when the recommendation comes from an unfamiliar system. Treatment support should make room for patient choice and documented exceptions. A departure from an algorithmic recommendation is not automatically an error; it may be the appropriate response to information, values, or circumstances that the system does not represent.

Generative Artificial Intelligence and Clinical Communication:

Generative artificial intelligence introduces a different interaction pattern because users can ask open-ended questions and receive fluent narrative answers. Possible roles include drafting notes, summarizing longitudinal records, translating technical language, and generating candidate differential diagnoses. These

tasks vary substantially in risk. Reformatting verified information is different from proposing an unverified treatment plan. Institutions should distinguish administrative assistance from recommendations that influence clinical action, while recognizing that even documentation errors can become clinically important when copied into later records or used as inputs to other systems. Language models may produce unsupported statements, fabricated references, or plausible explanations that omit a crucial alternative. A summary can change meaning by dropping uncertainty or presenting a historical diagnosis as current. Retrieval from approved sources can improve traceability, but retrieved material still requires checking for relevance, currency, and faithful interpretation. A citation is not sufficient if it does not support the associated claim. The WHO guidance on large multimodal models discusses risks associated with these systems and the need for governance (World Health Organization, 2024). Clinical users need processes that test outputs rather than relying on polished wording. Goh et al. (2024) examined language model access in a randomized clinical vignette study involving physicians and did not find a significant improvement in diagnostic reasoning compared with conventional resources. The result illustrates that providing a capable tool does not necessarily improve the performance of the combined clinician–system arrangement. Vignette results are not direct evidence of effects on patient morbidity or mortality. Their value is in revealing questions about user interaction, training, and task design that require further evaluation in clinical settings before broad claims of benefit are made. A practical safety approach is to assign verification according to the possible consequence of error. Names, dates, medications, allergies, and proposed actions require explicit checking before a generated draft enters an authoritative record. Users should be able to distinguish source text from model interpretation. External documents supplied to a system should be treated as untrusted input because embedded instructions can influence output in unintended ways. Version changes should trigger reassessment, since the same prompt may produce different answers after an update. These are proposed operational safeguards, and their effectiveness should itself be evaluated rather than assumed.

Data Quality Bias and Unequal Patient Outcomes:

Data quality is inseparable from clinical safety because records contain the traces of healthcare delivery, including its omissions. A dataset may underrepresent people who face barriers to attendance, communication, or specialist referral. Diagnostic labels may vary between clinicians, and recorded outcomes may depend on whether follow-up occurred. These issues cannot be repaired merely by increasing dataset size. Developers and hospitals need to understand how variables were collected, what labels mean, and whose experiences remain absent. The question is whether the data support the intended decision for the patients who will encounter the system. Obermeyer et al. (2019) demonstrated racial bias in a population health algorithm that used healthcare costs as a proxy for health need. Unequal expenditure meant that patients with comparable predicted costs could have different levels of illness. The study illustrates a problem in target selection: an apparently reasonable outcome can encode structural inequity. Removing a demographic variable does not necessarily solve this problem because other variables may preserve the same relationships. A safety review should therefore examine the clinical meaning of the prediction target before considering adjustments to the model. Performance assessment should consider clinically relevant subgroups, with attention to uncertainty when sample sizes are small. Aggregate results can conceal poor sensitivity in one population or excessive false alerts in another. Fairness metrics can also conflict, particularly when underlying outcome rates differ. Institutions should explain which disparities they intend to reduce and what trade-offs are acceptable for the use case. Patient representatives can help identify harms that technical teams might overlook, such as reduced access to services after an unfavorable risk classification. Such involvement should influence decisions rather than function as a ceremonial consultation. Distribution shift adds a temporal dimension to inequity. Changes in disease patterns, devices, referral pathways, coding, or treatment can alter model performance after deployment. A model may remain technically available while becoming clinically unreliable. Monitoring should include input quality, subgroup performance, and whether recommended actions are feasible across sites. Retraining is not automatically corrective; it can introduce new errors or change the system's behavior in unexpected ways. Any update should preserve a record of the previous version and should be evaluated against predefined clinical requirements before becoming the basis for patient care.

Human Factors Explainability and Clinical Accountability:

Patient safety depends on how people interpret and act on information. Automation bias can occur when users accept an output too readily or stop searching for contradictory evidence. The opposite response, habitual dismissal, can develop after repeated low-value alerts. Both reactions may arise from the interaction design and work environment rather than from an individual's lack of diligence. Training should therefore include uncertainty, failure cases, and appropriate disagreement. Institutions should observe actual use instead of assuming that a policy statement requiring professional judgment will ensure meaningful oversight. Explainability can support review, but it should be tailored to the decision. A clinician may need to know the input time, missing variables, supported population, and expected action more urgently than a technical account of the architecture. Feature-importance displays can be helpful but do not prove causality or validate a prediction. An explanation can also sound convincing when the underlying output is wrong. Safe communication should distinguish what the model calculated from what the clinical team has established. Where possible, interfaces should allow users to inspect source information and record why a recommendation was accepted or rejected. Accountability requires a clear division of responsibilities. Developers are responsible for describing intended use, limitations, changes, and known failure modes. Healthcare organizations are responsible for procurement, local integration, staffing, training, and surveillance. Clinicians remain responsible for decisions within their professional role, but they cannot compensate for undisclosed product limitations or an unsafe workflow through vigilance alone. Incident review should examine the entire system, including data pipelines and organizational incentives. Assigning all responsibility to the final user can obscure preventable design and governance failures. Patients also need intelligible information about material uses of artificial intelligence in their care. The appropriate communication depends on the task and local requirements, but it should enable questions and access to human review when decisions materially affect care. Explanations should avoid suggesting certainty beyond the evidence. Trust is supported when institutions disclose limitations, respond to concerns, and correct errors. The ethical principles discussed by the World Health Organization (2021) provide a foundation for this approach. Turning those principles into practice requires named responsibilities, usable escalation mechanisms, and evidence that concerns lead to action.

Validation Governance Privacy and Continuous Surveillance:

Validation should proceed through increasingly realistic stages. Technical testing can establish whether a model functions as specified, while external validation examines transportability beyond development data. A silent prospective phase can reveal input failures and workload implications without displaying recommendations to clinicians. Early live evaluation should then examine human interaction and unintended consequences before wider implementation. Kelly et al. (2019) discuss barriers to translating model performance into clinical impact. These stages are best understood as complementary questions rather than a single certificate of safety that permanently settles the matter. Reporting frameworks make evaluations more interpretable. SPIRIT-AI addresses protocols for clinical trials involving artificial intelligence, CONSORT-AI addresses reports of such trials, and DECIDE-AI addresses early clinical evaluation of decision support systems (Cruz Rivera et al., 2020; Liu et al., 2020; Vasey et al., 2022). These frameworks improve transparency about interventions and their use. Compliance does not itself prove effectiveness, eliminate bias, or replace appropriate study design. Investigators should prespecify meaningful outcomes and report failures and deviations alongside favorable results so readers can judge the evidence. Privacy and security controls belong within the safety case because data exposure, unauthorized modification, and service disruption can harm patients. Institutions should define what information leaves the clinical environment, who can access it, how long it is retained, and whether it can be reused for training. Access controls, audit records, and tested recovery procedures should match the sensitivity of the information and the importance of the service. Sending identifiable records to an unapproved service should not become an informal workaround for poor institutional tools. A usable authorized pathway reduces incentives for such workarounds. Post-deployment surveillance should connect technical monitoring with clinical incident reporting. Warning signs include growing missingness, unexpected output patterns, disproportionate errors, excessive overrides, and adverse events involving recommendations. Thresholds for investigation, restricted use, or suspension should be defined before deployment. Procurement agreements should support access to

necessary evidence and timely notification of updates. A shutdown plan should specify how ordinary care continues safely. Continuous surveillance is an organizational commitment requiring staff and resources, not merely an automated dashboard that no one is assigned to review.

Implementation Priorities and Future Research Directions:

Implementation should begin with a defined clinical need and a plausible pathway to improvement. A hospital considering an alert must establish whether earlier recognition is the main barrier or whether delays arise from staffing, bed capacity, or access to treatment. Adding predictions cannot resolve every downstream constraint. A small, well-supported application may produce more useful change than a broad assistant deployed without clear ownership. The appropriate implementation question is what decision will change, who will make it, and whether the service can deliver the response that the technology assumes. The table in this article summarizes proposed links between applications, possible harms, and safeguards. It is a conceptual comparison rather than a ranking of products or a synthesis of effect sizes. Each institution should adapt it to its own population and service. Implementation teams should include clinicians, nurses, pharmacists, data specialists, security staff, quality leaders, and patient representatives where appropriate. Their task is to identify the conditions under which the system adds value and the conditions under which it should defer, request additional information, or be withdrawn. Future research should prioritize patient outcomes, credible comparators, and explicit assessment of unintended effects. Studies can examine whether assistance changes avoidable harm, access, clinician workload, and the distribution of benefits across patient groups. Resource use and opportunity cost also matter: money and staff time spent maintaining an ineffective model cannot support other safety improvements. Evaluations should describe the complete intervention, including training and response pathways, so other institutions can understand what produced the result. Claims about artificial intelligence in general should remain narrower than evidence about a specific intervention. Research on clinician–system interaction should explore when independent initial judgment, structured verification, or selective display of recommendations improves decisions. Evaluation of generative tools should include unsupported content, omitted uncertainty, and the propagation of errors across records. Longer-term studies should investigate whether reliance changes professional skills and whether model updates preserve prior benefits. These questions are proposals from this review rather than established conclusions. A responsible research agenda will make negative and inconclusive findings accessible, supporting decisions to improve, limit, or discontinue systems as readily as decisions to expand them.

Table 1 Applications, plausible safety hazards, and proposed safeguards.

Application	Potential patient safety hazard	Proposed safeguard
Diagnostic imaging	Missed findings or unnecessary investigation	Validate locally; review discordant findings; ensure referral completion
Early warning alerts	False alerts, missed deterioration, delayed action	Measure alert burden; assign response ownership; monitor outcomes
Medication support	Recommendations based on incomplete clinical context	Check current medicines, allergies, renal function, and input timing
Generative assistance	Fabricated content or omission of uncertainty	Verify consequential statements against original records and sources
Risk-based service allocation	Unequal access caused by inappropriate proxy targets	Audit target choice, subgroup outcomes, and access to human review

Summary:

Artificial intelligence can support clinical decision-making through diagnostic assistance, prediction, treatment support, and communication, but its contribution to patient safety cannot be inferred from technical performance alone. The relevant intervention includes the data pipeline, interface, clinical user, response pathway, and organizational capacity. Documented examples involving external validation, outcome proxies, and physician interaction demonstrate why favorable development results need careful

interpretation. The hypothetical graph further shows that alert usefulness depends on disease prevalence as well as sensitivity and specificity. None of these observations supports a universal claim that clinical artificial intelligence is either safe or unsafe; the answer depends on a specified application and its evidence. A safety-oriented approach defines intended use, examines data provenance and subgroup performance, evaluates the combined clinical workflow, and sustains monitoring after introduction. Generative outputs require verification proportional to their consequences, while diagnostic and predictive applications require credible evidence that changed decisions improve care. Institutions should preserve patient involvement, meaningful clinical review, clear accountability, and dependable alternatives when systems fail. The aim is a learning healthcare service in which technological assistance remains open to scrutiny and correction. Patient benefit, equitable access, and avoidable harm should determine whether a system is adopted, expanded, revised, or withdrawn.

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>
- Cruz Rivera, S., Liu, X., Chan, A.-W., Denniston, A. K., & Calvert, M. J. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *Nature Medicine*, 26, 1351–1363. <https://doi.org/10.1038/s41591-020-1037-7>
- Goh, E., et al. (2024). Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, 7(10), e2440969. <https://doi.org/10.1001/jamanetworkopen.2024.40969>
- 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>
- Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., & Denniston, A. K. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nature Medicine*, 26, 1364–1374. <https://doi.org/10.1038/s41591-020-1034-x>
- Nagendran, M., Chen, Y., Lovejoy, C. A., Gordon, A. C., Komorowski, M., Harvey, H., Topol, E. J., Ioannidis, J. P. A., Collins, G. S., & Maruthappu, M. (2020). Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*, 368, m689. <https://doi.org/10.1136/bmj.m689>
- 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>
- Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25, 44–56. <https://doi.org/10.1038/s41591-018-0300-7>
- Vasey, B., et al. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine*, 28, 924–933. <https://doi.org/10.1038/s41591-022-01772-9>
- Wong, A., et al. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, 181(8), 1065–1070. <https://doi.org/10.1001/jamainternmed.2021.2626>
- World Health Organization. (2021). Ethics and governance of artificial intelligence for health: WHO guidance. <https://www.who.int/publications/i/item/9789240029200>
- World Health Organization. (2024). Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. <https://www.who.int/publications/b/70584>