Navigating Accountability and Oversight in AI-Integrated Clinical Systems

Mallesham Goli
Independent Researcher
mallesham.goli.research01@gmail.com

Abstract

AI holds the potential to transform healthcare delivery by improving decision-making and operational efficiency. However, it is important to address the ethical, governance, and operational issues associated with AI-enabled applications before they can be safely and effectively deployed. AI-enabled healthcare presents unique challenges for beneficence, non-maleficence, and patient autonomy. The AI development lifecycle is often not under the control of healthcare institutions, nor are the outputs of AI systems properly understood. Consequently, the true impact of AI on patient outcomes, equity, and justice cannot be adequately evaluated.

Governance frameworks play an essential role in establishing an initial level of assurance. A well-conceived but imperfectly implemented implementation governance framework can help reduce harm and increase public trust. IRBs, in combination with government-sponsored risk management and safety assurance measures, can address most of the requirements of the Safe and Effective Product Regulations. These agencies are best placed to prevent harm emanating from the use of AI-enabled interventions. The next operational development steps focus on building the evidence needed to inform and guide healthcare AI. Proactive information sharing, together with proper documentation and knowledge capes, can mitigate some of the consequences of working without feedback or clinical validation.

Keywords: Artificial intelligence, health, healthcare ethics, ethical theory, beneficence, non-maleficence, patient autonomy.

1. Introduction

Artificial Intelligence (AI) technologies are gradually penetrating the domain of healthcare. While the introduction of AI technologies promises to solve challenges in the field by leveraging patient data, it also raises socioeconomic, legal, moral, and ethical concerns. These concerns need to be formally addressed to ensure the safe and responsible use of AI technologies supporting patient-centred healthcare services. However, the application of AI technologies in healthcare is under-researched, suggesting a need for a thorough criminological analysis of the main forensic and criminological aspects of the diffusion of AI technologies in the industry and the definition of appropriate governance instruments and models capable of grappling with the issues raised by these technologies.

AI has been defined as the discipline aiming to study and make intelligent artefacts. Machine learning plays a prominent role in achieving AI, where machines adapt their behaviour to new situations based on the data they have. These algorithms can be further categorised as unsupervised and supervised algorithms, depending on the nature of the training data presented to the algorithm. Healthcare-related datasets can be classified as structured, semi-structured or unstructured. Structured datasets are defined as fixed field

ISSN: 3067-283X
AMERICAN JOURNAL OF ANALYTICS AND ARTIFICIAL INTELLIGENCE
VOLUME: 02 ISSUE: 01
RECEIVED: JANUARY 26 REVISED: FEBRUARY 09 ACCEPTED: FEBRUARY 22 PUBLISHED: MARCH 08

records, usually stored in relational databases and able to have their fields classified according to a limited number of types. A relational database consists of one or more tables, with rows representing records.The resultant evidence form enables the inclusion of all dimensions required to achieve the best decision by facilitating bias-free access to the perspectives of all parties impacted by that decision.

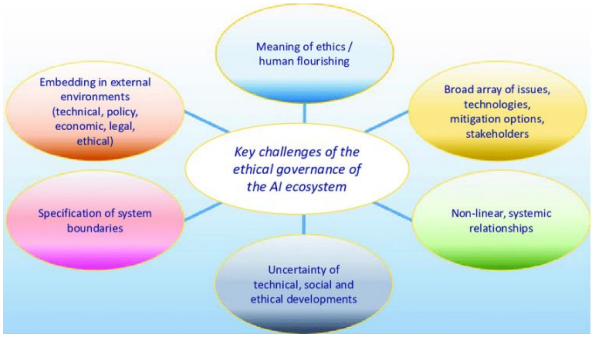


Fig 1: Challenges of ethical governance of AI ecosystems

1.1. Background and Significance

AI-enabled healthcare has been hailed as a medicine of the future. However, the considerable promises come with equally significant ethical and governance challenges, in both concrete and abstract categories. On an ethical level, artificial intelligence could violate beneficence and non-maleficence duties as an intelligent agent that behaves against the interest of humanity (e.g. AWS accident prediction). Moreover, the increasing dependency on decision support systems for clinical diagnosis and treatment makes patient autonomy a rising concern. The lack of human reasoning into determinants for disease predisposition, detection, and treatment may violate fairness in health and wellness. From a governance perspective, the suggestions from supportive communities and agencies are often conflicting and much less comprehensive than the technical expertise; indeed, they tend to overlook the quality-of-life aspect in addition to safety and enterprise.

A framework proposing best practices to mitigate ethical and governance challenges is described. The framework consists of three concrete and five abstract dimensions. The concrete constructs enable the quality taxation of risk and safety aspects. The foundation also supports the synthesis of the diverse ethical frameworks for AI in healthcare and medicine in general, focusing on beneficence, non-maleficence, and patient autonomy. The developed knowledge synthesis framework enhances decisions in medical AI by systematically guiding the extraction and aggregation of evidence relevant to a particular decision.

Equation 1: Validation metrics (derived step by step)

1.1 Confusion matrix definitions

Let the target condition be “disease present”.

- **TP**: true positives (disease present, model predicts present)
- **FP**: false positives (disease absent, model predicts present)
- **TN**: true negatives (disease absent, model predicts absence)
- **FN**: false negatives (disease present, model predicts absent)

Also define totals:

- Actual positives: $P = TP + FN$
- Actual negatives: $N = TN + FP$
- Total cases: $T = P + N = TP + FP + TN + FN$

1.2 Sensitivity (True Positive Rate, TPR)

Sensitivity is the probability the model predicts positive **given** the patient is truly positive.

$$\text{Sensitivity} = \Pr(\hat{Y} = 1 \mid Y = 1)$$

Among truly positive patients, there are $P = TP + FN$ patients total. Out of those, the model correctly identifies TP. Therefore:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

1.3 Specificity (True Negative Rate, TNR)

Specificity is the probability the model predicts negative **given** the patient is truly negative.

$$\text{Specificity} = \Pr(\hat{Y} = 0 \mid Y = 0)$$

Among truly negative patients, there are $N = TN + FP$. Out of those, the model correctly identifies TN. Therefore:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

1.4 Accuracy

Accuracy is the fraction of *all* predictions that are correct.

Correct predictions are TP and TN. Total is T. So:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

1.5 Precision (Positive Predictive Value, PPV)

Precision is the probability the patient is truly positive **given** the model predicted positive.

Predicted positives are $TP + FP$. Of these, correct positives are TP. So:

$$\text{PPV} = \frac{TP}{TP + FP}$$

1.6 Negative Predictive Value (NPV)

Predicted negatives are $TN + FN$. Of these, correct negatives are TN. So:

$$\text{NPV} = \frac{TN}{TN + FN}$$

1.7 F1-score

F1 is the harmonic mean of precision and recall (recall = sensitivity).

$$F1 = \frac{2 \cdot (\text{Precision})(\text{Recall})}{\text{Precision} + \text{Recall}}$$

Substitute $\text{Precision} = \frac{TP}{TP+FP}$ and $\text{Recall} = \frac{TP}{TP+FN}$:

$$F1 = \frac{2 \cdot \frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}}{\frac{TP}{TP + FP} + \frac{TP}{TP + FN}}$$

Factor out TP in denominator:

$$F1 = \frac{2 \cdot \frac{TP^2}{(TP + FP)(TP + FN)}}{TP\left(\frac{1}{TP + FP} + \frac{1}{TP + FN}\right)}$$

Cancel one TP:

:

$$F1 = \frac{2 \cdot \frac{TP}{(TP + FP)(TP + FN)}}{\frac{1}{TP + FP} + \frac{1}{TP + FN}}$$

Combine denominator terms:

$$\frac{1}{TP + FP} + \frac{1}{TP + FN} = \frac{(TP + FN) + (TP + FP)}{(TP + FP)(TP + FN)} = \frac{2TP + FP + FN}{(TP + FP)(TP + FN)}$$

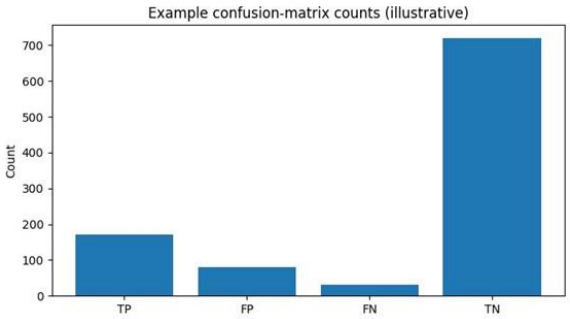
So:

$$F1 = \frac{2 \cdot \frac{TP}{(TP + FP)(TP + FN)}}{\frac{2TP + FP + FN}{(TP + FP)(TP + FN)}} = \frac{2TP}{2TP + FP + FN}$$

2. Conceptual Foundations of AI in Healthcare

Artificial Intelligence (AI) and its various forms have been a subject of foremost interest in scholarly, administrative, and industry circles for the past couple of decades. AI is a not entirely homogeneous category of computer applications using data-driven statistical methods to automatically learn patterns in any available data to predict and make recommendations. Such recommendations can either be qualitative or quantitative assessments that can be, but not necessarily are, used for decision-making by humans. The category also includes algorithms that match or classify rather than predict.

People working with AI speak about 'the intelligence of a machine and its ability to think, learn, reason, and act as a human at some level, often with more accuracy or speed'. In healthcare, Deep Learning



2.1. Research design

The development of AI-enabled healthcare tools is characterized by the exploitation of large amounts of clinical or medical data for the training of algorithms that support or enable clinical decision-making and healthcare delivery. While promising vast benefits in the form of improved patient outcomes, health systems efficiency, and patient-centredness, the introduction of these tools raises key ethical and governance challenges spanning the entire healthcare ecosystem. A methodological framework based on the AI impact convex (AIIC) enables a structured approach for tackling such AI-related concerns and lays the groundwork for the assessment and regulatory oversight of healthcare AI tools.

The AIIC encompasses five dimensions that facilitate an accurate, comprehensive understanding of the pertinent ethical issues and governance considerations, covering: (1) the ethical principles of beneficence, non-maleficence, and patient autonomy; (2) the regulatory landscape and compliance with relevant frameworks; (3) risk management and safety assurance during the validation, verification, and clinical evaluation phases; (4) societal and systemic implications in terms of equity, access, and health disparities; and (5) methodological and operational best practices for reliable evidence synthesis and informed decision-making.

3. Ethical Dimensions

Clinical applications of AIEH [AI-Enabled Healthcare] pose ethical dilemmas related to the principles of beneficence, non-maleficence, and patient autonomy. These standard ethical principles are commonly adhered to in medical ethics, which facilitate the development of an AI-based decision support system while considering the guidance for the welfare of the patient. The principle of beneficence requires that AI and machine-learning algorithms provide better predictions than previous classical prediction methods and possess a better diagnostic power than healthcare professionals, thereby aiding healthcare workers to function more accurately and efficiently. The principle of non-maleficence implies that AI algorithms should not harm patients. It should also be considered if an incorrect prediction leads to further investigations or treatment that do not affect the patient negatively. The question regarding potential negative side effects should be deliberated because possible damage through treatments based on AIEH-supported decisions could also be a factor. The principle of patient autonomy and informed consent requires patients to make their decisions themselves, including the possibility of refusing a treatment decision supported by AIEH. However,

AIEH systems that do not require any interpretation and can only provide solutions in limited cases without the intervention of healthcare professionals will not severely impede patient autonomy.

Equation 2: ROC curve and AUC (related to validation)

When the model outputs a *score* $s(x)$ (risk/probability), choose a threshold τ:

- Predict positive if $s(x) \geq \tau$
- Predict negative otherwise

Then:

$$TPR(\tau) = \frac{TP(\tau)}{TP(\tau) + FN(\tau)}, \quad FPR(\tau) = \frac{FP(\tau)}{FP(\tau) + TN(\tau)}$$

The **ROC curve** is the plot of $TPR(\tau)$ vs $FPR(\tau)$ as τ varies.

AUC (Area Under the ROC Curve) is:

$$AUC = \int_0^1 T\,PR(FPR)\,d(FPR)$$

Numerically (trapezoid rule) for sorted points (FPR_i, TPR_i):

$$AUC \approx \sum_{i=1}^{k-1} \frac{TPR_{i+1} + TPR_i}{2}\,(FPR_{i+1} - FPR_i)$$

3.1. Beneficence, Non-Maleficence, and Patient Autonomy

Three of the ethical principles established by Beauchamp and Childress (2013)—beneficence, non-maleficence, and respect for autonomy—provide a general way to assess the ethical implications of the introduction of AI into clinical routine. Physicians are obliged to provide healthcare to their patients that will have a clinical benefit (the principle of beneficence) and that will not cause avoidable harm (the principle of non-maleficence). Patients have the right to control what is done to their bodies; they should ideally be engaged in a decision-making process when a clinician recommends a procedure that will affect their bodies.

Many AI use cases that are now being realized have a direct relevance to these principles. However, these principles are framed in a context of honesty, justice, and fairness with respect to patient consent. A lack of reliable safety and efficacy data creates a social dilemma relative to the application of these principles. AI technologies with limited evidence—and even market approval—are reaching patients, in some cases affecting critical clinical areas, such as cardiology or pathology, where they cannot be assumed to be safe. They are being used in a situation that can be expressed as "dangerous exemption from judgment" (Wang et al., 2022): the decision to expose patients to an AI-enabled healthcare intervention rests with the clinician, who may not have the knowledge, the information, or the qualifications to appreciate the still limited results from validation and verification studies. In a sense, the clinician who does not take advantage of AI technologies is the one taking a risk, while allowing the patient access to these tools becomes the safe choice.

4. Governance Frameworks for AI in Healthcare

Investment in the software products, including AI algorithms, needs to be accompanied by policy frameworks that explicitly anticipate and address potential unintended consequences. These anticipated consequences are difficult to address and the regulatory landscape is still evolving. Existing regulatory guidance on AI-enabled products should be reviewed for applicability in the healthcare setting.

Regulatory pathways for AI-enabled technologies in healthcare are complex, especially in the United States with its model of pre-market approval and post-market surveillance. AI-enabled products deployed in a healthcare setting may be subject to direct regulation by healthcare authorities. Such regulation may include establishment of

validation datasets, requirements for explanation, safety assessments of data lifecycles, bias testing, transparency mandates, accessibility audits, and funding mechanisms. If the AI system fundamentally changes the safety or efficacy of the device it is integrated with, it will likely trigger pre-market regulatory scrutiny. If the model changes during clinical use (certain machine learning models are non-static), the product may fall under a distinct regulatory scheme specific to software as a medical device. The US Federal Trade Commission has a separate duty to prevent deceptive or unfair acts or practices in commerce, which may involve regulating AI in healthcare products for patients or users (e.g. hospitals). Other authorities may regulate aspects of AI beyond clinical performance (e.g. the US Equal Employment Opportunity Commission on discrimination). The consistency of disparate, sector-specific, public- and private-sector regulatory actions remains an open issue.

Equation 3: Risk management equations (ISO 14971 style)

3.1 Expected harm (basic risk)

If hazard i occurs with probability p_i and causes harm magnitude h_i, then the *expected harm* is:

$$\mathbb{E}[H] = \sum_i p_i \; h_i$$

This is the standard "probability × consequence" aggregation.

3.2 Risk Priority Number (common operational scoring)

A very common operational scoring approach uses ordinal scales:

- Severity $S \in \{1, \dots, 5\}$
- Occurrence / likelihood $O \in \{1, \dots, 5\}$
- Detectability $D \in \{1, \dots, 5\}$ (higher = harder to detect)

Then:

$$RPN = S \times O \times D$$

This is useful for prioritization (though not a perfect substitute for clinical risk evidence).

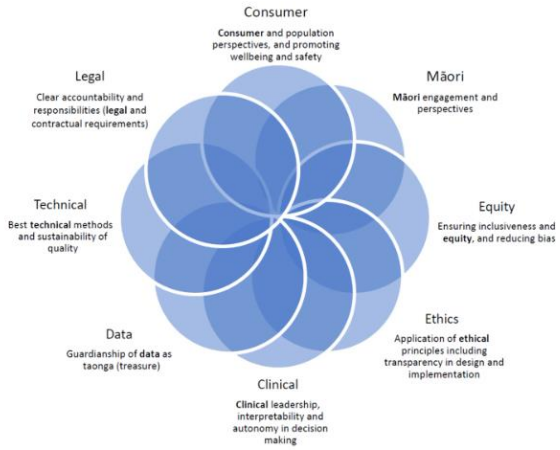


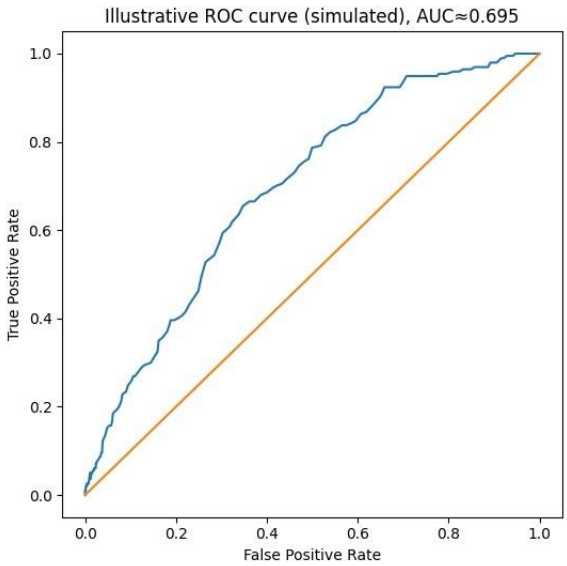
Fig 2: Governance Frameworks for AI in Healthcare

4.1. Regulatory Landscape and Compliance

The European Union's Artificial Intelligence Act, which is expected to come into force in 2024, constitutes the first comprehensive, domain-neutral, risk-based legislative proposal applicable to AI systems. Pursuant to the Act, a clear distinction is drawn between acceptable risk (e.g., social scoring systems, real-time biometric identification in public places), high risk (e.g., critical infrastructure, education, employment, essential services, migration, and law enforcement), limited risk (e.g., chatbots), and minimal risk (e.g., spam filters). Only specific pre-defined high-risk categories merit adherence to the regulatory requirements. At the time of writing, these include AI systems intended to be used in the healthcare sector for scientific research, medical devices, and medical imaging, together with systems intended to be used in other sectors with a mediating

role such as judicial sanctions or criminal conduct assessment.

Regulatory and compliance requirements differ between countries and continents. Furthermore, even within jurisdictions with similar approaches to regulation there can be considerable variation. The EU regulation is fulfilled by a self-assessment process, whereas in the USA external validation by a third party is necessary for medication development and safety before release of the marketing authorisation.



5. Risk Management and Safety Assurance

Ensuring the safety of AI-enabled technologies and instilling public trust in their clinical use are vital for patient safety and healthcare efficiency. A robust risk management process should be established for AI-enabled health technologies, considering their full lifecycle and associated risks. The identification of hazards and estimation of likelihood, severity, and detectability of incidents should be performed by following the principles laid down in ISO 14971. Safety assurance must extend beyond risk management to cover validation, verification, reliability testing, clinical evaluation, and continuous monitoring of actual use. Health technologies that fall under the definition of a medical device will also need to comply with the Medical Device Regulation.

Validation, verification, and testing of healthcare AI systems is important to meet accuracy and safety standards, such as the ISO/IEC 27001 and the ISO/IEC 27018. Validation follows the principle of testing the accuracy of the AI models by comparing predictions against real-life cases in the target population. Complete clinical testing of the AI model in a reasonable sample size is key to reducing clinical risk and increasing patient safety throughout the whole clinical life cycle.

5.1. Validation, Verification, and Clinical Evaluation

Risk management for AI-enabled medical devices must take special consideration of the techniques characteristic of ML systems. Regulatory authorities112 require that manufacturers develop and implement a systematic and integrated risk management process so as to achieve, maintain, and continuously improve the safety of these devices during their entire lifecycle. Such a process should address various risk factors including design, development, validation, production, storage, distribution, installation, use, servicing, disposal, and post-market activities.

Regulatory approval hinges on three key stages: validation (the assurance that a model does what it is intended to do), verification (the assurance that a model has been built correctly) and clinical evaluation. Validation captures the reality of real-world practice, defines the effectiveness and limits of the AI algorithm, establishes threshold levels for the intended use, and places bounds around extrapolation into a new domain. Verification builds confidence that the algorithm has not been corrupted through unintended alterations either in the software, hard-coded parameters, or the data.

ISSN: 3067-283X
AMERICAN JOURNAL OF ANALYTICS AND ARTIFICIAL INTELLIGENCE
VOLUME: 02 ISSUE: 01
RECEIVED: JANUARY 26 REVISED: FEBRUARY 09 ACCEPTED: FEBRUARY 22 PUBLISHED: MARCH 08

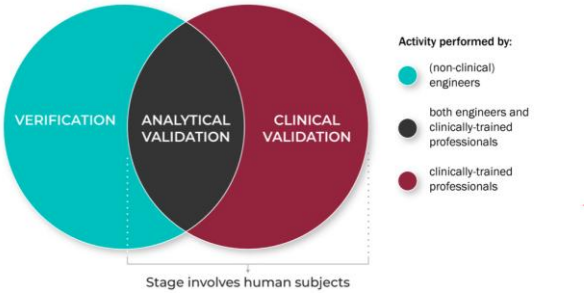


Fig 3: V3 (verification, analytical validation, and clinical validation)

6. Societal and Systemic Implications

Although healthcare is defined as the art of healing and helping others, traditional health systems worldwide continue to face inequitable access, exclusion, and disparities in health outcomes. AI is also being harnessed to enhance the welfare of population groups exposed to societal vulnerabilities, even if AI models can inadvertently introduce biases or propagate inequalities. Therefore, addressing AI health safety requires considering not only operational risks associated with specific products and services but also the broader implications for the health system as a whole and the public. In addition to traditional healthcare services, health systems offer high-quality preventive services and public health, which require substantial investment and collaboration among stakeholders.

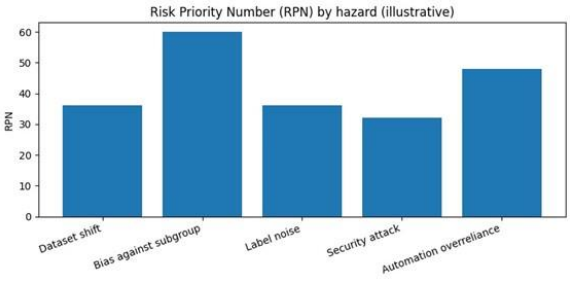
A governance framework for AI-enabled health safety must include measures to ensure that the deployment and integration of AI in healthcare services foster equitable access and equivalent or improved health outcomes. The development of AI models and systems employing datasets from specific demographic segments or populations may trigger additional vigilance, monitoring, scrutiny, or assessment activities. Moreover, the equitable delivery of the health benefits of AI must be a key aspect of decision-making by designers, developers, manufacturers, and users, as well as by oversight and regulatory bodies. All groups involved should agree on the expected balance of health benefits, risks, and unintended consequences at product, service, system, and population levels, and analyze the preventive measures that can be implemented to manage and mitigate those risks.

6.1. Access, Inclusion, and Health Disparities

A more comprehensive regulation of AI and support of national and regional projects is needed to prevent the creation of AI systems that could increase existing disparities in the provision of healthcare. Healthcare AI systems that maintain or increase health disparities are likely to violate the principles of justice, beneficence, non-maleficence, and even autonomy of patients. To reverse the trend of social injustice due to AI in healthcare, further enlargement of the regulation vertical and horizontal is needed. The principle of health for all must be evaluated as the precondition for AI health technology.

Inclusion and access to emerging technologies for all world population segments must be reassessed. The emergence of AI-enabled and AI-Aided technology has outpaced the guarantee of access to technology by society, especially in low-and middle-income. Society's response to the lack of inclusion must be guided by the achievement of distributional equity, that is, guaranteeing that all people and populations have the same opportunity to benefit from the implementation of new technology. The principle of just implementation of technology implies that the traffic of technology benefits, distributed at the moment of implementation, is positive and respects potential gains and losses for all stakeholders. Conversely, the process of specialization in hospitals and the reduction of professionals, the scourge of job degradation in the health area, and the installation of user-feeding overloads in the user-health system subsystem reinforce the debate about the just implementation of AI in medicine, provoking, therefore, the opposite and an indirect violation of the principle of justice.

ISSN: 3067-283X
AMERICAN JOURNAL OF ANALYTICS AND ARTIFICIAL INTELLIGENCE
VOLUME: 02 ISSUE: 01
RECEIVED: JANUARY 26 REVISED: FEBRUARY 09 ACCEPTED: FEBRUARY 22 PUBLISHED: MARCH 08



7. Methodological and Operational Best Practices

Formal methods provide the foundation for the assurance of reliable AI systems and, hence, derive recommendations relevant to validation, verification, and clinical evaluation. They offer a structured architectural approach to AI systems that aligns with the general breakdown of applications. In the simplest classification, AI-enabled tools may provide support to clinicians on a diagnostic or prognostic basis. Generally, a vendor claims performance in such areas based on a statistical analysis of generalisation error in the manner of traditional machine learning, typically with reference to a dataset with some form of clinical ground truth. These cheaper systems are not, however, likely to pass muster for AI safety approval, as they fail to demonstrate assurance and would also be vulnerable to attacks.

For tools advancing support on a treatment decision, the algorithmic underwriting may be less formally methodical, but the risk to patients is markedly higher. These offerings—where the AI recommendations are executed without doctor review—should proceed through a formal toolbox of formal proof and logical analysis. When the recommendations suggest treatment alternatives that differ outside acceptable bounds, they should trigger intuitive-dynamic simulations. For AI-enabled tools that easily process patient information and monitor clinical decision-making, a reactivity-AI underwriting toolbox is appropriate.

7.1. Evidence Synthesis and Decision-Making

Healthcare AI-enabled systems are deployed primarily as decision-support systems for clinicians. With this in mind, governing the design and usage of such systems calls for a practical paradigm, focusing on the evidence used for training or knowledge-generation and, ultimately, on validation and verification that the systems achieve their intended purpose. A clinical decision-support system for which the evidence sources for training or knowledge-generation (benchmarks) are not comprehensive, at least for the domain of greatest need, offers the prospect (barring specific counter-evidence) of inconclusive performance—neither confirmatory nor regulatory facilitating. Even such performance inhibits explanations by the users, whether operation of the AI-enabling models is explainable or not; in turn, lack of credible explanation limits user trust. Lack of comprehensive benchmarks during the clinical evaluation phase leads to an AI-enabled decision-support system that is a ‘black box’—capable of being reliable for some cases but dangerous elsewhere.

Numerous biases—present in the training and knowledge-generation evidence or acquired during testing and calibration of the AI-enabling models—can lead to some patients’ groups being overrepresented among the true positives or negative or the false negatives or positives of the AI-enabled decision-support system. Any AI-enabled declarative system, whether rule-based or using AI-enabling models for knowledge generation, that does not comply with medical-ethical tenets therefore raises the spectre of Griffin disease—also known as previous disease—where the disease at issue infects a population group that is subservient to a directly-affected population group. Group-based considerations become especially important when health services are scarce, with urgency due to catastrophic events, and the overriding criterion for effectiveness becomes the absolute number of cases treated and/or cured.,.

8. Conclusion

The incorporation of AI capabilities in healthcare systems has the potential to fundamentally impact the way healthcare is provided in relation to the cost, quality, and experience. Because of the growing interest in and use cases being deployed around the world, new avenues for research have begun to emerge, driving the need for greater methodological rigour and comprehensive evaluations. The responsible design, development, and deployment of AI technology for healthcare must consider a diverse range of factors. The continued emergence of large language models will also impact the industry.

Ethically sound AI in healthcare requires that fundamental principles—namely beneficence, non-maleficence, and patient autonomy—be embedded within the design and clinical deployment of technology. AI-driven systems enable the manipulation and analysis of large data sets in ways that were previously impossible. Development and deployment must therefore be considered from a systems-level perspective, acknowledging the health disparities that exist between communities and planning the dissemination of the innovation in a way that does not further exacerbate the inequalities that already exist. Methodological approaches must integrate the decision-making process throughout the development and design of the product.

8.1. Emerging Trends

Advances in large language models (LLMs) and diffusion models, among others, have allowed AI systems to generate language and images across diverse domains with little or no human oversight. These developments have opened up new areas for application, some of which were thought to lie over the horizon until now. The emerging capabilities of these systems, coupled with an exponential increase in their scalability, are also increasing the stakes with respect to the potential for misuse: only small spectrums of the user group—from those with malicious intents and substantial resources to those who just want to spread misinformation—are sufficient for exploitation.

Moreover, recent developments have sparked big debates about the claims of artificial general intelligence (AGI) and the risks of extinction-level catastrophe. Concerns relating to these claims emphasize the need for issuing warnings, debates about the proper philosophy of science, and ensuring that the development of highly capable LLMs is always accompanied by appropriate research into the implications of—and attendant prevention possibilities for—the emergence of LLMs that diverge from their desired purpose. These capabilities are clearly evident in the policy domain, where AI systems have a central role for addressing climate change and natural disasters, and where AI is contributing heavily to the production of what are usually considered the ‘traditional’ forms of misinformation, such as disinformation and fake news.

9. References

1. 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, Article 39.
2. Alderman, J. E., Arsenault, C., & Weng, W. H. (2024). Tackling algorithmic bias and promoting transparency in health datasets: The STANDING Together recommendations. The Lancet Digital Health, 6(2), e95–e104.
3. Allen, L. N., et al. (2025). Artificial intelligence in primary care: Frameworks for categorising applications and implications for practice. The Lancet Primary Care, 2(1), 14–26.
4. Bajwa, J., Munir, U., Nori, A., & Williams, B. (2021). Artificial

intelligence in healthcare: Transforming the practice of medicine and the delivery of healthcare. Future Healthcare Journal, 8(2), e188–e194.

5. Beauchamp, T. L., & Childress, J. F. (2019). Principles of biomedical ethics (8th ed.). Oxford University Press.
6. Bouderhem, R., et al. (2024). Shaping the future of AI in healthcare through ethics and regulation. Humanities and Social Sciences Communications, 11, Article 2894.
7. Calvert, M. J., et al. (2020). The CONSORT-AI extension for randomized trials involving artificial intelligence: Checklist and explanation. The Lancet Digital Health, 2(10), e537–e548.
8. Cerdá-Alberich, L., et al. (2023). MAIC-10 (Must AI Criteria-10): A brief quality checklist for publications using artificial intelligence in healthcare. International Journal of Environmental Research and Public Health, 20(3), 2808.
9. Collins, G. S., et al. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078378.
10. Cross, J. L., et al. (2024). Bias in medical AI: Implications for clinical decision-making. Frontiers in Medicine, 11, 1321145.
11. Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. Future Healthcare Journal, 6(2), 94–98.
12. Faes, L., et al. (2020). Automated deep learning design for medical imaging classification by healthcare professionals with no coding experience: A feasibility study. The Lancet Digital Health, 2(5), e232–e242.
13. Fehr, J., et al. (2024). A trustworthy AI reality-check: The lack of transparency of approved medical AI tools. NPJ Digital Medicine, 7, Article 10919164.
14. Flanagin, A., Bibbins-Domingo, K., Berkwits, M., & Christiansen, S. L. (2024). Reporting use of artificial intelligence in research and scholarly publication. JAMA, 331(11), 983–984.
15. Gerke, S., Minssen, T., & Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. Artificial Intelligence in Healthcare, 295–336.
16. General Medical Council. (2022). Regulating doctors in a digital world: Patient safety and the use of AI in clinical care. General Medical Council.
17. Ibrahim, H., Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., & Denniston, A. K. (2021). Guidelines for clinical trial protocols and reports involving artificial intelligence interventions: SPIRIT-AI and CONSORT-AI. Nature Medicine, 27(9), 1472–1477.
18. International Medical Device Regulators Forum. (2023). Good machine learning practice for medical device development: Guiding principles. IMDRF.
19. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17, Article 195.
20. Lekadir, K., et al. (2025). FUTURE-AI: International consensus guideline for trustworthy and deployable AI in

healthcare. BMJ, 388, bmj-2024-081554.

21. Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., & Denniston, A. K. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. The Lancet Digital Health, 2(10), e537–e548.
22. McGenity, C., et al. (2022). Reporting of artificial intelligence diagnostic accuracy study abstracts: An evaluation against reporting guidance. BMJ Open, 12(10), e060839.
23. McMahan, C. J., et al. (2022). Ethical oversight of clinical AI: Practical governance considerations for health systems. Journal of the American Medical Informatics Association, 29(9), 1601–1609.
24. Mennella, C., et al. (2024). Ethical and regulatory challenges of AI technologies in clinical practice. Heliyon, 10(5), e23284.
25. Mihan, A., et al. (2024). Mitigating the risk of artificial intelligence bias in cardiovascular healthcare. The Lancet Digital Health, 6(3), e174–e182.
26. Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2022). Governing data and artificial intelligence for healthcare: Developing an international strategy. JMIR Formative Research, 6(1), e31623.
27. Mongan, J., Moy, L., & Kahn, C. E., Jr. (2020). Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A guide for authors and reviewers. Radiology: Artificial Intelligence, 2(2), e200029.
28. Näher, A. F., et al. (2024). Measuring fairness preferences is important for artificial intelligence in healthcare. The Lancet Digital Health, 6(4), e240–e248.
29. National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce.
30. 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.
31. Panch, T., Mattie, H., & Celi, L. A. (2019). The inconvenient truth about AI in healthcare. NPJ Digital Medicine, 2, Article 77.
32. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine, 169(12), 866–872.
33. Singhal, A., et al. (2024). Toward fairness, accountability, transparency, and ethics in AI-enabled health information dissemination. JMIR Medical Informatics, 12, e50048.
34. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.
35. 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., Watkinson, P., Weber, W., Wheatstone, P., & McCulloch, P. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by

artificial intelligence: DECIDE-AI.
BMJ, 377, e070904.