Future Healthcare Journal 11 (2024) 100179

Contents lists available at ScienceDirect

Future Healthcare Journal

journal homepage: https://www.sciencedirect.com/journal/future-healthcare-journal

Moving beyond the AI sales pitch – Empowering clinicians to ask the right questions about clinical AI

Ibrahim Habli [a,*], Mark Sujan [a], Tom Lawton [b]

[a] Centre for Assuring Autonomy, University of York, United Kingdom
[b] Improvement Academy, Bradford Institute for Health Research, United Kingdom

A B S T R A C T

We challenge the dominant technology-centric narrative around clinical AI. To realise the true potential of the technology, clinicians must be empowered to take a whole-system perspective and assess the suitability of AI-supported tasks for their specific complex clinical setting. Key factors include the AI's capacity to augment human capabilities, evidence of clinical safety beyond general performance metrics and equitable clinical decision-making by the human–AI team. Proactively addressing these issues could pave the way for an accountable clinical buy-in and a trustworthy deployment of the technology.

Introduction

A team from a healthcare software company walks into a hospital. Their goal is to promote their latest AI-based decision support system for recognising deteriorating patients. The pitch, delivered by a sales officer, a clinician and an IT specialist, centres on the AI's impressive performance, regulatory approval and in-use evidence. Researchers at the company and partnering academic institutions have published peer-reviewed papers demonstrating the AI system's performance matching or exceeding that of human clinicians. The relevant regulatory approval has been obtained. More assuringly, the system has already been deployed in other hospitals. Early adopters are offered a significant first-year discount, with flexible cancellation options. To further assure the hospital, the pathways supported by the AI system remain clinically led, with clinicians making the final decision. Given the current pressures on an overstretched workforce, a key short-term benefit is the AI system's potential to reduce backlogs and long waiting lists, many of which were exacerbated by the COVID-19 pandemic.

On the face of it, this seems like an offer that the hospital cannot refuse!

The scenario is hypothetical, but probably not far from real situations that many clinicians and healthcare providers will have experienced in recent times with the extraordinary advancements in AI technologies.[1,2] However, despite the appeal of addressing some of the most pressing concerns, such as escalation of care and reducing backlogs, this scenario is an oversimplification and a technology-centric view of clinical practice and patient experience. While many AI systems perform well (ie with high accuracy) in retrospective evaluations,[3,4] few have been employed successfully in clinical practice, and the existing evidence base is weak.[5,6] Arguably, this problem of 'the last mile',[7] ie, of making the transition from AI development and testing into clinical practice, arises because clinical systems are complex socio-technical systems with inherent variability and uncertainty.[8] This disconnect illustrates a fundamental point: healthcare providers and clinicians must be empowered to ask the right questions about the AI's role within the wider clinical system, rather than allowing software development companies (SDCs) to dictate these questions without sufficient clinical oversight, which risks showcasing AI in isolation. Recent standards, such as BS 30440, which proposes an auditable validation framework for healthcare AI, represent helpful efforts to enable clinicians and healthcare providers to request meaningful assurance from AI developers.[9]

Here, we advocate a shift in the narrative from technology-centric questions to those that highlight the urgency of taking a systems perspective of how AI-based clinical systems could be safely and meaningfully used. We illustrate this shift through the three example questions listed in Table 1. We explore these questions in the rest of the article.

From substitution to augmentation

AI holds immense potential to revolutionise healthcare. However, current applications of AI in healthcare often reflect an unnecessarily narrow design approach based on the metaphor of substitution. From this perspective, AI is seen as a direct replacement for people, and the design challenge is to create AI algorithms that are better at doing something which was previously done by a person. Examples include algorithms that analyse mammograms for breast cancer, differentiate COVID-19 from pneumonia in chest X-rays, identify cardiac arrest from emergency calls, or large language models that sit medical exams.[10–13]

This article reflects the opinions of the author(s) and should not be taken to represent the policy of the Royal College of Physicians unless specifically stated.

* Corresponding author at: Centre for Assuring Autonomy, University of York, York YO10 5GH, United Kingdom.
E-mail address: ibrahim.habli@york.ac.uk (I. Habli).

https://doi.org/10.1016/j.fhj.2024.100179
Received 14 August 2024; Accepted 29 August 2024

Table 1
From technology centric to systems perspective questions.

Technology centric	Systems perspective
How can AI substitute for humans?	How can AI *augment* human performance?
How can we be confident that an AI is safe?	How can we be confident that the *use* of AI is safe?
How can we put a human in the loop to ensure that AI decisions are safe?	How can a *human/AI team* make decisions safely?

This technology-centric approach focuses on whether AI can equal or even surpass human performance in narrowly defined tasks. Evaluations typically look retrospectively at algorithmic accuracy, with few prospective studies examining AI embedded in real-world healthcare systems. The current evidence base raises concerns, as positive results from retrospective evaluations often fail to translate to real-world settings.[14]

Therefore, we propose reframing the design problem by employing alternative metaphors based on a systems perspective.[15] Instead of asking if AI can replace human tasks, we should consider how it can *augment* human capabilities, leading to a transformative improvement in overall task performance.[16]

The augmentation metaphor for healthcare AI holds significant promise. It leverages a systems perspective, which begins by analysing human work: people's capabilities and needs, the tasks involved, the existing tools and technologies, the physical workspace, and the organisational and external environments. We then identify challenges within these tasks and how people overcome them. This analysis can reveal opportunities for AI to augment task performance, rather than simply replicating and replacing human actions.

Box 1 provides an illustrative example of an AI tool designed to recognise cardiac arrest calls. Using the augmentation metaphor and looking at the challenges that people experience in their work, we can develop a range of alternative design options. These could include tools that enhance the intelligibility of unclear speech or improve audio quality from mobile phone calls. Such AI support would empower call handlers to manage calls more effectively, without replicating their existing tasks entirely.

In the deteriorating patient example, this leads to a question of where best to sit the AI in care processes. An algorithm that performs well against humans on electronic data may lead to suggestions that it can substitute for human judgement – but in the real world, the human has access to data, such as the 'end of the bed' or 'eyeball' test,[17] which the algorithm does not. By allowing people and AI to collaborate effectively, leveraging the augmentation metaphor, we can shift the focus from technology-centric substitution to a human-centred approach that utilises AI to empower healthcare professionals and ultimately improves patient care and staff wellbeing.

Design example: supporting ambulance service call handlers in the recognition of cardiac arrest calls
Out-of-hospital cardiac arrest (OHCA) is a major healthcare challenge, with low survival rates. Early detection by ambulance service call handlers is critical for timely intervention. Despite its importance, studies and audits show that call handlers miss more than 25% of OHCA cases.

Current AI design approach (substitution):
Existing AI for OHCA detection employs a substitution metaphor, aiming to replicate the task of call handers. While these AI systems might achieve high accuracy in retrospective evaluations, overall task performance hasn't demonstrably improved in real-world settings.[18]

Augmentation design metaphor:
An alternative approach could focus on designing AI that aims to augment the work of call handlers by focusing on what call handlers find challenging in the recognition of cardiac arrest calls.

This leverages AI to address specific challenges faced by call handlers, such as:
- unclear speech: AI can improve the intelligibility of slurred or difficult-to-understand speech from callers with speech impediments or callers who are in distress
- poor audio quality: AI can enhance audio quality from mobile phone calls with poor reception, allowing for clearer communication.

By addressing these challenges, AI can empower call handlers to more effectively manage calls and potentially improve patient outcomes during suspected OHCA events.

From safe technology to safe use of technology

The ethical principle of non-maleficence in clinical practice, along with related slogans like 'safety first' or 'safety is paramount', drives the clinical safety narrative in healthcare.[20,21] Clinical AI is no exception. The technology must not cause unnecessary harm to patients.[22] However, portraying safety as an inherent property of the technology itself is misleading. AI, like other health software, is merely a collection of 0s and 1s. AI safety becomes relevant only when the technology is integrated into the complexities of clinical settings, characterised by inherent uncertainties and constant change.

A major criticism of systems like the deteriorating patient AI in our example is that existing non-AI systems are already capable of identifying many deteriorating patients, but that this does not always trigger an escalation of treatment.[19] An AI with improved metrics at identifying deterioration may not therefore result in improved or safer care if the rate-limiting step is on the different limb of taking action. A whole-systems approach will ensure that we are solving the right problem.[23]

The transition from lab to bedside exposes several safety misconceptions.[24]

Confusing performance with safety is a classic one. While strong, or even superior, AI model performance is necessary for safe clinical use, it is not sufficient. Reporting only overall AI performance metrics can mask critical edge cases.[25] These edge cases, eg involving comorbidities or underserved groups typically underrepresented in AI training data, can expose patients to unacceptable physical and psychological harm.[26]

Another challenge is performance drift. AI models can degrade over time due to changes in patient demographics or pathologies, or variations across deployment sites. For AI tasks with broad clinical applications, personalised outputs adapting to inevitable clinical variability should be expected of 'intelligent' agents capable of functioning in diverse environments, a hallmark of resilient healthcare systems. For instance, an AI system for out-of-hospital cardiac arrest detection that fails to generalise across different accents might misclassify cases or hinder call handlers in making crucial decisions.

Safety of clinical AI can also be viewed, albeit in a blurry way, through the lens of medical device regulations.[27] A CE marking for an AI system does not guarantee safety. Like other devices, safety depends on the system's actual use and its integration into the actual clinical workflow. Safe deployment also hinges on the buy-in and readiness of the clinical, organisational and technological setting for AI's often distributive nature. This is a longstanding challenge. The 'type' approval nature of current medical device regulations for AI does not adequately

address the fluid nature of this software technology. The AI model itself might be subject to retraining.

Finally, the false sense of agency in AI models can interfere with clinical decision making in unpredictable ways. Consider the common tendency to discuss what AI 'thinks' about a particular case. The risk of overestimating AI capabilities is potentially underestimated. It may lead to scope creep, where AI medical devices approved for screening or triaging are used for diagnosis, potentially violating their intended purpose and conditions of use. To address this, we need to strike the right balance between regulating AI devices and regulating their use. In the UK, the Medicines and Healthcare products Regulatory Agency oversees the former, while the Care Quality Commission and other health regulators focus on the latter. Collaborative regulatory initiatives are emerging to bridge the approval gap between these authorities.[28]

From humans in the loop to human–AI teaming

Current NHS guidance states that where AI is used, the final decision must still be taken by a human.[29] Many SDCs therefore add a human clinician at the end of the decision-making chain, who could end up soaking up moral responsibility[30] and legal liability[31] for decisions taken. While this satisfies the guidance, it places the human in a very awkward position. Either they must check everything at the individual level, reproducing much of the work they would do without an AI and reducing its benefits, or they must give up some of that control, take a step back, and allow it to operate at some level on longer-term trust.[31] Even worse, when something goes wrong, it appears that we will blame the clinician even if their actions were to stop an AI doing something non-standard.[32]

This design pattern seems to result from SDCs developing AI tools in isolation, and then considering the human clinician as a late 'bolt-on' safety addition. As a result, the AI algorithm takes primacy and the humans occupy the remaining space – the so-called 'ironies of automation'[33] also apply to AI systems. People have likely been doing the job successfully for some time before AI came onto the scene, and may have access to information unavailable to the AI, including the ability to converse with patients and elicit signs, symptoms and thoughts.[34] Other approaches may avoid some of these issues.

Radiology is an area of healthcare that is already embracing AI, and some of the patterns here may translate well. The most frequent design pattern here is to support decision making by highlighting areas of interest and suggesting possible diagnoses, but to leave decision making to the human user. This may translate well to other contexts in terms of helping clinicians cope with the vast amounts of information now available in a modern electronic record system, by surfacing the most relevant data for the decision they are aiming to make. However, the pattern of 'decision referral' is also of interest, whereby an AI automates decisions that can be made with high certainty, and leaves the more complex ones to human clinicians.[35] Rather than teaming the human and AI, the AI's responsibilities are bounded by what can be assured to be safe – and the human occupies the rest of the decision space as they always have; decisions made by the AI are not rechecked by a human. While removing what were probably the easiest decisions may reduce the workload less than simple numbers might suggest, this approach gives room for AI to gradually expand as its abilities (and our abilities to assure it) improve.

In our deteriorating patient example, humans are already unavoidably in the loop as it is a system designed only to trigger review. But if it were taken a step further, for example to diagnose sepsis or recommend treatment, then these issues could come into play. Clinicians may anchor on an AI diagnosis or plan which risks frustration at the least, and may result in the discounting of the extra information available to the clinician but not the AI.[36]

It is important to be clear that human clinicians are generally used to accepting risk;[37] the issue is that accepting the risk entails having a degree of understanding and control,[22] and maintaining that understanding and control at an individual patient level likely means reproducing much of the work that they would have done without AI involvement. So the correct question to ask is how the whole system can best work for the benefit of the patient, and to ensure safety of the human–AI team whether they work in combination, or separately with clearly defined roles.

Conclusions

In order to fully realise the potential of healthcare AI and to ensure its suitability for purpose, we must empower clinicians and decision makers to see beyond headline-grabbing sales pitches, and to carefully frame their questions from a systems perspective to avoid hasty and overly simplistic conclusions. Acknowledging the power imbalance between technology companies, supported by influential policy makers and market forces, and the overburdened clinical workforce with outdated digital and organisational infrastructure is essential.

We should therefore pose questions that proactively uncover and meaningfully address the complexities and uncertainties inherent in healthcare delivery. This critical systems thinking mindset will advance a holistic and human-centred approach to AI design and deployment, ensuring its long-term sustainability. By considering how AI integrates into and supports the broader socio-technical system, we can better meet the actual needs of clinicians and, crucially, their patients.

Ethics information

None.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Ibrahim Habli: Conceptualisation, Writing – original draft, Writing – review & editing, Funding acquisition. **Mark Sujan:** Conceptualisation, Writing – original draft, Writing – review & editing. **Tom Lawton:** Conceptualisation, Writing – original draft, Writing – review & editing, Funding acquisition.

Funding

This work was supported by The MPS Foundation Grant Programme (2022-0000000206). The MPS Foundation was established to undertake research, analysis, education and training to enable healthcare professionals to provide better care for their patients and improve their own well-being. To achieve this, it supports and funds research across the world that will make a difference and can be applied in the workplace. The work was also supported by the Engineering and Physical Sciences Research Council (EP/W011239/1).

References

1. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. *N Engl J Med*. 2023;388:1201–1208.
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25:44–56.
3. Li L, Qin L, Xu Z, *et al*. Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. *Radiology*. 2020;296(2):E65–E71.
4. McKinney SM, Sieniek M, Godbole V, *et al*. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577:89–94.
5. Nagendran M, Chen Y, Lovejoy CA, *et al*. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ (Clinical Research Ed)*. 2020;368:m689.
6. Han R, Acosta JN, Shakeri Z, Ioannidis JPA, Topol EJ, Rajpurkar P. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *Lancet Digit Health*. 2024;6:e367–ee73.

7. Coiera E. The last mile: where artificial intelligence meets reality. *J Med Internet Res*. 2019;21:e16323.
8. Sujan M, Furniss D, Grundy K, *et al*. Human factors challenges for the safe use of artificial intelligence in patient care. *BMJ Health Care Inform*. 2019;26:e100081.
9. Sujan M, Smith-Frazer C, Malamateniou C, *et al*. Validation framework for the use of AI in healthcare: overview of the new British standard BS30440. *BMJ Health Care Inform*. 2023;30:e100749.
10. McKinney SM, Sieniek M, Godbole V, *et al*. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577:89–94.
11. Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, *et al*. Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. *Radiology*. 2020:200905.
12. Blomberg SN, Folke F, Ersbøll AK, *et al*. Machine learning as a supportive tool to recognise cardiac arrest in emergency calls. *Resuscitation*. 2019;138:322–329.
13. Ebrahimian M, Behnam B, Ghayebi N, Sobhrakhshankhah E. ChatGPT in Iranian medical licensing examination: evaluating the diagnostic accuracy and decision-making capabilities of an AI-based model. *BMJ Health Care Inform*. 2023;30:e100815.
14. Roberts M, Driggs D, Thorpe M, *et al*. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell*. 2021;3:199–217.
15. Sujan M, Pool R, Salmon P. Eight human factors and ergonomics principles for healthcare artificial intelligence. *BMJ Health Care Inform*. 2022;29:e100516.
16. Shneiderman B. *Human-Centered AI*. Oxford: Oxford University Press; 2022.
17. Arow Z, Gabarin M, Abu-Hosein H, *et al*. Eyeball test for the assessment of frailty in elderly patients with cardiovascular disease: a prospective study. *Am J Cardiol*. 2023;204:9–13.
18. Blomberg SN, Christensen HC, Lippert F, *et al*. Effect of machine learning on dispatcher recognition of out-of-hospital cardiac arrest during calls to emergency medical services: a randomised clinical trial. *JAMA Netw Open*. 2021;4:e2032320.
19. Smith D, Cartwright M, Dyson J, Hartin J, Aitken LM. Barriers and enablers of recognition and response to deteriorating patients in the acute hospital setting: a theory-driven interview study using the theoretical domains framework. *J Adv Nurs*. 2021;77(6):2831–2844. https://onlinelibrary.wiley.com/doi/full/10.1111/jan.14830.
20. Beauchamp TL, Childress JF. *Principles of Biomedical Ethics*. USA: Oxford University Press; 2001.
21. Porter Z, Habli I, McDermid J, Kaas M. A principles-based ethics assurance argument pattern for AI and autonomous systems. *AI Ethics*. 2023:1–24.
22. Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. *Bull World Health Organ*. 2020;98(4):251.
23. Sujan M, Bilbro N, Ross A, *et al*. Failure to rescue following emergency surgery: a FRAM analysis of the management of the deteriorating patient. *Appl Ergon*. 2022;98:103608.
24. Crosby D, Bhatia S, Brindle KM, *et al*. Early detection of cancer. *Science*. 2022;375(6586):eaay9040.
25. Jia Y, Burden J, Lawton T, Habli I. Safe reinforcement learning for sepsis treatment. *2020 IEEE International Conference on Healthcare Informatics (ICHI)* IEEE; 2020:1–7.
26. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*. 2019;28(3):231–237.
27. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *Lancet Digit Health*. 2021;3(3):e195–e203.
28. Understanding regulations of AI and digital technology in health and social care, https://www.digitalregulations.innovation.nhs.uk/[accessed 14 August 2024]
29. NHS England. Artificial Intelligence. NHS England - transformation directorate; [accessed 14 August 2024] https://transform.england.nhs.uk/information-governance/guidance/artificial-intelligence
30. Smith H, Birchley G, Ives J. Artificial intelligence in clinical decision-making: rethinking personal moral responsibility. *Bioethics*. 2024;38(1):78–86.
31. Lawton T, Morgan P, Porter Z, *et al*. Clinicians risk becoming 'liability sinks' for artificial intelligence. *Future Healthc J*. 2024;11(1):100007.
32. Tobia K, Nielsen A, Stremitzer A. When does physician use of AI increase liability? *J Nucl Med*. 2021;62(1):17–21.
33. Bainbridge L. Ironies of automation. *Automatica*. 1983;19:775–779.
34. Lauck SB, Achtem L, Borregaard B, *et al*. Can you see frailty? An exploratory study of the use of a patient photograph in the transcatheter aortic valve implantation programme. *Eur J Cardiovasc Nurs*. 2021;20(3):252–260.
35. Leibig C, Brehmer M, Bunk S, Byng D, Pinker K, Umutlu L. Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis. *Lancet Digit Health*. 2022;4(7):e507–e519.
36. Fogliato R, Chappidi S, Lungren M, *et al*. Who goes first? Influences of human-AI workflow on decision making in clinical imaging. *2022 ACM Conference on Fairness, Accountability, and Transparency*. Seoul Republic of Korea: ACM; 2022:1362–1374.
37. Lawton R, Robinson O, Harrison R, Mason S, Conner M, Wilson B. Are more experienced clinicians better able to tolerate uncertainty and manage risks? A vignette study of doctors in three NHS emergency departments in England. *BMJ Qual Saf*. 2019;28(5):382–388.