

ARTICLE OPEN



Quantifying the impact of AI recommendations with explanations on prescription decision making

Myura Nagendran^{1,2,3}, Paul Festor^{1,3,4}, Matthieu Komorowski², Anthony C. Gordon² and Aldo A. Faisal^{1,3,4,5}✉

The influence of AI recommendations on physician behaviour remains poorly characterised. We assess how clinicians' decisions may be influenced by additional information more broadly, and how this influence can be modified by either the source of the information (human peers or AI) and the presence or absence of an AI explanation (XAI, here using simple feature importance). We used a modified between-subjects design where intensive care doctors ($N = 86$) were presented on a computer for each of 16 trials with a patient case and prompted to prescribe continuous values for two drugs. We used a multi-factorial experimental design with four arms, where each clinician experienced all four arms on different subsets of our 24 patients. The four arms were (i) baseline (control), (ii) peer human clinician scenario showing what doses had been prescribed by other doctors, (iii) AI suggestion and (iv) XAI suggestion. We found that additional information (peer, AI or XAI) had a strong influence on prescriptions (significantly for AI, not so for peers) but simple XAI did not have higher influence than AI alone. There was no correlation between attitudes to AI or clinical experience on the AI-supported decisions and nor was there correlation between what doctors self-reported about how useful they found the XAI and whether the XAI actually influenced their prescriptions. Our findings suggest that the marginal impact of simple XAI was low in this setting and we also cast doubt on the utility of self-reports as a valid metric for assessing XAI in clinical experts.

npj Digital Medicine (2023)6:206; <https://doi.org/10.1038/s41746-023-00955-z>

INTRODUCTION

AI-driven medical decision support systems (AI-CDSS) could have a major impact on medical care due to their theoretically super-human performance. In practical settings however, a translation gap remains (especially and counterintuitively within the data-rich environment of critical care medicine) with few systems active in real-world hospital environments^{1–3}. This gap implies that the challenge of responsibly guiding clinicians to incorporate AI recommendations into their day-to-day practice might require more than AI suggestions alone. A key demand from clinicians, AI researchers and regulators alike is explainable AI (XAI) which aims to not only provide recommendations but also to justify or motivate the AI reasoning to experts^{4,5}. However, most studies that practically evaluate whether and how explanations affect expert decision-making focus on general problems with lessons that do not necessarily translate to high complexity tasks in the clinical sphere^{6,7}. In the few cases where medical XAI has been investigated with clinical experts, these have tended to focus on diagnostic scenarios for which a pre-existing gold standard exists with which to calculate accuracy^{8–10}. This is not the case for many non-diagnostic medical problems such as the haemodynamic management of sepsis that affects millions of patients worldwide¹¹.

Here, we use the flagship example of the AI Clinician system which addresses sepsis resuscitation¹², a topic fraught with uncertainty, wide variation in clinical practice and no clear optimal solution, at least to the human eye^{13,14}. This is despite both decades of research and the provision of international guidelines¹⁵. The ongoing prospective evaluation of our AI Clinician raises critical questions on how to best render the action recommendations explainable and trustworthy to clinicians who

may or may not choose to execute them. This is as much a problem of clinicians' cognition, influenceability and psychology as one of machine learning^{16–18}.

We address these issues in this study by assessing how clinicians' decisions may be influenced by additional information more broadly, and how this influence can be modified by either the source of the information (human peers or AI) and the presence or absence of an AI explanation (here using simple feature importance). Only by further understanding these critical building blocks for an AI-CDSS can we hope to achieve the end goal of improved outcomes for patients. We find that additional information (peer, AI or XAI) has a strong influence on prescriptions (significantly for AI, not so for peers) but simple XAI does not have higher influence than AI alone. There is no correlation between attitudes to AI or clinical experience on the AI-supported decisions and nor is there correlation between what doctors self-reported about how useful they found the XAI and whether the XAI actually influenced their prescriptions. Our findings suggest that the marginal impact of simple XAI is low in this setting and we also cast doubt on the utility of self-reports as a valid metric for assessing XAI in clinical experts.

RESULTS

86 ICU doctors were recruited (31 senior [consultant/attending], 42 intermediate [registrar/fellow], 13 junior [senior house officer/resident]). Median subject age was 37 years (interquartile range (IQR) 34–43). Median years of clinical experience was 11 years (IQR 9–19). All subjects completed the task successfully and there was no significant difference in per trial completion time between

¹UKRI Centre for Doctoral Training in AI for Healthcare, Imperial College London, London, UK. ²Division of Anaesthetics, Pain Medicine, and Intensive Care, Imperial College London, London, UK. ³Brain and Behaviour Lab, Imperial College London, London, UK. ⁴Department of Computing, Imperial College London, London, UK. ⁵Institute of Artificial & Human Intelligence, University of Bayreuth, Bayreuth, Germany. ✉email: a.faisal@imperial.ac.uk

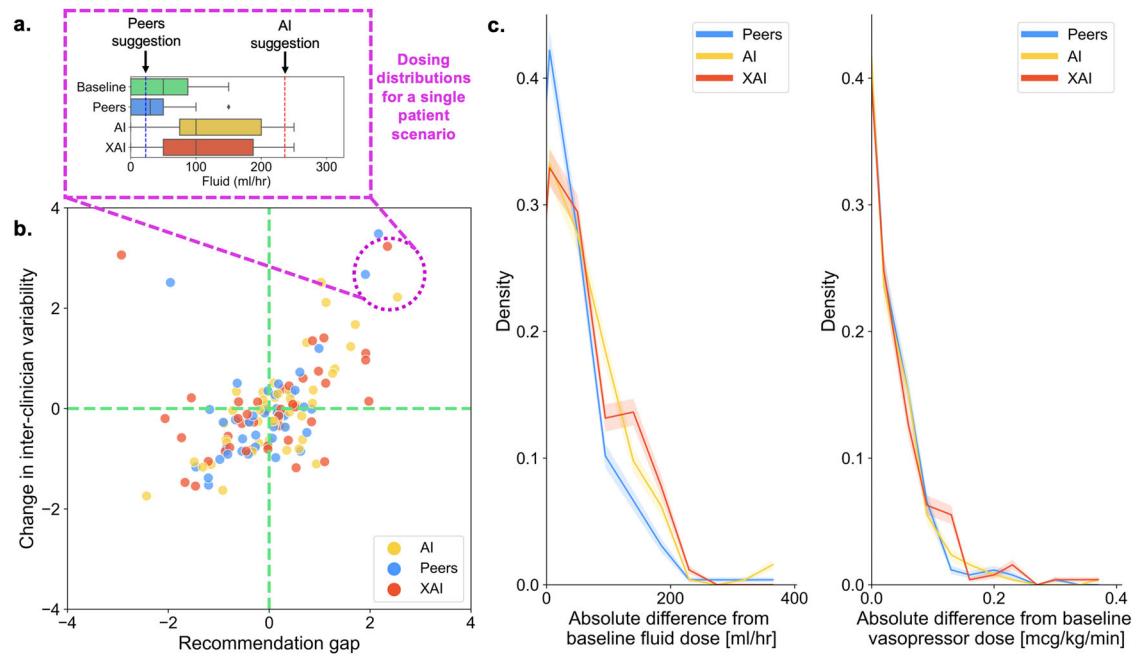


Fig. 1 Dose shift and variability by intervention arm. **a** The prescription distributions for a single patient scenario (and, for illustrative purposes only, mapped onto **(b)**). For each boxplot, the centre line represents the median, box edges represent upper and lower quartiles, whiskers represent 1.5 \times inter-quartile range and diamonds are outliers. Blue dashed line represents the median of the peer distribution data (only available to those in the 'Peers' arm). Red dashed line represents the AI suggested dose (only available to those in the 'AI' or 'XAI' arms). **2b**, change in inter-clinician variability by size of recommendation difference for peers/AI/XAI (i.e. was the recommended dose higher (positive recommendation gap) or lower (negative recommendation gap) than the baseline average (dashed green line) and how does this affect variability of clinicians (x- and y-axes scales are arbitrary units, normalised to allow fluid and vasopressor to be plotted together). **c** Absolute difference (i.e. 50 ml in either direction treated as +50 ml discordance) from dose in the baseline group, aggregated for all 24 patient scenarios. The error bars are formed by randomly taking 1000 bootstraps of the data (80% subset with replacement) and estimating a distribution for prescription doses (error bar is distribution mean \pm standard deviation).

arms (baseline 83 seconds (s), peers 83s, AI 79s, XAI 83s, $p = 0.574$ by Kruskal–Wallis test).

Impact of arms on dosing shift

An example of prescription shift for an individual patient scenario is shown in Fig. 1a. For the same patients in different arms, providing subjects with additional information from their respective arm led to an absolute prescription shift for fluid of 70 ml/hr (peers, standard deviation (SD) 86 ml/hr), 90 ml/hr (AI, SD 83 ml/hr) and 85 ml/hr (XAI, SD 60 ml/hr) relative to the baseline arm ($p = 0.872$ for peers, $p = 0.002$ for AI, $p = 0.007$ for XAI, all by independent T -test). For vasopressor, the prescription shift was 0.04 mcg/kg/min (peers, SD 0.06 mcg/kg/min), 0.05 mcg/kg/min (AI, SD 0.09 mcg/kg/min) and 0.05 mcg/kg/min (XAI, SD 0.09 mcg/kg/min) relative to the baseline arm ($p = 0.201$ for peers, $p = 0.010$ for AI, $p = 0.002$ for XAI, all by independent T -test). The aggregate prescription shifts are displayed in Fig. 1c. The individual patient scenario dosing shift figures for all 24 patients are shown in Supplementary Fig. 1.

Impact of arms on practice variation

Providing doctors with a recommendation (be it peer, AI or XAI) had a common effect: inter-clinician dose variability was differentially affected according to whether the recommendation was higher or lower than what subjects in the baseline arm did, i.e., when the recommendation was higher than baseline, the prescriptions of doctors in the peer/AI/XAI arms would be more variable across doctors; when it was lower than baseline, prescriptions were less variable across doctors. This can be seen in Fig. 1b.

Association of clinician factors with adherence to AI suggestions

Clinician attitude to AI was extracted as a principal component of the four pre-experiment AI enthusiasm questions subjects were asked (Fig. 2a). The first component explained 69% of the variance (Fig. 2b). Attitude to AI did not have a significant linear association to the difference between subject selected dose and AI recommended dose for either fluid ($r = -0.078$, $p = 0.075$ by LLSR (linear least-squares regression)) or vasopressor ($r = -0.074$, $p = 0.092$ by LLSR), see Fig. 2c. Similarly, years of clinical experience did not have a significant association to the difference between subject selected dose and AI recommended dose for either fluid ($r = 0.001$, $p = 0.862$ by LLSR) or vasopressor ($r = -0.086$, $p = 0.047$ by LLSR), see Fig. 3b, c. Practice variation and adherence to AI by grade of doctor are shown in Supplementary Fig. 2.

Clinician opinions on AI and the explanations

Post experiment, subject likelihood of using an AI system for sepsis prescriptions on a scale from 1 to 5 (higher more likely to use) was mean 2.55 for training doctors (which encompasses both junior and intermediate doctors, SD 0.96) versus 2.16 for non-training doctors (senior/consultants, SD 1.07), $p = 0.091$ by independent T -test (Fig. 4a). Subjects were asked to rate the usefulness of the explanations on a scale from 1 to 5 (higher more useful) with mean 2.22 for training doctors (SD 1.03) versus 1.97 for non-training doctors (SD 1.11), $p = 0.296$ by independent T -test (Fig. 4b). Self-reported usefulness of explanations did not correlate with adherence to XAI suggestions (Fig. 4d). Subjects were also asked to rate the usefulness of showing peer and AI suggestions together on a scale from 1 to 5 (higher more useful) with mean

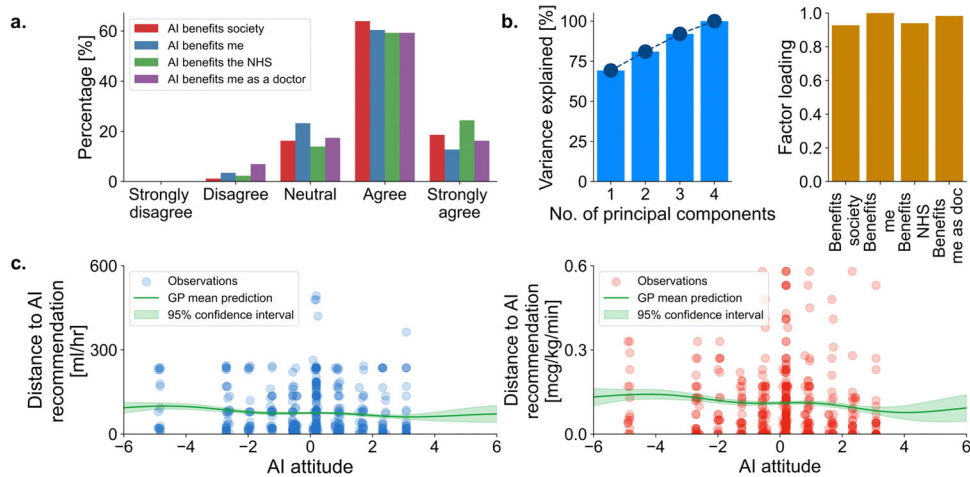


Fig. 2 Impact of AI attitude on adherence to AI suggestions. Four AI statements were presented to subjects pre-experiment who were asked for their agreement (a). Principal component analysis was applied to the results of these four questions with 69% of variance explained by a single component (b). This single component formed our composite for AI attitude (higher value, more positive AI attitude), which was then compared to absolute difference from the AI suggested dose, i.e., a proxy for AI adherence with lower value indicating greater adherence and vice versa in (c) for both fluid (blue) and vasopressor (red). Dot transparency in 3c represents density of points at any given location. ‘GP mean’ refers to a predicted Gaussian Process regression fit of the data with accompanying 95% confidence interval.

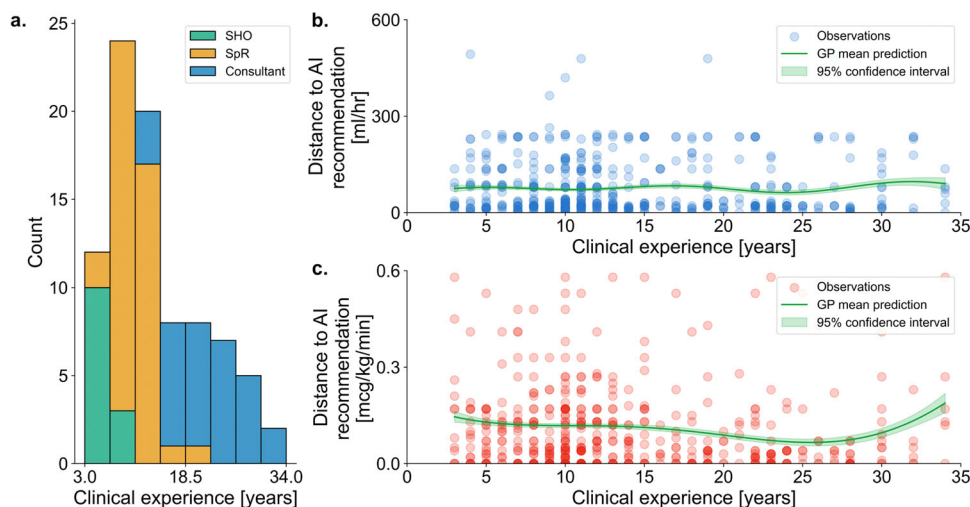


Fig. 3 Impact of duration of clinical experience on adherence to AI suggestions. The distribution of experience levels among the three categories of seniority is shown in (a) (Consultant, most senior and equivalent to attending in the United States (US); SpR, specialist registrar and equivalent to fellow in the US; SHO, senior house officer and equivalent to resident in the US). Experience level was compared to AI adherence (in the form of absolute difference from the AI suggested dose) for both fluid (b, blue) and vasopressor (c, red). Lower value indicates higher adherence. Dot transparency in (b) and (c) represents density of points at any given location. ‘GP mean’ refers to a predicted Gaussian Process regression fit of the data with accompanying 95% confidence interval.

2.98 for training doctors (SD 0.73) versus 2.39 for non-training doctors (SD 1.09), $p = 0.003$ by independent T -test (Fig. 4c). Finally, subjects were also asked to rate importance of evidence on a 1 to 5 scale (higher more important) for their use of an AI system with a mean rating of 3.01 (SD 0.85) for observational evidence versus mean 3.33 (SD 0.76) for randomised clinical trial evidence, $p = 0.011$ by independent T -test.

DISCUSSION

This study has several important findings that add to our understanding of how prescription decisions can be influenced by AI-driven decision support recommendations. First, additional information (peer, AI or XAI) has a strong influence (proxied by dose shift) on prescription decisions (significantly for AI, not so for

peers). However, whether the recommendation came in a plain form (AI alone) or garnished with an explanation (XAI, here simple feature importance) did not make a substantial difference. Second, inter-clinician dose variability was differentially affected according to whether the recommendation (whether peer, AI or XAI) was higher or lower than what subjects in the baseline arm did and this suggests that decision support systems might have a mixed impact on practice variation when deployed in a live setting. Third, there was no correlation between attitudes to AI or clinical experience on the AI-supported decisions suggesting a certain unvarying degree of AI acceptance in clinical experts, or one moderated more by variability in patient scenario than the clinician themselves. Fourth, there was no correlation between what doctors self-reported about how useful they found the XAI and whether the XAI actually influenced their decisions which

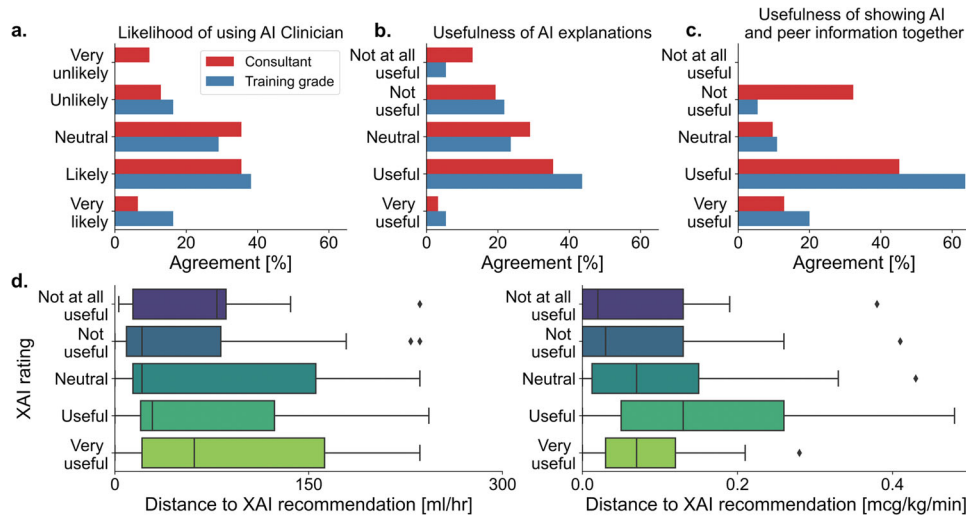


Fig. 4 Post-experiment questions. a–c Shows responses to post-experiment questions broken down by training status. Self-reported usefulness of explanations is also plotted against adherence to XAI recommendation for both fluid (d, left) and vasopressor (d, right). Lower values indicate greater adherence and vice versa. For each boxplot, the centre line represents the median, box edges represent upper and lower quartiles, whiskers represent 1.5× inter-quartile range and diamonds are outliers. Colour spectrum gradient over boxplots relates to how useful XAI was rated.

brings into question the reliability of using XAI self-reports as an outcome metric in clinical XAI studies.

These findings should be considered in the context of several limitations. First, the rendition of the XAI condition used feature importance. While this is a commonly used XAI modality (both within and outside of healthcare)^{12,19–23}, it is also on the lower end of the spectrum regarding what makes for a comprehensive explanation according to cognitive science²⁴. Indeed, the variability in modality of XAI can have different impacts on medical decisions depending on the complexity or ambiguity of the context as well as the user's experience of an AI system^{25,26}.

Second, while steps were taken to ensure a large and representative range of patient scenarios, this may still leave gaps in the patient state space. We intentionally developed a two-pronged selection strategy to ensure our scenario choices were as generalisable as possible (see Supplementary Methods 1 which suggests reasonable coverage of the MIMIC state space presented graphically as a function of three principal components for visual purposes). Third, the patient scenarios were low fidelity owing to the experimental vignette format. Whilst this allows for a high degree of standardisation of scenarios, the dynamic nature of evaluating a real patient over time and learning what effect a given treatment has (or does not have) is therefore missing. Fourth, although our sample of doctors was large compared to similar clinical studies, it nevertheless relied on convenience sampling and therefore may not be representative of the medical population as a whole. Fifth, the AI suggestion itself was an isolated message containing the recommended doses for fluid and vasopressor without any confidence bounds or ranges. It may be that adherence to AI suggestions would be higher when presented with estimates of certainty, when presented graphically or using other algorithmic approaches²⁷. Sixth, we do not have a detailed insight into the decision-making process that physicians internally performed. All we have observed in this study is self-reported data (which can be biased in many directions and from many causes) and final prescription data. Therefore, our conclusions are based on inferring that the input conditions (baseline, peer, AI, XAI) had a direct impact on the final prescription decisions via some internal cognitive effect on decision-making. Alternate methodologies to further explore this decision-making process might involve qualitative interviews and 'think-aloud'

studies in which the verbalisation of decision-making is encouraged and formally analysed at aggregate level using thematic analysis. As a result, our interpretation of the study findings and conclusions are based on a data-driven framework of values rather than a more cognitive theoretical framework about *how* exactly the decision-making is influenced.

Notwithstanding these limitations, reviewing our findings alongside existing literature provides important insights into how we can improve the design and deployment of AI-based medical decision support tools. Attempts to quantitatively evaluate AI recommendations in general (non-clinical) problems have demonstrated the sometimes counterintuitive nature of how explainability impacts on performance. In one experiment, the response shift (a marker for AI influence similar to our experiment) was greater when an explanation was provided⁶. However, the quality of the explanation (good vs. poor) did not affect this shift with the authors suggesting that subjects might have been reassured by the presence of an explanation when the AI performance itself was good rather than actually assessing explanation quality or fidelity⁶, a potential form of automation bias. In a clinical environment, the consequences of erroneously acting on poor advice can be considerable with a study among 50 general practitioners (GPs) demonstrating that only 10% were able to correctly disagree with incorrect AI advice on a dermatology problem²⁸ while in a radiology setting even task experts were not immune to the effects of poor AI advice (although they were considerably better at rejecting it than non-task experts)⁸.

Explanations have traditionally been posited as a means of rescuing users from poor AI advice though this has not clearly been borne out in clinical studies. For example, in a study investigating a psychiatric medication decision support tool, the presence of explanations did not provide rescue from intentionally poor AI recommendations suggesting a level of automation bias that could be problematic in a real-world clinical environment¹⁰. In our study, the explanation did not significantly increase adherence above and beyond the AI suggestion with several possible causes. It could be that trust and adherence was maximally achieved in most subjects by the AI suggestion alone leading to a ceiling effect. Or the very basic nature of the XAI was not persuasive. Or perhaps the nature of needing to make repetitive and cognitively burdensome decisions led users to quickly adopt a heuristic, one

way or the other, as to whether they used the explanation in their decision-making or not. Further still, there were several users who commented that the variables given by the AI as part of the feature importance explanation did not seem physiologically plausible (see Supplementary Note 1 for a selection of post-experiment subject comments). This poses a paradox to those designing XAI systems. On the one hand, some might argue that a more complex explanation could satisfy users and lead to higher adherence. However, the strength of AI in being able to identify patterns in large datasets that are imperceptible to human clinicians can also be a weakness with regards to developing XAI. If some clinicians associate the quality of an AI explanation with physiological plausibility, then is an AI explanation based on patterns that human clinicians don't usually see in their practice likely to be persuasive? Probably not. Ultimately, a mixed qualitative-quantitative approach might be critical for getting a better insight into the decision-making process that physicians use and where AI can be of most benefit and least harm during that process.

Other pertinent topics within medical AI-driven decision support research include the purported source of the advice as well as the experience level of the clinical audience receiving it. Gaube and colleagues studied both clinical experts and non-task experts for a chest radiograph diagnostic challenge with XAI suggestions. They found that experts rated their confidence as higher when advice was labelled as originating from an AI (although their accuracy was unchanged) while non-task experts had improved accuracy with the provision of explanations (compared to experts who did not)⁹. The peer suggestions in our study are not directly comparable as they were genuine rather than synthetic suggestions (and so differed in magnitude from the AI recommendations). However, we nonetheless also found that AI suggestions were more influencing than peers as a source of suggested advice. We did not, though, find any association with clinical experience level on adherence to AI.

Taken together, our findings on a comparatively large clinical expert population raise important questions for the meaning and design of medical XAI systems. Specifically, we show that the marginal impact of XAI was low in this experimental setting. The exact type, presentation and feedback loops for medical XAI systems that actually influence doctors remains unclear. It seems very likely that future research will need to more comprehensively consider social and cognitive aspects of decision-making alongside technical deployment of AI systems. We also cast doubt on the utility of self-reports as a valid metric for assessing XAI in clinical experts. Further work in this area could look to higher fidelity and more granular markers that assess the natural behaviour of clinicians when they interact with decision support tools. Answering these questions will be critical for bridging the translation gap between theoretical medical AI and real-world bedside implementation.

METHODS

Data source and AI clinical decision support system

The 'AI Clinician' is a reinforcement-learning based intensive care unit (ICU) clinical decision support system that provides semi-autonomous continuous dosing suggestions for intravenous (IV) fluid and vasopressors¹². The 'AI Clinician' was trained on the data of 17,083 ICU patients from the MIMIC-III database as previously described¹². MIMIC-III is an anonymised, open-access database of over 60,000 ICU admissions from 2001–2012 in six teaching hospital ICUs from Boston in the United States²⁹. Briefly, patients selected for training by the AI Clinician were adults with sepsis as defined by the sepsis-3 criteria³⁰. Each patient's data were split into 4-hour time blocks. For every 4-hour time block for each of the 17,083 patients, the AI Clinician clustered the patient into one of 750 states and produced a

suggested dose for intravenous fluid and noradrenaline (the most commonly used vasopressor agent in septic shock)³¹.

Twenty four patient scenarios were chosen for inclusion in the experiment. Twelve of these were 'expert selected' by trying to ensure representation from four broad categories: (i) three patients where both the fluid and vasopressor AI dose suggestions were similar to what human clinicians had done in MIMIC-III, (ii) three patients where only the AI vasopressor suggestion was similar to humans (iii) three patients where only the AI fluid suggestion was similar to humans and (iv) three patients where neither AI fluid nor vasopressor suggestions were similar to humans. These 12 patients also spanned scenarios where the patient was receiving anywhere from no vasopressor to a large dose (>0.5 mcg/kg/min of noradrenaline-equivalent), again to ensure a representative patient mix. The other 12 patients were chosen by clustering the entire MIMIC-III sepsis dataset of 17,083 patients into 12 clusters and then selecting a patient within the closest percentile to the cluster centroid. This resulted in 12 patients that were less sick (as defined by proportion on vasopressor support and APACHE score) than the initial 12 but that were more representative of the MIMIC-III septic cohort. The amount of fluid and vasopressor support is shown in Supplementary Methods 1, separated by whether the patient was 'expert-selected' or 'cluster-derived'.

Vignette experiment and conditions

We conducted an experimental human-AI interaction vignette study for doctors using a modified between-subjects design. There were four experimental arms. In every arm, subjects were provided with patient data in the form of a fixed variables table (e.g. age, gender, weight), an interactive graph displaying a limited set of time varying features and a second larger table showing all time varying features (see Supplementary Methods 2 for screenshots). This was designed to look similar to the way in which most ICU doctors in the UK encounter patient data on their respective electronic health records (EHRs).

Each subject (ICU doctor) performed 16 trials (Fig. 5a). The first four trials were identical for all subjects and were used as a pre-training period. The subsequent 12 trials comprised the main experiment. For each trial, subjects were asked to select a dose for fluid and a dose for vasopressor to be applied for the next hour. We used a multi-factorial experimental design with four arms, where each clinician experienced all four arms on different subsets of our 24 patients. The four arms were: baseline with no additional AI or peer human information (baseline); additional peer human clinical information (peer, see description below); additional AI decision support system information (AI); additional AI decision support system with explanation of the AI decision (feature importance, XAI). Examples of all four scenarios are available in the Supplementary Methods 2.

For the baseline scenario, subjects viewed only the patient data. For the peer human clinician scenario, subjects were also shown the probability density function of IV fluid and vasopressor doses prescribed by other doctors in the MIMIC-III dataset for patients in the same state. This can be thought of as a proxy for what peer clinicians have previously done for similar patients. This data was displayed as a violin plot (consisting of a conventional box plot with an overlaid distribution for the data derived via a kernel density estimation (KDE)). The rationale for including peer data as an experiment arm was to evaluate if clinicians merely want additional information or context to support their decision (regardless of source) or whether there is something specific about an AI suggestion that is more or less persuasive than simply knowing what their peers typically do.

For the AI scenario, subjects were also shown the AI Clinician suggested doses for fluid and vasopressor in text form. For the XAI scenario, subjects were shown the AI Clinician suggested doses as

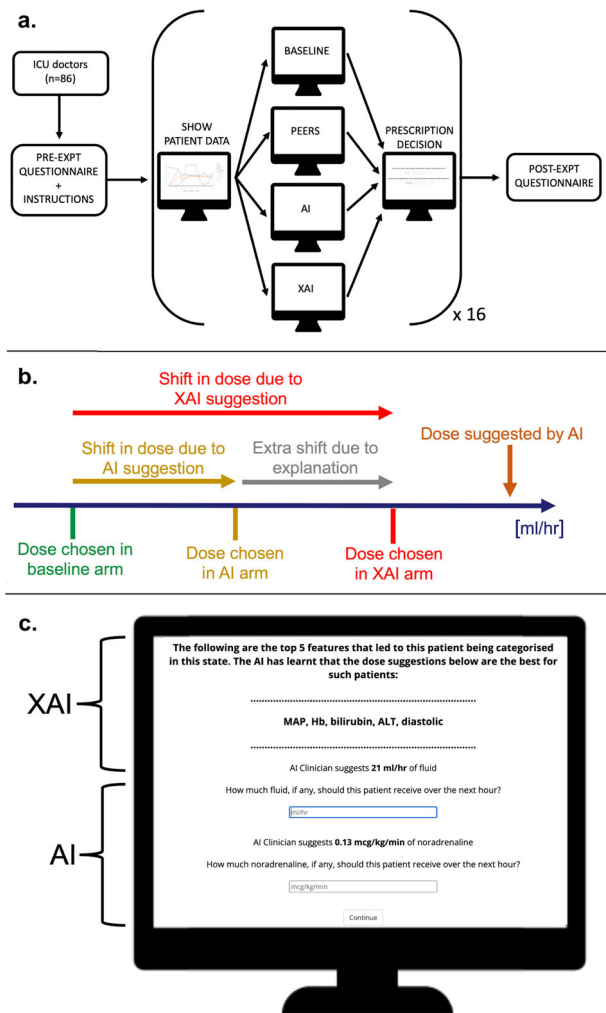


Fig. 5 Dose shift and experiment protocol. The experiment protocol is shown in (a). Dose shift relative to baseline occurring as a result of showing the AI suggestion is shown in (b). The extra shift between AI and XAI is the marginal shift attributable to the explanation. c Shows the AI and XAI components of the decision screen where participants input their prescription choice.

well as an explanation based on feature importance. The state space for the AI Clinician was constructed using a k-means clustering algorithm. After the algorithm converged, the cluster centroids represented the average feature values for patients in a particular state/cluster. A new patient would be assigned to the state/cluster that minimised the distance from their feature values to the respective cluster centroid. Intuitively, with over 40 features, some features will be closer to the cluster centroid value than others for any patient assigned to a given state. This is exploited to rank features in terms of their proximity to the cluster centroid (or average state feature values) given that the archetypal patient for whom an RL agent policy action most applies is a patient who is most typical of that state. Subjects were shown the top five ranked features contributing to state assignment (for details see Supplementary Methods 3). Although feature importance can be considered a basic form of XAI, it is nonetheless in widespread use within medical studies^{12,19–23}.

The trial design matrix (see Supplementary Methods 4) ensured that half the subjects saw a patient under one arm while the others encountered the same patient under a different arm, allowing estimation of between arm variability by controlling for the patient. Our primary measure of interest was the difference in

prescribed dose to the same patient across the four different arms—effectively measuring the shift in dose across arms as a measure of impact that the arm has on clinical decisions (Fig. 5b). The overall order of trials was varied to counterbalance any learning effects. Statistical analyses included two-sided *T*-tests for comparison of means (after confirming normality) and linear regression for assessing associations. Both were performed with no adjustment for multiple comparisons.

Subject recruitment and experiment conduct

The experiment was created as an interactive web page using HTML and JavaScript (jsPsych library) that could run locally on a laptop. Pre-cleaned data from MIMIC-III patients trained on by the AI Clinician were checked for consistency and then feature values were converted to standard clinical UK units.

Clinician demographics, experience and affinity to AI were collected using a questionnaire prior to completion of the main experiment (Fig. 5a). After the experiment, subjects further completed a short post-experiment questionnaire (see Supplementary Methods 5). Data collected for each patient scenario included: clinician's prescription doses for fluid and vasopressor per patient scenario as well as time taken per patient scenario.

A convenience sample of ICU doctors was recruited with the following inclusion criteria: (i) practising doctor, (ii) has worked for at least 4 months in an adult ICU, (iii) currently works in ICU or has worked in ICU within the last 6 months. Participants had the opportunity to participate remotely via Zoom or in person. Electronically recorded informed consent was obtained from all participants and each experiment lasted approximately 45 minutes. The study was approved by the Research Governance and Integrity Team (RGIT) at Imperial College London (ICREC reference 21IC7245). The institutional review board of the Massachusetts Institute of Technology (no. 0403000206) and Beth Israel Deaconess Medical Center (2001-P-001699/14) approved the use of MIMIC-III for research. Because our study made use of fully anonymised patient data, individual patient consent was not required.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The data (in CSV format) that support the findings of this study are available online at: <https://doi.org/10.6084/m9.figshare.23192624>.

CODE AVAILABILITY

The code (in the form of Jupyter notebooks) that support the findings of this study are available online at: <https://doi.org/10.6084/m9.figshare.23192624>.

Received: 23 May 2023; Accepted: 27 October 2023;
Published online: 07 November 2023

REFERENCES

- van de Sande, D., van Genderen, M. E., Huiskens, J., Gommers, D. & van Bommel, J. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. *Intensive Care Med.* **47**, 750–760 (2021).
- Olaye, I. M. & Seixas, A. A. The Gap Between AI and Bedside: Participatory Workshop on the Barriers to the Integration, Translation, and Adoption of Digital Health Care and AI Startup Technology Into Clinical Practice. *J. Med. Internet Res.* **25**, e32962 (2023).
- Academy of Medical Sciences and the Royal Academy of Engineering. Accelerating effective and safe adoption of artificial intelligence in the healthcare system. <https://acmedsci.ac.uk/file-download/92028281> (2023).

4. Mangold, C. Why EU regulators are pushing for more explainable AI. <https://www.fintechfutures.com/2021/05/why-eu-regulators-are-pushing-for-more-explainable-ai/> (2021).
5. Barredo Arrieta, A. et al. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*. **58**, 82–115 (2020).
6. Shafti, A., Derks, V., Kay, H. & Faisal, A. A. The response shift paradigm to quantify human trust in AI recommendations. Preprint at <https://arxiv.org/abs/2202.08979> (2022).
7. Benedikt, L., Andreas, H., Christina, H., Marc, S. & Martina, M. Explainable Artificial Intelligence improves human decision-making: Results from a mushroom picking experiment at a public art festival. Preprint at <https://osf.io/68emr/> (2022)
8. Gaube, S. et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digit Med*. **4**, 31 (2021).
9. Gaube, S. et al. Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. *Sci. Rep.* **13**, 1383 (2023).
10. Jacobs, M. et al. How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection. *Transl. Psychiatry* **11**, 108 (2021).
11. Rudd, K. E. et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *Lancet* **395**, 200–211 (2020).
12. Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C. & Faisal, A. A. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nat. Med.* **24**, 1716–1720 (2018).
13. Li, Y., Li, H. & Zhang, D. Timing of norepinephrine initiation in patients with septic shock: a systematic review and meta-analysis. *Crit. Care*. **24**, 488 (2020).
14. Marik, P. E., Linde-Zwirble, W. T., Bittner, E. A., Sahatjian, J. & Hansell, D. Fluid administration in severe sepsis and septic shock, patterns and outcomes: an analysis of a large national database. *Intensive Care Med.* **43**, 625–632 (2017).
15. Evans, L. et al. Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock 2021. *Crit. Care Med.* **49**, e1063–e1143 (2021).
16. Jussupow, E., Spohrer, K., Heinzl, A. & Gawlitza, J. Augmenting medical diagnosis decisions? an investigation into physicians' decision-making process with Artificial Intelligence. *Inf. Syst. Res.* **32**, 713–735 (2021).
17. Moussaïd, M., Kämmer, J. E., Analytis, P. P. & Neth, H. Social influence and the collective dynamics of opinion formation. *PLoS ONE* **8**, e78433 (2013).
18. Yaniv, I. Receiving other people's advice: influence and benefit. *Organ. Behav. Hum. Decis. Process.* **93**, 1–13 (2004).
19. Garriga, R. et al. Machine learning model to predict mental health crises from electronic health records. *Nat. Med.* **28**, 1240–1248 (2022).
20. Lam, J. Y. et al. A machine-learning algorithm for diagnosis of multisystem inflammatory syndrome in children and Kawasaki disease in the USA: a retrospective model development and validation study. *Lancet Digit Health* **4**, e717–e726 (2022).
21. Peine, A. et al. Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care. *npj Digit Med*. **4**, 32 (2021).
22. Laatifi, M. et al. Explanatory predictive model for COVID-19 severity risk employing machine learning, shapley addition, and LIME. *Sci. Rep.* **13**, 5481 (2023).
23. Deng, H. et al. Development and validation of an explainable machine learning model for major complications after cytoreductive surgery. *JAMA Netw. Open*. **5**, e2212930 (2022).
24. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **267**, 1–38 (2019).
25. Triberti, S., Torre, D. L. & Riva, G. The Artificial Face (ART-F) Project: Addressing the Problem of Interpretability, Interface, and Trust in Artificial Intelligence. *Cyberpsychology Behav. Soc. Netw.* **26**, 318–320 (2023).
26. Bienefeld, N. et al. Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals. *npj Digit Med*. **6**, 94 (2023).
27. Festor, P., Luise, G., Komorowski, M. & Faisal, A. A. Enabling risk-aware Reinforcement Learning for medical interventions through uncertainty decomposition. Preprint at <https://arxiv.org/abs/2109.07827> (2021).
28. Micocci, M. et al. Attitudes towards Trusting Artificial Intelligence Insights and Factors to Prevent the Passive Adherence of GPs: A Pilot Study. *J. Clin. Med.* **10**, 3101 (2021).
29. Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data*. **3**, 160035 (2016).
30. Singer, M. et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* **315**, 801–810 (2016).
31. Angus, D. C. & van der Poll, T. Severe sepsis and septic shock. *N. Engl. J. Med.* **369**, 840–851 (2013).

ACKNOWLEDGEMENTS

M.N. and P.F. are supported by the UKRI CDT in AI for Healthcare, <http://ai4health.io> (EP/S023283/1). A.G. is supported by an NIHR Research Professorship (RP-2015-06-018). A.F. is supported by a UKRI Turing AI Fellowship (EP/V025449/1). This work was also supported by the National Institute for Health Research (NIHR) Imperial Biomedical Research Centre (BRC). This work has been presented in poster form at the 2022 Conference on Cognitive Computational Neuroscience (San Francisco, USA).

AUTHOR CONTRIBUTIONS

M.N., A.G. and A.F. conceived the study. M.N. designed and built the experimental vignettes. M.N. recruited subjects. M.N. performed the initial data analysis. M.N., P.F., M.K., A.G. and A.F. contributed to subsequent interpretation of the data. M.N. drafted the initial version of the manuscript. M.N., P.F., M.K., A.G. and A.F. contributed to critical revision of the manuscript for important intellectual content, approved the final version, and take accountability for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work can be appropriately investigated and resolved.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41746-023-00955-z>.

Correspondence and requests for materials should be addressed to Aldo A. Faisal.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023