Quantifying the influence of an AI-based decision support system in the ICU: a high-fidelity simulation experiment

Introduction

Assuring the safety of AI-based systems in healthcare remains a challenge. In response to this, framework such as the Assurance of Machine Learning for Autonomous Systems (AMLAS) have been developed to provide guidance aiming to tackle this problem (Figure 1).

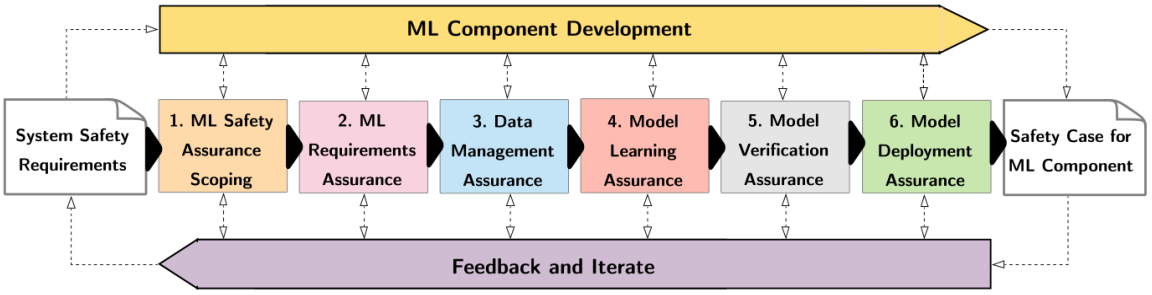

Figure 1: Overview of the AMLAS process.

The AMLAS framework includes a module for model verification assurance, prior to deployment in the target setting. Especially when dealing with high stakes applications such as healthcare, the model verification step of the assurance process can include the deployment of AI-based systems in simulated clinical settings, where the tools can be assessed without endangering human lives.

In a previous publication (Festor et al., BMJ HCI, 2022), we demonstrated that we could nudge the behaviour of an AI agent trained with reinforcement learning towards safer decisions using reward reshaping. Here, our objective was to study the integration of AI-based clinical decisions systems in the actual workflow of clinicians in the target environment, the intensive care unit (ICU). Specifically, we quantified the influence of disclosing AI recommendations to clinicians in a high-fidelity ICU simulation suite. We also recorded clinicians confidence and measured which factors were associated with the ability of clinicians to identify unsafe AI suggestions and to stop potential patient harm. This research adds to the growing literature on the safety assurance of autonomous and semi-autonomous AI-based systems in healthcare.

Methods

We conducted an observational human-AI interaction study in a high-fidelity ICU simulation suite, at the Clinical Skills laboratory of Imperial College London. Our primary objective was to quantify how AI influences ICU doctor prescriptions in terms of the variation in drug doses before versus after disclosing AI suggestions and whether this is affected by the AI suggestion being safe or unsafe. Secondary objectives included (i) assessing whether decisions vary according to subject-level factors (demographics, experience, affinity to AI as measured by a pre-experiment questionnaire) and the effect of peer advice on dealing with unsafe AI suggestions.

The main hypotheses tested in this study are the following:

- **H1:** Doctors stop significantly more unsafe than safe AI recommendations;
- **H2:** Unsafe AI recommendations increase requests to senior help;
- **H3:** The shift in administered treatments is larger with safe than unsafe recommendations.

We recruited ICU clinicians of any grade with at least 4 months of ICU experience into the study. The subjects were tasked with assessing six simulated ICU patients with sepsis as per Figure 2.

Each of the six patient scenarios could be encountered by a subject under one of three conditions: safe AI suggestion, unsafe AI suggestion, or peer-incentivised unsafe AI suggestion. The categorisation of suggestions as *safe* or *unsafe* was based on extreme over or underdosing of drug doses as per previous work (Festor et al., BMJ HCI 2022). The peer-incentivised suggestion involved the bedside nurse (an experimenter) trying up to three times to briefly persuade the subject to reverse their decision (i.e. if an unsafe suggestion was to be followed, the nurse would try to persuade the subject against it and if the unsafe suggestion was being rejected, the nurse would try to encourage the subject to follow the AI). The AI suggestions themselves were synthetic as the purpose of this experiment was to test interaction dynamics between humans and AI, rather than to test the safety of an actual system.

Within each of the six scenarios, they were tasked with conducting an assessment (to include a review of the available patient data and patient examination) before being asked by the bedside nurse for drug doses for the next hour of the patient's admission. They were also asked for their confidence on a 1 to 10 scale as well as whether they would like to request senior advice. Subjects were then shown the AI suggestion and asked to confirm or change their prescription doses and update their responses to the confidence and senior advice questions. They were further asked two AI-specific questions. First, their general agreement or disagreement with the AI suggestion on a 1 to 5 Likert scale. Second, whether they would override the AI suggestion were it to be automatically applied to the patient. The distinction between the two was intended to expose that a subject could disagree with an AI suggestion yet nonetheless regard it as not posing a large enough risk to warrant overriding.

The study was approved by the Imperial College Ethics Committee and the Health Research Authority (ref: 22/HRA/1610).

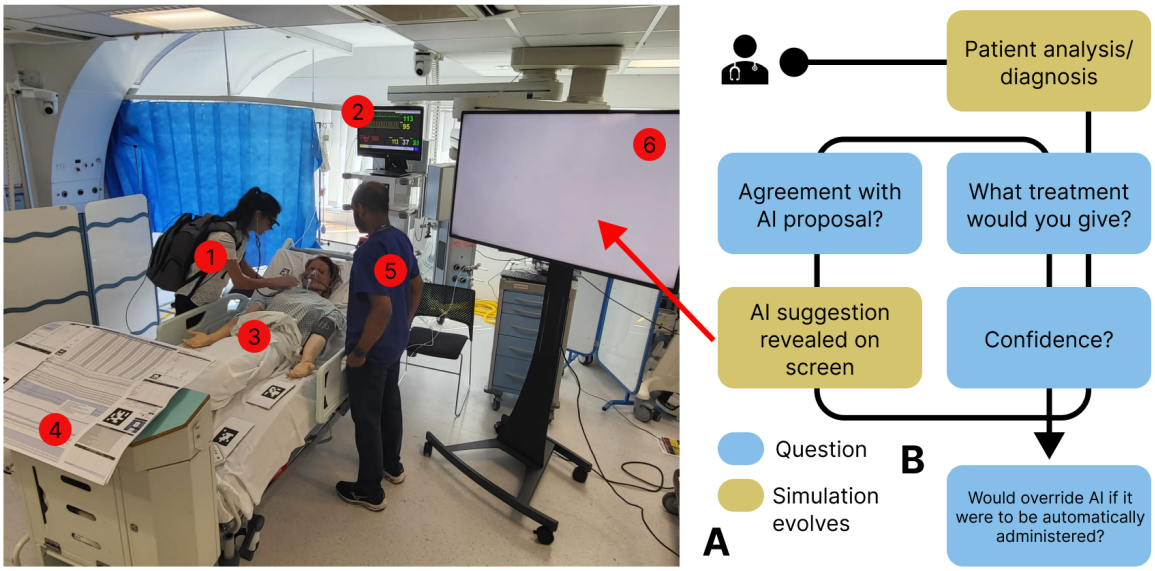

Figure 2: Experimental design - **A** Picture of the simulation suite with: (1) Subject (2) Bedside monitor (3) Patient mannequin (4) ICU bedside information chart (5) Bedside nurse (6) AI screen. **B** Experimental protocol diagram

Results

A total of 38 subjects took part in the experiment, with experience in intensive care ranging from less than a year to 28 years. Figure 3 presents a demographic summary of the recruited cohort.

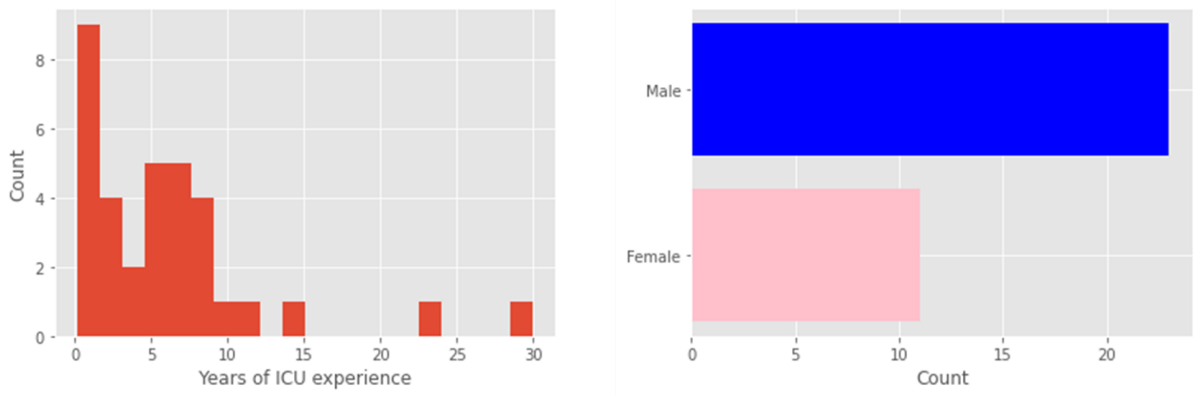

Figure 3: Simulation cohort overview - **A** Experience distribution. **B** Gender distribution.

Each subject reviewed a total of 6 simulated patients, the first one always with a safe AI recommendation and 2 of the remaining 5 were unsafe in a pseudo-randomized way. One of the rationales for developing an AI clinician for sepsis cardiovascular management is the uncertainty on the optimal doses of fluids and vasopressors to give to a specific patient. This uncertainty is visualised in figure 4 with the distributions of each drug given initially by subjects.

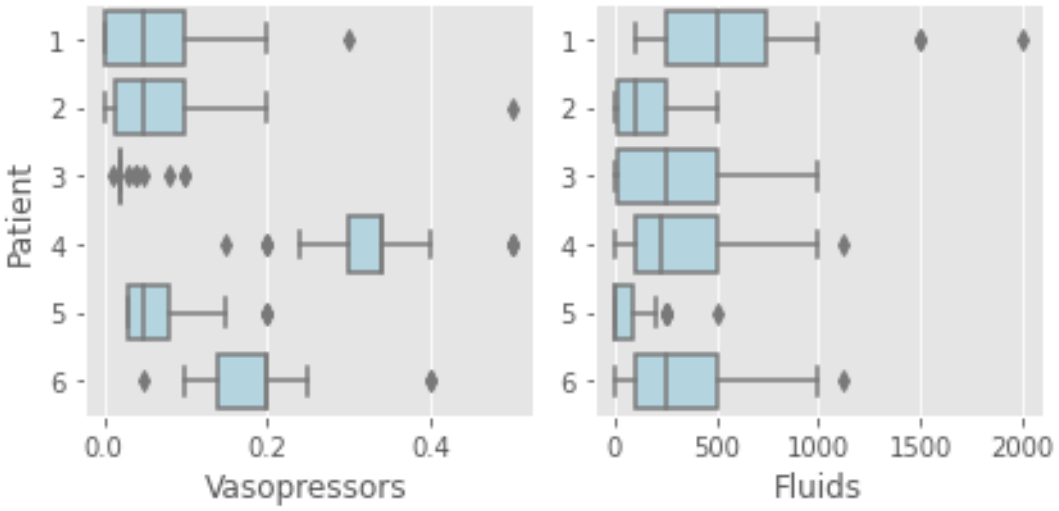

Figure 4: Variability in human treatment - Distribution of vasopressors (left) and fluids (right) doses given by clinicians after their initial examination of each of the six scenarios.

Figure 5 presents results relative to hypotheses H1 (subjects will stop more unsafe than safe recommendations) and H2 (unsafe recommendations will increase requests for senior help), alongside with comparison of confidence and agreement with the AI suggestion. On the whole cohort, 28.9% (± 3.7 %) of safe recommendations and 81.6% (± 4.48 %) of unsafe decisions were stopped. The latter increases to 92.1% (± 3.11 %) when adding subjects who requested senior advice on the case. Before seeing any recommendation, senior help was requested in 39.91 % (± 3.25 %) of trials compared to 36.18 % (± 3.91 %) and 56.58 % (± 5.72 %) after seeing a safe and an unsafe recommendation respectively. There was no tangible difference in reported confidence before and after seeing the recommendation, no matter its nature. According to the agreement distributions, clinicians were more positive towards the safe recommendations. Finally, leave one scenario out cross-validation shows the robustness of those conclusions against our choice of scenarios.

Looking at the influence of recommendations on treatment decisions, there was a statistically significant shift in dose distribution for vasopressors in 2 of the 6 scenarios. There were no such statistically significant results in fluid dose shift. However, when filtering the trials to keep only those where the subject did not request a second opinion unsafe recommendations did not have a significant influence anymore.

Finally, after the subjects had seen an unsafe recommendation, and one time only for each subject, the bedside nurse (played by an experimenter) tried to convince the subject to change their decision on the AI recommendation (do not override it if they would have or vice versa). Of the total 38 subjects, 2 were swayed by the challenge of the nurse and decided to not override a recommendation that they originally thought harmful: one subject changed their mind after the argument that the AI had seen more cases than them, the other one after being challenged on the legal implications for the medical team to act against the AI (“What if the family sues us and the discover we haven’t followed the AI?”).

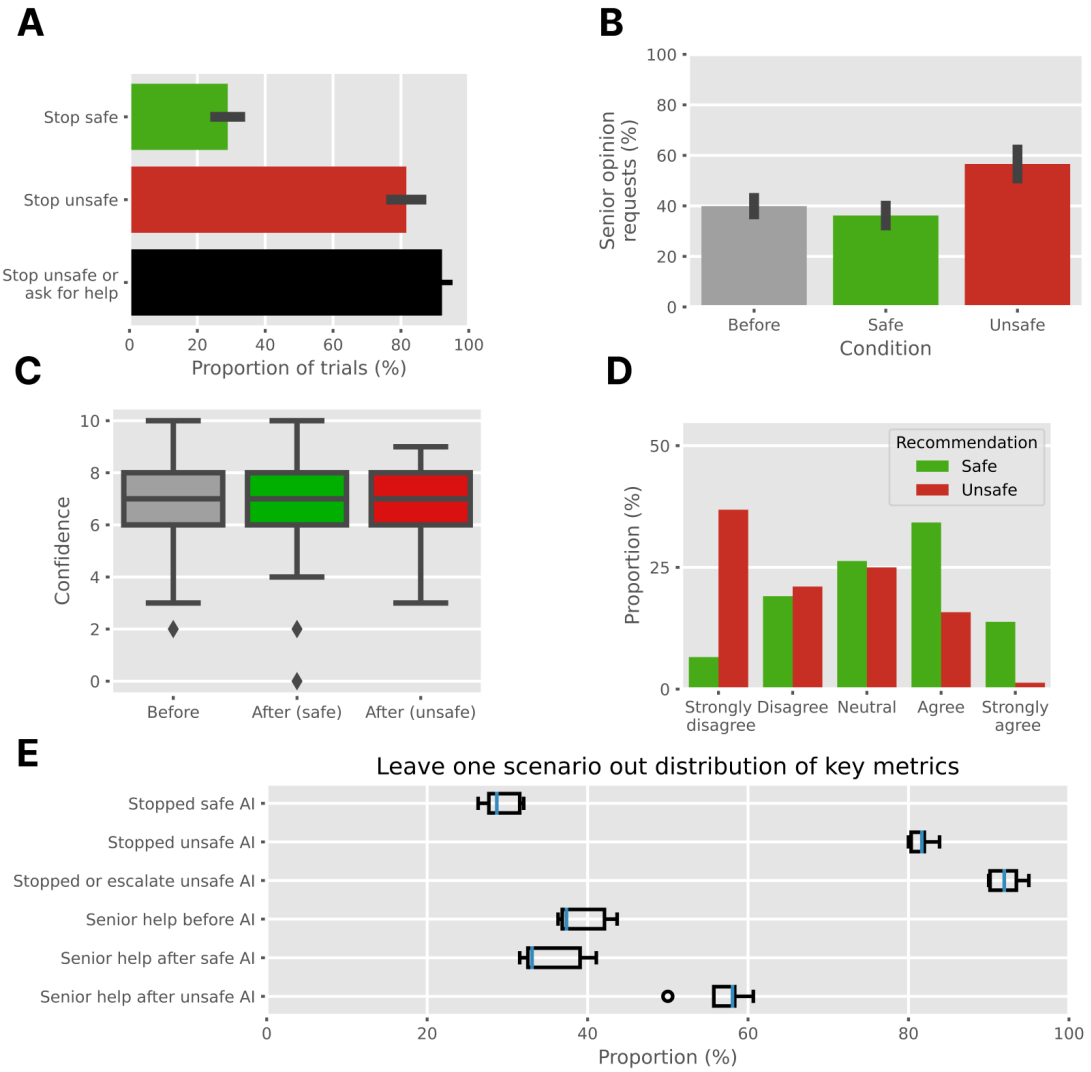

Figure 5: Comparison of clinicians' responses in safe and unsafe scenarios - **A** Willingness to interrupt the AI if it zas automatically administered to the patient. **B** Requests for senior help before and after seeing a safe or unsafe recommendation. **C** Reported confidence before and after seeing a safe or unsafe recommendation. **D** Distribution of agreement levels for safe and unsafe treatment recommendations. **E** Distribution of the 6 key experiment metrics on leave one scenario out cross-validation.

Discussion

This simulation experiment allowed us to verify that unreasonable AI recommendations would be stopped most of the time by the clinician in charge (around 82% of the time). Moreover, unsafe treatment recommendations increased the number of requests for senior opinion by around 16% in absolute value. Assuming that an unsafe recommendation discussed by a junior and a senior would eventually be stopped, the likelihood that hazardous suggestions would have been overridden increased to 92%.

Our work on retrospective data showed that situations prone to under or overdosing only represent 16% of the decisions (Festor et al., BMJ HCI 2022), that the AI recommendation is

hazardous in only 12% of those and that the clinical team only misses those 8% of the time, patients will get safe treatment in 99.8% of cases.

Further, attempts from the bedside nurse to challenge subjects on their decision to stop or not the AI were successful in only 2/38 subjects who have been convinced to not override an AI recommendation they otherwise would have. Those subjects changed their minds after challenges on the number of patients seen by the AI being way larger than theirs, or a question on the legal consequences of not following the system. Most subject defences orbited around 3 major axes: the mismatch with their biological understanding of sepsis, the lack of prospective evaluation for this clinical decision support tool, and the fact that they had never seen similar strategies being used before.

Therefore, to further improve the safety of the decision-making process, users should be briefed on the importance of biological reasoning when evaluating a recommendation, the extent to which the system has been validated and the legal framework around interacting with such a tool. Educating every user on how the tool works from an AI perspective would of course be ideal but it is also highly unrealistic. Providing a small introduction to the topics above should already take safety a long way.

This experiment however underlined yet again the challenge for clinicians of manipulating VP and IVF doses directly. In fact, while figure 2 shows clear variability in original treatment doses across subjects, ICU clinicians are not usually prescribing these doses directly, but rather give a Mean Arterial Pressure (MAP) target for the nurse to aim at. While these two tasks share similarities (aiming at a higher MAP is linked to increasing VPs), and doctors should be familiar with noradrenaline dose orders of magnitude, it would not be fair to say that clinicians usually directly give drug doses. Another important limitation of this study is the lack of interaction and dynamic response to treatments. In a real clinical context, clinicians titrate doses and adjust their behaviour to patient response, which was not possible here due to simulation limitations. On a similar note, the difficulty some subjects have shown in interpreting total fluid doses over the next hour could lead our team to modify the way information is presented to clinicians in later versions of the system to ensure proper communication of the quantities at play.

Conclusion

To evaluate the safety of the human-CDSS pair, we ran a simulation study of clinicians interacting with a treatment recommendation AI presented to them as retrospectively validated in the United States and other countries but not prospectively validated yet. We defined a set of six septic patient scenarios and tested how receiving AI recommendations would sway clinicians decisions. We demonstrated that 92% of unsafe treatment recommendations would be stopped by the clinical team.

This work brings the final stone to our work on safety analysis of AI for clinical decision support in the intensive care unit. After studying and improving the behaviour of the AI Clinician on retrospective data, we closed the loop and assessed the safety of the system in its context, or rather as close to it as possible in this simulation experiment. This work illustrates the pipeline required to evaluate the safety of a clinical decision support tool against pre-identified hazards. The work done on under and overdosing prevention for this

agent can serve as an example to quantify and prevent the transition to hazardous areas. This work also demonstrates the value of educating clinical teams about novel technology, including their validation, and ethical and legal implications of their utilisation in clinical practice.

References

Festor, P., Jia, Y., Gordon, A.C., Faisal, A.A., Habli, I. and Komorowski, M., 2022. Assuring the safety of AI-based clinical decision support systems: a case study of the AI Clinician for sepsis treatment. *BMJ health & care informatics*, 2022

Festor P, I Habli, Y Jia, AC Gordon, AA Faisal, M Komorowski, “Levels of Autonomy & Safety Assurance for AI-based Clinical Decision Systems”, 4th International Workshop on Artificial Intelligence Safety Engineering (WAISE). Proceedings in : International Conference on Computer Safety, Reliability, and Security, 2021, pp 291-296, June 2021

Jia Y, T Lawton, J Burden, J McDermid, I Habli. Safety-driven design of machine learning for sepsis treatment. Journal of Biomedical Informatics 117, 103762

Komorowski M, LA Celi, O. Badawi, AC Gordon, A Faisal, “The intensive care AI clinician learns optimal treatment strategies for sepsis.”, Nature Medicine, 24, 1716–1720, Nov 2018

McDermid JA, Y Jia, I Habli. Towards a framework for safety assurance of autonomous systems. Artificial Intelligence Safety 2019, 1-7, 2019.