Impact of XAI dose suggestions on the prescriptions of ICU doctors

Myura Nagendran[1,2,3], Anthony C. Gordon[2] & A. Aldo Faisal[1,3,4]

[1]UKRI Centre for Doctoral Training in AI for Healthcare, Imperial College London
[2]Division of Anaesthetics, Pain Medicine & Intensive Care, Imperial College London
[3]Brain & Behaviour Lab, Imperial College London
[4]Institute of Artificial & Human Intelligence, University of Bayreuth

Background.
The 'AI Clinician' *(Komorowski et al., 2018, Nat Med.)* is a reinforcement learning based intensive care unit (ICU) decision support system. It aims to provide semi-autonomous continuous dosing recommendations for intravenous (IV) fluids and vasopressors to treat patients with sepsis. Our AI Clinician has entered prospective evaluation in 4 UK ICUs and as part of this deployment critical questions arise 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 and psychology as one of machine learning. However, a systematic quantitative evaluation of how AI recommendations influence human decision makers has only recently begun in general problems *(Shafti et al., 2022, arXiv)*. Here we evaluate these in clinical experts to understand (i) how much an AI can influence an ICU doctor's prescribing decision, (ii) how much knowing the distribution of peer actions influences the doctor and (iii) whether or how much an AI explanation (here simple feature importance) influences the doctor's decision.

Methods.
We conducted an experimental human-AI interaction study with ICU doctors using a modified between-subjects design (N=85; 31 senior, 41 intermediate & 13 junior) with median 11 yrs clinical experience (IQR 9-19). We collected clinician demographics, experience and affinity to AI using a questionnaire. Then on a computer, doctors were presented for each of 16 trials with a patient case, potential additional information depending on the experiment arm, and then prompted to prescribe continuous values for IV fluid and vasopressor (to be applied to this patient for the next hour). Experiments lasted about 45 minutes each. We used a multi-factorial experimental design with 4 arms, where each clinician experienced all 4 arms on different subsets of our 16 patients. The 4 arms were (B1) baseline with no AI or peer human information; (B2) peer human clinician scenario showing the probability density function of IV fluid and vasopressor doses prescribed by other doctors (B3) decision support system scenario wherein the AI's suggestion was shown; (B4) explainable AI (XAI) scenario, as in B3 but also showing a list of the top 5 ranked features motivating this recommendation. The trial design matrix ensured half the subjects saw a patient under one arm while the others encountered the same patient under a different arm, with the overall order of patients varied to counterbalance any learning effects. The study was approved by the local ethics review board (ICREC Ref 21IC7245).

Results.
All subjects completed the task successfully and there was no significant difference in completion time for the different arms. Our primary measure was the difference in prescribed dose to the same patient across the 4 different arms - effectively measuring the shift in dose across arms as a measure of impact that the arm has on clinical decisions. For the same patients, providing clinicians with peer information (B2) did not lead to an overall significant prescription difference compared to baseline (B1). In contrast, providing AI/XAI (B3/B4) information led to significant prescription differences compared to baseline for IV fluid, but not for vasopressor. Importantly, the XAI condition (B4) did not lead to a larger shift towards the AI's recommendation than the AI condition (B3). 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 baseline doctors 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. Clinician demographics, experience and affinity to AI did not significantly impact their actions.

Discussion.
This study suggests that ICU clinicians are influenceable by dose recommendations. Knowing what peers had done had no significant overall impact on clinical decisions while knowing that the recommendation came from AI did make a measurable impact. However, whether the recommendation came in a "naked" form or garnished with an explanation (here simple feature importance) did not make a substantial difference - these findings are consistent with our studies in the non-expert population with general tasks. Among the correlations with demographics and views of our study population, the lack of impact of AI affinity and clinical experience on actions taken was noteworthy and suggests a certain unvarying degree of AI acceptance in clinical experts. In summary, our findings on a comparatively large clinical expert population (albeit with a small number of exemplar patients) raise important questions for the meaning and design of XAI systems in healthcare and the differential impact that recommendations may have on practice variation in healthcare.