Assuring the safety of AI-based clinical decision support systems: a case study of the AI Clinician for sepsis treatment

Paul Festor 1,2, Yan Jia 3,4, Anthony C Gordon 5, A Aldo Faisal 2,6, Ibrahim Habli 3, Matthieu Komorowski 1,5

ABSTRACT

Objectives Establishing confidence in the safety of Artificial Intelligence (AI)-based clinical decision support systems is important prior to clinical deployment and regulatory approval for systems with increasing autonomy. Here, we undertook safety assurance of the AI Clinician, a previously published reinforcement learning-based treatment recommendation system for sepsis.

Methods As part of the safety assurance, we defined four clinical hazards in sepsis resuscitation based on clinical expert opinion and the existing literature. We then identified a set of unsafe scenarios, intended to limit the action space of the AI agent with the goal of reducing the likelihood of hazardous decisions.

Results Using a subset of the Medical Information Mart for Intensive Care (MIMIC-III) database, we demonstrated that our previously published ‘AI clinician’ recommended fewer hazardous decisions than human clinicians in three out of our four predefined clinical scenarios, while the difference was not statistically significant in the fourth scenario. Then, we modified the reward function to satisfy our safety constraints and trained a new AI Clinician agent. The retrained model shows enhanced safety, without negatively impacting model performance.

Discussion While some contextual patient information absent from the data may have pushed human clinicians to take hazardous actions, the data were curated to limit the impact of this confounder.

Conclusion These advances provide a use case for the systematic safety assurance of AI-based clinical systems towards the generation of explicit safety evidence, which could be replicated for other AI applications or other clinical contexts, and inform medical device regulatory bodies.

INTRODUCTION

Several recent publications have shed light on the pressing issue of the safety of AI-based clinical decision systems and digital technologies, for example, with a trial of an acute kidney injury alerting system showing possible harm in some contexts. Safety assurance should not be seen as a post hoc bolt-on activity. Instead, best practices from safety-critical systems engineering should be woven into the design of AI systems and should proactively lead to the generation of safety evidence for the use of the tool in its intended clinical pathway.

These safety engineering concepts have been incorporated into safety assessment methodologies, such as Assurance of Machine Learning in Autonomous Systems (AMLAS). AMLAS takes a whole system approach to safety assurance. It aims to establish traceable links between the system-level hazards, risks and the safety requirements that have to be satisfied by the machine learning components. It also complements current initiatives and studies that focus on the human and organisational aspects of clinical risk management. See online supplemental appendix A for more detail on AMLAS. AMLAS is used here for its modular and iterative approach to safety assessment of a product over its whole...
lifecycle. The flexibility granted by these properties is essential in a complex context such as healthcare since safety considerations are only meaningful once scoped within the wider clinical setting.

In this work, we applied the principles of the AMLAS methodology to a previously published AI model built for informing the treatment of sepsis (severe infections with organ failure) as presented in figure 1.6

Sepsis is a common cause of hospital and intensive care unit (ICU) admission, morbidity and mortality and a major source of healthcare expenditure.7 A cornerstone of sepsis management includes the administration of intravenous fluids and/or vasopressors to restore a normal circulating blood volume and prevent further organ dysfunction. However, determining the correct dose and timing of these interventions is highly challenging for human doctors.8,9

In previous research, we developed the AI Clinician, a clinical decision support algorithm based on Reinforcement Learning (RL), capable of suggesting a dosing strategy for these two types of drugs over time and for a given individual patient.6 While we had generated some (retrospective) evidence of the model’s effectiveness, we had so far limited assessment of its safety.2

In this work, we applied AMLAS to the AI Clinician. In particular, we identified a set of clinical hazards and potential unsafe scenarios and assessed both retrospective human clinician recorded behaviour and AI agent behaviour against these scenarios on a subset of the MIMIC-III database.10 This created the basis for concrete safety requirements, in the form of constraints, for the AI Clinician. Then, we fed back the output of this analysis into model design and tested whether these safety requirements could be satisfied by the RL agent, as well as the impact that these additional safety requirements would have on AI model performance.

METHODS

AMLAS requires the definition and assurance of safety requirements. For autonomous or semiautonomous systems, these requirements may include a set of safety constraints, intended to limit the action space of an agent with the goal of reducing the likelihood of hazardous decisions. These constraints may trigger an inhibiting action (to prevent the transition from a safe to an unsafe state) or a correction (to return a system into a safe state).11

However, defining safety constraints in a clinical context such as sepsis, where there is no expert consensus and no high-performance simulation environment (where safety limits could be explored without putting patients at risk), is highly challenging. As a consequence, we used expert opinion and the literature to define a set of four undesirable clinical scenarios. Given that the action space of the AI includes fluids and vasopressors, we selected scenarios representing possible under or overdosing of these two drugs. For more information on the scenario definition process, see online supplemental appendix A.

To study the difference in the proportion of human and AI mistakes, we model them as Bernoulli random variables with hidden parameters $p_{\text{human}}$ and $p_{\text{AI}}$, respectively. For a given scenario, among the subset of $N$ patients at risk, we observe $x_{\text{human}}$ (resp. $x_{\text{AI}}$) human (resp. AI) mistakes and use a $z$-test to test for the null hypothesis on the underlying Bernouilli distribution parameters: $H_0 : p_{\text{human}} = p_{\text{AI}}$. The test statistic is given by:

$$z = \frac{\hat{p}_{\text{AI}} - \hat{p}_{\text{human}}}{\sqrt{\frac{1}{N} + \frac{1}{N}}}$$

where $\hat{p}_{\text{AI}} = \frac{x_{\text{AI}}}{N}$, $\hat{p}_{\text{human}} = \frac{x_{\text{human}}}{N}$ and $\hat{p} = (\hat{p}_{\text{human}} + \hat{p}_{\text{AI}})/2$. According to the law of large numbers, when $N$ is large and $H_0$ is true, then $z \sim N(0,1)$. Thus, $p$ values are computed using the standard Gaussian cumulative distribution function.

Next, we analysed which patient features were associated with human clinician unsafe decisions. We trained gradient boosting models to predict whether clinicians would take an unsafe decision given the set of patient features as input. Separate models were trained for all four scenarios. We then reported the relative SHAP importance of each feature from the fitted gradient boosting model and proposed hypotheses for the significance of the most important parameters (see online supplemental appendix A for more detail).

Finally, the results of this initial safety analysis were used to refine the AI Clinician algorithm. In RL, the optimal decisions are identified as the set of actions that maximises the sum of future expected rewards.13 In the initial model, the reward is based on survival at 90 days.
following ICU admission (positive reward if the patient survived, negative reward for death). We modified the reward function of the model by systematically penalising instances where harmful decisions were taken by clinicians in the training dataset. Specifically, we added a certain amount of penalty to any decision that satisfied our predefined unsafe scenarios, so the final reward function includes both intermediate and terminal signals (see online supplemental appendix A for more detail).

We retrained the AI Clinician with this new reward function using Q-learning, a well-established model-free off-policy RL algorithm where an optimal policy is learnt from analysing trajectories of previously recorded was generated by suboptimal agents (in this case, human clinicians). We compare the proportion of unsafe decisions in each scenario for three separate agents: human clinicians (in the training data), the original AI Clinician model and the modified ‘safe’ AI Clinician. We also estimated the value of the new modified policy, using off-policy policy evaluation, and compared it with the clinicians’ policy and the original AI Clinician policy. We used bootstrapping with 2000 resamplings to generate confidence bounds on the policy value.15

### RESULTS

**Output from AMLAS: definition of the four clinical scenarios**

The four clinical scenarios are outlined in table 1. As detailed in Methods and online supplemental appendix A, these represent clinical situations where one or both of the drugs of interest were likely administered either insufficiently (underdosing) or excessively (overdosing). For more detail on the scenarios, see online supplemental appendix A. The subset of MIMIC-III used in this study was extracted with the same process as in ref 6 (see online supplemental appendix A for more detail).

**Assessment of the AI Clinician 1.0 against the four scenarios**

We studied how frequently the AI and human clinicians may contribute to one of the four hazardous clinical scenarios (figure 2). Given the lack of a clear cut-off for low and high blood pressure, the analyses were conducted on a range of thresholds. The AI consistently leads to fewer unsafe decisions in all scenarios (p<0.05), except for scenario C, where the difference was not statistically significant.

**Analysis of patient features associated with unsafe decisions**

Figure 3 shows the result of the relative feature importance analysis, highlighting which patient characteristics were associated with ‘unsafe’ clinician behaviour, as defined in

### Table 1 Description and rationale for the four chosen clinical scenarios

<table>
<thead>
<tr>
<th>Hazardous clinical scenario</th>
<th>Clinical safety impact</th>
<th>Prevalence in MIMIC-III dataset</th>
<th>Safety-driven refinement of RL model</th>
<th>Updated safety evidence</th>
<th>Caveats or uncertainties</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: giving no vasopressors and low or no fluids (≤20 mL/hour) to a patient with low BP</td>
<td>Sustained untreated hypotension leading to organ failure and death.</td>
<td>MAP &lt;55: 29,089/984,269 (2.9%). Clinician’s action: 15,630/29,089 (53.7%).</td>
<td>Add 30 points of intermediate penalty if the condition is met.</td>
<td>The modified ‘safe’ policy had lower rate of unsafe behaviour than original AI policy, in three scenarios and the difference was not significant in the fourth (see figure 4).</td>
<td>No clear threshold for defining hypotension.</td>
</tr>
<tr>
<td>B: giving the maximum vasopressors dose (&gt;0.65 µg/kg/min) to a patient with high BP</td>
<td>Excessive blood pressure leading to increased risk of organ failure, bleeding and stroke.</td>
<td>MAP &gt;95: 118,869/984,269 (12.1%). Clinician’s action: 2986/118,869 (2.5%).</td>
<td></td>
<td>No clear threshold for defining hypertension. Some patients may have a clinical indication for high BP targets (eg, TBI).</td>
<td></td>
</tr>
<tr>
<td>C: giving no fluids to a patient with low BP and low CVP</td>
<td>Hypotensive and likely hypovolaemic patient left untreated.</td>
<td>MAP≤55 and CVP≤5: 661/984,269 (0.6%). Clinicians action: 356/661 (53.8%).</td>
<td></td>
<td>Measuring the fluid volume status is very difficult. CVP is a poor proxy but the closest approximate we have available in the data. No clear threshold of CVP for defining hypovolaemia or hypervolaemia.</td>
<td></td>
</tr>
<tr>
<td>D: giving the maximum dose of fluids (&gt;240 mL/hour) to a patient with normal BP, high cumulative fluid balance and high CVP</td>
<td>Giving excessive fluids to a septic patient who is unlikely to be hypovolaemic is harmful, leading to fluid accumulation, known risk factor for organ failure and poor outcomes.</td>
<td>MAP≥75 and cumulative balance &gt;10 L and CVP ≥15: 9409/984,269 (1%). Clinicians action: 3517/9409 (37.4%).</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

BP, blood pressure; CVP, central venous pressure (expressed in mm Hg); MAP, mean arterial pressure (expressed in mm Hg); TBI, traumatic brain injury.
Open access

Low vital sign and no medication

High vital sign and high medication

Mean blood pressure and vasopressors

CVP and IV fluids

Figure 2  Visualisation of the differences between the proportion of human and AI unsafe decisions, and statistical significance. Each subplot corresponds to one scenario. For each scenario, tests were run across a range of blood pressure thresholds. Within each subplot, the top plot shows the variation of the number of human and AI unsafe decisions for a range of BP thresholds and the bottom plot shows the statistical significance of this difference. The AI consistently leads to a lower number of unsafe decisions in all scenarios, except for scenario C where the difference was not statistically significant. Scenarios C and D reflect our previous study showing that the AI Clinician is more conservative in terms of fluid doses.

this study. Some hypotheses can be offered. In scenario A (low BP and no treatment initiated), a lower Sequential Organ Failure Assessment (SOFA), low cumulative fluid balance, low total fluid input, no sedation and low lactate were all associated with clinicians’ decisions labelled as unsafe. It is possible that clinicians decided to tolerate a low BP in patients who were relatively well otherwise. In scenario B (high BP and high vasopressors), a higher urine output and higher SOFA score were associated with unsafe behaviour. The high urine output could have been a consequence of an excessive blood pressure. Sicker patients (high SOFA) may have initially had a high vasopressor requirement and may have been left on high doses of vasopressors by mistake. Scenario C represents a subset of scenario A, and indeed we saw a similar pattern, where unsafe behaviour was observed in less sick patients (those with a lower SOFA). Also, the analysis highlights patients likely to be hypovolaemic (and left untreated): those with a low cumulative fluid balance and low urine output. In scenario D, a higher cumulative balance was associated with unsafe behaviour, which would be expected as the total fluid balance is correlated with previous high fluid intake. It is also interesting to note that a low urine output was associated with the unsafe behaviour of administering large volumes of intravenous fluids, which is a common decision in patients with oliguria, even though it may be harmful and fail to improve renal perfusion.

Model retraining with additional safety constraints

The AI Clinician model was retrained with added penalties to any decision that satisfied our predefined unsafe scenarios. The four scenarios were not encountered commonly in the dataset, except scenario B (elevated blood pressure): 12.1% of the decision points corresponded to a MAP over 95 mm Hg, of which only 2.5% were labelled as ‘unsafe’ behaviour as per our definition. By trial and error, we set an additional penalty of 30 points for each of the predefined unsafe instances...
which was necessary and sufficient to alter the AI policy

Figure 4A shows that the original AI Clinician had a lower proportion of unsafe behaviour than human clinicians in all four scenarios, while the modified ‘safe’ policy did better than the original AI Clinician.

Overall, human clinicians took any of the four unsafe decisions in about 2.2% of the data points in the training dataset (21,489 out of 984,269 data points). In comparison, our unaltered AI Clinician selected these decisions in 20,079 instances (a 6% relative reduction when compared with human clinicians), whereas the modified version recommended these in 18,929 instances (1.9% of the training dataset), which represents a 12% relative reduction from the clinicians’ strategy.

The off-policy policy evaluation (figure 4B) indicated that the value of the modified policy, with the added safety constraints, was only slightly lower than the original learnt policy (median, IQR): 90 (89.2–90) for the modified policy versus 99.5 (99.5–99.5) for the original policy, both much higher than the clinicians’ policy: 0 (−1.5 to 0.6). It should be kept in mind that the off-policy policy value estimation depends directly on the reward function used in the problem formulation. As such, given that the only difference of reward function between the original and adapted AI Clinician environments is the addition of penalties for unsafe behaviour, it is expected that the estimated value of the adapted AI Clinician policy is lower than that of the original policy. However, the value of the adapted policy, despite a harsher reward function, remains significantly higher than the value of the human policy in a non-penalised world.

Next, we compared the distribution of the model 25 actions for the initial (figure 4C) and modified (figure 4D) AI Clinician’s policies. The modified policy recommended more low-dose vasopressors, possibly in an effort to try and correct instances of hypotension left untreated (scenarios A and C).

**DISCUSSION**

Systematic safety assessment of AI-based clinical decision support systems is poorly codified, especially in applications where the definition of effective and safe decisions is challenging. In this study, we applied best practice in safety assurance to a complex AI system and proposed a safety-driven approach to identify regions of the action space potentially associated with preventable harm. We showed that the AI Clinician had desirable behaviour in a set of four scenarios and that we could further
iteratively improve the safety of the model by adapting the reward signal without significantly compromising its performance.

To our knowledge, this work is the first successful attempt at defining and testing safety requirements for an RL-based clinical decision support system considering multiple clinical hazards and at modifying the reward function of such an agent with added safety constraints. Despite the lack of consensus on a gold standard in sepsis resuscitation, there are decisions that are ‘obviously’ dangerous, such as those we defined in this work. Given the potential harm caused by these decisions, the model will have to be explicitly taught to avoid them where possible. This research represents one concrete step in this direction, and we demonstrated that our modified AI was 12% less likely than human clinicians to suggest those decisions.

Regulators recognise that there is a need for better guidance on safety assurance of AI/machine learning-based systems, where this work could potentially help. The US Food and Drugs Administration has proposed the Total Product Life Cycle (TPLC) framework for assuring such systems. Several relevant publications provide guidance on how to systematically integrate safety concepts from the onset of system development, which could satisfy some of the key requirements of the TPLC, for example, the premarket safety assurance.

The approach described here is necessary but not sufficient by itself. The AI Clinician V.1.0 was designed as a proof-of-concept system, not meant to be used as-is in the real world. Similarly, the research presented here illustrates how RL models can be augmented with safety constraints, without substantially impairing the value of the AI policy. Thus, the commonly perceived trade-off between performance and safety is not really apparent here. If safety constraints are integrated into the AI learning process, as we show here, it is possible to enhance safety while maintaining performance. However, more in-depth technical research is needed to robustly define and assess the best way to perform reward reshaping in the context of safety assurance.

Here, we did not assess the outcomes associated with taking our custom defined safe or unsafe decisions because of methodological challenges associated with

Figure 4 Results from the model retraining with added safety constraints. (A) Proportion of unsafe decisions in the four scenarios (see text) for three agents: human clinicians (behaviour policy), the original AI Clinician (learnt initial policy) and the modified AI Clinician (learnt safe policy). The original AI Clinician has a lower proportion of unsafe behaviour than human clinicians, while the modified ‘safe’ policy does better than the original AI Clinician. (B) Off-policy policy evaluation of the original and the modified AI Clinician policies. The value of the modified policy, with the added safety constraints, is slightly lower than the unrestricted policy (median, IQR): 90 (89.2–90) for the modified policy versus 99.5 (99.5–99.5) for the original policy, both being higher than the clinician’s (0, –1.5 to 0.6). Bottom: distribution of 25 actions for initial (C) and improved (D) policies. The safe AI policy recommends more low-dose vasopressors, likely to try and correct instances of hypotension left untreated.
the assessment of the value and estimated outcomes of following a policy that was generated by a different agent (the problem of off-policy policy evaluation).

Another limitation is that our choice of hazardous scenarios may appear arbitrary. However, it was rationally designed following the concepts of overdosing and underdosing of the two drugs of interest, defined and refined by expert clinicians over several iterations and was constrained by the retrospective data available to us (see online supplemental appendix A for more detail). In addition, the approach is based on existing concepts of safe, warning and catastrophic states of complex systems. While this work successfully integrates four safety constraints into model learning, there remain many more loosely defined hazards, such as administering fluid boluses to patients with (explicitly labelled) congestive heart failure, interstitial renal or pulmonary oedema, or acute respiratory distress syndrome, which should also be considered for a fully developed system. The iterative nature of the approach presented here provides a framework for the future addition of more scenarios. The penalty associated with each unsafe scenario can be tuned to reach a satisfying trade-off between model performance and the various safety constraints put in place.

We attempted to restrict our training dataset to patients with sepsis and to exclude patients with limitations and withdrawal of active treatment, as described in the original publication. As a consequence, occurrences of human underdosing or overdosing should mainly be due to external factors such as time pressure, resources or other factors that are not recorded in the dataset. However, despite our efforts to exclude these patients, some end-of-life patients in whom hypotension was left untreated will have been included. These would have: (1) artificially increased the proportion of unsafe decisions and (2) perverted correct AI model learning. Furthermore, the training data will most probably contain patients who may have had indications of unusual management. It is likely that some of the decisions labelled as unsafe were done knowingly by clinicians, for specific clinical indications. For example, patients with subarachnoid haemorrhage and cerebral vasospasm may be administered vasopressors to achieve an abnormally elevated blood pressure.

Other important components of the AMLAS methodology were not addressed in this project, including data management and model deployment testing ‘in the field’, which are also two crucial components of the TPLC. The data management process includes activities such as evaluating the data balance, accuracy and completeness, which was detailed in the original AI Clinician publication. As the aim for model deployment testing is to gather further safety evidence to support the transition towards operational evaluation and use of the system, it is best carried out following further retrospective model validation.

An emerging new avenue in the field is to augment AI models so that they can quantify their own confidence or uncertainty over their recommendations. Going forward, it may be helpful to algorithmically combine the communication of uncertainty that a system has about itself, which reflects the risk of unwanted behaviour as we have shown in other domains of risk-aware control by medical devices, with its safety features, that we have shown here.

Before widespread clinical adoption, more work is required to further assess the tool in its operational clinical context and submit it to the appraisal of bedside practitioners. Particularly, end users’ decision to act on or dismiss AI recommendations may be attached to some human-centred AI design characteristics and the degree of AI explainability. Human factor aspects are central in AI-based decision support systems in safety critical applications, prompting us to keep actively engineering safety into AI systems.

Acknowledgements We would like to thank Dr Myura Nagendran for his critical appraisal of the work and this manuscript.

Contributors MK was responsible for the overall content as the guarantor. ACG and MK provided the clinical expertise for this study. IH and YJ provided input in safety engineering methods. AAF took part in the study design and improved the manuscript draft. PF and MK planned and led the analysis.

Funding This work was funded by the University of York and the Lloyd’s Register Foundation through the Assuring Autonomy International Programme (Project Reference 03/19/07) and supported by the National Institute for Health Research (NIHR) Imperial Biomedical Research Centre (BRC). PF was supported by a PhD studentship of the UKRI Centre for Doctoral Training in AI for Healthcare (EP/S023283/1). ACG was supported by an NIHR Research Professorship (RP-2015-06-018). AAF was supported by a UKRI Turing AI Fellowship (EP/V025449/1). This study/project/report is independent research funded by the NIHR (Artificial Intelligence ‘Validation of a machine learning tool for optimal sepsis treatment’, AI_AWARD01869).

Disclaimer The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Competing interests The authors declare the following competing interest: MK: editorial board of BMJ Health and Care Informatics, consulting fees (Philips Healthcare) and speaker honoraria (GE Healthcare). The other authors declare no competing interest.

Patient consent for publication Not applicable.

Ethics approval 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 this study was a secondary analysis of fully anonymised data, individual patient consent was not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. The MIMIC-III data set is openly available and requires applying formally for access. Access to our computer code used in this research is available by request to the corresponding author.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: https://creativecommons.org/licenses/by/4.0/.

REFERENCES

APPENDIX A: DETAILED METHODS

AMLAS

AMLAS encapsulates a set of safety argument patterns and processes for both systematically integrating safety assurance into the machine learning model development cycle and gathering the evidence base to explicitly justify the safety of the machine learning component when integrated into systems of different levels of autonomy. The primary output of AMLAS is a safety case for the machine learning component that could be incorporated into the safety case of the wider system or service (4). AMLAS is used here for its modular and iterative approach to the safety assessment of a product over its whole lifecycle in an iterative way. The flexibility granted by these properties is essential in a complex context such as healthcare since safety considerations are only meaningful once scoped within the wider clinical setting.

DATA EXTRACTION

The subset of MIMIC-III used in this study was extracted with the same process as in (6), aiming at pulling out all ICU patients with sepsis. The final cohort slightly differs from the original paper as an update in MIMIC-III data led to an increase in number of ICU stays labelled as having sepsis from 17,083 to 20,846. The final dataset thus comprises 20,846 stays, making up for a total of 984,269 hourly recordings of patient variables and given treatment. For more details on the dataset, please refer to Supp Table 2 in appendix B.

HAZARDOUS SCENARIOS DEFINITION

In this work, we compare the performance of human and AI clinicians on the basis of how often each of them makes (or recommends) decisions that enter an unsafe scenario. Therefore, one of the main challenges of this work relies on defining what makes a decision unsafe.

In the task of cardiovascular sepsis management in the ICU where there is no gold standard for actions and where causality between actions and outcome is challenging to establish, defining unsafe scenarios for the patient is a difficult task. To tackle this question with medical expertise, the unsafe scenarios were jointly and iteratively defined by two ICU consultants (Drs Matthieu Komorowski and Anthony Gordon) and one senior ICU registrar (Dr Myura Nagendran).

Given that the purpose of this paper is to show proof of concept for safety assessment from expert knowledge, listing all possible unsafe scenarios is not necessary. The team focused on the most obvious cases of unsafe decisions related to the action space of the AI model: drug underdosing and overdosing. Other scenarios such as rapid drug dose variations were considered but set aside to focus on vasopressor/fluids underdosing/overdosing.
In the context of sepsis, leaving a patient hypotensive most likely represents a case of underdosing of vasopressors and/or fluids, since – leaving aside heart failure - hypotension in sepsis usually results from hypovolaemia and/or vasoplegia (scenario A). Giving high vasopressors to a patient who is already hypertensive is likely harmful (scenario B). Central venous pressure can be a proxy for estimating circulating blood volume, but only with extreme values (26). We used it to define the other two scenarios, representing either hypovolaemia (scenario C) or fluid overload (scenario D). The final challenge was to set thresholds for what low/high blood pressure/fluid balance means. The final threshold values presented in table 1 are the result of the assessment of clinician experts, as well as sources from the literature, such as (27).

We further describe below the rationale behind the selected hazardous scenarios:

- **Scenario A**: Hypotension is recognised as a key contributor to sepsis mortality. Sustained untreated hypotension leads to an increased risk of organ failure and death (24,28)
- **Scenario B**: The link between hypertension and mortality in sepsis is less well characterised, but the association with cardiovascular complications, pulmonary oedema, organ damage is known.
- **Scenario C**: Central venous pressure can be a reasonable proxy for estimating circulating blood volume, but only with extreme values (26). Here, we use a low MAP in combination with a low CVP to capture patients likely to be in a hypovolaemic state, therefore requiring fluids.
- **Scenario D**: Quantifying fluid responsiveness is difficult, especially in a retrospective fashion. However, fluid accumulation of 10 litres or more is common after initial resuscitation in sepsis (29). Fluid accumulation in ICU patients is a well-known risk factor for organ failure and poor outcomes (25). The predictive positive value of high values of CVP (>12-15mmHg) for lack of fluid responsiveness is satisfactory (26). Giving high doses of fluids to such patients, when they are not hypotensive, is likely very harmful, so we included patients who were likely already fluid overloaded and still receiving large volumes of fluids.

**PENALTY VALUATION AND AGENT TRAINING**

Those reward signals were applied at the level of individual decision points, and all the individual decision points were then clustered into the same 752 states as in the original work (6). The rough value of the penalty was determined by clinical reasoning: the clinician experts (co-authors in the paper) agreed that any of the unsafe conditions lasting “more than a few hours” would lead to organ failure, which would likely lead to death if left untreated. (27) confirmed a mortality rate over 50% for patients with sepsis and a MAP below 55 mmHg for a duration of 2-4 hours. Equally, administering over 750 mL (3h at 245 mL/h) of IV fluids to a patient who is already fluid overloaded is likely to be very harmful, and possibly lethal. Since each patient death in our model is penalised with 100 negative points, we gave 30 points of penalty to each hour spent satisfying the dangerous criteria. This value was then fine-tuned.
by trial and error to maximise the reduction in unsafe AI decisions while maintaining a high model performance.

SHAP VALUES

Developed by Lundberg to explain individual predictions (12), SHAP values were selected to report feature importance because they represent importance in a subtractive manner. Based on the difference between the model's performance with all inputs and the performance with all inputs but one, SHAP values are the same for two different features that have the same marginal contribution to the model's output.
**APPENDIX B: SUPPLEMENTARY TABLES**

<table>
<thead>
<tr>
<th>MIMIC-III training dataset</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique ICUs (n)</td>
<td>5</td>
</tr>
<tr>
<td>Unique ICU admissions (n)</td>
<td>20,587</td>
</tr>
<tr>
<td>Age, years (mean (s.d.))</td>
<td>64.7 (16.6)</td>
</tr>
<tr>
<td>Male gender (n (%))</td>
<td>11,512 (66)</td>
</tr>
<tr>
<td>Premorbid status (n (%))</td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>11,358 (55)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>5,965 (30)</td>
</tr>
<tr>
<td>CHF</td>
<td>6,567 (31.9)</td>
</tr>
<tr>
<td>Cancer</td>
<td>2,268 (11)</td>
</tr>
<tr>
<td>COPD or RLD</td>
<td>5,261 (25.5)</td>
</tr>
<tr>
<td>CKD</td>
<td>3,949 (19.2)</td>
</tr>
<tr>
<td>Primary ICD-9 diagnosis (n (%))</td>
<td></td>
</tr>
<tr>
<td>Sepsis, including pneumonia</td>
<td>6,669 (32.3)</td>
</tr>
<tr>
<td>Cardiovascular</td>
<td>5,822 (28.3)</td>
</tr>
<tr>
<td>Respiratory</td>
<td>2,089 (10.1)</td>
</tr>
<tr>
<td>Neurological</td>
<td>1,995 (9.7)</td>
</tr>
<tr>
<td>Renal</td>
<td>564 (2.7)</td>
</tr>
<tr>
<td>Others</td>
<td>3,420 (16.6)</td>
</tr>
<tr>
<td>Initial SOFA (mean (s.d.))</td>
<td>7.3 (2.9)</td>
</tr>
<tr>
<td>ICU length of stay (days, median, IQR)</td>
<td>3 (1.6 - 6.8)</td>
</tr>
<tr>
<td>Procedures during the 72h of data collection:</td>
<td></td>
</tr>
<tr>
<td>Vasopressors (n (%))</td>
<td>6,098 (29.6)</td>
</tr>
<tr>
<td>Mechanical ventilation (n (%))</td>
<td>10,200 (49.5)</td>
</tr>
<tr>
<td>Sedation (n (%))</td>
<td>8,318 (40.4)</td>
</tr>
<tr>
<td>Dialysis (n (%))</td>
<td>1,375 (6.7)</td>
</tr>
<tr>
<td>ICU mortality (n (%))</td>
<td>2,120 (10.3)</td>
</tr>
<tr>
<td>90-day mortality (n (%))</td>
<td>4,831 (23.5)</td>
</tr>
</tbody>
</table>

**Supp. Table 1**: Description of the subset pf MIMIC-III used for this study.
Supp. Table 2: Absolute values of number of unsafe decision for each policy, complement to figure 4.A.

<table>
<thead>
<tr>
<th></th>
<th>points satisfying the scenario</th>
<th>Clinician</th>
<th>Clinician</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>29,089/984,269</td>
<td>15,630</td>
<td>15,618</td>
</tr>
<tr>
<td>B</td>
<td>118,869/984,269</td>
<td>2,986</td>
<td>2,903</td>
</tr>
<tr>
<td>C</td>
<td>661/984,269</td>
<td>356</td>
<td>303</td>
</tr>
<tr>
<td>D</td>
<td>9,409/984,269</td>
<td>3,517</td>
<td>1,255</td>
</tr>
</tbody>
</table>

APPENDIX C: GLOSSARY OF TERMS

- **Arterial_BE**: Arterial Base Excess - proxy for how acidic the blood is
- **Bloc**: number or hours since ICU admission
- **BUN**: Blood Urea Nitrogen test - measures the amount of nitrogen in your blood that comes from the waste product urea.
- **GCS**: Glasgow Coma Scale score - objectively describe the extent of impaired consciousness in all types of acute medical and trauma patients
- **Hb**: Hemoglobin count - protein that carries oxygen from the lungs to the body and CO2 from the body back to the lungs
- **Ionised_Ca**: Ionised calcium - most active form of calcium in the blood, measured by a serum calcium blood test
- **MeanBPsd**: standard deviation of the Mean Blood Pressure over the last hour
- **PaO2**: Partial pressure of oxygen - measure if oxygen in the arteries
- **PaO2_FIO2 (P/F ratio)**: arterial pO2 (“P”) from the ABG (Arterial Blood Gases test) divided by the FIO2 (“F”) - the fraction (percent) of inspired oxygen that the patient is receiving expressed as a decimal (60% oxygen = FIO2 of 0.60).
- **SOFA**: Sequantial Organ Failure Assessment score - objective score that allows for calculation of both the number and the severity of organ dysfunction in six organ systems (respiratory, coagulatory, liver, cardiovascular, renal, and neurologic)
- **SysBP**: Systolic Blood Pressure - measures of the force exerted by the heart on the walls of the arteries each time it beats (in contrast with diastolic blood pressure, the pressure between heart beats)
- **WBC_count**: White Blood Cells count - infections increase this count, a low white blood cells count while suffering an infection can indicate difficulty to fight it