Article:
Jia, Yan, Lawton, Tom, Burden, John et al. (2 more authors) (2021) Safety-Driven Design of Machine Learning for Sepsis Treatment. Journal of Biomedical Informatics. 103762. ISSN 1532-0464

https://doi.org/10.1016/j.jbi.2021.103762

Safety-Driven Design of Machine Learning for Sepsis Treatment

Yan Jia[1], **Tom Lawton** [2], **John Burden**[1,3], **John McDermid**[1], **Ibrahim Habli**[1]

[1]Department of Computer Science, University of York, York, UK

[2] Bradford Royal Infirmary and Bradford Institute for Health Research, Bradford, UK

[3] The Centre for the Study of Existential Risk, University of Cambridge, Cambridge, UK

Abstract. *Machine learning (ML) has the potential to bring significant clinical benefits. However, there are patient safety challenges in introducing ML in complex healthcare settings and in assuring the technology to the satisfaction of the different regulators. The work presented in this paper tackles the urgent problem of proactively assuring ML in its clinical context as a step towards enabling the safe introduction of ML into clinical practice. In particular, the paper considers the use of deep Reinforcement Learning, a type of ML, for sepsis treatment. The methodology starts with the modelling of a clinical workflow that integrates the ML model for sepsis treatment recommendations. Then safety analysis is carried out based on the clinical workflow, identifying hazards and safety requirements for the ML model. In this paper the design of the ML model is enhanced to satisfy the safety requirements for mitigating a major clinical hazard: sudden change of vasopressor dose. A rigorous evaluation is conducted to show how these requirements are met. A safety case is presented, providing a basis for regulators to make a judgement on the acceptability of introducing the ML model into sepsis treatment in a healthcare setting. The overall argument is broad in considering the wider patient safety considerations, but the detailed rationale and supporting evidence presented relate to this specific hazard. Whilst there are no agreed regulatory approaches to introducing ML into healthcare, the work presented in this paper has shown a possible direction for overcoming this barrier and exploit the benefits of ML without compromising safety.*

Key Words: Machine learning, sepsis treatment, safety assurance.

1. Introduction

Machine learning (ML) has received a lot of attention recently due to its rapid development and promising applications in many areas, particularly in healthcare. ML has the ability to process huge datasets beyond the scope of human capability and use the analysis of that data to produce meaningful insights and enable timely responses. For example, it is believed that ML can assist clinicians

in planning and providing care, ultimately leading to better outcomes with lower costs of care [1]. Indeed, recent research [2] shows how ML can be used to help pathologists to identify and localise cancerous tumours from images with promising results. ML has also been used to discover antibiotics which are structurally different from known antibiotics [3] and the same group at MIT is working on the use of ML to discover treatments for COVID-19 [4]. However, before such applications can be deployed, it is necessary to demonstrate their safety.

Healthcare regulators have developed standards for assuring the safety of digital systems [5], e.g. DCB0160 from NHS Digital [6]. However, these standards and the associated regulatory approaches assume that software is developed in a "conventional way" and thus are not well-suited to ML applications, where systems are produced without explicit programming but by automatically learning from complex data sets. Although these issues are starting to be addressed, e.g. by the US Federal Drug Administration (FDA) [7], there is still a disconnect between regulatory practices and the processes for assuring ML in healthcare. Indeed, one of the key findings of a recent study by the UK Care Quality Commission (CQC) was "the need for more assurance about the clinical aspects of the algorithms in machine learning, and more clarity on how hospitals should implement machine learning devices within clinical pathways to ensure high-quality care" [8]. This indicates the need for more focused effort on practical methods of safely translating ML from research into clinical practice. One of the problems to be addressed is that development of ML is often undertaken in "silos", e.g. focusing on particular data analysis challenges [9], without addressing the broader issues of clinical adoption. To overcome this problem it is necessary to bring together expertise and stakeholders from many disciplines including clinical practice, ML and safety engineering.

The paper provides a concrete clinical use case for sepsis treatment using ML, specifically deep Reinforcement Learning (RL) in this case. Sepsis is a life-threatening condition and a major cause of fatalities in hospitals. It is hard to detect the onset of the condition and the optimal treatment is as yet unclear [10]. RL is well-suited to decision-support problems and several researchers have already applied RL to the problem of recommending optimal sepsis treatment, e.g. [11]. We have also adopted RL, as the existing work both gives a baseline on which to build and to demonstrate how to achieve safety-driven design of the RL model.

In particular, we developed and applied a novel methodology that incorporates safety engineering processes to support development and refinement of the clinical workflow and the ML model. The safety engineering process identifies hazards (i.e. sources of potential patient harm), hazard causes and requirements for hazard controls. The design of the ML model is then enhanced to sat-

isfy the relevant safety requirements and a rigorous evaluation is undertaken to provide evidence that these requirements are met. The evidence feeds into a safety case which presents the safety rationale, including showing the completeness of the controls. This work provides a process for assuring the safety of the ML model in its clinical context of use thus supporting regulators in assessing the acceptability of introducing an ML model into a healthcare setting.

The rest of the paper is structured as follows. Section 2 discusses the background and related work, including the safety of ML in healthcare. Section 3 describes the methodology we have used in this work covering the clinical, safety and ML elements outlined above. Section 4 presents our detailed clinical use case on the treatment of sepsis, focusing on mitigating a major clinical hazard: sudden change of vasopressor dose. A discussion of the role of the work and the possible future directions is presented in Section 5. Section 6 presents conclusions.

2. Background and Related Work

It is common to categorise ML algorithms into three types according to the way they are trained, viz: supervised learning, unsupervised learning and RL. All three types have been explored in healthcare. Supervised learning involves training using data points with known outcomes and the learning algorithms are a form of optimisation which seeks to minimise loss or error. There is a lot of work using supervised learning for classification problems in healthcare, e.g. for breast cancer screening [12, 2]. Unsupervised learning identifies previously unknown correlations in data with the minimum of human supervision. A typical application in healthcare is to try to identify phenotypes – that is groups of patients who are homogeneous in how the specific medical condition is presented. Examples include Acute Respiratory Distress Syndrome (ARDS), identifying hypo- and hyper-inflammatory phenotypes [13] and sepsis, identifying four novel phenotypes [14]. RL is an ML technique that is often used in complex decision making tasks to find an optimal strategy [15]. It has been applied to identify optimal treatments in healthcare very recently, e.g. determining treatment regimes in chronic disease and automated medical diagnosis [16]. It involves an agent seeking to maximise its reward through interaction with its environment. A more focused discussion of RL and its application to sepsis can be found in Section 4.

Although there are many research activities investigating how to exploit the potential benefits of ML in healthcare, few studies have progressed to deployment in clinical care [17]. Thus, researchers are now beginning to realise that more effort needs to be put into safe deployment of ML in healthcare. For example, “sepsis watch”, has reported on the work of a multi-disciplinary team

including statisticians, data scientists and clinicians introducing a deep-learning based sepsis detection and management system into clinical care [18]. In this work, front-line clinical staff were highly engaged in the design and development of the workflow, ML model, and its application. Several iterations occurred throughout the product lifecycle to improve the ML model to suit its clinical context of use. Rigorous evaluation was carried out with external partners to assess the possible inequality and bias introduced by ML and they conducted operational impact evaluation to demonstrate safety and efficacy. They emphasised the importance of multi-disciplinary working and early involvement of all stakeholders in order to successfully integrate ML technologies into routine clinical care.

In [17] the authors took a broad view of the issues, providing an overview of the barriers to deployment of ML and translating research into practice. The work focuses on developing a "roadmap" for accelerated translation of ML based interventions into healthcare, which includes choosing the right problems, developing a useful solution, carrying out rigorous evaluation, and deploying responsibly, by first undertaking "silent mode" operation, i.e. running the system but not using its results, to evaluate the technology. They then suggest undertaking a clinical trial but they think a randomised control trial (RCT) might not be feasible as it requires a different workflow compared with the control group, which might lead to confusion, and suggest that other forms of trial might be more appropriate, e.g. a pre-post study. Similar to "sepsis watch" they emphasise the importance of multi-disciplinary teams, although no actual deployment was reported.

When it comes to effectiveness research in healthcare, the "gold standard" is RCTs [19]. However, only a few projects have carried out RCTs for ML-based applications. For example, an RCT was conducted on an ML-based severe sepsis prediction algorithm finding reductions in average length of stay and in-hospital mortality in the group using the ML-based tool as opposed to the control group [20]. Another project studied a deep learning-based polyp detection system. Evaluation of its use during colonoscopy showed increases in polyp adenoma detection rates against the control group [21]. A third example is an AI-based decision-support tool used to aid anaesthetists in controlling hypotension [22]. Like the polyp detection system, this decision-support tool operates in real-time and was shown to be effective, i.e. to reduce periods of intraoperative hypotension. Despite these successes, there remains a debate about the practicality and effectiveness of RCTs for ML-based tools. For example [23] discusses the cost and difficulty of conducting RCTs, including the effort involved, e.g. clinician training, and the problem of evaluation where the ML-based systems continue learning from operational data, an issue which the Federal Drug Administration (FDA) is currently investigating [7], proposing an Algorithm Change Process (ACP) for updating the deployed ML model.

Both “sepsis watch” [18] and the work on “roadmaps” [17] provide useful insights and guidance into the successful translation of ML applications into clinical practice. However, despite their emphasis on multi-disciplinarity neither considers the early involvement of safety engineers nor a proactive approach to managing patient safety, although patient safety is mentioned in both papers. The work described here extends the notion of multi-disciplinarity to include safety engineering thus enabling proactive management of safety when introducing ML-based systems in healthcare, whether an RCT is used or not.

3. Methodology

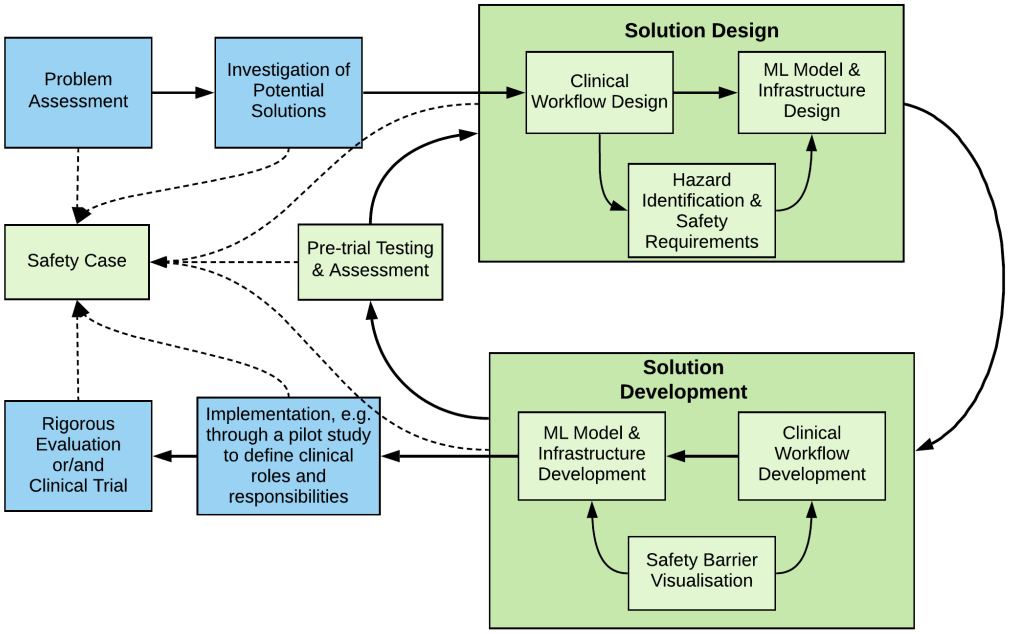

Figure 1. Framework for integrating ML system into clinical care

Our methodology is shown in Fig. 1 and incorporates concepts from “sepsis watch” [18] and the work on “roadmaps” [17], extended to enable the proactive incorporation of patient safety into the development of ML models. The rectangular boxes describe the activities performed while the solid arrows show the flow of the activities. The dashed arrows represent the information that pertains to assurance rationale and evidence, which is captured in the safety case. The “flow” starts at the top left, iterates through *Solution design*, *Solution development* and *Pre-trial testing & assessment*, ending with *Rigorous Evaluation or/and Clinical Trial*.

In this paper, we are mainly concerned with the *Solution design*, *Solution development*, *Pre-trial testing & assessment* and the *Safety Case*, which are all marked in green in Fig. 1. Infrastructure is an important element to enable the deployment of ML models in healthcare but is out of scope for

this paper. The elements marked in blue have a clinical focus and are largely outside the scope of this paper, although an overview of sepsis is given in Section 4 to provide context for the safety work and ML model development.

Solution design comprises *Clinical workflow design*, *Hazard identification & safety requirements* and *ML Model design*. In order to deploy ML models effectively in healthcare, it is important to ensure they fit into the clinical context. *Clinical workflow design* defines the integration of the ML model into the socio-technical clinical setting to address the healthcare problem, supporting the clinicians in their work. Thus, it is necessary to involve the front-line staff at this step to identify potential constraints or requirements to ensure that the clinical workflow is feasible and efficient for the end-users of the ML model. Additionally, the clinical workflow will serve as the basis for proactive safety analysis, including identifying hazards and deriving safety requirements for the ML model design. Hazards are situations which, if not controlled, could lead to harm [24]. *Hazard identification* is central in safety risk management as it gives us a focus for assessing risks and defining safety barriers. Traditionally, this has been done using systematic analysis techniques and documented in the form of tables. *ML Model design* includes identifying the set of input features that will be used in training the model so that it is effective in its clinical setting and for the problem being addressed. Although there are various techniques to help to select the relevant features, it is important to incorporate clinical domain expertise to identify the right set of features. Once the input features have been identified, it is time to extract the right source of data for the model development because the quality and quantity of the data will directly determine how good the ML model can be [25]. In addition, it is necessary to identify the performance metrics that are most suitable and informative to evaluate the ML model, given the problem being addressed [26].

Solution development comprises *Clinical workflow development*, *Safety barrier visualisation* and *ML Model development*. *Clinical workflow development* includes developing user interfaces to support the implementation of the clinical workflow which would help the front-line staff to use the ML model effectively. The front-line staff would be particularly involved in testing and validating the functions, information, control, and visual components of the interface. *Safety barriers* are means of controlling the potential hazards that we identified previously based on the clinical workflow to reduce the risk that they will compromise patient safety. In this paper, we especially focus on the barriers that can be implemented in the ML model itself. This may include altering the input features used by the ML model to ensure it takes into account safety-relevant information or improving the interpretability of the ML model to help clinicians make informed decisions. The *ML Model development* involves

training the model using the data identified during the *Solution design* augmented if necessary to implement the defined *Safety barriers*.

Pre-trial testing & assessment mainly concerns the technical issues of the ML model's readiness for use, e.g. predictive accuracy based on the previously defined performance metrics. The ethical and other challenges could be evaluated later [27], e.g. in the rigorous evaluation through clinical trials. In practice, there is no clear cut distinction between the activities shown in Fig. 1. In fact, the activities often overlap and iterate. Ideally the safety activities occur in conjunction with the clinical and ML model design & development activities. The iteration between *Solution Design*, *Solution development* and the *Pre-trial testing & assessment* is the basis for developing the ML model to be safe enough to go on to a pilot study or a "silent mode" use prior to rigorous evaluation, e.g. clinical trials.

The use of *safety cases* is a long-established practice in many safety-critical domains. Particularly in the UK, the development of a safety case is a mandatory requirement in key sectors such as defence, nuclear and railways [28]. In the National Health Service (NHS) in England, compliance with the clinical safety standards DCB0129 and DCB0160 requires a safety case for Health IT systems [5]. A safety case for clinical risk management "is a structured argument which is supported by a body of relevant evidence that provides a compelling, comprehensible and valid case that a system is safe for a given application in a given operating environment" [6]. In our methodology, the safety case draws evidence from all the phases in Fig. 1 and documents the safety rationale for the integrated workflow including the ML model at all stages in its development.

Next, we apply the methodology to a clinical use case involving treatment of sepsis patients.

4. Clinical Use Case: Sepsis Treatment

The clinical use case focuses on the treatment of sepsis. Sepsis is a life-threatening organ dysfunction which is caused by a dysregulated host response to infection [29]. It is estimated that one in five deaths worldwide are due to sepsis [30]. Evidence suggests that current practices in the administration of intravenous fluids and vasopressors for treating sepsis are suboptimal [31]. Consequently, researchers have harnessed RL to learn the "optimal" treatment strategy for recommending intravenous fluids and vasopressors, e.g. [32] [11].

Basic Concepts of Reinforcement Learning (RL)

RL consists of an agent interacting with its environment by performing actions and receiving feedback from the environment. The environment is often represented by a Markov Decision Process (MDP) in which an assumption is made that the future state of the process depends only on the current state; that is, given the current state, the future state does not depend on the cumulative history of past states. An MDP is defined by $M = \langle S, A, P, R \rangle$, where S is the state space, A is the action space, P is the transition function with $P(s'|s, a)$ denoting the probability of reaching state s' if taking action a in state s. R is the reward function such that $R(s, a, s')$ is the immediate reward given to the agent for transitioning between states s and s' via action a. A policy is a function defining the agent's behaviour and maps a perceived state of the environment to an action for the agent to take.

In this work, we use a previously published deep RL model [32] for sepsis treatments which recommends intravenous fluids and vasopressors. In particular, we apply our methodology and show how to integrate the ML model into clinical care in a way that enables proactive management of patient safety. The clinical use case shows the iteration round the *Solution design*, *Solution development* and *Pre-trial testing & assessment* "loop" in our methodology. The main work products of this iteration, e.g. the *Clinical workflow* and the *Hazard analysis and safety requirements* as well as the the *Safety case* are shown in the following subsections. For ease of presentation we combine the design and implementation of the *Clinical workflow* and *ML model* in the following section.

4.1. Clinical workflow design and development

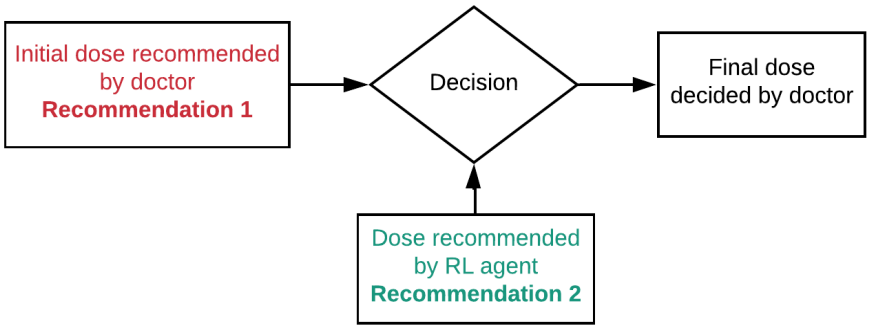

Figure 2. High-level workflow design

There are two main ways of introducing ML models into healthcare: either replacing clinicians or assisting them. For example, [12] explained that an ML model for breast cancer screening can be used in the standard double-reading process to replace the second reader while maintaining

an equivalent performance. In our work, the ML model serves as a decision support tool, assisting clinicians in sepsis treatment as shown in Fig. 2. First, a doctor recommends initial doses of intravenous fluids and vasopressors for the sepsis patient. Then the doctor is shown the recommendations from an RL agent for the same sepsis patient. Afterwards, the doctor makes the final decision on the recommended dosage for intravenous fluids and vasopressors, reflecting the role of the ML model as a decision aid. This is different from most current advisory systems in healthcare in that they make recommendations first, then doctors choose to accept them or to modify them, without the explicit initial recommendation. The reason for designing the workflow this way is to support later pilot studies and/or clinical trials to evaluate the ML model not only in a technical sense, but also to see how it affects the clinicians' behaviour in the socio-technical context, e.g. due to automation bias. After evaluation, if confidence and trust has been built in the ML model, then it would be appropriate to alter the workflow to allow the clinicians to use the ML model like a normal advisory systems, i.e. without the explicit initial recommendation.

The detailed workflow that integrates the ML model is shown in Fig 3. This workflow shows a broader view of sepsis treatment including the screening activities. There are often two distinct phases: the initial resuscitation and the more stable period thereafter. However, the workflow intentionally doesn't distinguish these two phases, but is intended to give guidance for both, as appropriate.

The workflow starts by screening the patient for (suspected) sepsis. The screening criteria are based on published NHS improvement protocols [33]; if necessary, it can also be altered to suit the local hospital screening protocol. Here, early warning score (EWS) [34] is used and sepsis is suspected if EWS is greater than 3 and at least one sepsis red flag criterion, e.g. newly altered mental state, is present. The rest of the workflow shows both the initial resuscitation for sepsis and septic shock and the treatment afterwards, i.e. the stable period. It is mainly based on the sepsis 6 pathway from the Sepsis Trust [35] and the Hour-1 Bundle from the Surviving Sepsis Campaign [36]. The Hour-1 bundle is designed for initial resuscitation but intravenous fluids and vasopressors will continue to be given in the stable period, most likely for several days. Specifically, when it comes to recommending intravenous fluids and vasopressors, the workflow integrates the ML model, i.e. the RL agent, into the clinical workflow. This is shown as recommendation 2, which matches the high-level workflow design in Fig. 2. Recommendation 1 is the doctors' initial recommendation based on current clinical practice. If necessary, recommendation 1 can also be altered to suit the local hospital protocol. The final decision is made by the doctors after they are informed about the RL agent's recommendation. As noted above, we designed the clinical workflow this way to reflect its role as a

decision aid, and to enable us to assess how much the RL agent influences the behaviour of the doctors and whether the RL model could indeed improve clinical results, e.g. reducing in-hospital mortality. Importantly, the approach helps to ensure that an accountable doctor makes the final decision [27].

The workflow concludes with the nurses administering the intravenous fluids and vasopressors as advised by the doctors. It is important to recognise the role of nurses in this clinical workflow as they usually are the ones at the bedside actually making the adjustments according to more general guides set by the doctors. This also needs to be considered in the hazard analysis especially deriving the causes and controls of the hazards (as detailed in the next section).

After the iteration on the designs of the clinical workflow and ML model (model design is discussed in Section 4.3), development begins. Implementing the clinical workflow involves integrating tools and providing appropriate user interfaces for clinical staff. Integration requires data exchange with the electronic health record, particularly to transfer the features that the RL model needs to process in order to recommend the doses for the patient. This work is primarily the responsibility of IT specialists, including those working for vendors of Health IT systems that are integrated into the clinical workflow. User interfaces will be needed for clinicians both to provide them with information, e.g. recommended doses from the RL model, and to enable them to input information, including recording decisions they have made [37]. It is good practice to employ "user-centred design" [38] where specialists in user interface design work with all the different classes of user, including nurses and doctors, to produce appropriate systems. Generally the design process will be iterative, to define and refine functions, information, control, and visual components of the system. These capabilities need to be provided in compliance with relevant standards and guides, to allow the hospital to comply with audit requirements – in general to support management processes as well as clinical ones. Finally, staff need to be trained to understand the new workflow and to work effectively with the tools. Using the clinical staff who were engaged in design and development to train other users may prove effective, as they will understand and be able to explain the systems from a user perspective.

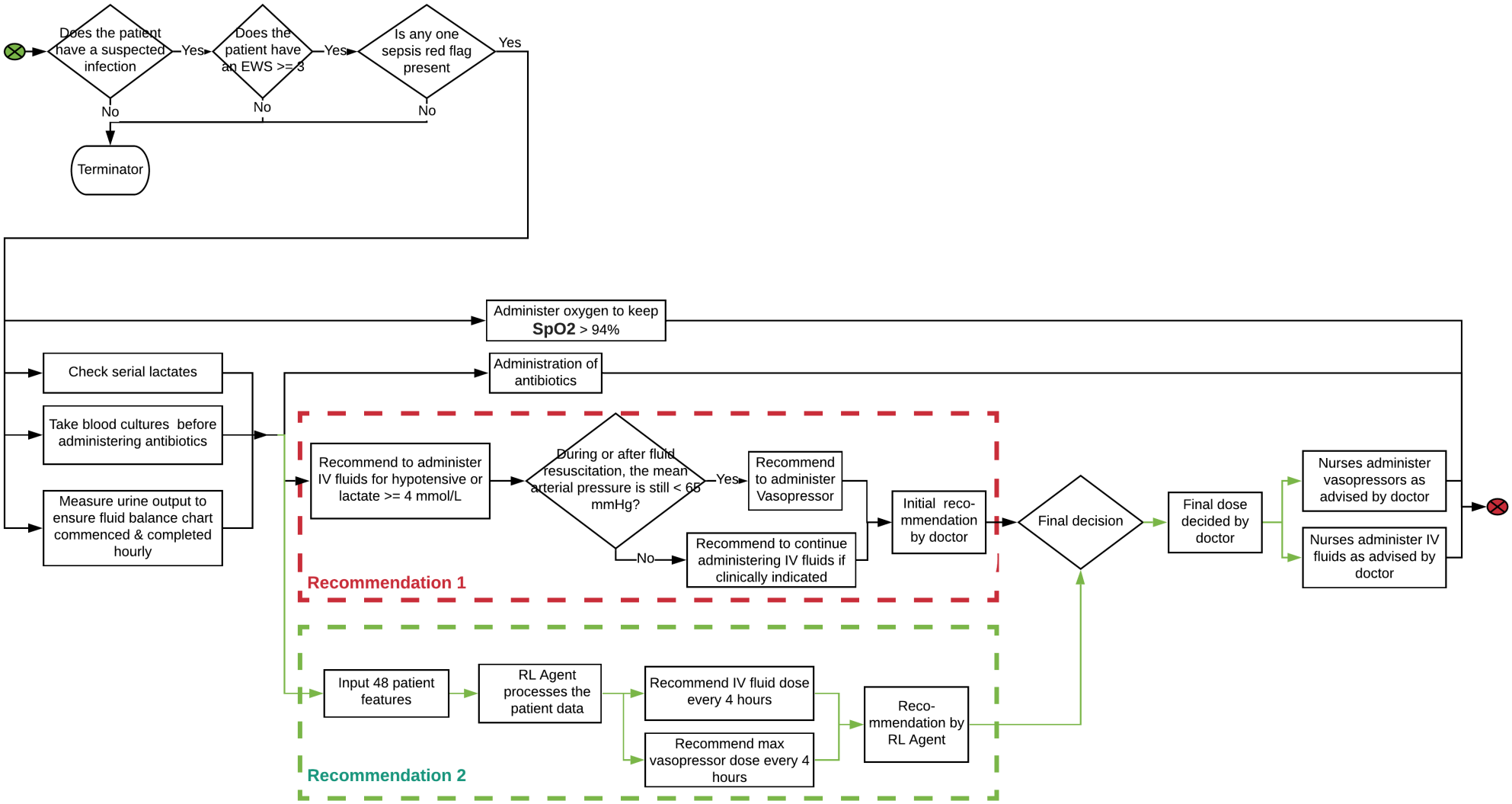

Figure 3. The detailed workflow integrating ML model to treat sepsis patient

4.2. Hazard identification & safety requirements

In safety management it is common to organise risk analysis and control around the notion of a hazard [24], however it needs to be interpreted in the context of a particular system or situation [39]. Once hazards have been identified, the system or situation is analysed to determine potential causes of the hazards and the potential clinical effects. For each identified hazard the associated risk is estimated. Typically, the risk reflects the severity and the likelihood of the hazard's consequences. Sometimes the likelihoods can be quantified; other times estimates are qualitative, based on domain knowledge. In addition, the risk of the hazards will determine the priority for the introduction of safety barriers (means of preventing the causes of hazards or reducing the impact of hazards if they do arise). Once safety barriers have been identified and introduced, then the risk associated with the hazards can be re-evaluated.

Hazard and safety analysis of computer-based systems often uses variants of Hazard and Operability Studies (HAZOP) [40] from the chemical industry. The HAZOP approach is based on flows (of chemicals, etc.) through process plant. The variants used for computer-based systems, e.g. SHARD [41], consider information flows through systems. SHARD is suitable for identifying both hazards and causes of hazards, as it focuses on deviations from intent that could be hazardous. It provides a structured approach to the identification of deviations from intent by systematically applying the guidewords (omission, commission, early, late and incorrect) to each flow. In this context, commission means doing something that was not intended.

In this paper, we only show the hazard identification for the delivery of vasopressors; intravenous fluid can be analysed in a similar way. The analysis is carried out by a multidisciplinary team comprising two safety engineers, one Intensive Care Consultant and two ML engineers prompted by the guidewords in the SHARD method. The resulting hazards are as follows:

- *Omission* – No vasopressor administered;
- *Commission* – Unnecessary vasopressor administered;
- *Incorrect* – Wrong vasopressor administered;
- *Incorrect* – Wrong dose administered (this hazard concerns a single dose);
- *Incorrect* – Sudden change of vasopressor dose administered (this hazard concerns two consecutive doses);
- *Late* – Delay in administering vasopressor.

The guideword *early* is not considered, as there is ongoing clinical research about whether or not to deliver vasopressor earlier to increase mean arterial pressure (MAP) for sepsis treatment. The

guideword *incorrect* results in three potential hazards: one concerns administering the wrong vasopressor; another concerns administering a single wrong vasopressor dose; the third concerns a sudden change of vasopressor dose between two consecutive doses. Current clinical practice is to change the dosage of vasopressors gradually as a sudden major change in the dose can be dangerous to some patients, e.g. resulting in acute hypotension (arising from rapidly decreasing doses), hypertension or cardiac arrhythmias (arising from rapidly increasing doses) [42] [43] [44]. Because the half life (the period of time for the concentration of a drug in the body to reduce by 50%) of Norepinephrine (a commonly used vasopressor) is measured in seconds or minutes [45], changes in Norepinephrine can have rapid effects on patients.

After the identification of the potential hazards, we applied SHARD analysis to the clinical workflow to identify the causes of the hazards. This is done by going through each activity (the rectangular boxes) in Fig. 3 with a focus on recommendation 2, i.e. the part of the workflow marked in green. The Table below shows a fragment of the SHARD analysis with a focus on one hazard -– *sudden change of vasopressor dose administered* – identified above. The analysis for the other hazards can be found in the supplementary material. Table 1 is a high-level summary of the analysis. The full analysis is also included in the supplementary material but a brief summary of the approach is presented here.

The SHARD analysis works "backwards" through the workflow, starting with the identified hazards then considers each activity in the workflow in turn, following the process outlined in [46]. Each hazard, e.g. "No vasopressor administered" is an *output deviation* from the final activity – "administer vasopressor as decided by doctor" in this case. The hazard can have many causes. First, it can arise within (an *internal deviation*) in the final activity in the workflow; *internal deviations* are identified using the SHARD guidewords. For example, omission by the nurse responsible for vasopressor administration, perhaps due to a heavy workload, leads to the hazard "no vasopressor adminstered". Second, the hazard can be caused by deviations in activities earlier in the workflow which propagate from earlier activities to the final activity. Specifically, *input deviations* of the final activity arise from *output deviations* of the preceding activity, and so on through the workflow. For example, in this case, the *input deviation* for the final activity "administer vasopressor as decided by doctor" can be omission of the final dose recommendation, which ultimately contribute to the hazard "no vasopressor adminstered".

In this way we can identify how deviations from intent for each activity can combine and propagate through the complete workflow to give rise to hazards, noting that the deviation of one

Table 1. Fragment of SHARD analysis showing a single hazard

Guide word	Deviation (Hazards)	Possible Causes	Effects	Severity
Incorrect	Sudden change of vasopressor dose is administered (concerns two consecutive doses)	1 Kink of line 2 The pump fails, e.g. due to electrical problem or bag/syringe not installed correctly 3 The delivery line might not be connected to patient's central line, e.g. due to the patient pulling out the central line 4 The drug might not be added to the diluent, so the syringe/bag just contains saline (a problem when bags/syringes are being changed over) 5 Nurse prepared wrong dose (e.g. due to calculation error) 6 Inappropriate titration of dose by nurse 7 Doctor fails to check current dose 8 Initial recommendation by doctor has a sharp change in dose and doctor carried through the recommendation (not considered in this paper) 9 RL agent recommends a sharp change in dose and doctor accepts the advice, e.g. due to automation bias	Acute Hypotension, Strokes, Renal failure, Heart attack could occur from a sharp drop in the dose	Major/ considerable
		10 Features in state space of the RL model are not sufficient to represent the patient conditions for sepsis decision making 11 Reward function used for RL model is coarse 12 Cost function used for RL model development is not appropriate 13 Hyperparameters used for RL model development are not optimised 14 Training data for RL model development is not appropriate	Hypertension, Cardiac Arrhythmia, Strokes, Raised intracranial pressure, Pulmonary oedema could occur from a sharp rise in the dose	
		15 Data corruption (e.g. invalid or wrong data produced by over-writing patient's features) 16 Features for wrong patient entered 17 Wrong patient feature values entered (e.g. due to unit difference) 18 Test results for wrong patient received 19 Incorrect test results received		

class, e.g. *omission*, can lead to the deviation of another class, e.g. *incorrect*. This process enables us to produce a summary of possible hazard causes, taking into account the complex interdependencies between the activities, as illustrated in Table 1. The severity classification used in the table is based on the standard DCB160 developed by NHS digital [6].

As indicated above, Table 1 summarises the detailed analysis in the supplementary material, combining the results from analysing all the different activities in the workflow in Figure 3. The possible causes of most interest in this paper are numbers 10-14, which are highlighted in the table, as they directly affect the RL recommendation, i.e. recommendation 2 in the workflow. In addition, causes 1 to 6 can arise from the administration phase, which is the final activity in the workflow. Causes 7 to 9 can arise from the final decision phase which is the activity before administration in the workflow, where cause 9 is a combination of an RL agent failure (a potential consequence of numbers 10-14) and a human error (automation bias). Causes 15 to 19 can affect the quality of the input data to the RL model, which is the beginning activity in recommendation 2 in the workflow. The possible causes in Table 1 can arise from different types of failure, e.g. technical failure and human errors. However, a single cause can trace back to multiple different sources. For example, cause 2 can arise from a technical failure, but also a human error. Although our focus is mainly on the ML components in this paper, the visualisation of controls in Section 4.5 addresses some of the other possible causes identified in Table 1.

Safety requirements are derived from the hazard analysis to control the hazard causes identified in Table 1. To produce a set of requirements for the ML components in the workflow it is helpful to identify the *interfaces* in the workflow that bound those components. The key interface is between "Recommendation by RL agent" and the "Final decision' in Fig. 3' which shows the interface between the ML model and the clinicians. Given this, we can identify that the hazardous interface failure is "RL agent recommends a sharp change in dose" (an *output deviation* from the ML model) which contributes to the clinical hazard "Sudden change of vasopressor dose administered". Thus the requirements derived from controlling the hazardous interface failure help guide the design of the ML model which falls within the scope of "Recommendation 2" in the clinical workflow.

The resultant requirements are set out in Table 2. R0 follows directly from the definition of the hazardous interface failure. Requirements R1 to R5 are lower level design and development requirements necessary to support R0. R1 relates to cause 10 and is concerned with input feature issues. Defining the features in the state space for the RL model is a design issue, so R1 is allocated accordingly. R2 relates to cause 11 in Table 1. Similarly, it is allocated to "RL model design" as this

Table 2. Safety Requirements for RL model derived from Hazard analysis

ID	Description	Type	Allocation
R0	Sudden changes in recommended dose shall be close to clinician practice	Performance & Safety	RL model development
R1	Feature representation in the state space shall be sufficient to allow the control of sudden changes in recommended dose	Performance & Safety	RL model design
R2	An appropriate reward function shall be defined to allow the recognition of desired clinical outcome	Performance & Safety	RL model design
R3	An appropriate cost function shall be defined to penalise hazardous behaviours	Performance & Safety	RL model development
R4	Hyperparameters shall be optimised based on the validation dataset	Performance & Safety	RL model development
R5	Patient cohort shall be defined using recognised criteria, i.e. sepsis-3	Performance & Safety	RL model design

is the phase in the methodology where reward functions are defined. Requirements R3, R4 and R5 relate to causes 12, 13, and 14 respectively; they are all allocated appropriately. Thus, Table 2 covers all the RL agent-related causes in Table 1 and if the requirements are satisfied, this should reduce the likelihood of the hazardous interface failure arising – "RL agent recommends a sudden change in dose". The requirements have to be produced using specialist knowledge of ML, reinforcing the need for a multi-disciplinary team. Causes 15 to 19 in Table 1 should be addressed in the user interface design in that it can reduce the likelihood that such causes arise.

4.3. Model design & development

In this paper, we have adapted the RL model in [32] to train an agent to learn the optimal policy for sepsis treatment; from now on we refer to this as the original policy. The adapted RL model used 47 features to represent the state space (as against 48 in the original work), including patients' demographics, Elixhauser premorbid status, vital signs, laboratory values, fluids and vasopressors received to satisfy safety requirement R1 in Table 2. The action space includes 25 possible actions with five discretised choices for the dose of intravenous fluids and five for vasopressors respectively, shown in Table 3. The terminal reward is based on 90-day mortality (as against hospital-mortality

in the original work) with +15 for survived and -15 otherwise. The intermediate reward uses SOFA (Sequential Organ Failure Assessment) score and Arterial Lactate (the level of lactate from arterial blood) as they did in the original work to satisfy safety requirement R2. The SOFA score is a measurement of organ failure with high values associated with poor outcomes; similarly, high levels of lactate suggest stress or inadequate organ perfusion and are associated with poor outcomes in sepsis treatment. A well-established and widely-used RL algorithm – Double Deep Q-networks (DQN) [47] is used to determine the policy (a brief introduction to DQN is given in the box below). Therefore, the cost function used a standard double DQN loss function plus one regularisation term, as indicated in the original work to satisfy safety requirement R3.

Table 3. Dosage actions

		Dose of vasopressor (mcg/kg/min)				
		No.: 0 **Range:** 0 **Median:** 0	1 (0.002, 0.079) 0.04	2 (0.08, 0.2) 0.135	3 (0.201,0.449) 0.27	4 (0.45, 1.005) 0.786
Dose of IV fluid	0	0	1	2	3	4
	1	5	6	7	8	9
	2	10	11	12	13	14
	3	15	16	17	18	19
	4	20	21	22	23	24

Principles of Deep Q-Networks (DQN)

DQN is a widely-used modern RL algorithm, which combines Q-learning [48] with a deep artificial neural network. It learns a policy by employing the same core update rules and operating principles as Q-learning but using a neural network in order to represent its Q-function. DQN uses the experiences or samples $\langle s, a, r, s' \rangle$ generated by interaction with the environment to train the neural network, where r is the observed immediate reward. A common implementation uses a squared error loss of the difference between the output of the so called prediction network, $Q(s, a, \theta)$ and the desired target $Q_{target} = r + \gamma \max_{a'} Q(s', a', \theta)$ to update the neural network's weights.

Simple DQNs have some shortcomings and there are various ways of refining them to improve their performance. One way to improve algorithmic stability is to use double DQN which introduces a second network — the target network. The purpose of the target network, pa-

rameterised by θ', is to provide a stationary target upon which the Q-function can converge. Periodically, the target network is updated to match the prediction network.

An additional improvement of using double DQN is that the target network is used to select the action for the prediction network to evaluate. The standard double DQN loss is shown in equation (1).

$$L(\theta) = E[(Q_{double-target} - Q(s, a; \theta))^2], \tag{1}$$

where $Q_{double-target} = r + \gamma Q(s', argmax_{a'} Q(s', a'; \theta); \theta')$.

The data used for model development is based on the same data set and the same patient cohort taken from MIMIC III – a large publicly available database [49] – as in the original work. Patients are included in the cohort when they meet the sepsis-3 criteria [29] – suspected infections combined with SOFA score ≥ 2. Exclusion criteria are: 1. not adult, 2. intravenous fluid intake not documented, 3. possible withdrawal of treatment, 4. erroneous intake or output data. The detailed MIMIC III data pre-processing can be found in the supplement to [11]. This satisfies safety requirement R5. The resulting patient cohorts were divided into a training dataset (80%, 20,938), a validation dataset (10%, 2149) and a testing dataset (10%, 2160). For detailed patient features included in the state space, see the supplement to [11]. The hyperparameters are manually tuned and optimised using the validation data to satisfy safety requirement R4. By satisfying requirements R1 to R5, we could state that this will also satisfy requirement R0, but it is necessary to evaluate the RL model after training to see if this is the case, see Section 4.4.

The RL model was developed in Python and uses the TensorFlow library [50]; the code developed is available at: https://github.com/Yanjiayork/sepsisRL. As the MIMIC III data set was generated by recording the real clinicians' actions, we refer to it as the clinician policy in contrast with the (learnt) original policy. We evaluated the original policy and compared it against the clinician policy, i.e. the real patient trajectories in the test data set, including whether or not they show the sudden major change related to the hazardous interface failure "RL agent recommends a sudden change in dose" when recommending vasopressor dosage for each patient.

4.4. Pre-trial testing & assessment

As indicated above, this phase of the methodology mainly concerns the technical issues of the ML model's readiness for use. Evaluation of *performance* is standard in ML after the training of the model. In the original work [32] they carried out evaluation to check the "sanity" of their learnt policy. In addition, in our work, we evaluate the original policy from the safety perspective – specifically in

Table 4. Summary of max dose change between consecutive doses for the three policies

	Dose of vasopressor (mcg/kg/min)	
	Small-Medium Dose Change (0-0.75)	Large Dose Change (>0.75)
Clinician Policy	97% (2,100)	3% (60)
Original Policy	65% (1,404)	35% (756)
Modified Policy	92% (1,990)	8% (170)

terms of sudden changes in the recommended vasopressor dosage by the RL agent, given our focus on this hazardous interface failure.

According to [51], doses of Norepinephrine over 0.5 mcg/kg/min are usually considered to be "high" and suggest the need for rescue or second-line therapy. Doses over 1.0 mcg/kg/min are rarely used. In the action space, shown in Table 3 in Section 4.3, moving from action 0 to action 4 in the following step for the same patient, or *vice versa*, gives a dose change > 0.75 mcg/kg/min, as 0.786 mcg/kg/min is the median of action 4 and the median for action 0 is 0. This is clearly in a dangerous range and it is considered hazardous, i.e. "RL agent recommends a sudden change in dose".

We evaluated the maximum vasopressor dose change for the clinician policy and the original policy on the test data set, which has 2,160 patients, by calculating the max absolute vasopressor dose change in one step for each patient during their treatment, see Table 4. In the clinician policy, we found 3% (60 patients) among 2,160 patients have a dose change > 0.75 mcg/kg/min. In contrast, in the original policy, we found 35% (756 patients) among 2,160 patients have this sudden change. The max absolute vasopressor dose change following the original policy is substantially higher than that of following the clinician policy. This implies that the original policy gives rise to the hazardous interface failure, because of the prevalence of these sudden major dose changes.

In response to the above clinical safety concerns, we have modified the model in order to further satisfy safety requirement R0 in Table 2, which is to reduce the rate of sudden major vasopressor dose changes close to clinician policy. We made two alterations to enable the RL agent to learn a safer policy. Firstly, we added an extra feature in the state space, which is the relative dose change compared with the previous vasopressor dose for each patient. This enables the agent to take account of the difference between the current step and the previous step in terms of vasopressor dose while learning the policy, rather than merely using the current step state features. Secondly, we have also altered the cost function used for training. We have added a second regularisation term to penalise the output Q-values when the recommended dose is higher or lower than the previous dose by 0.75

mcg/kg/min (i.e., a jump from action 0 to action 4 or *vice versa* in one step when recommending vasopressor doses for the patients). These changes are summarised in Table 5.

Table 5. Major changes in the modified RL model

	Features in state space (R1)	**Cost Function(R3)**
RL model in [32]	48	$L(\theta) = E[(Q_{double-target} - Q(s,a;\theta))^2]$ + $\lambda_1 max(\|Q(s,a;\theta)\| - Q_{thresh}, 0)$
Modified RL model	48 (Removed one feature – timestep, added an extra one – relative dose change)	$L(\theta) = E[(Q_{double-target} - Q(s,a;\theta))^2]$ + $\lambda_1 max(\|Q(s,a;\theta)\| - Q_{thresh}, 0)$ + $\lambda_2 max(\|V_{change}\| - 0.75, 0)$ V_{change} is the agent recommended dose (argmax of $Q(s,a;\theta)$) minus the vasopressor dose in the previous step; λ_1 and λ_2 are the tuning parameters that decide how much to penalise the flexibility of the model.

This reflects the importance of iteration of the model design and development in order to meet safety requirement R0, through further refinement to meet the lower level requirements, specifically R1 and R3.

After the implementation of these two alterations we have learnt a new modified policy. The maximum absolute vasopressor dose change in one step for each patient for the modified policy is also shown in Table 4, providing a comparison with the clinician and original policies. Table 4 shows that the modified policy gives a clear reduction in sudden major dose changes. Particularly, we found that there are 8% (170 patients) amongst the 2,160 patients in the test data set found with the maximum dose change, i.e. > 0.75 mcg/kg/min in the modified policy. Thus, the modified policy has reduced the rate of such sudden major changes of vasopressor dose by 77.5% when compared with the original policy. Therefore, we consider this modified policy meets requirement R0 through satisfying the lower-level requirements (R1 to R5). For detailed implementation of the modified policy, refer to our previous publication [52].

A further important aspect of assessment is to understand the interpretability of the modified policy, i.e. the extent to which the recommendations made by the RL agent reflect clinical understanding. In ML it is common to train a surrogate model to approximate a complex ML model [53]. Often a simpler ML model is used as the surrogate. In this case we trained two random forest classifiers as surrogate models to understand the relative importance assigned to the features when recommend-

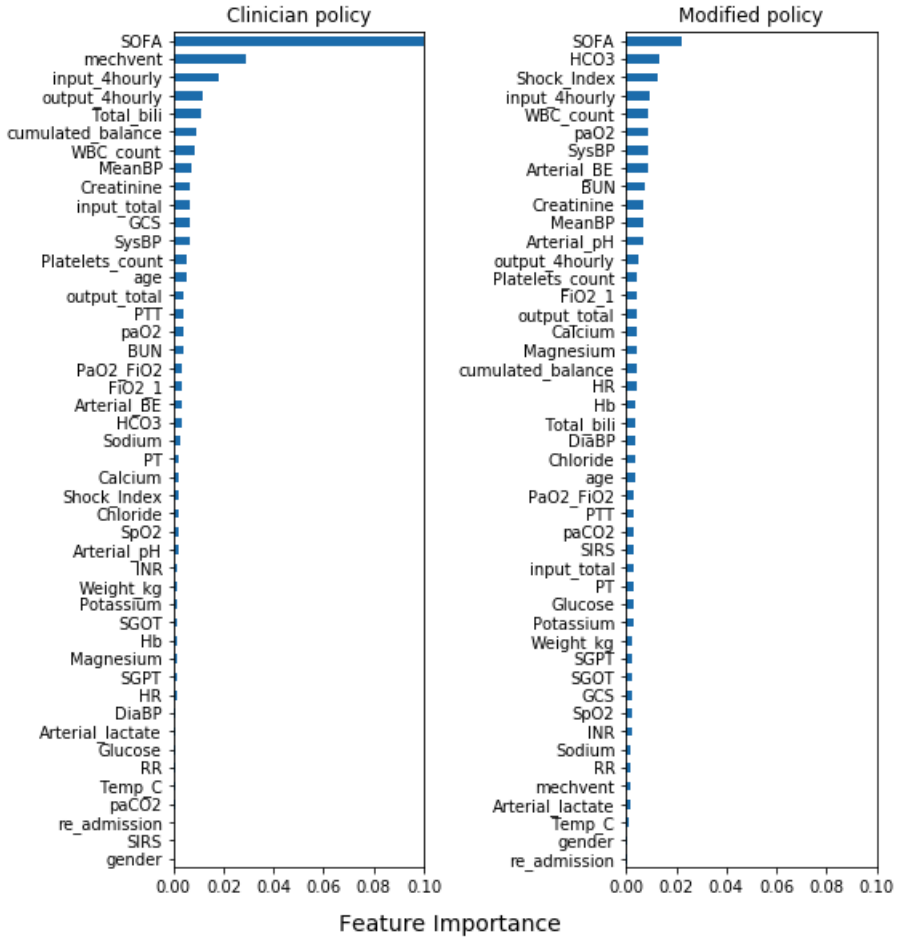

Figure 4. Feature importance (from out of bag score) for clinician policy and the modified policy

ing vasopressor in the modified policy and the clinician policy, see Fig.4. Note the clinician policy is the dose decided by clinicians and extracted from MIMIC III. When training these two random forest classifiers, the classes are binarised in the same way where 0 means no vasopressor prescribed (action 0) and 1 means vasopressor prescribed (action 1, 2, 3, 4). In other words, the current dose of vasopressor was discarded for both random forest classifiers (clinician and modified policy) as the concern here is what features influence whether or not vasopressor is recommended, not the size of the recommended dose.

The relative importance of each feature was estimated using an out-of-bag score on the whole dataset, by permuting the values of each feature, which is also called permutation feature importance [54]. Note that the clinician policy can only represent what was recorded in MIMIC III, not necessarily what was in the clinicians' minds when they made their decisions, thus Fig. 4 shows the relative importance of the clinical features for the classification, rather than directly comparing decision-making. With this caveat, in both policies, SOFA plays the most important role, which is as expected as SOFA describes sepsis-related organ failure. The two policies also give high importance to mean blood pressure and white blood cell count (WBC_count). Gender and re-admission are of low importance in both policies; this is unsurprising as these parameters would not be expected to affect the decision to recommend vasopressor (or not). However, compared with the clinician policy, the modified policy is more balanced rather than having such a heavy focus on SOFA. And by comparison with the clinician policy, the modified policy places emphasis on other important factors, e.g. shock index, which has been shown to indicate the need for vasopressor therapy [55]. Thus the feature importance assessment has confirmed that the decisions suggested by the modified policy rely primarily on sensible clinical parameters, and it is not dominated by a single factor, i.e. SOFA.

4.5. Safety barrier visualisation

Our understanding of the hazards, potential causes of hazards, safety requirements and means of satisfying the requirements does not arise all at once. Instead, this understanding emerges and is refined by iteration around the "Solution design", "Solution implementation" and "Pre-trial testing & assessment" phases shown in Fig. 1. We use Bow Tie Diagrams (BTDs) to consolidate this emerging understanding. BTDs represent a barrier model of safety, where barriers are a collection of related controls, and provide a graphical view of how hazards are controlled [56]. Through the visualisation of the safety barriers and controls, it can help to expose the weak points in the system and identify the need for new barriers and controls if necessary. This implies that there are two types of barriers and controls: pre-existing and newly introduced that arising from the safety analysis. The visualisation

of the safety barriers and controls also helps in the development the safety case by showing how the risks associated with the system or situation are being managed.

Here, we use AdvoCATE [57] to produce the BTDs and safety case (see Section 4.6). AdvoCATE is an advanced Assurance Case Automation Toolset developed by NASA. Two linked BTDs are presented as follows: Fig 5 presents the BTD for the hazardous interface failure "RL agent recommends a sharp change in dose" and Fig. 6 presents that for the hazard "sudden change in vasopressor dosage adminstered" which also shows the role of the hazardous interface failure, and its patient safety impact within the clinical pathway (as modelled in Fig. 3).

We start with Fig. 5. The elements in the figure as are follows:

- Context (square with the black and yellow border) – an activity or condition that is part of normal operation, but which can be a source of harm when control is lost, in this case the activities related to the RL agent in the workflow, grouped together as "Recommendation 2" in Fig. 2;
- Top event (orange circle) – the occurrence of an undesirable event, in this case the hazardous interface failure "RL agent recommends a sudden change of vasopressor dose";
- Threats (round-cornered blue box) – a cause that contributes to the top event, in this case arising from the design and development of the RL agent, i.e. causes 10 to 14 in Table 1;
- Barriers (round-cornered box with yellow heading) – a group of related controls that reduce the likelihood that a threat causes the top event. For example, "design considerations" includes different controls over the way the RL agent is designed and developed;
- Controls (associated with a barrier) – a specific control for a threat, in this case the controls address all the threats that can give rise to the interface hazard.

To further illustrate how the safety barriers in Fig. 5 are linked to the previous sections, we consider one of the threats at the bottom left of the figure that "Cost function for RL model development is not appropriate". This threat is cause 12 in Table 1 and it is addressed by safety requirement R3 in Table 2. There are in total three controls for this threat with two under the "design considerations" barrier and one under the "Evaluation" barrier. Among them, the control "Add a second regularisation term ..." was newly introduced in "Pre-trial testing & assessment" (see Section 4.4) in order to further satisfy requirement R3 also shown in Table 5. This illustrates how the BTD draws together the safety work done at different phases of the workflow to provide a consolidated visualisation of hazards, threats, controls, etc. The BTD also provides extra information in terms

of temporal dependencies, showing how the threats can combine to result in the hazardous interface failure or the ultimate hazard if the controls fail.

Fig. 6 presents a partial BTD for the hazard "Sudden change of vasopressor dose administered" (Fig. 5 and Fig. 6 link to form a more complete BTD). The events in Fig. 6 link directly to the causes in Table 1, for example, one of the threats "kink of line" is cause 1 in the table. The completeness of the BTDs in terms of coverage of threats can be checked by inspection against Table 1. In addition, the hazardous interface failure is also shown as a threat in Fig. 6, which helps us to see how the design and development of the RL model can contribute to patient harm. In other words, the BTD in Fig. 6 enables us to understand the role of the RL model in its clinical context and to proactively and systematically address patient safety in its design. The main entities in the BTD in 6 are:

- Context – the final activity in the workflow in Fig. 2;
- Top event – the hazard "sudden change in vasopressor dosage adminstered";
- Threats – causes from the SHARD analysis in Table 1 that contribute to the top event, e.g. "kink of line" and the hazardous interface failure;
- Barriers – clinician and other barriers, e.g. "infusion pump" which addresses the "kink of line" threat;
- Controls – for example "infusion pump alarm" is part of the "infusion pump" barrier.

The assemblage of new and pre-existing controls are presented in Fig. 6, e.g. "Infusion pump alarm" and "Nurses refer back to the doctor if they have a concern" are pre-existing controls. The "Interpretability" barrier is newly introduced in order to support the doctor to make an informed final decision as shown in the top-level workflow, see Fig. 2. The implementation of this control is explained in Section 4.4 and illustrated in Fig. 4 by showing the feature importance for the modified policy.

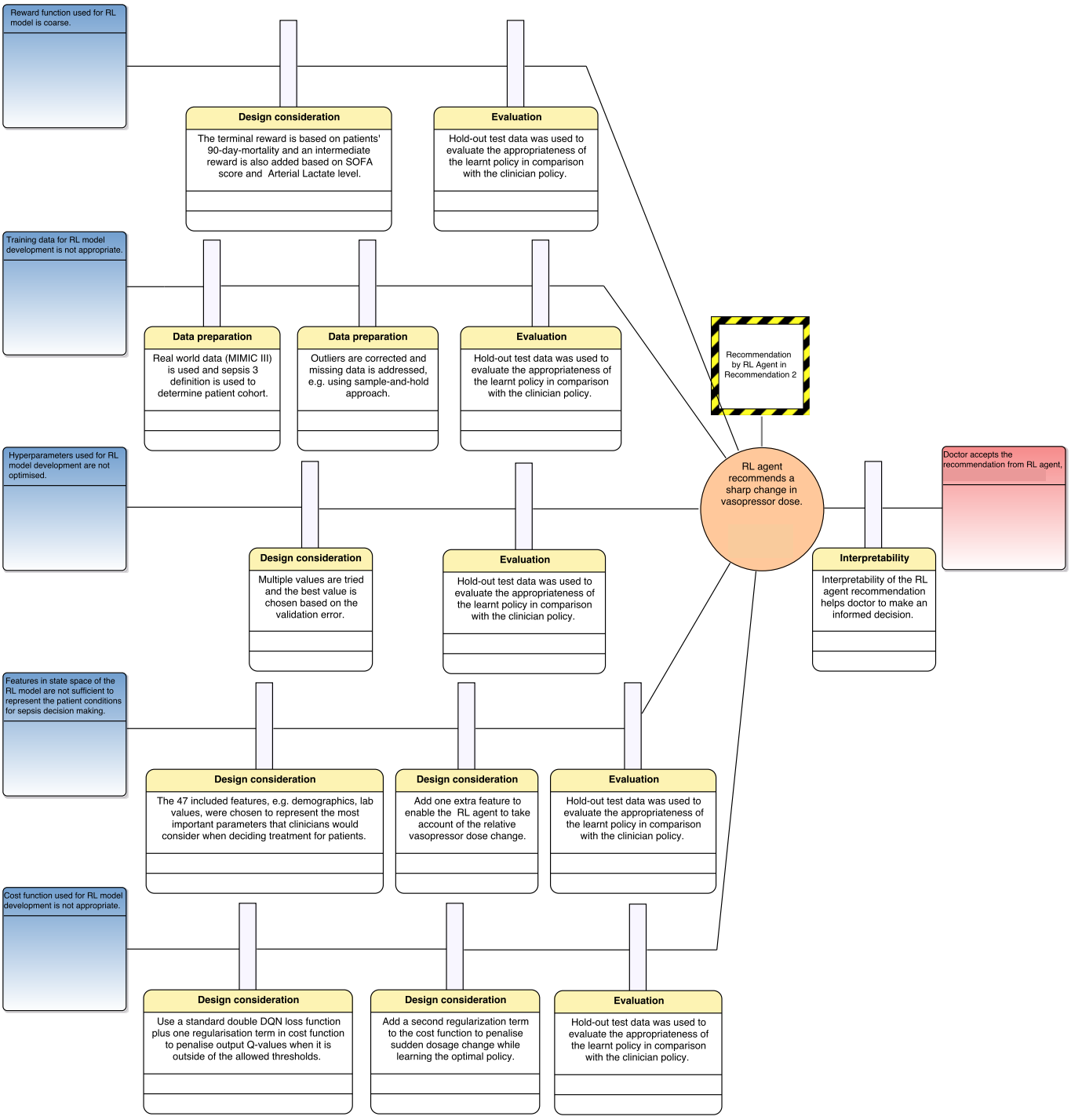

Figure 5. Bow Tie Diagram for interface hazard "RL agent recommends a sharp change in dose"

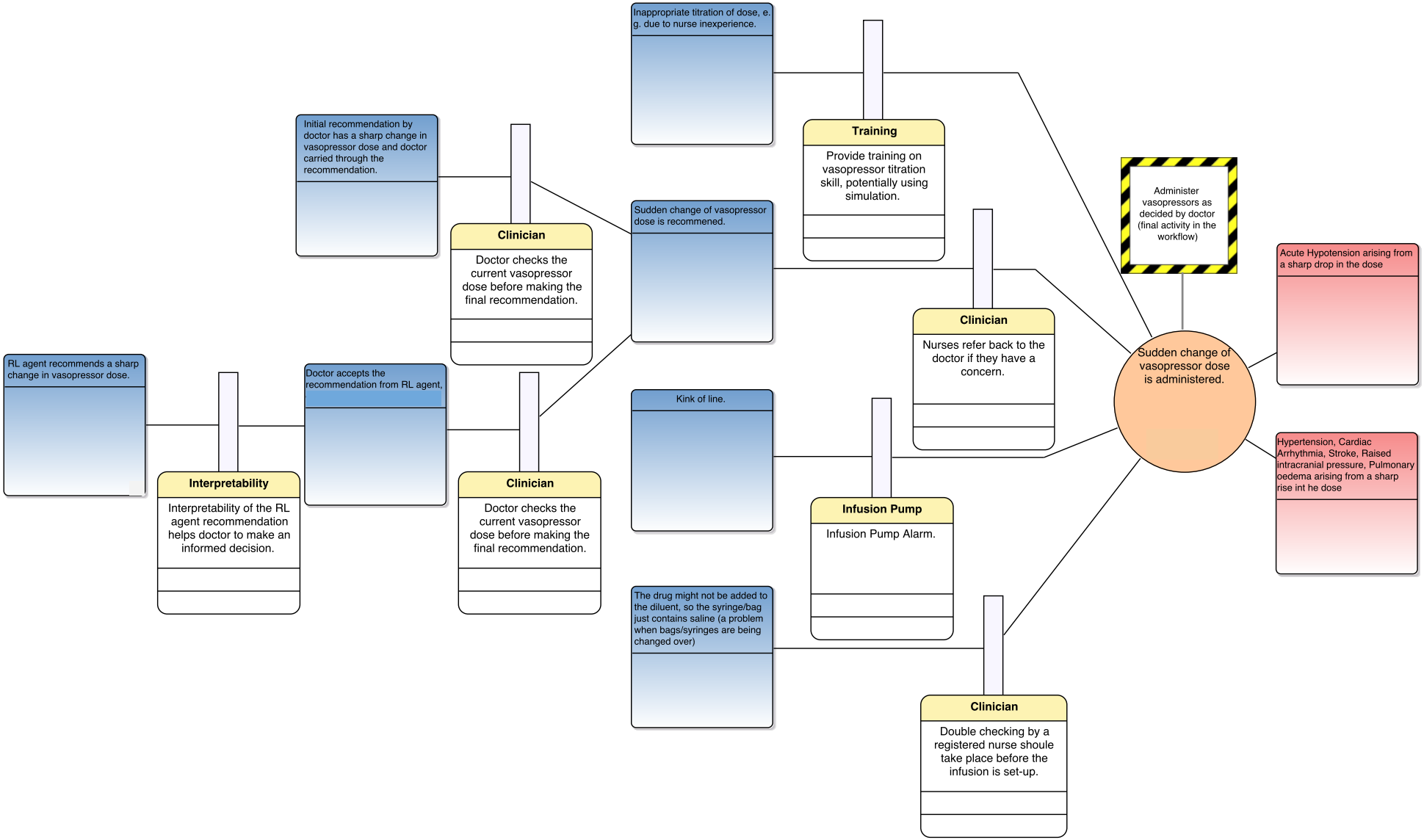

Figure 6. Partial Bow Tie Diagram for ultimate hazard "Sudden change of vasopressor dose is administered"

The BTDs are an important result of iteration through the framework shown in Fig. 1. The phases are not linear and may be visited multiple times, e.g. as is shown in Section 4.4 where model design and development is revisited, responding to the propensity of the initial RL agent to produce sudden vasopressor dose changes. The resultant modification of the RL agent is reflected in the BTD by adding a new control under the "Design considerations" barrier. Further, as mentioned above, developing safe clinical applications of ML requires a multi-disciplinary team, at least including clinicians, ML experts, human factors specialists and safety engineers. However, these disciplines are not necessarily all involved at the same time and the BTDs provide a platform for integrating and visualising information arising from the different specialisms in a way that could support communication and gaining a shared understanding of the issues across disciplines.

4.6. Safety Case

All the phases of the methodology in Fig. 1 feed into the safety case. The safety case draws together and integrates the work in the different phases of the workflow, showing and critically evaluating how the information produced might demonstrate the safety of the "system" which, in this paper, is taken to mean the complete clinical workflow presented in Fig. 3. Before we describe the safety case we have developed, we briefly introduce the notation.

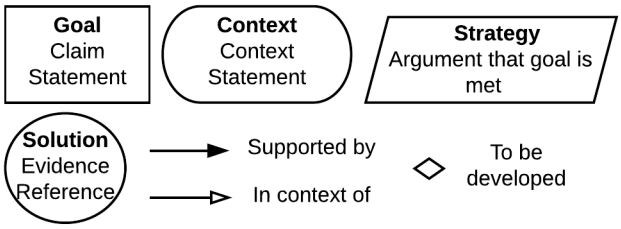

Figure 7. Goal Structuring Notation

In this work we present the safety argument using the Goal Structuring Notation (GSN) [58]; a legend showing the key elements of the notation is presented in Fig. 7. The *goals* – claims that we wish to make and support – are shown as rectangles and they can be decomposed into sub-goals thus forming a tree. *Goals* are understood in a *context* – for example, the operating environment for the system, which is analogous to the context in the BTD. Where the decomposition of goals is not obvious this is explained through a *strategy*, represented as a rhombus. In a complete safety case all leaf-level goals are supported by *solutions*, represented as circles; the solutions provide evidence references to support the argument. Incomplete parts of the argument are shown with a diamond, meaning that part of the argument is to be developed. The detailed description of the notation can be

found in [59].

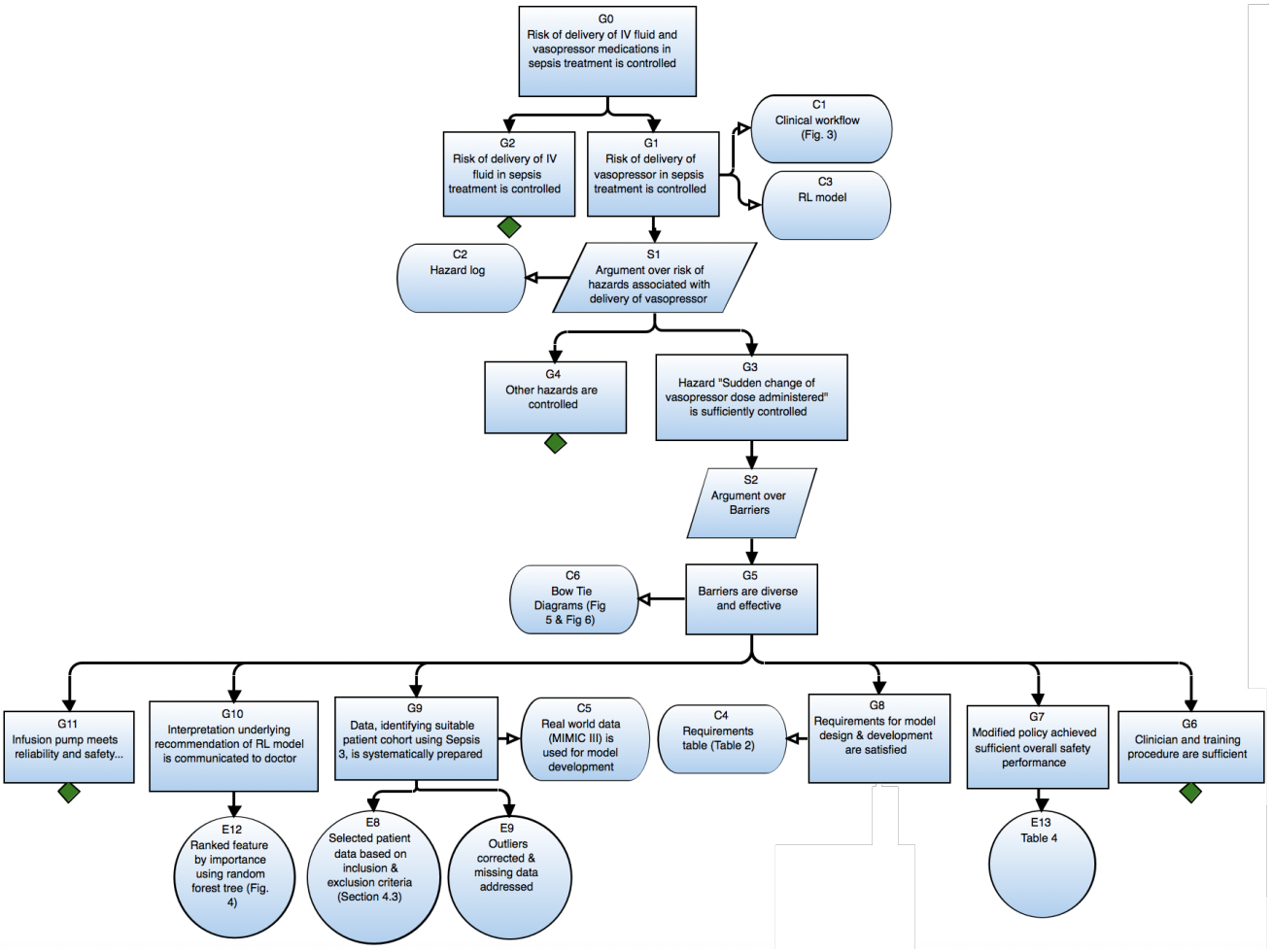

Figure 8. Top Safety Argument

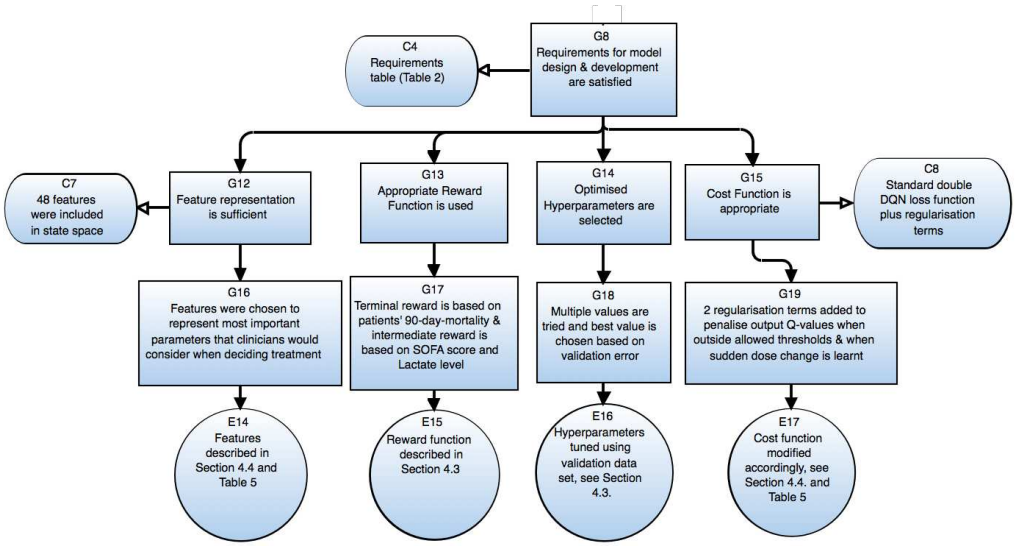

Figure 9. G8 Safety Argument

Here, we present two linked safety arguments in Fig 8 and Fig 9 with the top *goal* G0 "Risk

of delivery of IV fluid and vasopressor medications in sepsis treatment is controlled". The term "controlled" is used as it is unrealistic to assume that the risk in sepsis treatment can be eliminated, given the dependence on individual patient characteristics and circumstances, including comorbidities. This goal decomposes naturally into the IV and vasopressor treatment; as our focus in this paper is on vasopressors, the IV *goal* (G2) is left undeveloped.

The *goal* G1 "Risk of delivery of vasopressor in sepsis treatment is controlled" is stated in the *context* of the clinical workflow in Fig. 3 and the RL model which is set out in the context (C2) of the hazard log. A hazard log summarises information about all hazards including severity, causes and controls. In this case the hazards are identified through the SHARD analysis in Section 4.2. The *strategy* (S1) is an argument over the hazard risks. By showing how the hazards are controlled, this supports G1. For brevity here we focus on showing how a single hazard "Sudden change of vasopressor dose administered" is controlled, i.e. *goal* G3. The remaining hazards can be addressed in a similar way, through *goal* G4 as indicated in GSN using the diamond symbol, i.e. *to be developed*. The *strategy* for meeting *goal* G3 is an argument over the barriers showing that they are diverse and effective, see *goal* G5.

In Fig. 8, *goal* G5 is further decomposed across the six barriers shown in Figs. 5 and 6. Some of these *sub-goals* relate to the pre-existing barriers and controls, e.g. clinician and training procedure, G6 and infusion pump, G11. The rest of the *sub-goals* are all related to the RL model with G7 relating to the evaluation of the model, G8 relating to the "design considerations" described in Section 4.3, G9 relating to the "data preparation to identify a suitable patient cohort" and G10 relating to the "interpretability" described in Section 4.4. G7 is supported by Table 4 which compare the original and modified learnt policies with the clinician policy. G9 is supported by the selected patient data, see section 4.3 and correction of outliers. G10 is supported by the *solution* Fig. 4 showing the "Ranked feature by importance using the random forest tree".

Goal G8, "Requirements for model design and development are satisfied", is further decomposed, in Fig. 9, into four *goals* that are consistent with the safety requirements R1 to R4 (R5 is addressed in G9). These four *goals* all have a single *sub-goal* that is more "concrete" and thus identifies how the higher-level *goal* is met. For example, *goal* G16 defines the broadening of the set of features in the state space for the RL model to reduce the occurrence of the hazardous interface failure "RL agent recommends a sharp change in dose", by including the relative dose change in the state space (see Section 4.4) and thus meeting *goal* G12. The other *goals* G17-G19 have a similar role with respect to *goals* G13-G15. The solutions for *goals* G16-G19 summarise the relevant information in

Sections 4.3 and 4.4. The process of developing the safety case for the overall clinical workflow shows how the different phases in the methodology link together and support each other demonstrating the safety of the RL model in its clinical context.

5. Discussion

The best way to safely and justifiably deploy ML in clinical care remains an open issue. Some work has compared the route of introducing ML into clinical deployment with the process of drug discovery [60], which highlights the difficulties being faced. Our work made an initial attempt to address this issue by integrating safety into the design and development of the ML model in order to minimise the risk of patient harm without compromising its potential benefit. We illustrated our methodology through a concrete clinical use case which concerns sepsis treatment. The clinical use case we show is important and also challenging as sepsis is a major cause of fatalities worldwide and its optimal treatment remains uncertain. The use of RL is suitable given that the problem is to find the optimal treatment. The results show the feasibility and promise of our methodology. Therefore, we review and reflect on the work presented to give insight into the steps that could potentially lead to wider use of ML in healthcare including acceptance by regulators.

First, in healthcare, technology needs to be developed and assured in its clinical context. We believe that this is true in general, but particularly important for ML due to its complex and subtle nature. We demonstrated the merit of doing so by first modelling a clinical workflow which explicitly shows the role of ML in its clinical context. This helps us to understand how the ML model is intended to be used and thus to determine the risk associated with it. We call this "safety-driven design", which proactively manages patient safety by identifying the potential hazards, evaluating the ML model against the hazards, and finally finding ways to improve its safety in a systematic way if any weaknesses of the model are exposed. The work here focuses on a major clinical hazard within a safety case that considers wider socio-technical patient safety factors. However, to gain further confidence in the utility of the methodology we intend to test it in different clinical settings and for different clinical conditions.

Second, ML design & development and safety work must proceed in parallel – there is no simple linear ordering of development and analysis tasks, and the safety work needs to be contemporaneous with design in order to "drive it". Further, a multi-disciplinary approach is essential to safely introduce ML into healthcare [61]. As indicated previously, ML models are often developed in isolation and a culture change will be required to overcome this. Our methodology is intended to support

this multidisciplinary approach but also including safety engineers, in contrast to earlier work, e.g. [18]. The BTDs in particular provide an effective way of integrating and visualising the relationships between the work of the different disciplines.

Third, as our methodology and clinical use case have shown, there is iteration between design, development, safety and assessment activities prior to pilot studies. As a result, safety artefacts, e.g. BTDs and the safety case, evolve during this iteration. However, changes will also occur in the operational phase of the system as clinical understanding evolves, working practices adjust to the new technology and the behaviour of the ML model becomes better understood. Thus, the BTDs and safety case should continue to evolve during operation and the associated risks need to be reassessed. We have previously shown how ML can be used on data from operations to inform operational updates to safety cases [62] taking a step towards dynamic safety cases [63]. Although neither our methodology nor the clinical use case extend into operations at this stage, it is essential that safety and risk continues to be monitored in operation so we are seeking opportunities to integrate the approach set out here with our earlier work on operational monitoring of safety [62]. For example, in this case, we could collect the operational data and use statistical methods to measure whether the change in the mortality rate due to sepsis is statistically significant.

Finally, for ML models to be deployed in healthcare, it is essential to involve and influence regulators. As explained earlier, a report from the UK Care Quality Commission (CQC) [8] has emphasised the importance of safety and assurance of ML and the clarity of its use in the clinical context. We believe that our methodology can provide advantages in practice by assuring the safety of the ML in a clearly defined clinical workflow in a way that enables effective communication between the developers and users of ML models and regulators, thus facilitating their safe introduction.

6. Conclusion

We have developed a methodology for "safety-driven design" and shown how it can be used to guide design & development to improve safety of ML models. It is proactive in that it leads to improvements of the ML models as they are being produced. In contrast, a "design-first, assess safety later" approach can result in expensive rework or even deployment of unsatisfactory systems. This paper has presented a novel methodology that can be used for development of ML models systematically incorporating patient safety considerations. It has integrated key aspects of clinical workflow design, ML design and development, and safety analysis to provide a pragmatic and integrated approach to safely introducing ML into a healthcare setting. It has built on leading research on the use

of RL for sepsis treatment – and shown how the "safety-driven design" methodology can result in safety-significant improvements. In particular, the clinical use case concerns using an RL model to recommend vasopressors and IV fluids for the treatment of sepsis, which showed that "safety-driven design" can identify unsafe behaviour of the RL model, specifically sudden changes in vasopressor dose, and guide the model learning to reduce this undesirable behaviour. It also provided an interpretation of the learnt model to help clinicians to make informed decisions. The results of this iterative and multidisciplinary work were integrated and visualised through the use of BTDs and a safety case showing the rationale for believing that the RL model is acceptable for use in its clinical context.

Finally, we have shown a possible direction for regulators to undertake the assessment of ML models. We believe that it could help satisfy the CQC's stated need for "more assurance about the clinical aspects of the algorithms in machine learning" [8]. We have not conducted an RCT for the ML models developed here. The intent is that our analysis approach could serve as a risk-reduction step, prior to conducting a clinical pilot study and an RCT, as indicated in Figure 1. It is not intended to replace these evaluation methods but to help meet the safety preconditions for rigorous clinical evaluation. In this way, our work may enable healthcare to gain the benefits of ML without compromising patient safety.

Acknowledgement

This work is funded by Bradford Teaching Hospitals NHS Foundation Trust and supported by the Assuring Autonomy International Programme at the University of York. The views expressed in this paper are those of the authors and not necessarily those of the NHS, or the Department of Health and Social Care.

References

[1] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[2] Y. Liu, K. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P. Q. Nelson, G. S. Corrado, *et al.*, "Detecting cancer metastases on gigapixel pathology images," *arXiv preprint arXiv:1703.02442*, 2017.

[3] J. M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, S. French, L. A. Carfrae, Z. Bloom-Ackerman, *et al.*, "A deep learning approach to antibiotic discovery," *Cell*, vol. 180, no. 4, pp. 688–702, 2020.

[4] MIT, "AI Cures." https://www.aicures.mit.edu, 2020. Accessed: 2020-05-21.

[5] I. Habli, S. White, M. Sujan, S. Harrison, and M. Ugarte, "What is the safety case for health it? a study of assurance practices in england," *Safety Science*, vol. 110, pp. 324–335, 2018.

[6] NHS Digital, "DCB0160: Clinical risk management: its Application in the Deployment and Use of health IT Systems," 2018.

[7] US Food and Drug Administration, "Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SAMD)—discussion paper and request for feedback. 2019," 2019.

[8] Care Quality Commission and Medical and Healthcare products Regulatory Agency, "Using machine learning in diagnostic services: A report with recommendations from CQC's regulatory sandbox," 2020.

[9] M. Hutson *et al.*, "Even artificial intelligence can acquire biases against race and gender," *Science Magazine, Science AAAS*, vol. 13, 2017.

[10] P. Marik, "The demise of early goal-directed therapy for severe sepsis and septic shock," *Acta Anaesthesiologica Scandinavica*, vol. 59, no. 5, pp. 561–567, 2015.

[11] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal, "The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care," *Nature medicine*, vol. 24, no. 11, pp. 1716–1720, 2018.

[12] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. C. Corrado, A. Darzi, *et al.*, "International evaluation of an ai system for breast cancer screening," *Nature*, vol. 577, no. 7788, pp. 89–94, 2020.

[13] K. R. Famous, K. Delucchi, L. B. Ware, K. N. Kangelaris, K. D. Liu, B. T. Thompson, and C. S. Calfee, "Acute respiratory distress syndrome subphenotypes respond differently to randomized fluid management strategy," *American journal of respiratory and critical care medicine*, vol. 195, no. 3, pp. 331–338, 2017.

[14] C. W. Seymour, J. N. Kennedy, S. Wang, C.-C. H. Chang, C. F. Elliott, Z. Xu, S. Berry, G. Clermont, G. Cooper, H. Gomez, *et al.*, "Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis," *Jama*, vol. 321, no. 20, pp. 2003–2017, 2019.

[15] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[16] C. Yu, J. Liu, and S. Nemati, "Reinforcement learning in healthcare: a survey," *arXiv preprint arXiv:1908.08796*, 2019.

[17] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V. X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed, *et al.*, "Do no harm: a roadmap for responsible machine learning for health care," *Nature medicine*, vol. 25, no. 9, pp. 1337–1340, 2019.

[18] M. Sendak, W. Ratliff, D. Sarro, E. Alderton, J. Futoma, M. Gao, M. Nichols, M. Revoir, F. Yashar, C. Miller, *et al.*, "Sepsis watch: A real-world integration of deep learning into routine clinical care," *JMIR Preprints*, vol. 15182, 2019.

[19] E. Hariton and J. J. Locascio, "Randomised controlled trials—the gold standard for effectiveness research," *BJOG: an international journal of obstetrics and gynaecology*, vol. 125, no. 13, p. 1716, 2018.

[20] D. W. Shimabukuro, C. W. Barton, M. D. Feldman, S. J. Mataraso, and R. Das, "Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial," *BMJ open respiratory research*, vol. 4, no. 1, p. e000234, 2017.

[21] P. Wang, T. M. Berzin, J. R. G. Brown, S. Bharadwaj, A. Becq, X. Xiao, P. Liu, L. Li, Y. Song, D. Zhang, *et al.*, "Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study," *Gut*, vol. 68, no. 10, pp. 1813–1819, 2019.

[22] M. Wijnberge, B. F. Geerts, L. Hol, N. Lemmers, M. P. Mulder, P. Berge, J. Schenk, L. E. Terwindt, M. W. Hollmann, A. P. Vlaar, *et al.*, "Effect of a machine learning–derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the hype randomized clinical trial," *JAMA*, vol. 323, no. 11, pp. 1052–1060, 2020.

[23] D. C. Angus, "Randomized Clinical Trials of Artificial Intelligence," *JAMA*, vol. 323, pp. 1043–1045, 03 2020.

[24] N. G. Leveson and J. P. Thomas, "Stpa handbook," *Cambridge, MA, USA*, 2018.

[25] P.-H. C. Chen, Y. Liu, and L. Peng, "How to develop machine learning models for healthcare," *Nature materials*, vol. 18, no. 5, p. 410, 2019.

[26] C. Picardi, R. Hawkins, C. Paterson, and I. Habli, "A pattern for arguing the assurance of ma-

chine learning in medical diagnosis systems," in *International Conference on Computer Safety, Reliability, and Security*, pp. 165–179, Springer, 2019.

[27] I. Habli, T. Lawton, and Z. Porter, "Artificial intelligence in health care: accountability and safety," *Bulletin of the World Health Organization*, vol. 98, no. 4, p. 251, 2020.

[28] R. Bloomfield and P. Bishop, "Safety and assurance cases: Past, present and possible future–an adelard perspective," in *Making Systems Safer*, pp. 51–67, Springer, 2010.

[29] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith, *et al.*, "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *Jama*, vol. 315, no. 8, pp. 801–810, 2016.

[30] J. Gallagher, "'Alarming' one in five deaths due to sepsis." https://www.bbc.co.uk/news/health-51138859, 2020. Accessed: 2020-03-01.

[31] J. Waechter, A. Kumar, S. E. Lapinsky, J. Marshall, P. Dodek, Y. Arabi, J. E. Parrillo, R. P. Dellinger, and A. Garland, "Interaction between fluids and vasoactive agents on mortality in septic shock: a multicenter, observational study," *Critical care medicine*, vol. 42, no. 10, pp. 2158–2168, 2014.

[32] A. Raghu, M. Komorowski, I. Ahmed, L. Celi, P. Szolovits, and M. Ghassemi, "Deep reinforcement learning for sepsis treatment," *arXiv preprint arXiv:1711.09602*, 2017.

[33] NHS Improvement, "Sepsis is a medical emergency!." https://improvement.nhs.uk/documents/652/Sepsis_Ae_Easy_Guide.pdf. Accessed: 2020-05-21.

[34] Royal College of Physicians, "National early warning score." https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2. Accessed: 2020-05-21.

[35] The UK Sepsis Trust, "ED/ AMU Sepsis Screening & Action Tool." https://sepsistrust.org/wp-content/uploads/2018/06/ED-adult-NICE-Final-1107.pdf. Accessed: 2020-05-21.

[36] Surviving Sepsis Campaign, "Hour-1 Bundle." https://www.sccm.org/getattachment/SurvivingSepsisCampaign/Guidelines/Adult-Patients/Surviving-Sepsis-Campaign-Hour-1-Bundle.pdf?lang=en-US. Accessed: 2020-05-21.

[37] M. Sujan, S. White, D. Furniss, I. Habli, K. Grundy, H. Grundy, D. Nelson, M. Elliott, and N. Reynolds, "Human factors challenges for the safe use of artificial intelligence in patient care," *BMJ Health and Care Informatics*, 2019.

[38] A. J. Abugabah and O. Alfarraj, "Issues to consider in designing health care information systems: A user-centred design approach," *electronic Journal of Health Informatics*, vol. 9, no. 1, p. 8, 2015.

[39] I. Habli, Y. Jia, S. White, G. Gabriel, T. Lawton, M. Sujan, and C. Tomsett, "Development and piloting of a software tool to facilitate proactive hazard and risk analysis of health information technology," *Health informatics journal*, p. 1460458219852789, 2019.

[40] T. A. Kletz, *HAZOP and HAZAN: identifying and assessing process industry hazards*. IChemE, 1999.

[41] D. J. Pumfrey, *The principled design of computer system safety analyses.* PhD thesis, University of York, 1999.

[42] K. L. Fadale, D. Tucker, J. Dungan, and V. Sabol, "Improving nurses' vasopressor titration skills and self-efficacy via simulation-based learning," *Clinical Simulation in Nursing*, vol. 10, no. 6, pp. e291–e299, 2014.

[43] Hospira UK Ltd, "Noradrenaline (Norepinephrine) 1 mg/ml Concentrate for Solution for Infusion." `https://www.medicines.org.uk/emc/product/4115/smpc`, 2018. Accessed: 2020-03-01.

[44] J. M. Allen, "Understanding vasoactive medications: focus on pharmacology and effective titration," *Journal of Infusion Nursing*, vol. 37, no. 2, pp. 82–86, 2014.

[45] H. Beloeil, J.-X. Mazoit, D. Benhamou, and J. Duranteau, "Norepinephrine kinetics and dynamics in septic shock and trauma patients," *British journal of anaesthesia*, vol. 95, no. 6, pp. 782–788, 2005.

[46] P. Fenelon, J. A. McDermid, M. Nicolson, and D. J. Pumfrey, "Towards integrated safety analysis and design," *ACM SIGAPP Applied Computing Review*, vol. 2, no. 1, pp. 21–32, 1994.

[47] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[48] C. J. C. H. Watkins, "Learning from delayed rewards," 1989.

[49] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific data*, vol. 3, p. 160035, 2016.

[50] "An end-to-end open source machine learning platform." `https://www.tensorflow.org`. Accessed: 2020-05-21.

[51] E. Bassi, M. Park, and L. C. P. Azevedo, "Therapeutic strategies for high-dose vasopressor-dependent shock," *Critical care research and practice*, vol. 2013, 2013.

[52] Y. Jia, J. Burden, T. Lawton, and I. Habli, "Safe reinforcement learning for sepsis treatment," in *8th IEEE International Conference on Healthcare Informatics*, IEEE, 2020.

[53] C. Molnar, *Interpretable Machine Learning*. Lulu.com, 2020.

[54] L. Breiman, "Random forests machine learning 45 (1), 5-32 (2001) 10.1023," *A: 1010933404324*.

[55] C. R. Wira, M. W. Francis, S. Bhat, R. Ehrman, D. Conner, and M. Siegel, "The shock index as a predictor of vasopressor use in emergency department patients with severe sepsis," *Western Journal of Emergency Medicine*, vol. 15, no. 1, p. 60, 2014.

[56] E. Denney, G. Pai, and I. Whiteside, "The role of safety architectures in aviation safety cases," *Reliability Engineering & System Safety*, vol. 191, p. 106502, 2019.

[57] E. Denney and G. Pai, "Tool support for assurance case development," *Automated Software Engineering*, vol. 25, no. 3, pp. 435–499, 2018.

[58] T. Kelly and R. Weaver, "The goal structuring notation–a safety argument notation," in *Proceedings of the dependable systems and networks 2004 workshop on assurance cases*, p. 6, Citeseer, 2004.

[59] Assurance Case Working Group [ACWG], "Goal Structing Notation Community Standard version 2." `https://scsc.uk/r141B:1?t=1`, 2018. Accessed on 11/13/2018.

[60] M. Komorowski, "Clinical management of sepsis can be improved by artificial intelligence: yes," 2019.

[61] S. Burton, I. Habli, T. Lawton, J. McDermid, P. Morgan, and Z. Porter, "Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective," *Artificial Intelligence*, vol. 279, p. 103201, 2020.

[62] J. A. McDermid, Y. Jia, and I. Habli, "Towards a framework for safety assurance of autonomous systems," in *Artificial Intelligence Safety 2019*, pp. 1–7, CEUR Workshop Proceedings, 2019.

[63] E. Denney, G. Pai, and I. Habli, "Dynamic safety cases for through-life safety assurance," in *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, vol. 2, pp. 587–590, IEEE, 2015.