

Embracing Machine Learning in Safety Assurance in Healthcare

Yan Jia

Doctor of Philosophy

University of York
Computer Science

August 2021

Dedication

To my family and many friends. A special thanks to my parents who have been extremely generous to me and have supported me financially and morally throughout my education, even though I have been so far from home for so long. The good examples they set will always influence my life. I also thank my wider family including my three brothers and their families for the encouragement and care when I needed it. Finally, I thank my mentor for the inspiration, unstinting support and help.

Abstract

Machine learning (ML) is becoming more widely used in many different sectors, including automotive, aviation and healthcare. ML has a great potential to change society and to improve peoples' lives. However, the prospect of ML also poses many challenges; one of the biggest challenges is safety. Thus, there are two important questions that require urgent answers: (1) Are well-established safety engineering methods still appropriate and effective in assuring the safety of ML in some representative healthcare scenarios? (2) Are there new opportunities for well-established safety engineering methods with the development of ML and why are they specifically good for safety in this domain?

In this thesis, the first question is explored from the viewpoint of *designing* ML models. The second question is explored from two perspectives: explainability of ML models in support of safety assurance; and using ML to update safety analysis. Both these questions are addressed in the context of healthcare. In other words, this thesis investigates how ML can be embraced in the safety assurance of healthcare applications.

Through exploration of three concrete clinical case studies, the thesis demonstrates that well-established safety engineering methods can be applied to ML systems to integrate safety into their design process in healthcare. It further identifies different ways in which ML can assist well-established safety engineering methods, and concludes that there are many opportunities for greater synergy between ML and safety engineering in healthcare and, potentially, in other domains.

Contents

| | |
|---|-------------|
| Abstract | iii |
| List of figures | vi |
| List of tables | viii |
| Acknowledgements | xi |
| Declaration | xiii |
| 1 Introduction | 1 |
| 1.1 Context | 1 |
| 1.2 Research Questions and Contributions | 4 |
| 1.3 Structure of the Thesis | 5 |
| 2 Background & Literature Review | 7 |
| 2.1 Machine Learning | 7 |
| 2.2 Safety Engineering and Safety Cases | 14 |
| 2.3 Healthcare | 20 |
| 2.4 Regulation and Safety Assurance | 28 |
| 2.5 Conclusion | 40 |
| 3 Overview of the Thesis | 41 |
| 4 Safety-Driven Design of Machine Learning | 45 |
| 4.1 Introduction | 46 |
| 4.2 Background and Related Work | 48 |
| 4.3 Methodology | 50 |
| 4.4 Clinical Case Study: Sepsis Treatment | 52 |

| | | |
|----------|--|------------|
| 4.5 | Discussion | 78 |
| 4.6 | Conclusion | 79 |
| 5 | The Role of Explainability in Assuring Safety of Machine Learning | 81 |
| 5.1 | Introduction | 82 |
| 5.2 | Background and Related Work | 84 |
| 5.3 | Explainability in the ML Life-cycle | 90 |
| 5.4 | Clinical Case Study: Weaning from Mechanical Ventilation | 96 |
| 5.5 | Discussion | 111 |
| 5.6 | Conclusion | 115 |
| 6 | Using Machine Learning to Update Safety Analysis | 117 |
| 6.1 | Introduction | 118 |
| 6.2 | Background and Related Work | 119 |
| 6.3 | Methodology | 120 |
| 6.4 | Clinical Case Study: Beta-Blocker Delivery | 122 |
| 6.5 | Discussion | 136 |
| 6.6 | Conclusion | 138 |
| 7 | Conclusions | 141 |
| 7.1 | Research Question 1 | 141 |
| 7.2 | Research Question 2 | 144 |
| 7.3 | Summary and Directions for Future Work | 148 |
| | Appendices | 153 |
| A | SHARD analysis for the clinical workflow in case study 1 | 153 |
| B | Feature correlation matrix for the RL model in case study 1 | 163 |
| C | Feature correlation matrix for the weaning model in case study 2 | 169 |
| | Abbreviations | 171 |
| | References | 175 |

List of Figures

| | | |
|------|---|----|
| 2.1 | A simple illustration of Machine Learning | 10 |
| 2.2 | Goal Structuring Notation Legend | 17 |
| 2.3 | Examples of existing clinical applications of ML in diagnostics where darker purple boxes indicate that more care is needed when using these solutions in live clinical services. Taken from [118]. | 36 |
| 3.1 | Overview of the Thesis | 41 |
| 4.1 | Overview for the Case Study | 45 |
| 4.2 | Framework for integrating ML system into clinical care | 50 |
| 4.3 | High-level workflow design | 53 |
| 4.4 | The detailed workflow integrating ML model to treat sepsis patient | 56 |
| 4.5 | Original Policy: Comparison of max absolute vasopressor dose change in one step for each patient in the test dataset between the clinician and the learnt optimal policy | 65 |
| 4.6 | Modified Policy: Comparison of max absolute vasopressor dose change in one step for each patient in the test dataset between the clinician and the learnt modified policy | 65 |
| 4.7 | Feature importance (from out of bag score) for clinician policy and the modified policy | 68 |
| 4.8 | Bow Tie Diagram for interface hazard “RL agent recommends a sharp change in dose” | 73 |
| 4.9 | Partial Bow Tie Diagram for ultimate hazard “Sudden change of vasopressor dose is administered” | 74 |
| 4.10 | Top Safety Argument | 76 |
| 4.11 | G8 Safety Argument | 77 |

| | | |
|-----|--|------|
| 5.1 | Overview for the Case Study | 82 |
| 5.2 | Process for development and operation of an ML System | 91 |
| 5.3 | Patient inclusion diagrams in MIMIC-III | 99 |
| 5.4 | Performance of ML models | 101 |
| 5.5 | Top 30 most influential training instances | 104 |
| 5.6 | Distribution of influential instances | 104 |
| 5.7 | Feature Importance for the CNN Model | 107 |
| 5.8 | Feature Importance for a Single Patient | 109 |
| 5.9 | Partial Safety Argument for Weaning ML Model emphasising Explainability | 110 |
| | | |
| 6.1 | Overview for the Case Study | 118 |
| 6.2 | Framework for using ML to update Safety analysis | 121 |
| 6.3 | Pathway for Nutrition and Medication following oesophagectomy | 124 |
| 6.4 | Decision-making flowchart for prescription and administration of medica- tion | 125 |
| 6.5 | Learnt Bayesian Network Structure based on Safety analysis | 132 |
| 6.6 | Safety Argument for Prevention of AF (with Emphasis on Omission of BBs) | 135 |
| | | |
| B.1 | Feature correlation matrix for case study concerning sepsis treatment, Part | 1165 |
| B.2 | Feature correlation matrix for case study concerning sepsis treatment, Part | 2166 |
| B.3 | Feature correlation matrix for case study concerning sepsis treatment, Part | 3167 |
| | | |
| C.1 | Feature correlation matrix for case study concerning ventilator weaning . . | 170 |

List of Tables

| | | |
|-----|--|-----|
| 2.1 | Approaches to ML | 10 |
| 2.2 | Examples of Severity Classes derived from the use of medications | 24 |
| 2.3 | The similarities & differences of medical device regulation in USA, EU and UK | 37 |
| 4.1 | Fragment of SHARD analysis showing a single Hazard | 59 |
| 4.2 | Safety Requirements for RL model derived from Hazard analysis | 60 |
| 4.3 | Dosage actions | 63 |
| 4.4 | Summary of max dose change between consecutive doses for the three policies | 66 |
| 4.5 | Major changes in the modified RL model | 66 |
| 4.6 | Performance comparison for different policies | 67 |
| 5.1 | Categorisation of Explainable AI Methods with examples | 90 |
| 5.2 | Legend for Figure 5.4 | 101 |
| 5.3 | Performance comparison with different ML classifiers | 102 |
| 5.4 | CNN architecture | 103 |
| 5.5 | Counterfactual examples for a given original instance | 108 |
| 5.6 | Role of Explainable AI Methods in the development and operation phases | 112 |
| 6.1 | SHARD results of the decision-making model (* assumes correct medicate) | 127 |
| 6.2 | Variables extracted from MIMIC-III Dataset | 130 |
| 6.3 | Predictive accuracy of estimation methods | 134 |
| 6.4 | Effects of post_beta on AF for patients with pre_beta and undergoing thoracic surgery | 134 |
| 6.5 | Effects of hypotension on post_beta for patients with pre_beta and undergoing thoracic surgery | 134 |
| A.1 | Overall Administration of Vasopressors, Part 1 | 154 |

| | | |
|-----|--|-----|
| A.2 | Overall Administration of Vasopressors, Part 2 | 155 |
| A.3 | Nurses Administer Vasopressors as Advised by Doctor | 157 |
| A.4 | Final Decision/Final Dose Decided by Doctor | 158 |
| A.5 | Recommendation by RL Agent | 159 |
| A.6 | RL Agent processes the patient data | 159 |
| A.7 | Input Patient Features | 160 |
| A.8 | Interface between RL agent and Clinical care, Part 1 | 161 |
| A.9 | Interface between RL agent and Clinical care, Part 2 | 162 |
| B.1 | List of features used in the RL model | 163 |

Acknowledgements

I wish to thank my supervisors Drs Ibrahim Habli and Tom Lawton for their encouragement, support and guidance throughout my PhD. Their complementary skills enabled me to carry out inter-disciplinary research spanning safety, machine learning and healthcare. In particular, Dr Lawton gave freely of his clinical expertise enabling me to ground my research in a range of challenging healthcare problems and Dr Habli motivated me to take on this interdisciplinary challenge and gave me the confidence and the support, especially on safety issues. Finally, special thanks to Dr Ewen Denney from NASA for sharing the AdvoCATE tool with me.

This work is funded by Bradford Teaching Hospitals NHS Foundation Trust and supported by the Assuring Autonomy International Programme at the University of York.

Declaration

This thesis is a presentation of original work and was undertaken during my study at the University of York from 2018 - 2021 and has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

The content of some of the chapters has already been published in conference proceedings and journals, shown as below ordered by date of publication:

- **Yan Jia**. Improving medication safety using machine learning. In *AIME Doctoral Consortium: Artificial Intelligence in Medicine in Europe*, 2019
- **Yan Jia**, Tom Lawton, Sean White, and Ibrahim Habli. Developing a safety case for electronic prescribing. In *Studies in Health Technology and Informatics: MED-INFO2019*, volume 264, pages 629–633, August 2019
- John McDermid, **Yan Jia**, and Ibrahim Habli. Towards a framework for safety assurance of autonomous systems. In *Artificial Intelligence Safety 2019*, pages 1–7. CEUR Workshop Proceedings, 2019
- Ibrahim Habli, **Yan Jia**, Sean White, George Gabriel, Tom Lawton, Mark Sujjan, and Clive Tomsett. Development and piloting of a software tool to facilitate proactive hazard and risk analysis of health information technology. *Health informatics journal*, 26(1):683–702, 2020
- **Yan Jia**, John Burden, Tom Lawton, and Ibrahim Habli. Safe reinforcement learning for sepsis treatment. In *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–7, 2020
- John McDermid and **Yan Jia**. Safety of artificial intelligence: A collaborative model. In *AI Safety@IJCAI*, 2020

-
- **Yan Jia**, Tom Lawton, John Burden, John McDermid, and Ibrahim Habli. Safety-driven design of machine learning for sepsis treatment. *Journal of Biomedical Informatics*, 117:103762, 2021
 - **Yan Jia**, Chaitanya Kaul, Tom Lawton, Roderick Murray-Smith, and Ibrahim Habli. Prediction of weaning from mechanical ventilation using convolutional neural networks. *Artificial Intelligence in Medicine*, 117:102087, 2021
 - **Yan Jia**, John McDermid, and Ibrahim Habli. Enhancing the value of counterfactual explanations for deep learning. In *AIME 2021: Artificial Intelligence in Medicine in Europe*. Porto, 2021
 - **Yan Jia**, Tom Lawton, John McDermid, Eric Rojas, and Ibrahim Habli. A framework for assurance of medication safety using machine learning. *arXiv preprint-arXiv:2101.05620*, 2021
 - John McDermid, **Yan Jia**, Zoe Porter, and Ibrahim Habli. AI Explainability: The Technical and Ethical Dimensions. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 2021
 - **Yan Jia**, John McDermid, Tom Lawton, and Ibrahim Habli. The role of explainability in assuring safety of machine learning in healthcare. *IEEE Transactions on Emerging Topics in Computing (Revised)*, 2022

Copyright © 2021 by Yan Jia

The copyright of this thesis rests with the author. Any quotations from it should be acknowledged appropriately.

Chapter 1

Introduction

Machine Learning (ML) incorporates a family of methods that enable computer systems to derive general capabilities from training data, hence they can be applied in situations beyond their initial training dataset. The power of modern ML methods means that their applications are very varied, including autonomous driving and medical image analysis.

There is a growing interest in the use of ML in healthcare due to the complexity of the problems faced by clinicians and the availability of data on which to train systems. There is evidence that ML-based systems can undertake some tasks more effectively than humans [1] and this leads to a desire to migrate such systems from research into clinical practice, thus contributing to patient safety. However, it is also necessary to show that the systems do not adversely impact safety, which is the classical role of safety engineering.

This thesis explores the ways in which ML and safety engineering can be combined to provide benefits in healthcare. It is hoped that this work will help to enable the *safe* introduction of ML into clinical practice.

1.1 Context

ML is a branch of Artificial Intelligence (AI) based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. It is born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks. Alpaydin [2] describes ML as “optimising a performance criterion using example data and past experience”. Unlike traditional methods of software development, ML is capable of analysing high-dimensional data, identifying and exploring unknown patterns in data, so it has wide applicability.

ML is a natural outgrowth of the intersection of computer science and statistics, involving the underlying theories and techniques from both fields. From computer science, it exploits the key ideas of using efficient algorithms for optimisation, model representation, and performance evaluation. From statistics, it inherits the basic concept of learning and inferring the statistical properties of a given dataset. However, computer science has focused primarily on how to program computers manually, whereas ML focuses on the question of how to get computers to “program themselves”. Statistics has focused primarily on what conclusions can be inferred from data, whereas ML incorporates additional questions about what computational architectures and algorithms can be used to most effectively capture, store, index, retrieve and merge the data, how multiple learning sub-tasks can be orchestrated in a larger system, and questions of computational tractability [3] to ensure the utility of the resultant systems.

Safety engineering is an established discipline, having its roots in the space and nuclear industries in the mid 1900s. Safety engineering focuses on hazards – situations which, if not controlled, could lead to harm, e.g. injury or loss of life. Loss of brakes on a car and delivering drugs at an incorrect rate are both examples of hazards. A key aspect of safety engineering is safety assurance which aims to influence the design of the system (“design for safety”) and/or to provide evidence of its safety both pre-deployment and post-deployment (“demonstrate safety”).

There are many safety engineering methods. Some are exploratory, asking “what if?” questions; these tend to be used to identify hazards. For example, asking “what if drug X was delivered at too high a rate?” might get the response “it could lead to atrial fibrillation”. Deductive methods are used to investigate potential causes of hazards, e.g. “incorrect entry of delivery rate on the infusion pump” or “infusion pump motor runs over-speed”, in the example above. Deductive methods are complemented with inductive methods which identify effects of the hazards or the causal chain from low-level problems, e.g. “line from infusion pump to patient blocked” to hazards. Deductive and inductive methods are complementary; inductive methods can confirm the results of deductive analysis but also identify problems (including hazards) missed by deductive analysis. One of the most common deductive methods is Fault-Tree Analysis (FTA) [4] which builds cause-effect models of system component failure. Failure Modes and Effects Analysis (FMEA) [5] is probably the most commonly used inductive method and is typically used to explore the impact of component failures on the wider system. Event Tree Analysis (ETA) [6] is

an inductive method often used to explore the consequences of hazards.

Many safety engineering methods have been adapted to apply to software or software-intensive systems. The exploratory methods are of most interest in this thesis. A well-known exploratory method is Hazard and Operability Analysis (HAZOP) [7] initially developed in the chemical industry. HAZOP and its derivatives analyse flows in the system using guidewords to prompt analysis of possible deviations from intent. One of the adaptations of HAZOP to software, known as Software Hazard Analysis and Resolution in Design (SHARD) [8], considers deviations including omission (flow not provided when intended), commission, early/late and incorrect value. As well as considering the potential impact of these deviations, the methods also consider whether or not the deviations are credible (can occur in the system design). If they are credible and undesirable (potentially hazardous), then ways are identified to make the deviation less likely to occur or to mitigate its effects if it does arise; these controls are often seen as means to satisfy Derived Safety Requirements (DSRs). This helps to design for safety.

A range of methods are used to provide evidence of safety, including the use of inductive methods. For example, FMEA might be used to show that no single point of failure gives rise to a hazard. More generally, tests and other forms of analysis are used to show that the DSRs are met. In many industries, the results of the analyses and tests are drawn together into a safety case, which is a “structured argument, supported by evidence, that a system is safe to use in a given context” [9]. Safety arguments are often presented graphically and provide the rationale that explains why the evidence is sufficient to show that applicable safety requirements are met. In industries where there are formal regulatory schemes it is quite common to require a safety case in support of regulatory approval, see for example the Federal Drug Administration (FDA) regulations on infusion pumps [10].

Safety cases are typically prepared pre-deployment but safety engineering does not stop when a system starts operation. Rather, safety engineering includes monitoring of system operation to provide ongoing assurance of safety or to identify problems that need to be rectified.

The same safety engineering and assurance principles apply when the systems employ ML but, as we shall see in this thesis, some adaptation is needed to deal with the particular characteristics of ML-based systems.

Healthcare can be viewed as “the organised provision of medical care to individuals or a community”. Archaeological records show that medicine has a long history, for

example surgery being carried out on fractured bones in Iron Age Britain [11]. However, healthcare, as we now know it, has more recent origins. The “organised provision” was in part driven by concerns for patient safety and included the licensing of practitioners. Licensing started around 500 years ago in the United Kingdom (UK) and the rules and standards have evolved over time. More recently, there has been a broadening of safety concerns beyond licensing of practitioners to include:

- Patient safety – the absence of preventable harm to a patient during the process of healthcare and reduction of risk of unnecessary harm associated with healthcare to an acceptable minimum [12];
- Medication safety – use of medicines to achieve the desired outcomes and improve quality of life, while minimising risks of accidental injury due to errors in the medication process [13];
- Safety of medical devices – designing and manufacturing devices so that, when used as intended they will not compromise the safety of patients [14].

Of course, nowadays, many medical devices contain software and some stand-alone software applications (apps), e.g. on-line triage systems, are now available. This has led the regulatory community to address the safety of Software as a Medical Device (SaMD) [15].

In the last few years there has been a growing interest in developing SaMD using ML and a number of ML-based SaMD are now available for clinical use [16]. The regulatory community is now addressing the safety of ML-based SaMD, see Chapter 2 for a discussion, but there remain many research questions which are made all the more challenging by the fact that the field is evolving so fast. This growing interest in the use of ML in healthcare and the importance of assuring safety of patients provides the context for the research presented in this thesis.

1.2 Research Questions and Contributions

There is now a nexus between healthcare, ML and safety engineering and this gives rise to two important questions: (1) *Are well-established safety engineering methods still appropriate and effective in assuring the safety of ML in some representative healthcare scenarios?* (2) *Are there new opportunities for well-established safety engineering methods with the*

development of ML and why are they specifically good for safety in this domain? Due to the rate of development of ML-based SaMD and the potential benefit from deploying such systems, these questions require urgent answers in the healthcare context.

This thesis explores these questions, for example identifying safety challenges posed by ML from a general and a regulatory perspective. The thesis then presents technical solutions which go some way towards answering those questions. More specifically, the contribution of this thesis is three-fold:

1. Showed how to use well-established safety engineering methods to proactively incorporate patient safety in the design of ML, supported by a clinical case study employing Reinforcement Learning (RL) to aid sepsis treatment. This addresses question 1.
2. Demonstrated how explainability can help to improve the safety assurance of ML, supported by a clinical case study employing Convolutional Neural Networks (CNNs) to determine when to wean a patient from mechanical ventilation. Given the demand for mechanical ventilators during the COVID-19 pandemic, this is especially important and timely. This addresses question 2.
3. Demonstrated how to use ML to update and enhance well-established safety engineering methods, supported by a clinical case study employing Bayesian Network (BN) structure learning to understand the correlations of different factors concerning the delivery of Beta-Blockers (BBs). This also addresses question 2 but from a different perspective.

1.3 Structure of the Thesis

The rest of the thesis is structured so as to highlight the three major contributions. Chapter 2 provides background on healthcare, ML, safety assurance and the regulatory framework for safety related systems in healthcare, setting the context for the case studies in Chapters 4 to 6. The healthcare issues, ML methods and safety methods used in the three case studies are quite different, so additional background and a survey of the relevant literature is provided in each of these three chapters to make them self-contained. Chapter 3 presents an overview to illustrate the different emphases of these three contributions and to show how they relate to each other.

Chapter 4 presents a case study which shows how well-established safety engineering methods can be adapted and applied to an ML-based system which uses RL to make recommendations for sepsis treatment. At a more detailed level, this chapter shows the use of an exploratory safety engineering method, SHARD, to identify hazards and hazard causes, including those relating to ML, and how DSRs can be produced so that they can be used to influence the ML learning process. It also illustrates the roles of different risk mitigation methods in overall system safety and culminates with the presentation of a safety case. Thus, this chapter addresses question 1.

Chapters 5 and 6 provide complementary answers to question 2, with Chapter 5 focusing on the use of explainability to support safety assurance and Chapter 6 showing how ML can be used to update and enhance safety analysis.

Chapter 5 uses CNNs to make predictions about patients' readiness for weaning from mechanical ventilation. It shows three uses of explainable AI methods in support of safety assurance. First, it uses influential instances to determine the right training dataset for the problem concerned. Second, it uses feature importance to show the validity of the learnt model and support the clinical decision making in operation. Third, it uses counterfactual explanations both to provide assurance about model robustness and to make the predictions more actionable in operation.

Chapter 6 considers the use of ML to validate and refine the results of safety engineering methods. The case study concerns clinical practice for medication management, specifically the delivery of BBs following thoracic surgery. It presents a safety analysis of the clinical practice identifying *potential* causes of hazardous effects on patients, such as Atrial Fibrillation (AF). It then employs ML methods, specifically BN structure learning, on data generated from the clinical practice to identify the *actual* causes. For the most part, this confirmed the exploratory analysis, i.e. SHARD, but also identified some important differences, which are used to update the safety analysis.

The case study in Chapter 5 is primarily a patient safety issue, whereas the other two case studies, in Chapter 4 and 6, illustrate medication safety (as well as patient safety) issues. The case study in Chapter 4 also directly illustrates the issues of safety of ML-based SaMD. Thus the three case studies contribute to answering the two research questions from different and complementary perspectives. Chapter 7 reflects on the results of this work as a whole, identifies areas for future work, and presents overall conclusions.

Chapter 2

Background & Literature Review

Part of this chapter is based on my previous publications [17] [18]. This chapter first gives a broad overview of ML, safety engineering and healthcare to set the context for the rest of this thesis. Then it provides a literature review on assuring AI/ML-based medical devices in healthcare, mainly from a regulatory perspective. This is particularly relevant because AI/ML-based systems used in healthcare are currently mainly regulated as medical devices and the safety engineering discipline both influences the way regulators assure safety and, in turn, is influenced by regulators. The literature review presented in this chapter provides a deeper understanding of the challenges posed to regulators by the emergence of ML and how the different legislative frameworks respond to the challenges. This highlights the importance and the value of conducting this work, specifically with the two research questions outlined in Chapter 1.

For ease of presentation, the rest of chapters are largely self contained with a focus on one specific research question in each chapter. Therefore, the details of the particular healthcare concern, the related work, the specific ML methods and the safety engineering methods used are introduced in the each individual chapter.

2.1 Machine Learning

This section introduces the basic concepts of ML and gives an overview of the different categories of ML methods. This will give the basis for understanding the specific ML methods used in the later chapters. More details on the specific ML methods used for the three case studies are presented in Chapters 4 to 6.

2.1.1 Basic Concepts of Machine Learning

In ML, data plays an indispensable role, and learning algorithms are used to discover and learn knowledge or properties from the data without relying on rule-based programming. The quality and quantity of the dataset have a fundamental effect on the learning and prediction performance. There are two general dataset types:

- **Unlabelled dataset** D: $X = \{x^{(n)} \in R^d\}_{n=1}^N$
- **Labelled dataset** D: $X = \{x^{(n)} \in R^d\}_{n=1}^N, Y = \{y^{(n)} \in R\}_{n=1}^N$

Where X denotes the **feature set** containing N samples. Each sample is a d -dimensional vector $x^{(n)} = [x_1^{(n)}, x_2^{(n)}, \dots, x_d^{(n)}]^T$ and called a feature vector or feature sample, while each dimension of a vector is called an attribute, feature, variable or element. Y stands for the **label set**, recording what label a feature vector corresponds to. Another form of labelled dataset is described as $\{x^{(n)} \in R^d, y^{(n)} \in R\}_{n=1}^N$, where each $\{x^{(n)}, y^{(n)}\}$ is called a data pair.

Given a sample space X and a label space Y , there exists some target function $y = f(x)$, so that for any $x \in X$, this function outputs the correct y in the label space. In ML, we want to find a function $g(x)$ that is as close as possible to $f(x)$ when we don't know what f is. The dataset used to learn the function $g(x)$ is called the **training set** (training data). The dataset reserved for testing the performance of $g(x)$ is called **test set** (test data) [19].

2.1.2 Categories of Machine Learning

Based on the given dataset and the problem being addressed, there are generally three types of ML, (1) supervised learning, (2) unsupervised learning, and (3) RL. It is worth mentioning that in this thesis, the ML methods we used are supervised learning and RL, unsupervised learning is out of the scope of this thesis.

- **Supervised learning:** the training set given for supervised learning is the labelled dataset. Supervised learning tries to find the relationships between the feature set and the label set, which is the knowledge and properties we can learn from labelled dataset. If each feature vector x corresponds to a label $y \in L, L = \{l_1, l_2, \dots, l_c\}$ (where c usually ranges from 2 to a hundred), the learning problem is denoted as **classification**. On the other hand, if each feature vector x is corresponding to a real

value $y \in R$, the learning problem is defined as **regression** problem. The knowledge extracted from supervised learning is often utilized for prediction and recognition.

- **Unsupervised learning**: the training set given for unsupervised learning is the unlabelled dataset. Unsupervised learning doesn't figure out the "right answer" based on the input data, but it explores the data and can draw inferences from the dataset to describe hidden structures from unlabelled data. Unsupervised learning is also used for clustering [20], probability density estimation, finding association among features, and dimensionality reduction [21]. In general, an unsupervised algorithm may simultaneously learn more than one of the properties listed above, and the results from unsupervised learning could further be used for supervised learning.
- **Reinforcement learning** [22]: RL is a learning method that interacts with its environment by producing actions and discovering errors or receiving rewards. Trial and error search and delayed reward are the most relevant characteristics of RL. In this type of learning, there are three primary components: the agent (the learner or decision-maker), the environment (everything the agent interacts with) and actions (what the agent can do). The environment gives the agent a state s_t . Next, the agent takes an action a_t . Then the environment gives back a reward r_t , as well as the next state s_{t+1} . This loop continues until the environment gives back a terminal state, which ends the episode. The objective is for the agent to automatically determine the ideal behaviour within a specific context in order to maximise its performance. Reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal. The agent will reach the goal much faster by following a good policy.

2.1.3 Key Elements of Machine Learning

No matter what type of ML is chosen (supervised, unsupervised, reinforcement), ML could be considered to consist of a combination of three components [23]. The components are:

- **Representation**. From a practical standpoint, a key step in the development of a ML system is how to represent the knowledge. Conversely, choosing a representation for a learner is akin to choosing the set of all possible hypotheses it can possibly learn. This set is called the **hypothesis space** or hypothesis set H , which contains

several hypotheses h (a mapping function or distribution). The goal of the learning is to find the best h , called the final hypothesis, approximating the target function. For example, neural networks form one type of representation, as do decision trees, probabilistic graphical models and support vector machines.

- **Evaluation.** Evaluation is essentially the way to “score” candidate hypotheses. An evaluation function, also called an **objective function**, utility function, loss function, scoring function or fitness function in some contexts, is needed to distinguish one hypothesis h from another. Mean squared error or likelihood are examples of different evaluation functions that will imply somewhat different preference in the hypothesis space.
- **Optimisation.** Finally, optimisation is the way to search the hypothesis space to obtain the highest-scoring one, i.e. the final hypothesis, which either minimises or maximises the objective function. The choice of optimisation techniques is key to the efficiency of the learning process. For example, stochastic gradient descent and greedy search are two different ways of optimising a model class. Note that once a model has been trained, it may not be possible to recover exactly how it was optimised.

Based on these concepts, a simple illustration of ML is shown in Figure 2.1.

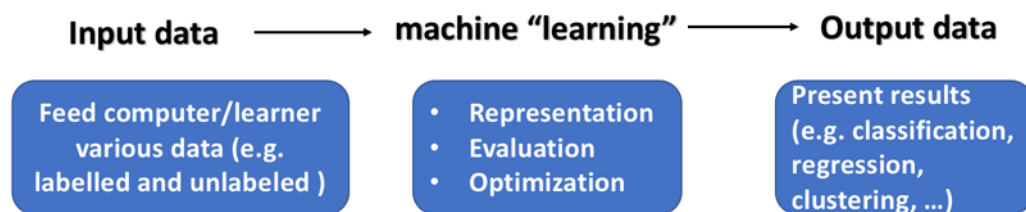
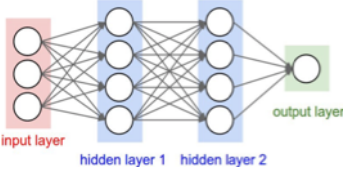


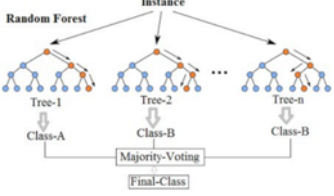
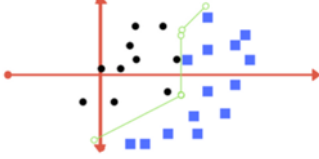
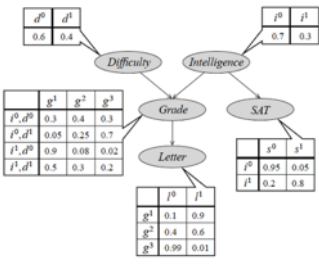
Figure 2.1: A simple illustration of Machine Learning

2.1.4 Approaches to Machine Learning

This section introduces some of the important and popular approaches to ML. These are summarised in Table 2.1 with a brief description of each approach.

Table 2.1: Approaches to ML

| Model | Description |
|--|---|
| <p data-bbox="309 801 600 835">Neural Network [24]</p>  <p>The diagram shows a neural network with four layers of nodes. The first layer on the left is the 'input layer' with three nodes, highlighted in a red box. The next two layers are 'hidden layer 1' and 'hidden layer 2', each with four nodes, highlighted in blue boxes. The final layer on the right is the 'output layer' with one node, highlighted in a green box. Arrows indicate connections between nodes in adjacent layers.</p> | <p data-bbox="683 293 1406 701">A Neural Network (NN) is a non-linear network of neurons inspired by the human brain. It is configured for applications such as pattern recognition and data classification through a learning process. The learning process involves adaptive adjustments to the connections between the neurons. Typically, there are three different layers in an NN, including an input layer, hidden layers, and an output layer.</p> <ol data-bbox="719 750 1406 1153" style="list-style-type: none"> 1. Input layer: all the inputs are fed in the model through this layer. 2. Hidden layers: there can be more than one hidden layer which are used for processing the inputs received from the input layers. 3. Output layer: the data after processing is made available at the output layer. <p data-bbox="683 1200 1406 1440">The representations used by NNs are generally opaque to humans, therefore are useful only in the context of learning from input data and can't be integrated with domain knowledge. The learning algorithm for a neural network can either be supervised or unsupervised.</p> |

| | |
|--|---|
| <p>Random Forest [25]</p>  | <p>A Random Forest (RF) aggregates thousands of decision trees. Each decision tree in the forest considers a random subset of features in the training dataset. This increases diversity in the forest, which leads to more robust overall predictions. By identifying the most important predictors of an outcome, a random forest would have a better performance than a decision tree. In addition, it also corrects for overfitting that can occur with a single decision tree. The random forest algorithm is often supervised learning.</p> |
| <p>Support Vector Machine [26]</p>  | <p>A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given a labelled training set, the algorithm outputs an optimal hyperplane, which categorizes new examples. In two-dimensional space this hyperplane is either a straight line or non-linear boundary dividing a plane into two parts where each class lies in either side. A non-linear boundary can be defined using kernels. SVM are a supervised learning method.</p> |
| <p>Probabilistic Graphical Model [27]</p>  | <p>Probabilistic Graphical Models (PGMs) represent complex domains using probability distributions. The graphical models bring together graph theory and probability theory, and provide a flexible way for modelling large collections of random variables with complex interactions. The nodes (or ovals) correspond to the variables in the domain, and the edges correspond to direct probabilistic interactions between them. Therefore, a graphical model is understandable by humans and domain knowledge can more readily be integrated than with NNs. Markov networks and BNs are the most common form of PGMs. There are both supervised and unsupervised uses of the algorithms.</p> |

2.1.5 Performance Metrics for ML Models

It is important to evaluate the performance of ML models. This can help in choosing amongst the many different types of ML model to find the one that is best suited to a given problem. Here we introduce some of the more commonly used metrics: accuracy, precision, recall, F1 score, as well as the AUC-ROC performance measure, which plots true positives against false positives. This is the generally accepted set of evaluation metrics for deep learning.

The accuracy of a model is calculated as the ratio of the number of correct predictions to the total number of predictions. Formally:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

where, TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives in the predictions. Similarly, specificity is given by:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

and can be interpreted as the true negative rate. Further, precision is given by:

$$\text{Precision} = \frac{TP}{TP + FP}$$

and can be interpreted as the proportion of the positive predictions that were correct.

Recall is given by:

$$\text{Recall} = \frac{TP}{TP + FN}$$

and can be interpreted as the true positive rate (i.e. the number of true positives divided by the total number of elements that actually belong to the positive class). A model can have a high precision or recall and do badly on the other metric. An F1 score takes both scores into account so as to better evaluate the model's performance. It is given by:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The Receiver Operating Characteristic Curve (ROC) curve demonstrates the model's ability to provide predictions at various decision thresholds. It assesses how well a model can distinguish between the classes and is a plot of the True Positive Rate (TPR) of a model against its False Positive Rate (FPR). The Area Under the Curve (AUC) denotes the probability of the classifier ranking a random positive sample in the data higher than a random negative sample. For a "random" model the AUC-ROC would be 0.5 and for a "perfect" model it would be 1. Section 5.4.4 illustrates the use of the ROC and presents a comparison of a set of ML models using AUC-ROC in Figure 5.4.

2.2 Safety Engineering and Safety Cases

This section covers both safety engineering methods and safety cases. As indicated in Chapter 1, safety engineering originated around 70 years ago motivated by perceived risks in space flight and nuclear power. In practice, many advances in safety engineering have been prompted by major accidents and incidents and the excellent safety record in some domains is the result of effective processes of learning from incidents and accidents. There are many safety engineering methods, and this chapter illustrates some of the more widely used methods, including introducing SHARD which is used in two of the case studies.

The concept of Safety Cases originated from Lord Cullen’s report [28] into the accident on the Piper Alpha platform in 1988. The Offshore Installations (Safety Case) Regulations came into force in 1992, requiring safety cases for all installations. Since then, safety cases have become widely used for justification of system safety in many other domains such as aviation and nuclear power. Recently safety cases have been increasingly introduced into healthcare. The term “assurance case” is also often used as a generalisation of the safety case concept. In this thesis, the focus is on safety cases, but the term assurance case will be used where appropriate, e.g. to be consistent with the literature.

2.2.1 Safety Engineering Methods

Healthcare is often encouraged to consider practices in other safety-critical industries, e.g. aviation, which adopt systematic safety analysis to support safety management see, for example [29]. We identify the key principles of safety engineering as interpreted in healthcare, then introduce one of the methods from well-established safety engineering which we use in our case studies.

First, and most fundamental, is the need for hazard and risk analysis. A hazard is a defined as a “potential source of harm” in ISO 14971 [30] which is recognised both by the FDA and in Europe (where it is referred to as EN ISO 14971). It is necessary to identify hazards in normal operation, under fault conditions and arising from human error. Risk is normally considered as the combination of the severity and likelihood of harm arising from the hazard. Some definitions of risk in healthcare also include duration of harm, but we view this as one element to assess severity [31]. ISO 14971 defines processes for hazard and risk analysis for medical devices, and guidance is provided in ISO/TR 24971 [32]. Results of hazard and risk analysis should be used to inform design, e.g. to prompt redesign to eliminate hazards or to minimise risk associated with the hazards. There is also a need to

ensure that design changes do not introduce new hazards.

As well as the guidance in ISO/TR 24971, there are methods from well-established safety engineering which are relevant to healthcare. For example, checklists can be used where a new system is similar to older ones (this may be relevant when clearing systems through the 510(k) pathway, see Section 2.4). There are also flow-based methods, for example HAZOP [7] from the chemical industry. All of these methods have relevance in healthcare, but variants of HAZOP are potentially most relevant for SaMD as the focus is on functions and information flow.

Hazard and safety analysis of computer-based systems often use variants of HAZOP that consider information flows through systems, e.g. SHARD [8]. SHARD is suitable for identifying both hazards and causes of hazards as it focuses on deviations from intent that could be hazardous. One advantage of the method is that it can be applied to any form of information flow so it can not only be used on computer-based systems, but also on clinical workflows. 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:

- *Omission* – no flow provided when intended;
- *Commission* – flow provided when not intended;
- *Incorrect* – wrong information;
- *Early* – flow is earlier than intended;
- *Late* – flow is later than intended.

The application of the guidewords requires judgement. For example, commission is not meaningful for a flow that is provided continuously. We have shown how to use SHARD on a clinical workflow in Chapter 6, where we first developed a decision-making model concerning the delivery of beta-blockers for patients undergoing thoracic surgery who are at risk of atrial fibrillation, then applied SHARD on the decision model and the results were recorded in Table 6.1. Once hazards have been identified they can be assessed for severity, using clinical knowledge. We introduce the World Health Organisation (WHO) categories for risk severity in Section 2.3.2 and show how they might be interpreted in the context of medication safety in Section 2.3.3.

Analysis methods such as FTA and FMEA mentioned in the introduction are often used to evaluate the likelihood of the hazard occurring. This information can be combined with the severity determined in hazard analysis to enable risk to be determined. The use of safety engineering methods informs the development of the safety case and provides some of the evidence.

2.2.2 Development of Safety Cases

Early safety cases, e.g. those developed in response to the Offshore Installations (Safety Case) Regulations, were complex and often multi-volume textual documents that could be very hard to understand and to maintain. Work to address these problems led to development of tools, e.g. the Safety Argument Manager (SAM) toolset [33] and definition of a systematic approach to safety case construction [34]. This work resulted in a clear definition of the concept: “*A Safety Case is a structured argument, supported by a body of evidence that provides a compelling, comprehensible and valid case that a system is safe for a given application in a given operating environment.*”, which is now widely adopted.

2.2.2.1 Arguments

The term “argument” is used in the safety case context to mean the rationale or reasons for believing something, not a dialogue that presents point and counter-point. In this thesis we use Goal Structuring Notation (GSN) for structuring and presenting safety arguments in a graphical manner, although there are other argument notations such as Claims Argument Evidence (CAE) [35]. GSN is based on work on the structure of natural language arguments [36], but it was simplified in order to make it easy to use, resulting in four key concepts: Goals, Strategies, Context, Solutions. The legend showing these elements of the notation is presented in Figure 2.2.

- Goals – these elements represent the claims that we wish to make and support, which are shown as rectangles in Figure 2.2. Normally a safety case has a top-level goal which is concerned with safety of a particular system in its context of use. Goals can be broken down into sub-goals until the goals can be proved (see Solutions below);
- Strategies – these explain the reason/rationale for decomposing goals into sub-goals where this is not obvious. This is represented as a rhombus in Figure 2.2;
- Context – these elements help to constrain goals or strategies. For example, the

operating environment for the system, which gives the context for the system itself. This is represented as a stadium in Figure 2.2;

- Solutions – these elements represent evidence that support or prove the leaf goals in the GSN, which are represented as a circle in Figure 2.2;
- Supported by – these represent the *main argument flow* from top goal through sub-goals and strategies to solutions, which are represented as solid-headed arrows in Figure 2.2;
- In context of – these represent the links from the main argument flow to contextual elements, which are represented as open-headed arrows in Figure 2.2;
- To be developed – these annotations indicate that part of the argument is yet to be developed, i.e. it is deferred for now, which is represented by a diamond under another argument element (usually a goal) in Figure 2.2.

If satisfactory solutions can be provided for each leaf goal, then the top-level goal and all the intermediate goals are proven within the constraints or assumptions provided by the contextual elements.

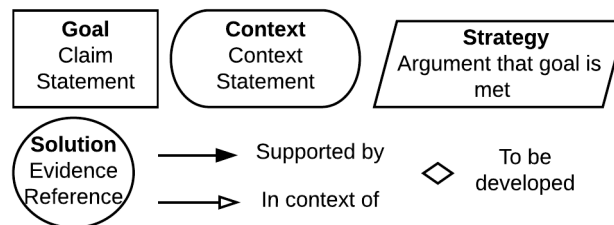


Figure 2.2: Goal Structuring Notation Legend

The main benefit of using GSN is that the argument structure is much more explicit and visible than in a textual safety case. In practice, safety cases are often large collections of documents, but an effective (GSN) argument can help in navigating the information. Figure 4.10 gives an example of how to present a safety argument using GSN. The top goal in Figure 4.10 concerns the safe delivery of intravenous fluid and vasopressor medications for sepsis treatment and the strategies include arguing over the hazards identified for the delivery of these two medications in the context of clinical practice for treating sepsis.

2.2.2.2 Evidence

What kind of evidence is needed to prove the leaf goals will depend on the technology used in the system and the system properties of concern. In practice, it is common to provide evidence that all hazard-related risks are mitigated to an acceptable level. Thus evidence will come from the use of safety engineering methods, such as FTA which allows the likelihood of the hazard occurring and hence the risk to be estimated.

For most systems, the safety case will also contain evidence drawn from testing. Testing is often focused on showing that the system works as intended. This has a role in a safety case, but more importantly, tests can provide evidence about the absence of hazardous behaviour. Therefore, tests may need to be focused on abnormal situations, such as overload of a computer system, to show that it “fails safely” under such conditions. For software-intensive systems, the required evidence is often determined by the standards in the domain, such as IEC 62304 for medical devices [37]. There are many different standards, but there is little agreement on what is appropriate evidence [38].

2.2.3 Safety Case Limitations

There have been some criticisms of safety cases. Wassying *et al* [39] have produced quite a philosophical review of safety cases; their focus is on software, but the ideas seem much wider. They make comparisons between safety cases and the established disciplines, e.g. civil engineering, saying that: “*the safety case approach lacks the highly prescriptive and domain specific nature that can be seen in other engineering disciplines*”. However, when the safety case is used for a system, it would be sensible to keep the relevant prescription or standards for the normal design of the system and leaving the flexibility of goal-based approaches for radical design where such standards are not available. Therefore it seems appropriate to use goal-based approaches for AI/ML based SaMD as such standards are not yet available.

Leveson has stated that a key flaw of safety cases is that they are prone to confirmation bias [40], i.e. tend to emphasise the evidence that supports the safety claims, and overlook contradictory evidence. This is valid, and is consistent with the idea that safety cases present a rationale, not point and counter-point. However, it can be seem to reinforce the need for effective safety cases rather than undermining the safety case concept.

Another concern is that safety cases are often “static”, i.e. not updated once the system is put into service. This means that if a system is updated then the evidence

should be updated and the associated argument should be reviewed to see whether or not it is still valid.

Finally, most safety cases are system-focused, but to be valid they must also consider the environment of use. For example, if part of the safety argument rests on the fact that users are given alerts, but it is known that clinicians suffer from alert fatigue, then this aspect of the safety case is not valid. It is always important to consider the safety of a system in context, and the safety case must address enough of the context of use to be valid and compelling.

2.2.4 Safety Cases in the Healthcare Domain

Safety cases have been used widely for many years in the aerospace, nuclear-energy and transportation domains. However, their use in healthcare has been fairly recent and is primarily due to their inclusion in a recent guidance document under a FDA pilot program [10]. This guidance, issued in 2014, was developed to assist industry in preparing pre-market submissions for infusion pumps and to identify device features that manufacturers should address, in which it requires safety case to organise the information of infusion pump. The definition of safety case given by FDA in this guidance is “*The safety assurance case (or safety case) consists of a structured argument, supported by a body of valid scientific evidence that provides an organised case that the infusion pump adequately addresses hazards associated with its intended use within its environment of use. The argument should be commensurate with the potential risk posed by the infusion pump, the complexity of the infusion pump, and the familiarity with the identified risks and mitigation measures.*”

This is consistent with the general definition given above and was motivated by the level of problems seen with infusion pumps [41]. There is growing interest in adopting safety cases in healthcare, especially for digital innovations [42] and medical devices [43]. The evolving approach to regulation of medical devices is considered in more detail in Section 2.4.

2.3 Healthcare

2.3.1 Nature of Healthcare

Healthcare is deeply rooted in the conditions of life, the conditions of work and the social relations of society. As noted in Chapter 1, healthcare has a long history but our focus here is on the approaches and challenges arising in “modern healthcare”, i.e. in the 21st Century. Over the past 20 years from the release of “To Err is Human” by the Institute of Medicine (IoM) [44], healthcare has increasingly focused on improving patient safety and healthcare quality. However, although many initiatives aim to improve patient safety by imposing or encouraging a range of quality improvement strategies, healthcare has evolved slowly [45] by comparison with other safety critical industries, e.g. aerospace. Thus, it is essential to understand the nature of healthcare first, then to draw parallels between healthcare and other industries, highlighting their similarities and differences, before concluding what learning and experience can be transferred from these other domains.

Healthcare is a complex socio-technical system and is very different from other safety-critical industries:

- First, healthcare contains an extraordinarily diverse set of activities. On the one hand, it encompasses a lot of routine practices, such as infection control strategies with standards designed to prevent the transmission of germs in all healthcare settings [46]. On the other hand, it also involves highly unpredictable and hazardous activities. For example, in hospital medicine, clinical staff often face very high levels of uncertainty as the patient’s disease may be masked by other similar conditions, difficult to diagnose, the symptoms or treatment are complicated by multiple comorbidities and so on [47]. For example, there might be a “trade-off” in treating a patient who has low platelets (bleeding tendency) but also has a clot (for which clinicians would normally use a blood thinning agent).
- Second, in healthcare, automation of procedures is relatively low with many of the healthcare tasks being “hands on” [48]. A lot of surgical procedures have to be done by surgeons in a broad spectrum of situations. Although it is possible to define overall processes, the exact procedures or solutions will have to be determined by surgeons and may need to be modified “on the fly” if a situation arises that wasn’t anticipated. Thus such procedures are more liable to error. By comparison, in other industries, many procedures are automated and humans are more often monitoring

activities rather than undertaking them.

- Third, healthcare, even in national systems such as the NHS in England, is decentralised and fragmented [49] [50]. This makes it very difficult to standardise and regulate design of both medical equipment, including SaMD, and the associated procedures. For example, it is highly desirable to standardise the design of infusion pumps as several accidents have occurred because of them [51], but it is very difficult to achieve standardisation in practice [47]. In addition, if nurses are constantly having to work with different medical devices when delivering care, unnecessary variability will be introduced, thus increasing the potential for errors.
- Fourth, healthcare is extremely heterogeneous, using a mix of old-fashioned technology and state-of-the-art devices. For example, even now, paperwork and fax are still commonly used. On the other hand, a lot of advanced diagnostic devices are also deployed in healthcare, and there is a growing interest in making use of AI/ML in diagnosis and even in recommending treatments.
- Finally, in healthcare, major adverse events are usually investigated locally, although they may be subject to wider investigation or reported in the media if they are deemed of particular significance [48]. In the healthcare reporting culture, an individual doctor is usually blamed or ascribed responsibility for the accident [52], which is in striking contrast with pilot near-immunity in aviation. This, to some extent, contributes to the reactive (and defensive) attitude [53] in healthcare when dealing with adverse events (in contrast with the learning that has happened in aviation).

These factors need to be taken into account if seeking to adopt and adapt practices from other safety-critical industries.

2.3.2 Patient Safety

Patient safety began as a discipline, emphasising prevention of harm to patients, in response to evidence that adverse medical events are widespread and preventable and that there is “too much harm” [54]. Patient safety can also be treated as a property that emerges from healthcare systems design to achieve high reliability under conditions of uncertainty and risk. Illness brings the first condition of risk in healthcare and patient safety applies to the second condition of risk, which is the therapeutic intervention to combat the illness. Patient safety has been increasingly recognised as an issue of global

importance, which aims to make risky interventions reliable through the design of safe healthcare systems, including clear policies, organisational leadership capacity, data to drive safety improvements, skilled healthcare professionals and effective involvement of patients in their own care [54].

Currently, there is no standardised definition of patient safety. The IoM (which has now changed its name to The Health and Medicine Division) defined patient safety as “*the prevention of harm to patients*” [55], which is quite abstract and doesn’t give concrete guidance. The WHO defined patient safety as “*the absence of preventable harm to a patient during the process of healthcare and reduction of risk of unnecessary harm associated with healthcare to an acceptable minimum*” [12]. The emphasis in this definition is placed on healthcare systems that, first reduce risk to an acceptable minimum, and second are free from preventable harm. Here, an acceptable minimum must be interpreted given current knowledge, resources available and the context in which care was delivered weighed against the risk of non-treatment and the benefits/risks of other treatment(s) [12]. This principle of balancing benefits and risks also applies to the introduction of medical devices.

Preventing harm is another important aspect of patient safety. Every point in the process of care-giving contains a certain degree of inherent risk, although this is not well-appreciated by patients. A number of countries have published studies showing that significant numbers of patients are harmed during healthcare, either resulting in permanent injury, suffering due to increased length of stay in healthcare facilities, disability or even death. Having a systematic approach to severity classification would help in making international comparisons and assessing the outcome of any patient safety improvement programme. Whilst there are merits in classifying type, severity and duration of harm separately, most practical harm scales conflate these elements when assigning a degree of harm (as noted in Section 2.2.1). For example, the WHO’s conceptual framework for the international classification for patient safety categorises the degree of harm as follows: none, mild, moderate, severe, fatal [56]. We illustrate how this severity scale can be interpreted in the context of medication safety in Section 2.3.3.

While a goal of zero harm is desirable, this may not always be feasible as some of the harm might not be preventable. But eliminating preventable harm is certainly a much more reasonable goal to achieve in the context of patient safety. This is also what the WHO implies when defining patient safety. Currently, there is also no clear and agreed definition of preventable harm. Most working definitions include the idea that the harm

is “identifiable” in that it can be attributed to medical care and “modifiable” in that it is possible to avoid [57]. According to a systematic review of preventable harm in healthcare [58], the most prevalent preventable harms cited in the included studies were medication adverse events, defined as errors in prescribing, delivering or monitoring the effects of a drug, but this does not include regular side effects. This emphasises the importance of medication safety.

2.3.3 Medication Safety

Medications are the most common treatment intervention employed in healthcare around the world. When used safely and properly, they can significantly improve patient wellbeing. However, inappropriate medications, e.g. giving a medicine to which a patient is allergic, or giving the wrong dose of an appropriate medication, could also cause patient safety incidents despite the good intention of care providers. For this reason, medication safety has become a priority for improving patient safety in healthcare organisations [59]. For example, “Tall Man Lettering” is used to try to reduce the likelihood of administering the wrong medication [60].

Medication safety is defined as *freedom from preventable harm with medication use* by The Institute for Safe Medication Practices Canada (ISMP Canada) [61]. Medication safety has a critical impact on patient outcomes, e.g. readmission rate, length of stay, post-acute referral and organisational outcomes, and consequently increased overall costs to the healthcare system. As medication errors continue to be a leading cause of patient harm in hospitals, with an estimate that one in every five doses administered to patients is liable to medication error in the typical US hospital [62], both government and regulatory agencies have paid close attention and have revised their standards to place a strong emphasis on a systematic approach to assure medication safety. For instance, NHS England in 2014 circulated *Patient Safety Alert, Stage Three: Directive* [63] to reinforce the importance of learning from medication errors to improve medication safety.

To support effective learning it is important to have a clear mapping of the consequence of medication errors to severity of harm. From our literature review, we discovered little information to help produce such a mapping. Studies such as [18] and [19] either just give a severity classification without a detailed description of how they mapped their patient results to the severity, or they focus on error types and their causes, but do not identify the severity of the patient outcome. Further work is needed on how to categorise the severity

Table 2.2: Examples of Severity Classes derived from the use of medications

| Severity | None | Mild | Moderate | Severe | Fatal |
|----------|--|--|---|---|--|
| Summary | No symptoms (detected) | Symptoms short-term, requiring minimal intervention | Harm or loss of function may be long-term, requiring intervention | Life-saving intervention needed; long-term harm or permanent loss of function | Death caused or brought forward by the incident |
| Examples | Substitution of ceftriaxone for cefotaxime Use of antibiotics to treat viral infections (NB reduces utility of antibiotics) | Nausea, vomiting or diarrhoea from overdose of epirubicin Forgetting to specify maximum daily dosage for an “as required” drug Accidental sedation due to prescribing diazepam not diltiazem | Digestive problems including ulcers and internal bleeding (caused by non-steroidals) Hypotension due to overdose of lisinopril Dyspepsia and ulcers from overuse of non-steroidal anti-inflammatory drugs for arthritis | Blindness due to prescribing a diuretic to patients with low blood pressure Renal failure from diuretics Hearing loss from gentamicin Lung damage and possible sepsis by giving oral treatment to patient with dysphagia | Weekly dose of methotrexate given daily Ten times overdose of insulin Haemorrhage from incorrect use of warfarin |

of harm and giving concrete examples of how to map the consequences (patient outcomes) to different severities. In order to illustrate this, we present examples of patient harm in Table 2.2 using the WHO severity classification introduced in Section 2.3.2. This table is intended to be illustrative but refining and expanding it, e.g. by considering different aspects of human function such as vision and respiration, might aid in future hazard and risk assessment.

2.3.3.1 Medication Error

As medication errors are important indicators of medication safety, an understanding of what defines medication error and how to classify medication errors is important.

However, the definition of medication error varies between studies and there is no consensus. Lisby *et al* [64] conducted a systematic literature review of medication error and found 26 different terminological frameworks.

The United States National Coordinating Council for Medication Error Reporting and Prevention [65] defines a medication error as “*any preventable event that may cause or lead to inappropriate medication use or patient harm while the medication is in the con-*

trol of the healthcare professional, patient, or consumer. Such events may be related to professional practice, healthcare products, procedures, and systems, including prescribing, order communication, product labelling, packaging, and nomenclature, compounding, dispensing, distribution, administration, education, monitoring, and use.” This definition is broadly stated and underscores that errors arise in, and are preventable at, different levels or phases in the medication process.

Bates *et al* [66] defined medication error as “*error in the process of ordering, dispensing, or administering a medication, regardless of whether an injury occurred or whether the potential for injury was present.*” This definition emphasises the divorce between an error and its consequence.

Aronson *et al* [67] defined medication error as “*a failure in the treatment process that leads to, or has the potential to lead to, harm to the patient.*” This definition is similar to the definition of a hazard, which implies that what should be counted as medication errors have a potential to cause harm to a patient. Therefore, those minor errors which don’t have the potential to cause harm should be excluded. Medication error has also been defined as an unintentional reduction in the probability of treatment being timely and effective, or an increase in the risk of harm relating to medicines and prescribing [68].

As with definitions, there are also a number of different approaches to classifying medication errors. Among them, three commonly used classifications for medication errors are:

- By medication process;
- By type & modality;
- By psychological theory.

When classifying by medication process, errors are linked to the stages in the sequence of the medication use process, usually prescribing, transcribing, dispensing, administering and monitoring [69] [70]. This approach is particularly useful for management purposes, as different phases of the medication process often happen in different places, with different actors. For example, prescribing normally will occur in a hospital or in a GP’s surgery, accordingly the dispensing stage might happen in a community pharmacy or in a pharmacy affiliated with a hospital. Thus, this classification could help managers to control issues arising within their sphere of responsibility.

When classifying by type & modality, the focus is on the administration elements, which are often referred to as “five rights”: right patient, right drug, right dose, right route, right timing (frequency and duration) [71]. This is normally complemented by Modality. Modality examines the way in which errors occur, including omission, commission, early, late and value which bears strong similarity to categorisation used in exploratory safety analysis methods, such as SHARD. Thus, combining the modality with the type will give a precise characterisation of the error, e.g. wrong drug, duplicated drug, wrong dose, omitted dose. A lot of studies [72] [73] adopted this classification as it could give a richer context to explain what goes wrong around the five rights for patient administration. My previous work has shown how to link the five rights to modality in order to better understand risk as part of a study on medication safety [17] [18].

When classifying by psychological theory, errors are associated with the way they happen. In consequence, it yields four broad types of medication errors [74]: knowledge-based errors, rule-based errors, action-based errors and memory-based errors. This approach not only describes the errors, but also gives a hint about how to help reduce their occurrence based on how they occur. For example, knowledge-based errors, which can be related to any type of knowledge, general, specific or expert, can obviously be prevented (at least in principle) by improving knowledge, e.g. by ensuring the basic principles of therapeutics are taught properly and the procedures are aligned with the best practices. Memory-based errors can be tackled by putting in place computer systems that detect such errors or using checklists that could prompt the healthcare professionals.

These different approaches to classifying medication errors are neither mutually exclusive nor orthogonal. There is no strong evidence to identify which method is more effective than others [75]. The approach which should be taken will depend on the setting and the purpose of the classification. However, the literature shows that there is a common problem in that the definitions and classifications of medication errors are often applied inconsistently in the studies or even mixed up, which makes it hard to understand whether the classifications are complete or not and whether errors are counted fairly.

2.3.4 Medical Devices

A medical device is defined as follows by Council Directive 93/42/EEC [76], “*any instrument, apparatus, appliance, material or other article, whether used alone or in combination, including the software necessary for its proper application intended by the manufac-*

turer to be used for human beings for the purpose of:

- *diagnosis, prevention, monitoring, treatment or alleviation of disease,*
- *diagnosis, monitoring, treatment, alleviation of or compensation for an injury or handicap,*
- *investigation, replacement or modification of the anatomy or of a physiological process,*
- *control of conception;*

and which does not achieve its principal intended action in or on the human body by pharmacological, immunological or metabolic means, but which may be assisted in its function by such means”.

Historically, medical devices were often stand-alone systems with embedded software. However, recently, the Medicines & Healthcare products Regulatory Agency (MHRA) and others [77] have expanded the scope of medical devices so that stand-alone software applications (apps), e.g. Clinical Decision Support System (DSS), are also considered as medical devices. MHRA have produced a medical device (app) determination flow chart [78] to help judge whether or not a particular app is a medical device, which we will discuss further in Section 2.4.

A growing number of ML-based DSS are being developed to provide guidance on the safe prescription of medicines, guideline adherence, diagnostic decision support and prognostic scoring [79]. ML-based DSS can be found in clinical domains such as radiology, using algorithms that learn from training data to classify images [80] [81] [82]. Significant examples of such usage of ML include the identification of malignant lesions and cancers from skin photographs [83] [84], analysis of echocardiograms to detect heart problems, e.g. hypertrophic cardiomyopathy and pulmonary artery hypertension [85], and prediction of sight-threatening diseases from eye scans using optical coherence tomography [86].

Outside of diagnostic support, ML systems are being developed to provide other kinds of decision support, such as treatment recommendations [87] [88]. Other work on decision support provides risk predictions where many complex and interacting factors have to be taken into consideration. For example, one project has used BN to predict the risk of developing coronary heart disease [89], based on life-style data collected over many years.

Another project has used ML to predict the risk of suicide attempts [90], again based on complex data.

Research has also focused on automatic triage for patients or prioritising individual access to clinical services by screening referrals. Babylon Health has an ambitious mission: “*to put an accessible and affordable health service in the hands of every person on earth*”. They have a growing range of services [91], including an on-line triage tool which analyses data provided by patients to advise them on a course of action, e.g. to consult a physician or to visit a pharmacist. The tool uses a range of ML technologies, e.g. recurrent NNs for processing and analysing the text input by the user. It also employs an extensive medical knowledge base to assist in interpreting the information on symptoms provided by the user [92]. Whilst Babylon see their technology as helping address some of the problems of clinician shortages, their work is not without its critics [93].

2.4 Regulation and Safety Assurance

This section reviews the current standards and regulatory practices for assuring the safety of AI/ML based software systems used in healthcare in the USA, UK and EU. It also considers insights from the wider research community that might inform the safety assurance of AI/ML based SaMD in healthcare.

AI/ML-based software is defined as a medical device when it is intended to diagnose, treat or prevent health problems under the Food, Drug, and Cosmetic Act (in the USA) [15] as we indicated in section 2.3.4. There are common concepts and principles relating to safety and regulation of medical devices, even though the approval and regulation of such devices are handled differently in the USA, the EU and the UK. Therefore, we first consider these widely applicable concepts and principles, then discuss each jurisdiction in turn and finally summarise the similarities and differences in a table.

2.4.1 General Concepts and Principles

The approaches in the three jurisdictions all seek to balance benefit against risk – this is somewhat different to traditional safety engineering which is much more focused on risk. Benefits are typically assessed in terms of type, magnitude, likelihood and duration [31] [94]. As noted above, risk typically reflects severity (which might include duration specifically) and likelihood of the harm, although the precise risk classifications for

medical devices are different in the three jurisdictions (see below for details). Any residual risks associated with the intended use of a medical device should be acceptable when weighed against the benefits to the patient. More specifically, the benefit/risk profile in the intended target groups for use of the medical device and any undesirable side-effects must be acceptable when weighed against the intended performance of the device. This evaluation should be carried out according to the state-of-the-art in the relevant medical field.

Further, at a more detailed level, all medical devices must be designed and manufactured so that they achieve the performance intended by the manufacturer and will not adversely affect the clinical condition or the safety of the patients, when used for their intended purpose in the specified conditions. This also applies to the safety and health of users of the medical devices or, where applicable, other persons who might be affected [95]. Achievement of the clinical performance should be supported by sufficient clinical evidence.

These concepts and principles can be seen across the three jurisdictions even though there are some differences in the details of the approaches, e.g. in the approach to risk classification of medical devices (at least between the USA and Europe).

2.4.2 Regulation in the USA

In the USA, medical devices are regulated by a centralised agency, i.e. the FDA. The process for approving medical devices, including AI/ML-based systems, varies according to their risks. Device classification depends on the intended use of the device, and on its indicated use, i.e. the way its use is described on labels or verbally by the device vendor [96]. In addition, a major factor in classification is the risk to the patient or the user. Class I is for the lowest risk devices, Class II for medium risk and Class III for the highest risk. There is a process for assessing risk of medical devices by consulting a product classification database or by making a request to the FDA. This would also be the case for AI/ML-based systems.

In the USA, the FDA “clears” medical devices, including software, through one of the following four pathways:

- Pre-market approval – the most stringent review for high-risk devices which requires that safety and effectiveness is demonstrated by providing extensive scientific evidence [97];

- 510(k) – demonstrating that an algorithm is at least as safe as another, legally marketed, algorithm [98];
- *De novo* – for novel low and moderate risk devices (where there are no existing counterparts so the 510(k) pathway cannot be used) and safety and effectiveness can be assured through general controls [99];
- Humanitarian device exemption – used for devices intended for diagnosis or treatment of a rare disease where it is hard to get enough clinical evidence to meet FDA’s normal standards for safety and effectiveness [100].

The clearance pathways introduced above are meant to assess the safety of the device as a whole, rather than to clear or certify specific algorithms that the device uses. Thus, there is no specific, or separate, pathway for AI/ML-based systems and they are often approved through one of the first three pathways. According to [101], the number of approved AI/ML-based devices has increased substantially since 2015. Most of them are approved through the 510(k) pathway and only a few are approved through pre-market approval. The 510(k) pathway is used for the approval of evolutionary medical devices and such devices are generally exempt from the rigorous pre-market approval. Although some AI/ML-based systems employ advanced ML methods, e.g. deep learning, which is often considered to be revolutionary in *techniques* compared with other rule-based software, it is not necessarily viewed as a revolutionary medical device in terms of the FDA framework (if the application is not new). So as long as it is possible to find a similar medical device that has already been approved even if it uses conventional software, the AI/ML-based system can be approved through the 510(k) pathway by showing that the hazards identified for the previous system have also been sufficiently controlled in the new AI/ML based system. There have been criticisms of the 510(k) pathway [102], for example that it doesn’t require either pre-market or post-market assessment of safety and effectiveness. Further, allowing claims of “substantial equivalence” over a chain of products can lead to approval of a product that is radically different to from the original device which might have been introduced decades ago. No AI/ML-based systems have used the humanitarian exemption; this is unsurprising as there is unlikely to be sufficient data on rare diseases to make ML a viable approach.

More recently, the FDA has been exploring new approaches to regulation of AI/ML-based systems, recognising the rapid evolution of medical devices and that they can con-

tinue learning in operation, meaning that it is impractical to assess every iteration of the device. The FDA has proposed a framework for dealing with modifications to AI/ML-based systems for public discussion and consultation, using the term SaMD [15]. This proposal builds on previous work, including risk classification principles [103] and software modification guidance [104]. The FDA has also started a pilot programme to pre-certify developers of SaMD with the intent that this will enable streamlining certification of an actual product [105]. In other words, the focus is on assessing and certifying the organisation developing the medical device, rather than on the device itself. Based on this work, and the feedback from the consultation, the FDA has proposed an action plan for AI/ML-based SaMD, with five key action areas [106]:

- Change control plan – documentation of ways in which the algorithms will change over time, yet remain safe and effective;
- Good ML practice – definition of good practices for developing AI/ML-based systems, e.g. data management, interpretability, akin to software engineering good practices;
- Patient-centred approach – providing transparency to users of the systems and to patients more generally, e.g. describing the data used to train the algorithms;
- Bias and robustness – improving the scientific basis for assessing algorithms, e.g. where biases relating to race or ethnicity might impact safety or effectiveness;
- Real-world performance – adopting a Total Product Life-Cycle (TPLC) approach involving collecting and monitoring real-world data.

These action areas, e.g. the TPLC approach, help to address one of the criticisms of the 501(k) pathway as they provide a means for post-market assessment. The FDA proposals explicitly distinguish “locked” algorithms that do not change over time from those that adapt, e.g. continue learning in operation. Typically, AI/ML-based systems approved to date by the FDA have been “locked”. Under current policy, changes to approved AI/ML-based systems could require a further pre-market submission, but the TPLC approach is intended to allow devices to “continually improve while providing effective safeguards” [15]. The third case study, see Chapter 6, might provide one way of doing this.

2.4.3 Regulation in the European Union

The European Union (EU) is in a period of transition in how it deals with medical devices, replacing three Directives including Council Directive 93/42/EEC on Medical Devices (MDD), with two Regulations [107]. The new Regulation (EU) 2017/745 on medical devices (MDR) [108] came into effect on May 26 2021 replacing the MDD. Separate Directives and Regulations addressing *in vitro* devices are not covered here.

The original definition of medical device in the MDD 93/42/EEC [109] included software only where it was part of other medical devices, although later guidance [77] for the Directive included standalone software. This change in scope has now been formalised in the MDR [108], which includes standalone software with a medical purpose in the definition of an active medical device; this is what the FDA refer to as SaMD.

As with the USA, the clearance process for medical devices also varies with the associated level of risk [94]. There are four risk classes for medical devices: Class I, Class IIa, Class IIb and Class III with Class I the lowest risk and Class III the highest risk. The medical device classification depends on the intended purpose of the device, its length of use and how invasive the device is. All active implantable medical devices, e.g. prosthetic heart valves, fall into Class III. Class IIa includes active diagnostic devices when they are intended for direct diagnosis or monitoring of vital physiological processes. However, diagnostic devices that monitor physiological processes where the nature of variations, e.g. in cardiac performance, is such that they could result in immediate danger to the patient fall into Class IIb [94]. Active therapeutic devices that administer or exchange energy to or from the human body are in Class IIa, but if they can do so in a potentially hazardous way they will fall into Class IIb. Other devices, e.g. non-invasive tubing to evacuate bodily fluids, fall into Class I. Associated control or monitoring devices inherit the class of the primary medical devices.

In the USA medical devices are approved by a centralised agency, the FDA; this is not the case in the EU despite the existence of a central body, the European Medicines Agency (EMA). For Class I devices the manufacturer usually bears the responsibility for ensuring that their products comply with the regulations, and there is no formal approval process. Medical devices in the higher risk classes are handled by private organisations, known as Notified Bodies, that have been accredited to carry out a conformity assessment and issue a Conformité Européenne (CE) mark. Although issued in a single country, CE marks are recognised throughout the EU and the manufacturer can select the country, and

Notified Body, to which they submit their product for approval. For a higher risk medical device, a clinical investigation might be necessary [110]. The EU doesn't have an explicit 510(K) pathway for approving medical devices as in the USA, however it is still possible to use information on equivalent devices to facilitate the approval process for the clinical evaluation.

Turning to AI/ML-based medical devices, the EU does not seem to have addressed such systems directly, but has been working on the related issue of "big data" with the National Heads of Medical Agencies (HMA) [111]. However, the EU has posted a web page [112] that provides a link to the US FDA 2019 proposed framework for AI and ML in SaMD [15] and this is presumably intended to encourage EU citizens to contribute to the US FDA consultation. The European Commission has been active in AI ethics and related issues. It published a white paper on excellence and trust in AI [113] and a report on the safety and liability aspects of AI [114] which, although general, are likely to be applied to AI/ML-based medical devices.

Despite the lack of specific rules on AI/ML-based medical devices, it seems likely that the new MDR will have an impact on such systems because of the way it now treats software. A medical device is now defined more broadly and specifically includes software for the "prediction" and "prognosis" of diseases not only for diagnosis and treatment. It introduces a new classification rule for software. As a consequence, risk classification of many software products and apps will be up-classed, i.e. move into a higher classification. For example, Rule 11 of the MDR [108] states that: "Software intended to provide information which is used to take decisions with diagnosis or therapeutic purposes is classified as class IIa, except if such decisions have an impact that may cause:

- Death or an irreversible deterioration of a person's state of health, in which case it is in class III; or
- A serious deterioration of a person's state of health or a surgical intervention, in which case it is classified as class IIb."

This means that AI/ML-based devices can be assessed as being in class III whereas, previously the highest was Class IIb [77]. Another major change is that manufacturers have to appoint at least one person responsible for regulatory compliance. In addition, there is now a heightened requirement for post-market monitoring and traceability of the devices as in the USA although the detailed requirements vary between the jurisdictions [115].

2.4.4 Regulation in the UK

In the UK, the MHRA is responsible for regulating medical devices. Since the UK has recently left the EU, a new UK Conformity Assessed (UKCA) marking has been introduced to apply to medical devices from 1st January 2021 [116]. However, CE marking will continue to be recognised until 30 June 2023 [117]. UK Notified Bodies are no longer able to issue CE marks and the EU no longer recognises the UK Notified Bodies. Note that the situation in Northern Ireland is different, so strictly the following applies to Great Britain (GB) not the UK, despite the use of the term UKCA.

The three EU Directives mentioned previously are given effect in UK law through the Medical Devices Regulations 2002 (SI 2002 No 618, as amended) (UK MDR 2002) [117]. These Regulations (in the form in which they existed on 1 January 2021) continue to have effect in Great Britain after the transition period. This means that since 1 January 2021, the Great Britain route to market and UKCA marking requirements is still based on the requirements derived from current EU legislation and that the changes to introduce the MDR in the EU will not automatically apply in the UK.

All medical devices placed on the market in GB must be registered with the MHRA from 1st January 2021. All suppliers need to register with the MHRA and, if they are based outside the UK, appoint a UK Registered Person to carry out registration on their behalf. As in the EU, compliance of Class I devices is based on their manufacturers' self-declaration. All other UKCA marked devices must have compliance assessed by an UK approved body.

Turning to AI/ML-based medical devices, the Care Quality Commission (CQC) and the MHRA conducted a regulatory sandbox (experimental study) on the use of ML in diagnostic devices. The sandbox involved a range of stakeholders including developers of AI/ML-based systems for healthcare and NHS trusts where these systems might be used. The sandbox report [118] includes five main findings and recommendations which are summarised below:

- Governance – the CQC needs to work with service providers using AI/ML-based systems to identify good governance in relation to the clinical, information, technical and human aspects of the system;
- Registration – suppliers of AI/ML-based systems who also deliver clinical activity (services) will have to register with the CQC;

- Technical standards – standards to assure the public that services provided by registered suppliers are safe and effective will need to be developed by other National bodies (presumably standards organisations such as the British Standards Institution (BSI));
- Close assurance gaps – 1) provide guidance on validation of algorithms both at the CE marking stage (presumably now UKCA marking stage) and when implementing at a new site; 2) provide clarity on how hospitals should implement ML devices within clinical pathways to ensure high-quality care;
- Clear communication – encourage medical device suppliers to provide greater clarity on how their devices perform, including whether or not they employ ML.

The report refers to, and builds on the FDA’s proposed regulatory framework for AI/ML-based SaMD and proposes a slightly different approach to classifying systems, and gives some examples of SaMD to illustrate the classes, which is shown in Figure 2.3.

The report also highlights the complexity of the regulatory framework, identifying around a dozen bodies, even treating all the Medical Royal Colleges as a single organisation. This shows the organisational, as well as technical, complexity in making regulatory changes to enable safe introduction of AI/ML-based devices into clinical practice. The government has already recognised the regulatory complexity, and has produced a blog post [119] on regulating AI in healthcare to clarify the roles of the major regulatory and advisory bodies. And it also envisioned a joined-up approach to set up a single platform to bring all the regulatory strands together and a joined-up regulatory sandbox for AI where innovators can find all of the sandbox initiatives from different regulators. REFORM [120] has also produced resources on data-driven healthcare which seeks to provide an easily understood introduction to the “full regulatory pipeline for data-driven technologies in healthcare” including AI/ML-based medical devices.

2.4.5 Summary of Regulation in the Three Jurisdictions

The similarities and differences of medical devices regulations across the three jurisdictions have been summarised in Table 2.3. The comparison is conducted from six perspectives. Some studies suggested that medical device regulation in the EU is less rigorous compared with that of the USA due to the decentralised nature of medical devices regulation in Europe and different Notified Bodies might not always work to the same standards [101],

| | | Level and scope of clinical risk | | |
|-------------------|---|--|---|--|
| | | Narrow: low clinical risk | Narrow: high risk pathway | Broad: high risk pathway |
| Level of autonomy | Autonomous clinical software | | IDx-DR – diabetic retinopathy screening (FDA approved for direct to consumer activity) | Plans for Behold.AI / Dartford and Gravesham NHS Foundation Trust to rule out high normal chest X-rays Plans for Oxipit to perform chest x-ray reporting Qure’s delivery of AI for TB and other lung disease in Indian systems |
| | Clinical software working independently, under a clinician’s supervision for each patient | Zebra Medical’s fatty liver detection algorithm | Rapid diagnostic analysis of whether a stroke is ischaemic or haemorrhagic, where a clinician retains full responsibility prior to starting treatment Kheiron / EMRAD proposals for Breast Screening pathway | Oxipit and Behold.AI in their current pilots with NHS trusts, which prioritise patients for radiologist review ResApp, using cough as a biomarker to differentially diagnose lung-based presentations |
| | Computer aided detection (CAD) ^c | A lot of software is in this category, and it has been for several decades. However, there are now also machine learning-based products on the market, such as Aidence and Veye’s lung nodule check, ZebraMed’s heart calcification detection algorithm, and Healthy.IO’s AI based urine dipstick analysis for chronic kidney disease screening. | | |

Figure 2.3: Examples of existing clinical applications of ML in diagnostics where darker purple boxes indicate that more care is needed when using these solutions in live clinical services. Taken from [118].

although it is not straightforward to draw that conclusion from the table. In addition, there is no centralised publicly available database of approved medical devices in Europe in contrast with the USA, this will also make it more difficult to assess the comparative rigour of the approval processes.

In terms of AI/ML-based medical devices, it is clearly that the FDA is taking the lead in tackling the challenges associated with such systems. For “locked” AI/ML algorithms, the traditional paradigm of medical device regulation seems to be less problematic, although the impact of AI/ML-based system can be broader than the traditional SaMD, which is not taken into account in the current regulations. For adaptive AI/ML-based medical

devices, i.e. continuous learning, it seems that the new proposed framework for TPLC might be promising. But it is still too early to say that TPLC will be the final solution. In addition, the new risk categorisation for AI/ML-based medical devices proposed by CQC & MHRA based on the level of autonomy of the device and scope of the influence might be also a good direction to carry forward the regulation paradigm.

Table 2.3: The similarities & differences of medical device regulation in USA, EU and UK

| | Similarities | Differences |
|-------------------------------------|---|--|
| Medical device classification | Higher risk to the patient and/or the user associated with higher classification | Three classification in the USA Four classification in the EU & UK |
| Regulatory bodies | N/A | Centralised in the USA (FDA) Decentralised in the EU & UK (National Notified Bodies) |
| Regulatory pathway | Higher risk classes require more rigorous approval, e.g. Pre-market approval in the USA and clinical investigation in the EU No specific pathway for AI/ML-based devices Data related to equivalent devices can assist the approval processes | Four specific pathways in the USA Self-declaration for low risk devices in the EU & UK; General requirements for higher risk devices, e.g. safety, performance and reliability in the EU&UK |
| Public access to approval documents | N/A | Available in the USA Limited availability in the EU & UK |
| Approval documentation | N/A | Recorded by the FDA CE mark in Europe UKCA mark in the UK (strictly GB) |
| Post market requirements | Requirements vary with classification Adverse incidents involving medical devices must be reported to the relevant authorities Requirements to withdraw or recall non-conformal devices until the problems are rectified | Medical device tracking and surveillance for certain Class II and Class III devices in the USA A voluntary program of third party inspections of devices in the USA Post-market surveillance report required for Class I device; periodic safety update report required for Class IIa, IIb, III in the EU&UK |

2.4.6 Safety Assurance and Regulation in Other Communities

There has been a growing interest in AI/ML-based medical devices not only from regulatory bodies but also in academia and in industry. This section focuses on the academic and industrial activities.

First, a discussion paper reviewing the FDA’s current approach to regulating AI/ML-based software identifies a number of limitations, e.g. in dealing with AI/ML software that continues to learn in operation [121]. The paper proposes a lifecycle-based framework for regulation and makes three key recommendations:

- Provision of adequate evidence of safety and effectiveness prior to product introduction, ideally gained in clinical trials;
- Identify allowable (safe) changes to the software that would not necessarily require further pre-market assessment using the term “safe harbor”;
- Periodical review of accumulated changes to demonstrate that the risk/benefit profile remains acceptable.

Second is work on reporting for clinical trials involving AI/ML-based medical devices. The CONSORT-AI extension of these reporting guidelines [122] considers clinical trials involving AI/ML-based systems. In particular it identifies the need for:

- Explanations of the intended use of the AI intervention in the context of the clinical pathway;
- Identification and analysis of performance errors;
- Recognition of the limits to generalisability.

The first point is very similar to one of the recommendations of the CQC sandbox [118].

Third is an analysis of trends in AI and the use of such technology in a clinical context, which considered known AI problems, e.g. bias, and their potential impact on clinical safety [79]. It draws on well-known analyses of concrete problems in AI [123] and other sources providing an interpretation of issues such as distributional shift in a clinical context. The paper also proposes a set of tests which help to identify whether or not distributional shift, or the other general AI problems, are of concern for a particular system. Detailed mitigation strategies are outside the scope of the paper but it does suggest that

an adaptation of the Standards for Reporting of Diagnostic Accuracy initiative [124] could help to address bias in ML.

Fourth, the Topol Review [125] focused on the healthcare workforce to deliver the digital future including the educational changes needed for the safe and effective introduction of AI and robotics. Three of the recommendations are relevant to AI/ML-based systems:

- Patients should be involved throughout the design and implementation of AI software (co-design) so their needs and preferences are reflected in the system;
- Resources should be developed to educate and train healthcare professionals in all relevant aspects of AI development including health data provenance, the ethics of AI and critical appraisal and interpretation of AI;
- A national programme of “Industry Exchange Networks” should be established to enable the NHS to access skills in industry.

Fifth, a review of the AI/ML-based systems approved by the FDA in 2020 [126] produced a database which summarises the capabilities of each system, the approval pathway, etc. They found most of the approved AI/ML-based medical devices are cleared through the 510(k) pathway, with a few through the *de novo* pathway and only one through the pre-market assessment pathway. They also found that some medical devices are not announced as AI/ML-based in the FDA database but found that they are claimed to have such technology in other online resources, which suggested that greater clarity is needed to improve the ability to analyse and track the introduction of AI/ML-based medical devices.

Finally, a recent review of current regulatory approaches argues for moving from evaluation of AI/ML-based medical devices to taking a systems view [127]. This is somewhat analogous with the recommendations above to consider an AI/ML-based system in the context of the clinical workflow, but rather broader in that it might ultimately lead to regulation of the practice of medicine. The authors claim that the systems approach would better deal with systems that learn in operation, i.e. adaptive AI/ML-based system, and with decision-making involving collaboration between the AI/ML-based systems and humans. Whilst the paper does not present a concrete model of how to move towards the systems approach, it does recognise the need to define a transition path from current regulatory practices and provides some insights that might assist in the safe and progressive introduction of AI/ML-based medical devices into clinical use. Some of the case studies

in this thesis reflect this systems view, for example showing how to assess the role of ML-based SaMD in the context of a clinical workflow.

2.5 Conclusion

This chapter has provided background for the rest of the thesis, presenting introductions to healthcare, safety engineering and ML. The case studies in Chapters 4 to 6 introduce additional detail where appropriate, e.g. on clinical conditions being addressed, or on ML methods that are used. The discussion of regulation shows the complexity of introducing ML systems into clinical practice. This thesis does not address regulation directly but aims to provide a set of contributions that could help to realise technically sound approaches to support regulatory practices for ML-based SaMD and for supporting safety through life, consistent with the notion of TPLC introduced by the FDA.

Three findings are highlighted from the literature survey, particularly from the analysis of regulation and safety assurance in Section 2.4:

- Most of the approved AI/ML-based medical devices are cleared through the 510(k) pathway, with some through the *de novo* pathway and only a few through the pre-market approval pathway;
- Technology suppliers do not always accurately state whether their products use ML and how their devices perform;
- There is no centralised publicly available database of approved medical devices in EU in contrast with the USA.

Similarly, three over-arching recommendations for assuring ML-based SaMD are given:

- Provide more clarity on how hospitals should implement ML devices within clinical pathways to ensure high-quality care;
- Provide for more assurance about the clinical aspects of the ML algorithms, e.g. providing more guidance to support clinical validation of algorithms;
- Establish an international publicly accessible database for approved AI/ML medical devices.

This thesis makes contributions, especially based on the first two recommendations as will be seen in Chapters 4 and 5.

Chapter 3

Overview of the Thesis

Chapter 1 gave a brief outline of the three main contributions of this thesis which are presented in Chapters 4-6 respectively. The aim in this Chapter is to give an overview of the work that has been done and to give greater clarity on the three main contributions and on their relationships.

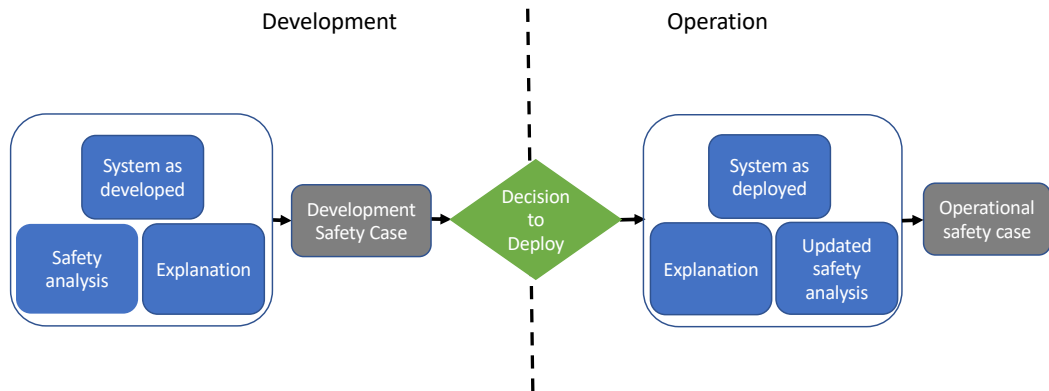


Figure 3.1: Overview of the Thesis

We do this in relationship to an overview diagram (see Figure 3.1) that shows the developmental and operational phases for a system, with the decision to deploy. Although the decision to deploy marks the transition from development to operation, these phases are strongly related and the constituent elements are mirrored across the two phases. These elements are: the *system* itself, *safety analysis*, and *explanations*, all of which support the *safety case*. We made the distinction between the *development safety case* and the *operational safety case*, where the *development safety case* is intended to support the decision to deploy whilst the *operational safety case* builds on the *development safety*

case but is intended to be updated based on the operational information. This also shows that safety is an ongoing issue which needs to be managed throughout the system’s life. In particular, updating the safety case might, in extreme cases, lead to a decision to revoke approval for use of a system where the risks are deemed to be too high.

The *system* in this thesis mainly means ML models but we also include clinical practice. In safety critical domains, *safety analysis* is often used to influence the design of the *system* and to provide assurance of safety by producing evidence to feed into the *safety case*. ML models, by comparison with conventional software, are often viewed as “black boxes”. Explainable AI methods have been developed aiming to provide human interpretable representations of the ML models. These *explanations*, e.g. showing feature importance, can provide evidence for the *safety case* thus helping to assure safety of the ML models. Thus, where the *system* employs ML models, the *system* itself, the *safety analysis* and the *explanations* have a symbiotic relationship and jointly underpin the *safety case*.

Safety analysis is subject to uncertainty, e.g. assumptions about the operational context for the system. When data is available from operation, we can apply ML on the data to update *safety analysis* based on the insights gained. Therefore, we refer to *updated safety analysis* in the operation phase.

We use this overview diagram to show the contributions of the individual case studies, and how we can embrace ML in the safety assurance of healthcare applications. The three contributions are illustrated in the following Chapters 4 to 6, where the scope of each contribution is highlighted using this overview figure.

The first contribution, set out in Chapter 4, addresses question 1 and encompasses the *system as developed*, *safety analysis* and *development safety case*. The case study employs RL in support of sepsis treatment, and is based on a previously published RL model; the RL model and the clinical pathway form the *system as developed*. The case study shows how to use well-established safety engineering methods, specifically SHARD, to proactively incorporate patient safety in the design of the RL model. The results of the SHARD analysis are used to identify DSRs that are used to “drive” the design of the RL model and to produce a new RL model which has better safety properties than the previously published work. Safety controls including, but not limited to, the DSRs for the RL are summarised using bow-tie diagrams and this summary of hazard causes and controls is used to guide the construction of the *development safety case*. The key

contribution in this case study is to show how well-established safety engineering methods can be applied to ML-based SaMD thus showing that they are appropriate and effective in assuring the safety of ML in this healthcare application. In addition, it briefly illustrates the role of explanations, but this is addressed in much more detail in Chapter 5.

The second research question is addressed in different ways by Chapters 5 and 6. The second contribution, set out in Chapter 5, encompasses the *system as developed*, *explanations* and the *development safety case*. The case study employs CNN to assist clinicians in determining when to wean a patient from mechanical ventilation; the CNN model forms the *system as developed*. The focus of Chapter 5 is on *explanations* and explainability, so it incorporates an overview of explainable AI methods and shows how various explainable AI methods can support safety assurance in the context of the ML development process. First, it uses influence functions (a way to identify which input instances have a strong effect on the trained model without retraining the model) to guide the model learning process to meet safety requirements (mainly related to timing of weaning). Second, it uses feature importance (a way to rank or score the input features based on their contribution to the model prediction) to help make the learnt model interpretable by clinicians and hence contribute to assurance of its clinical validity and thus to patient safety. Third it uses counterfactual examples (informally, “what is not, but could have been”) to shed light on model robustness. The chapter explicitly shows the role of these explainable AI methods in contributing to the development safety case for the DSS, and hence their role in contributing to safety assurance. It also illustrates the potential of counterfactual and feature importance *explanations* in operation to support clinical decision-making.

The third contribution, set out in Chapter 6, encompasses *safety analysis*, *updated safety analysis*, and the *operational safety case* (which implicitly draws on the *development safety case*). The case study employs Bayesian Network structure learning to understand the correlations of different factors concerning the delivery of Beta-Blocker for patients who have undergone thoracic surgery and who are potentially at risk of atrial fibrillation; the clinical practice for medication management for such patients forms the *system*. It shows how to use ML to update and enhance the results of safety analysis and the operational safety case, providing a different perspective on research question 2. It shows how a Bayesian Network can extract information from operational data both to confirm (validate) aspects of the safety analysis and to update it to reflect what actually occurs in operation as opposed to what is predicted during safety analysis. This case study illustrates how a

clinical practice might deviate over time, i.e. evolve from development phase to operation phase, so it shows the role of ML in improving safety analysis in general, not just when an ML-based SaMD is used.

The three case studies cover all the artefacts identified in Figure 3.1, with individual contributions as presented above. The limitations of the individual case studies are acknowledged in the relevant chapter. Whilst the individual contributions are worthwhile in their own right, taken together, they illustrate the potential for a rich and supportive interplay between ML and safety engineering in healthcare, as we show in more detail in the following chapters.

Chapter 4

Safety-Driven Design of Machine Learning

This chapter is based on my previous publication [128] [129] [130] and contributes to answering research question 1. ML is becoming more widely employed in healthcare and there are a growing number of ML-based SaMD. The challenges in assuring safety of ML-based SaMD include showing that the learnt model respects relevant safety requirements. Indeed we might characterise the first question as treating ML as “a problem”. At a technical level, answers to this question depend on the ML method used. The case study in this chapter addresses this question in the context of an SaMD which employs RL. The other two case studies have a different focus and show the role of ML as “a solution”.

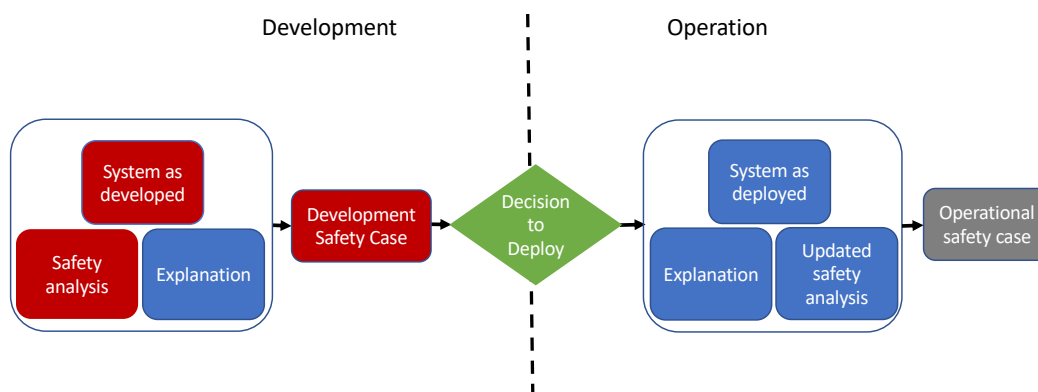


Figure 4.1: Overview for the Case Study

The contribution of the chapter is summarised in Figure 4.1; the elements highlighted in red show the focus of the chapter. 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 chapter tackles the 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 chapter considers the use of deep RL for sepsis treatment. The methodology starts with the modelling of a clinical workflow that integrates the RL model for sepsis treatment recommendations; the clinical workflow and the RL-based DSS form the *System as developed*. Then *Safety analysis* is carried out based on the clinical workflow, identifying hazards and safety requirements for the RL model. The design of the RL 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 *Development safety case* is presented, providing a basis for regulators to make a judgement on the acceptability of introducing the RL 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 (see the discussion in Chapter 2), the work presented in this chapter has shown a possible direction for overcoming this barrier and exploit the benefits of ML without compromising safety.

4.1 Introduction

As outlined in Chapter 2 ML has considerable potential in healthcare, but 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 [131], e.g. DCB0160 from NHS Digital [9]. 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 datasets. Although these issues are starting to be addressed, e.g. by the US Federal Drug Administration (FDA) [15], 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” [118]. 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 [132], 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 chapter provides a concrete clinical case study for sepsis treatment using ML, specifically deep 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 [133]. 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. [88]. 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 satisfy 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 chapter is structured as follows. Section 4.2 discusses the background and related work, including the safety of ML in healthcare with a particular emphasis on sepsis. Section 4.3 describes the methodology we have used in this work covering the clinical, safety and ML elements outlined above. Section 4.4 presents our detailed clinical case study 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 4.5. Section 4.6 presents conclusions.

4.2 Background and Related Work

As discussed in Chapter 2 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. There is a lot of work using supervised learning for classification problems in healthcare, e.g. for breast cancer screening [1, 134]. Comparatively, there is less work employing the other two types of ML in healthcare.

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 [135] and sepsis, identifying four novel phenotypes [136]. RL is an ML technique that is often used in complex decision making tasks to find an optimal strategy [22]. It has been applied to identify optimal treatments in healthcare very recently, e.g. determining treatment regimes in chronic disease and automated medical diagnosis [137]. 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 4.4.

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 [138]. 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 [139]. 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 [138] 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 [140]. 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 [141]. 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 [142]. A third example is an AI-based decision-support tool used to aid anaesthetists in controlling hypotension [143]. 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 [144] 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 FDA is currently investigating [15], proposing a TPLC approach for updating the deployed ML model.

Both “sepsis watch” [139] and the work on “roadmaps” [138] 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.

4.3 Methodology

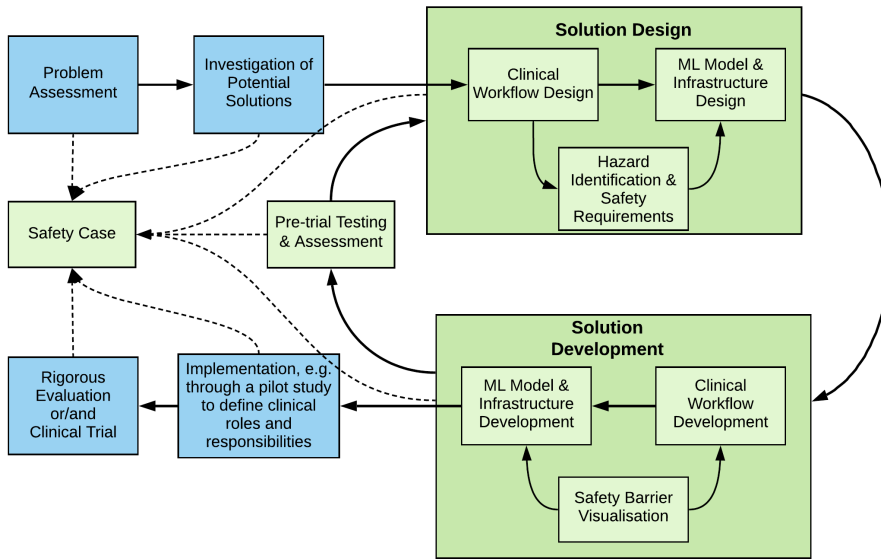


Figure 4.2: Framework for integrating ML system into clinical care

Our methodology is shown in Figure 4.2 and incorporates concepts from “sepsis watch” [139] and the work on “roadmaps” [138], 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 chapter, 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 Figure 4.2. Infrastructure is an important element to enable the deployment of ML models in healthcare but is out of scope for this chapter. The elements marked in blue have a clinical focus and are largely outside the scope of this chapter, although an overview of sepsis is given in Section 4.4 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. 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. *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 data for the model development because the quality and quantity of the data will directly determine how good the ML model can be [145]. 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 [146].

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 chapter, 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 [147], e.g. in the rigorous evaluation through clinical trials. In practice, there is no clear cut distinction between the activities shown in Figure 4.2. 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.

In our methodology, the safety case draws evidence from all the phases in Figure 4.2 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 case study involving treatment of sepsis patients.

4.4 Clinical Case Study: Sepsis Treatment

Basic Concepts of 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.

The clinical case study focuses on the treatment of sepsis. Sepsis is a life-threatening organ dysfunction which is caused by a dysregulated host response to infection [148]. It is estimated that one in five deaths worldwide are due to sepsis [149]. A major challenge is early detection of sepsis since the earlier the treatment begins the greater the chance of patient recovery. Once the condition has been detected, treatment normally involves administration of antibiotics and infection source control. When it turns into septic shock, administration of intravenous (IV) fluids and vasopressors will be necessary, but deciding on the treatment strategy for IV fluids and vasopressors is often difficult. Different fluid and vasopressor treatment strategies have been tested leading to quite different results in terms of patient mortality [150]. Further, many healthcare agencies and communities

have devoted significant efforts to sepsis management, e.g. the Surviving Sepsis Campaign [151]. Despite such efforts, the optimal strategy for the administration of IV fluids and vasopressors remains unclear. Consequently, researchers have harnessed RL to learn the “optimal” treatment strategy for recommending IV fluids and vasopressors, e.g. [152] [88].

We start this work by using a previously published deep RL model [152] for sepsis treatments which recommends IV 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 case study 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.

4.4.1 Clinical Workflow Design and Development

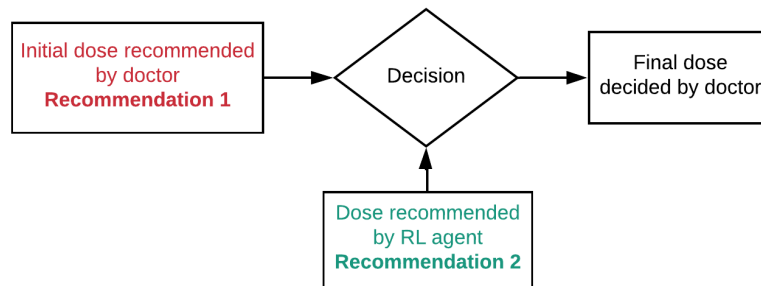


Figure 4.3: High-level workflow design

There are two main ways of introducing ML models into healthcare: either replacing clinicians or assisting them. For example, [1] 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 DSS, assisting clinicians in sepsis treatment as shown in Figure 4.3. First, a doctor recommends initial doses of IV 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 IV 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. Mosier and Skitka [153] defined automation bias as resulting from people's using the outcome of the decision aid "as a heuristic replacement for vigilant information seeking and processing". In this case, we can measure pre- and post-advice decision accuracy from the clinicians as indicators of the influence of automation on decision-making. For example, a negative consultation occurs if a pre-advice decision is correct but changed to an incorrect post-advice decision. 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 Figure 4.4. 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 [154]; if necessary, it can also be altered to suit the local hospital screening protocol. Here, Early Warning Score (EWS) [155] 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 [156] and the Hour-1 Bundle from the Surviving Sepsis Campaign [151]. The Hour-1 bundle is designed for initial resuscitation but IV fluids and vasopressors will continue to be given in the stable period, most likely for several days. This is shown as recommendation 1 in Figure 4.4, and is the doctors' initial recommendation based on current clinical practice. If necessary, recommendation 1 can also be altered to suit the local hospital protocol. Further, the ML model, i.e. the RL agent, is integrated into the clinical workflow shown as recommendation 2, which matches the high-level workflow design in Figure 4.3. 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 [147].

The workflow concludes with the nurses administering the IV 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 4.4.3), 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 (EHR), 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 Healthcare IT (HIT) 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 [157]. It is good practice to employ “user-centred design” [158] 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.

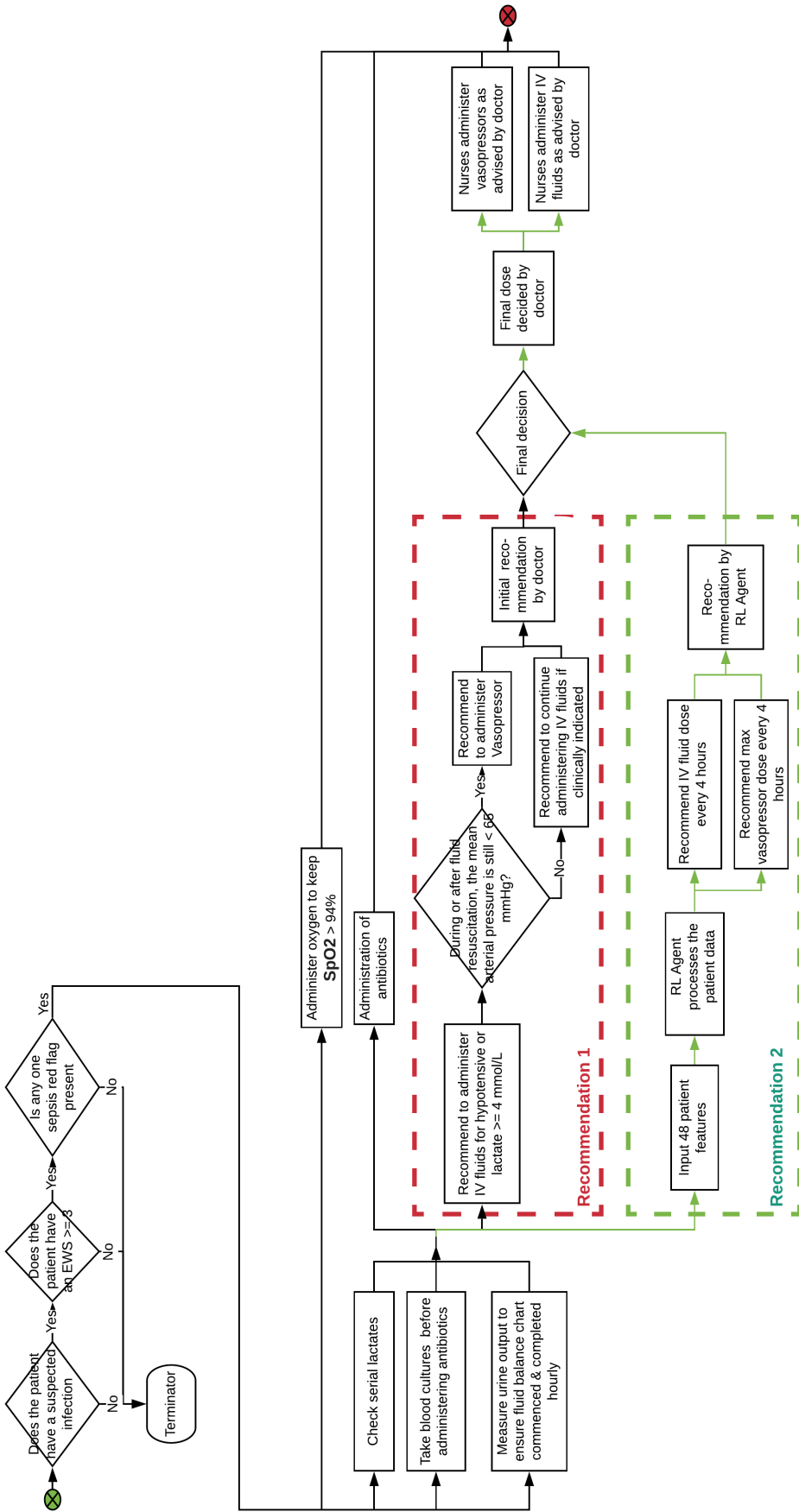


Figure 4.4: The detailed workflow integrating ML model to treat sepsis patient

4.4.2 Hazard Identification and Safety Requirements

The safety engineering process starts by identifying hazards, then 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 and used to 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.

In this chapter, we only show the hazard identification for the delivery of vasopressors; IV fluid can be analysed in a similar way. The analysis was carried out by a multi-disciplinary team comprising two safety engineers, one Intensive Care Consultant and two ML engineers, using the SHARD method (which was introduced in Chapter 2). The hazards identified, prompted by the use of the SHARD guidewords, 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) [159] [160] [161]. 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 [162], changes in Norepinephrine can have rapid effects on patients.

After the identification of the potential hazards, we applied SHARD to the clinical workflow to identify the causes of the hazards. This is done by going through each activity (the rectangular boxes) in Figure 4.4 with a focus on recommendation 2, i.e. the part of the workflow marked in green. Table 4.1 shows a fragment of the SHARD analysis with a focus on one hazard – *sudden change of vasopressor dose administered* – identified above. The full SHARD analysis can be found in Appendix A. Table 4.1 is a high-level summary of the analysis. The full analysis is also included in the material at the link above, 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 [163]. 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*) 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 administered”. 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 an omission of the final dose recommendation, which ultimately contributes to the hazard “no vasopressor administered”.

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 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 4.1. The severity classification used in the table is based on the DCB160 standard developed by NHS digital [9].

As indicated above, Table 4.1 summarises the detailed analysis, combining the results from analysing all the different activities in the workflow in Figure 4.4. The possible causes

Table 4.1: 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 | Acute Hypotension, Strokes, Renal failure, Heart attack could occur from a sharp drop in the dose | Major/ considerable |
| | | 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 | | |
| | | 10 Features in state space of the RL model are not sufficient to represent the patient conditions for sepsis decision making | Hypertension, Cardiac Arrhythmia, Strokes, Raised intracranial pressure, Pulmonary oedema could occur from a sharp rise in the dose | |
| | | 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 | | |
| | | 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 | | |

of most interest in this chapter 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 initial activity in recommendation 2 in the workflow. The possible causes in Table 4.1 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 chapter, the visualisation of controls in Section 4.4.5 addresses some of the other possible causes identified in Table 4.1.

Table 4.2: 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 |

Safety requirements are derived from the hazard analysis to control the hazard causes identified in Table 4.1. 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 Figure 4.4 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 4.2. 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 4.1. Similarly, it is allocated to “RL model design” as this 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 4.2 covers all the RL agent-related causes in Table 4.1 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 4.1 should be addressed in the user interface design to reduce the likelihood that such causes arise.

4.4.3 Model Design and Development

In this chapter, we have adapted the RL model in [152] 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 pre-morbid status, vital signs, laboratory values, fluids and vasopressors received to satisfy safety requirement R1 in Table 4.2. A definition of the features used in the RL model together with a feature correlation matrix is presented in Appendix B. The action space includes 25 possible actions with five discretised choices for the dose of IV fluids and five for vasopressors respectively, as shown in Table 4.3. The terminal reward is based on 90-day mortality (as against hospital-mortality in the original work) with +15 for survival and -15 otherwise. The intermediate reward uses Sequential Organ Failure Assessment (SOFA) score and Arterial Lactate (the level of lactate from arterial blood) as they did in the original work to satisfy safety requirement R2. The detailed intermediate reward function is shown in equation 4.1. 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) [164] 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.

$$r(s_t, s_{t+1}) = C_0 \mathbb{1}(s_{t+1}^{SOFA} = s_t^{SOFA} \ \& \ s_{t+1}^{SOFA} > 0) + C_1(s_{t+1}^{SOFA} - s_t^{SOFA}) + C_2 \tanh(s_{t+1}^{Lactate} - s_t^{Lactate}) \quad (4.1)$$

Principles of Deep Q-Networks (DQN)

DQN is a widely-used modern RL algorithm, which combines Q-learning [165] 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, parameterised 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.

In double DQN, the prediction network θ is used to select the greedy action $a' = \operatorname{argmax}_{a'} Q(s', a'; \theta)$, while the target network θ' is used to estimate its Q -value. The standard double DQN loss is shown in Equation (4.2).

$$L(\theta) = E[(Q_{double-target} - Q(s, a; \theta))^2], \quad (4.2)$$

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

The data used for model development is based on the same dataset and the same patient cohort taken from MIMIC-III – a large publicly available database [166] – as in the original work. Patients are included in the cohort when they meet the sepsis-3

Table 4.3: Dosage actions

| | | Dose of vasopressor (mcg/kg/min) | | | | |
|---------------------------|---|----------------------------------|----------------|-------------|---------------|---------------|
| | | No.: 0 | 1 | 2 | 3 | 4 |
| | | Range: 0 | (0.002, 0.079) | (0.08, 0.2) | (0.201,0.449) | (0.45, 1.005) |
| | | Median: 0 | 0.04 | 0.135 | 0.27 | 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 |

criteria [148] – suspected infections combined with SOFA score ≥ 2 . Exclusion criteria are: 1. not adult, 2. IV fluid intake not documented, 3. possible withdrawal of treatment, 4. erroneous intake or output data. Patients’ data were coded as multidimensional discrete time series with 4 hour time steps. There is no “right” time step but 4 hours was chosen as this is long enough that changes in vasopressor dose could arise. In addition, using this time step allows direct comparison with the original work. The detailed MIMIC-III data pre-processing can be found in the supplement to [88]. This satisfies safety requirement R5. The resulting patient cohorts were divided into a training dataset (80%, 20,938), a validation dataset (10%, 2,149) and a testing dataset (10%, 2,160). For detailed patient features included in the state space, see the supplement to [88]. 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 4.4.4.

The RL model was developed in Python and uses the TensorFlow library [167]; the code developed is available at: <https://github.com/Yanjiayork/sepsisRL>. As the MIMIC-III dataset 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 dataset, 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.

4.4.4 Pre-trial Testing and 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 our work, we first evaluate the original policy from the safety perspective – specifically in terms of sudden changes in the recommended vasopressor dosage by the RL agent, given our focus on this hazardous interface failure. Then, we evaluate the policies from both performance and explanation perspectives.

4.4.4.1 Safety Evaluation

According to [168], 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 4.3 in Section 4.4.3, 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 for 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 dataset, which has 2,160 patients, by calculating the max absolute vasopressor dose change in one step for each patient during their treatment. Figure 4.5 shows the comparison of max absolute vasopressor dose change between the clinician policy and the original policy for these 2,160 patients. 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. The apparent “noise” in Figure 4.5 arises because the patients are sorted (ordered) first by the maximum change in the test data (i.e. the clinician policy), then by the maximum change in the original policy and, for some patients the clinician policy gave a higher maximum change than the original policy. Table 4.4 shows the exact number of patients for the different dose changes. 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. This is consistent with Figure 4.5.

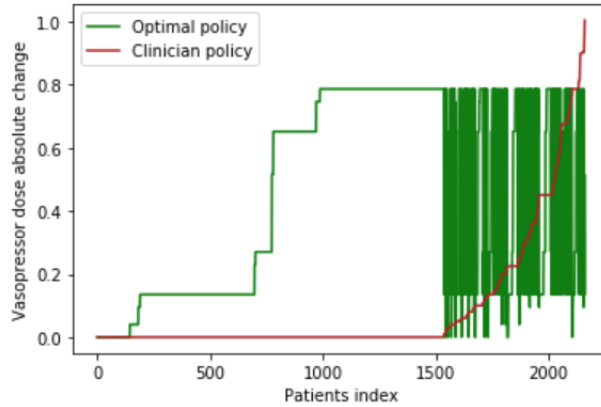


Figure 4.5: Original Policy: Comparison of max absolute vasopressor dose change in one step for each patient in the test dataset between the clinician and the learnt optimal policy

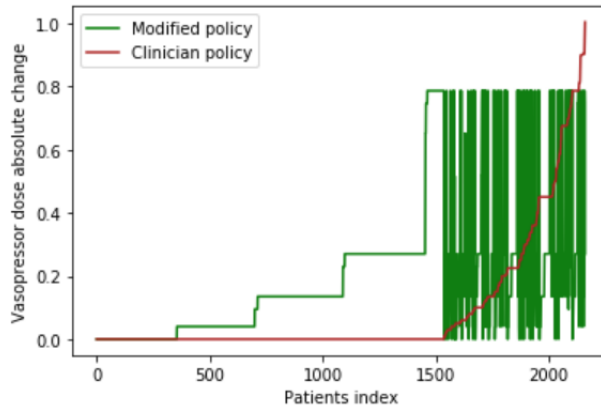


Figure 4.6: Modified Policy: Comparison of max absolute vasopressor dose change in one step for each patient in the test dataset between the clinician and the learnt modified policy

In response to the above clinical safety concerns, we modified the model in order to further satisfy safety requirement R0 in Table 4.2, which is to reduce the rate of sudden major vasopressor dose changes to be closer to the 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 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 4.5.

Table 4.4: 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) |

Table 4.5: 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. Figure 4.6 shows the maximum absolute vasopressor dose change in one step for each patient between the clinician policy and the modified policy. It shows a clear reduction in sudden major dose changes and the absolute change is much more reasonable compared to Figure 4.5. The exact number of patients for the different dose changes in the modified policy are also shown in Table 4.4. Table 4.4 shows that there are 8% (170 patients)

amongst the 2,160 patients in the test dataset 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 my previous paper [128].

4.4.4.2 Performance Evaluation

Table 4.6: Performance comparison for different policies

| Policy | Estimated Discounted Reward |
|------------------|-----------------------------|
| Clinician policy | 7.16 |
| Original policy | 10.9 |
| Modified policy | 8.07 |

It is not feasible to evaluate the policy on real patients because of ethical, legal and risk issues. Instead, we have carried out off-policy evaluation to assess the performance of the original policy and the modified policy by fitting an MDP model \widehat{M} from the current data to approximate the environment. The fitted \widehat{M} can then be used to estimate what the next state is if the agent follows a different policy from the clinician policy, which is what recorded in the dataset, at a specific state. Once the next state is estimated, we then could estimate the reward \widehat{R} for the agent following original policy and the modified policy respectively using equation 4.1. Finally the H-step discounted value of the original policy and the modified policy can be computed using the estimated reward \widehat{R} recursively following the equation $v = E[\sum_{t=1}^H \gamma^{t-1} r_t]$, where γ is the discount factor and r_t is the observed immediate reward at step t. The final estimated value averaged the resulting value function across all the observed trajectories in the test dataset (refer to [169] [170] for a detailed description of the method). The average discounted reward of the chosen actions under the clinician policy across all of the trajectories in the test dataset is also calculated as the benchmark, as shown in Table 4.6. It shows that the original policy has a higher value than our modified policy. However, our modified policy is still higher than the clinician policy and in terms of vasopressor delivery, it is safer in the sense of avoiding

sudden vasopressor changes and its dangerous effects on patients.

4.4.4.3 Model Explanations

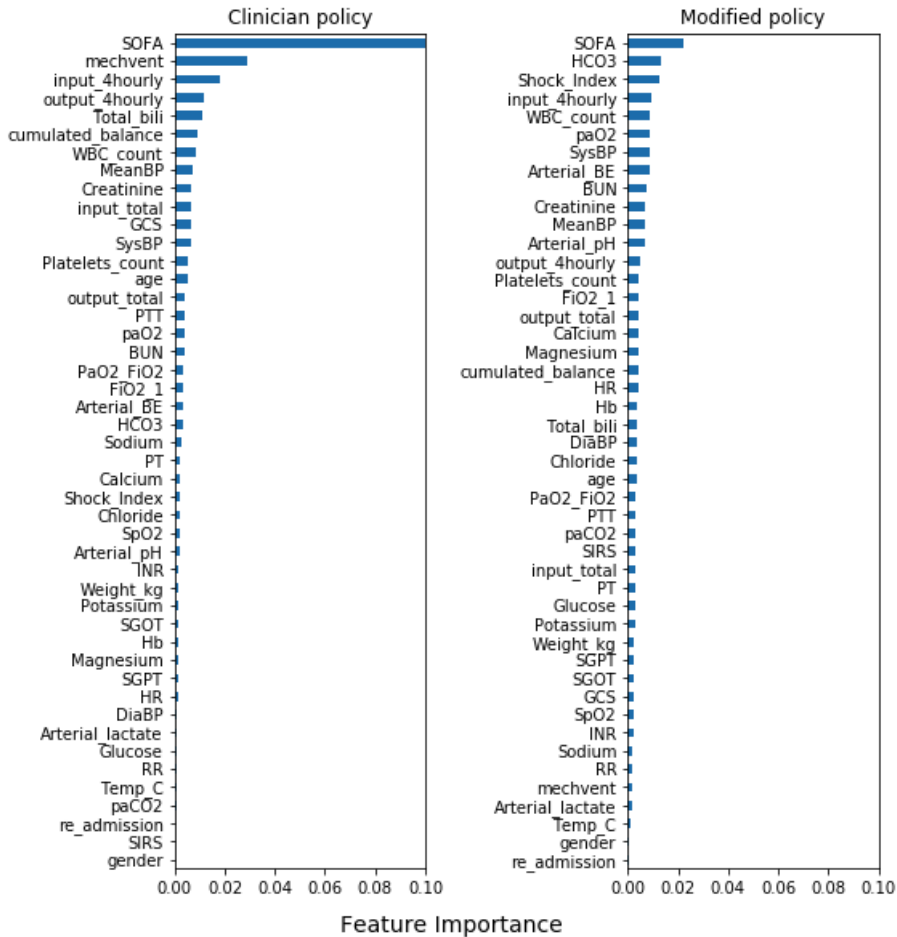


Figure 4.7: Feature importance (from out of bag score) for clinician policy and the modified policy

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 [171]. 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 recommending vasopressor in the modified policy and the clinician policy, see Figure 4.7. Note that 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 va-

sopressor 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. Note that the label for training these two random forest classifiers is generated based on the different policies of interest and there is no ground truth in this case.

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 [172]. 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 Figure 4.7 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 [173]. 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.

Further discussion of ML model explanations is presented in Chapter 5.

4.4.5 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 Figure 4.2. 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 [174]. 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 arise 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 [175] to produce the BTDs and safety case (see Section 4.4.6). AdvoCATE is an advanced Assurance Case Automation Toolset developed by NASA. Two linked BTDs are presented as follows: Figure 4.8 presents the BTD for the hazardous interface failure “RL agent recommends a sharp change in dose” and Figure 4.9 presents the BTD for the hazard “sudden change in vasopressor dosage administered” which also shows the role of the hazardous interface failure, and its patient safety impact within the clinical workflow (as modelled in Figure 4.4).

We start with Figure 4.8. 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 Figure 4.3;
- 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 4.1;
- 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 Figure 4.8 are linked to the previous sections, we consider one of the threats at the bottom left of the figure, specifically “Cost

function for RL model development is not appropriate”. This threat is cause 12 in Table 4.1 and it is addressed by safety requirement R3 in Table 4.2. There are a total of 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 4.4.4) in order to further satisfy requirement R3 also shown in Table 4.5. 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.

Figure 4.9 presents a partial BTD for the hazard “Sudden change of vasopressor dose administered” (Figure 4.8 and Figure. 4.9 link to form a more complete BTD). The events in Figure 4.9 link directly to the causes in Table 4.1, 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 4.1. In addition, the hazardous interface failure is also shown as a threat in Figure 4.9, 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 Figure 4.9 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 Figure 4.9 are:

- Context – the final activity in the workflow in Figure 4.3;
- Top event – the hazard “sudden change in vasopressor dosage administered”;
- Threats – causes from the SHARD analysis in Table 4.1 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 Figure 4.9, 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 Figure 4.3. The implementation of this control is explained in Section 4.4.4 and illustrated in Figure 4.7 by showing the feature importance for the modified policy.

The BTDs are an important result of iteration through the framework shown in Figure 4.2. The phases are not linear and may be visited multiple times, e.g. as is shown in Section 4.4.4 where model design and development is revisited, responding to the propensity of the original 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.

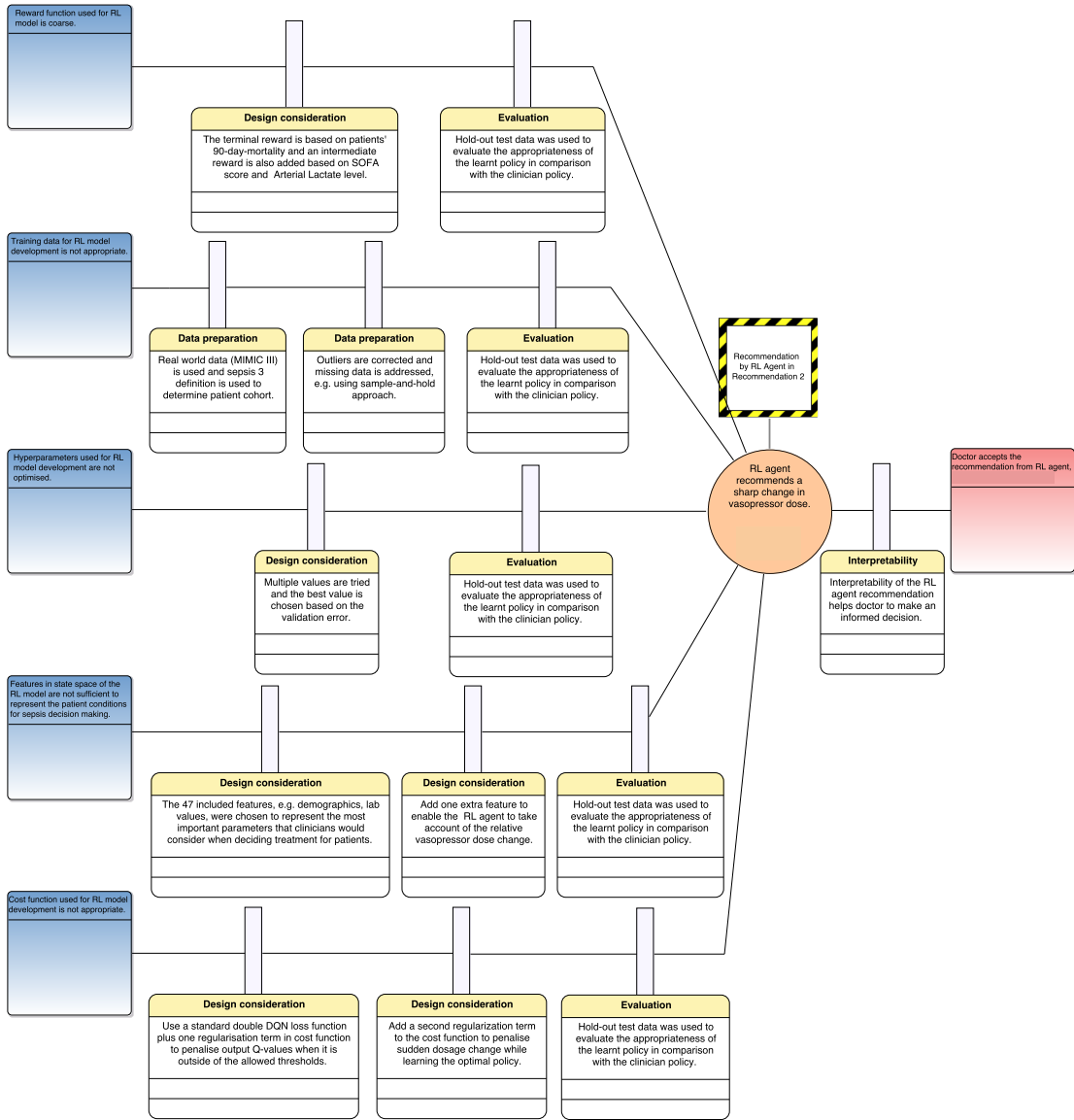


Figure 4.8: Bow Tie Diagram for interface hazard "RL agent recommends a sharp change in dose"

4.4.6 Safety Case

All the phases of the methodology in Figure 4.2 feed into the safety case. The safety case is generated using AdvoCATE, as we mentioned earlier; the goals are automatically numbered by the tool, so the numbers are not always obvious.

Here, we present two linked safety arguments in Figure 4.10 and Figure 4.11 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 chapter 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 Figure 4.4 and the RL model. This goal is then supported by *strategy* S1 which is to argue over the hazards and 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 study the hazards are identified through the SHARD analysis in Section 4.4.2. G1 is supported by showing how the hazards are controlled. 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 Figure 4.10, *goal* G5 is further decomposed across the barriers shown in Figures 4.8 and 4.9. 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 overall performance of the model, G8 relating to the safety requirements in Table 4.2, which includes data preparation and “design considerations” described in Section 4.4.3 and G10 relating to the “interpretability” described in Section 4.4.4. G7 is supported by Table 4.6, which shows the overall performance of the modified policy. G10 is supported by the *solution* Figure 4.7 showing the “Ranked feature importance using the random forest tree”. Data preparation is included as a barrier in Figure 4.8 and it is integrated with G8 as it is one of the safety requirements in Table 4.2.

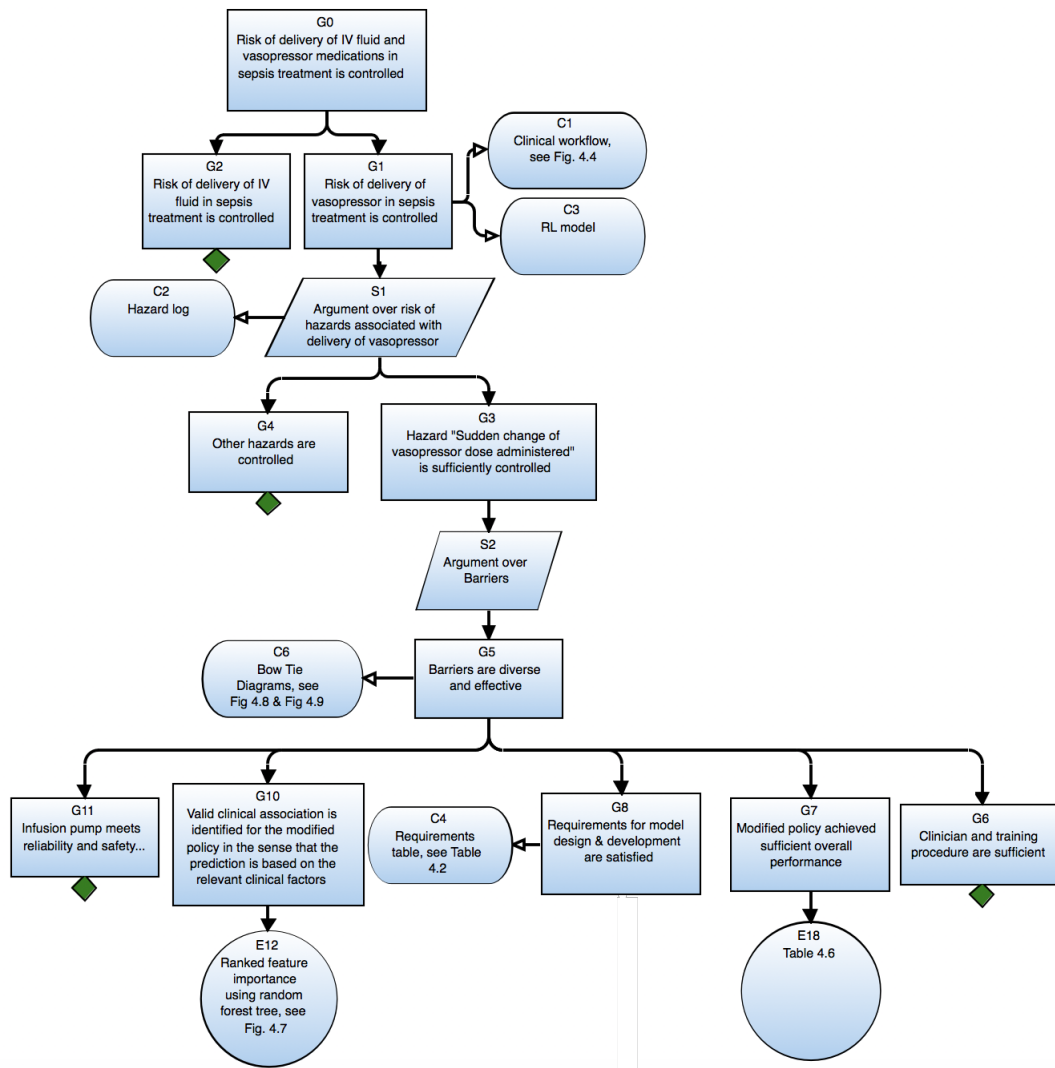


Figure 4.10: Top Safety Argument

Goal G8, “Requirements for model design and development are satisfied”, is further decomposed, in Figure 4.11, into six goals that address the safety requirements R0 to R5. Goal G26 and solution E13 provide direct support for requirement R0, in terms of the evaluated safety performance. Evidence E13 includes Table 4.4 and Figures 4.5 and 4.6 which compare the original and modified learnt policies with the clinician policy. Requirement R0 is further supported by the other goals which relate to the five more detailed requirements, R1 to R5.

The four goals G12 to G15 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

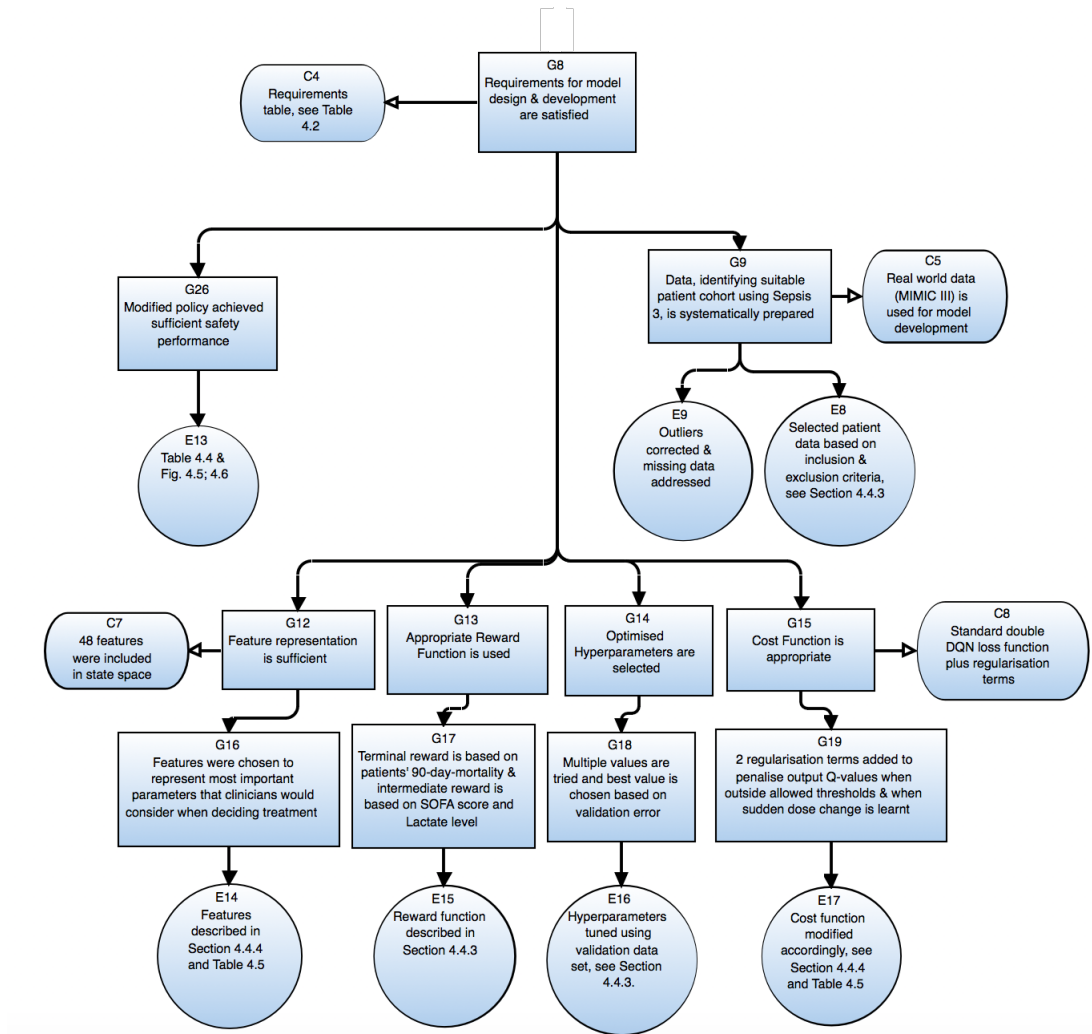


Figure 4.11: G8 Safety Argument

the relative dose change in the state space (see Section 4.4.4) and thus meeting *goal* G12. The other *sub-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 4.4.3 and 4.4.4. Finally, *goal* G9 “data preparation to identify a suitable patient cohort” is solved directly, reflecting the removal of outliers (*solution* E9) and the patient inclusion criteria set out in section 4.4.3 (*solution* E8). 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.

4.5 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 [176], which highlights the difficulties being faced. The work reported in this chapter has 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 case study which concerns sepsis treatment. The clinical case study 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 it would need to be tested 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 [177]. 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 multi-disciplinary approach but also including safety engineers, in contrast to earlier work, e.g. [139]. 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 the clinical case study 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 from time to time. Although neither our methodology nor the clinical case study extend into operations at this stage, it is essential that safety and risk continues to be monitored in operation. In Chapter 6 we show how ML can be used on operational data to inform updates.

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) [118] 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.

4.6 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 chapter 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 recent 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 case study 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 multi-disciplinary 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.

Further, 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” [118]. 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 4.2. 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.

Returning to the research questions, this case study provides a positive answer to question 1: *are well-established safety engineering methods still appropriate and effective in assuring the safety of ML in some representative healthcare scenarios?* They are appropriate – the case study has demonstrated that they are applicable and give sound results which are credible in the real world. They are effective – the case study shows that they contribute to assuring safety. Specifically, DSRs R1 & R3 improve the design of the ML model and the safety analysis results provide evidence for the safety case.

This chapter provides a positive answer to question 1 in a specific context. However, since the scope and focus of this thesis is healthcare, we have no evidence that the methodology introduced here would generalise to other domains. Because we used RL in this case study we would have more confidence that the methodology would apply to other applications in healthcare using RL for making treatment recommendations. Exploration of the methodology in other domains, or with other ML models, would be a major undertaking which is outside the scope of this thesis.

Chapter 5

The Role of Explainability in Assuring Safety of Machine Learning

This chapter is based on my previous publication [178] [179] [180] [181] and contributes to answering research question 2. Established approaches to assuring safety-critical systems and software are difficult to apply to systems employing ML. In many cases, ML is used on ill-defined problems, e.g. optimising sepsis treatment, where there is no clear, pre-defined specification against which to assess validity. This problem is exacerbated by the “opaque” nature of ML where the learnt model is not amenable to human scrutiny. Explainable AI methods have been proposed to tackle such issues by producing human-interpretable representations of ML models which can help users to gain confidence and build trust in the ML system. However, there is not much work explicitly investigating the role of explainability for safety assurance in the context of ML development. This chapter identifies ways in which explainable AI methods can contribute to safety assurance of ML-based systems.

The contribution of the chapter is summarised in Figure 5.1; as in Chapter 4, the elements highlighted in red show the focus of the chapter. The case study is based on a concrete ML-based clinical DSS, concerning weaning of patients from mechanical ventilation. This DSS was developed for the case study and constitutes the *System as developed*. A range of explainable AI methods were employed to produce *Explanations* which, in turn, provide evidence to support safety assurance. The results are also presented in a

Development safety case to show where, and in what way, explainable AI methods can contribute to a safety case prior to deployment. Further, the case study briefly explores the role of *Explanations* to provide assurance in operation.

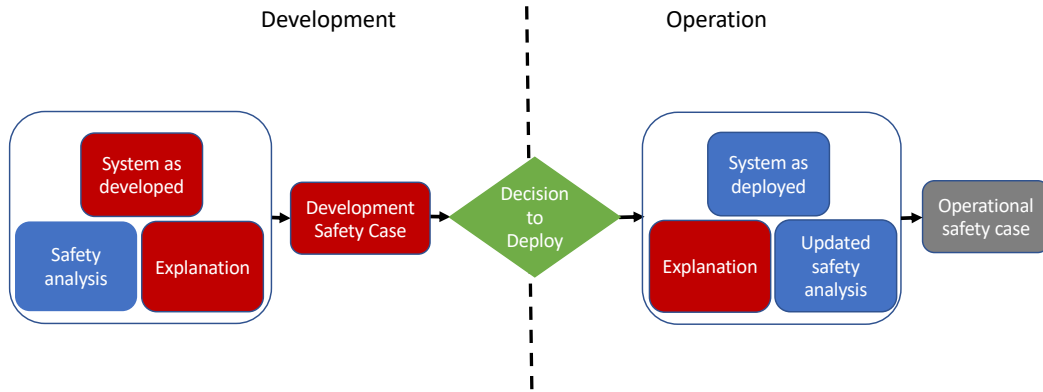


Figure 5.1: Overview for the Case Study

5.1 Introduction

In healthcare, ML is used on various problems, e.g. learning optimal treatments, or to detect abnormalities in radiology images, where it has achieved outstanding performance. However, assuring safety for such systems employing ML remains a challenge. In many domains there are well-established approaches and standards for assuring safety-critical systems and software. Assurance means establishing *justified confidence* in the system for its intended use. The assurance principles underlying these standards include validating that the system works as intended and verifying that the system meets explicit *safety* requirements. These assurance principles remain essential for systems employing ML. However, the details of these approaches and standards can be difficult to apply where systems use ML.

First, the established approaches are based, implicitly or explicitly, on the *V life-cycle model* moving from requirements, through design onto implementation then testing. In contrast, the development of ML-based systems follows a very different, much more iterative, life-cycle with four main phases: data management, ML algorithm selection, model learning, and model verification & validation, which makes it hard to apply established methods. Some emerging standards and guidance better reflect the ML life-cycle, e.g. the US Federal Drug Administration (FDA) proposed regulatory framework on AI/ML-

based Software as a Medical Device (SaMD) [15] and Assurance of Machine Learning for Autonomous Systems (AMLAS) [182].

Second, because of the “black box” (opaque) nature of the ML models [183], it is hard to assess what has been learnt, which exacerbates the challenging of defining concrete requirements for the safety of SaMD in its clinical context. Instead, human performance is often used as a “gold standard” and the current practice is often to (seek to) achieve performance that is better than humans. This makes validation difficult as human performance is variable both from individual-to-individual and over time for a single individual. Also, performance will vary from patient-to-patient, e.g., with comorbidities, and clinicians might not agree on the best treatment strategy.

To overcome such problems, the ML community is actively studying “explainability”, which is intended to “peek inside the black box” and to illuminate the underlying workings of the ML models. Explainability is often equated with producing explainable artificial intelligence methods, which seek to provide human interpretable representations of ML models [79]. Although there is considerable variation in the definition of terms such as explainability, interpretability and transparency, in this thesis we adopt the view from the FDA AI/ML-enabled Medical Devices Transparency Workshop [184] that explainability is one component of transparency. Transparency is a much broader concept in their definition and we see interpretability as a necessary facet of explainability, as suggested by Gilpin et al [185].

Explainable AI methods produce explanations which can be local, i.e. relate to a specific output or prediction of an ML model or global, i.e. explain the ML model as a whole. Explainable AI methods can therefore, in principle, have a role in validation by giving stakeholders, including clinicians, assurance that the ML model will produce valid predictions beyond the data used in development. We will consider the role of various explainable AI methods in safety assurance. Our focus is on development activities and deployment decisions for ML-based systems. Operations and incident investigation are outside the scope of this chapter, although we briefly consider the potential role of explainability in operations.

The chapter identifies the ways in which explainable AI methods can contribute to safety assurance of ML-based systems and demonstrates the role of various explainable AI methods using a clinical case study concerned with predicting readiness for extubation from mechanical ventilation. Further, it shows the potential use of these methods in supporting

a safety case for ML systems. The rest of the chapter is structured as follows. Section 5.2 presents the background and related work with Section 5.2.1 amplifying on the challenges of assuring ML-based systems and the limitations of established safety assurance methods when addressing ML and Section 5.2.2 giving an introduction to the different types of explainable AI methods. Section 5.3 considers the need for explainability in the ML life-cycle, including the potential roles of explainable AI methods. These potential roles are then illustrated in Section 5.4 using an clinical case study of a DSS for weaning patients from mechanical ventilation; the section also contains details on the development of the DSS. This is followed by a discussion and conclusions in Sections 5.5 and 5.6, respectively.

5.2 Background and Related Work

5.2.1 Challenges of Assuring ML-based Systems

This section discusses established approaches to assurance of safety-critical systems and identifies their limitations when dealing with systems employing ML.

We use the term *assurance* to mean *confidence that the system behaviour is as intended in the environment of use*, where *as intended* includes being safe. In this context, we are interested in assurance of patient safety when ML-based systems are used in a healthcare context.

Most approaches to assurance emphasise verification and validation, although the definitions of the terms can vary. The International Medical Device Regulators Forum (IMDRF) define the terms as follows:

- Verification – confirmation through provision of objective evidence that specified requirements have been fulfilled [14];
- Validation – confirmation through provision of objective evidence that the requirements for a specific intended use or application have been fulfilled [14].

To interpret these definitions we can say that validation is concerned with *building the right system*, including defining requirements that meet our intent and that verification is concerned with *building the system right* by verifying that the system meets these requirements. Verification & Validation (V & V) should encompass safety requirements and, as previously explained, safety engineering methods can be used to identify hazards and to assess risks. Where risks are deemed too high, DSRs are identified to reduce the

likelihood of hazard occurrence, e.g. by controlling hazard causes, or to mitigate the consequences of the hazard should it arise. Chapter 4 has shown how these established safety engineering methods can be applied to ML-based systems in healthcare. However, where requirements are not stated explicitly, explainable AI methods can help by providing explanations that enable direct validation of the ML model as a whole, e.g. predictions are based on valid clinical factors and consistent with clinical knowledge. In this chapter we will show how explainable AI methods can contribute to safety assurance and provide evidence to feed into a safety case.

There are a number of initiatives concerned with the assurance of ML in safety-critical systems both in healthcare and more generally. For example, AMLAS defines a process for assurance of the safety of ML-based systems to reflect the ML development life-cycle, which identifies both evidence artefacts and argument patterns (standard forms of argument that can be instantiated for a particular system) in GSN. AMLAS also considers issues of the robustness of ML-based systems, e.g. response to unexpected inputs. The FDA also proposed a total life-cycle regulatory approach for ML-based SaMD [15]. However, these approaches are evolving in that they provide good high-level guidance and objectives, but how to meet such objectives is not sufficiently detailed. The work we present here is intended to be complementary to, and build on, these approaches and shows how XAI methods can provide evidence to meet these objectives, and thus contributes to improving their maturity.

In addition, it is always desirable to consider assurance “through life”, as proposed by the FDA [15], not just as an activity undertaken prior to deployment. This includes getting feedback from operations to check whether or not the assumptions made in pre-deployment assurance activities are sound. This is even more important for ML-based systems than it is for “conventional” systems because of the opacity of ML models. However, there are other important aspects of using explainable AI methods for operational assurance for ML-based systems including the need to show compliance with legal frameworks such as the General Data Protection Regulations (GDPR) [186]. Further, as performance criteria for ML models tend to give only statistical assurance, e.g. 93% accuracy, explainable AI methods can have an important role in giving concrete insights to system users, e.g. clinicians, related to a specific prediction. Explainable AI methods might also have a role in accident and incident investigation, see [181] for a discussion, but this is outside the scope of this chapter.

5.2.2 Explainable AI Methods

The study of explainable AI methods seeks to provide insight into how and why ML models make their predictions. Work on explainable AI includes formalising definitions of explainability [187] [188], development of explainable AI methods themselves and establishing evaluation methods. In this section we provide a brief introduction to some relevant explainable AI methods. There are many different ways to categorise explainable AI methods, e.g. local or global based on the scope of the explanation. Here we present explainable AI methods in three different classes based on the explanation generating mechanism, as shown in Table 5.1.

Some ML models are perceived as intrinsically interpretable to the user, so we refer to these as **interpretable models**. This includes linear/logistic regression, decision trees, K-nearest neighbours, decision rules, Bayesian models, general additive models (GAMs), etc. Note that, although normally these models are viewed as intrinsically interpretable, when the number of input features are beyond human ability to grasp, e.g. when a decision tree is very deep, it is still difficult for humans to interpret the model. In addition, these interpretable ML models are often used as a surrogate to approximate other complex ML models giving insight into the more complex ML model [171]. Figure 4.7 is an example of using a simpler ML model, RF to approximate the complex RL model.

When it comes to explaining more complex ML models, e.g. NNs, which are not intrinsically interpretable, a *post-hoc* explanation can be used to provide insights without knowing the mechanisms by which the model works (e.g. by showing feature importance). There are two main *post-hoc* explanation classes: feature importance and example-based explanations. Feature importance is the more widely researched method [185], which can be model-agnostic (explainable ML methods that work for any class of ML model) or model-specific (explainable AI methods that work only for a given class of ML model). Example-based methods were relatively recently proposed and are often model-agnostic. We now describe each of these two classes in more detail.

5.2.2.1 Feature Importance Methods

Generally, feature importance methods for complex ML models try to build a simpler model than the original one (sometimes known as the “explanation model”), as the original model is hard to interpret. Lundberg [189] has pointed out that many current feature importance methods use the same explanation model, which is a linear function summing

the effects of all feature attributions to approximate the output of the original model; methods that match this definition are called additive feature attribution methods.

Feature importance methods rank or score the input features based on their influence on the model prediction. There are two main ways to obtain the feature importance score, one is perturbation-based and the another is gradient-based.

Perturbation-based methods observe the difference between the original model prediction and the prediction after perturbation by removing, masking or altering an input feature or set of input features. This approach has wide applicability and can be used on image, tabular, or textual data [190] [191]. For example, perturbation was implemented for image classification by occluding different segments of an input image and observing the change in the predicted probability of the classification [192]. There are several popular perturbation-based methods.

Local Interpretable Model-Agnostic Explanations (LIME) [193] provides local explanations by approximating a complex ML model with an interpretable model which can then be used to explain the prediction. LIME is based on the assumption that it is possible to fit an interpretable model around a single input sample that mimics the local behaviour of the complex ML model.

Several perturbation-based explainable AI methods are based on **Shapley values** from cooperative game theory [194], which provide a way to assign the gain from a cooperative game to its players. Shapley values are used to explain a model prediction by treating input features as the players and the model prediction as the gain resulting from the cooperative game. Computing Shapley values is exponential in the size of the model input features, hence approximate methods have been proposed, e.g. aggregation based methods [195] and Monte Carlo sampling [196]. There are also approaches for graph-structured data such as natural language text and images [197].

SHapley Additive exPlanations (SHAP) [189] is another approximation for Shapley values. KernelSHAP is a model agnostic weighted linear regression approximation of the exact Shapley value inspired by LIME. TreeSHAP [198] is an efficient estimation approach for tree-based models and is model-specific. The work on SHAP has wider significance as it has defined a new class of additive feature importance measures, unifying several existing explainable AI methods [189].

Perturbation-based explainable AI methods tend to be very slow since they perturb a single input feature or set of features at a time, so the computational cost increases as

the number of input features in the ML model grows. Further as complex ML models are typically nonlinear, explanation is heavily dependent on the (size of the) set of features that are perturbed at the same time. In contrast, gradient-based methods are potentially more efficient.

In essence, gradient-based methods calculate the gradient of the output with respect to the input. For example, in an image classification task a “saliency map” is produced by calculating the gradient of the output with respect to the input, identifying pixels that have a significant influence on the classification [199]. There are a number of variants of gradient-based methods. **Gradient * Input** multiplies the gradient (strictly the partial derivative) by the input value to improve the sharpness of feature importance [200]. Similarly, **Integrated Gradients** computes the average gradient of the output with respect to each input feature by integrating from a baseline to the current feature value [201]. It is one of the popular additive feature attribution methods. **DeepLIFT** (Deep Learning Important Features) [202] works with deep NNs and it is a good approximation to the **Integrated Gradients** method [203]. Similar to integrated gradients, it also defines a “reference activation” which is often viewed as “uninformative” in context, e.g. a totally black image for image classification. It works by comparing the activation of each neuron to its “reference activation” and uses the difference to determine an importance score for each input.

5.2.2.2 Example-Based Methods

Example-based explanations explain the ML model by selecting particular instances from the dataset or creating new instances. It comprises counterfactual explanation, adversarial examples and influential instances, see Table 5.1.

Counterfactual explanations for ML models were introduced by Wachter et al [204] but bear similarities to earlier work in psychology [205]. Counterfactuals can be thought of as “what is not, but could have been”. Counterfactual explanation is intended to produce a sparse human-interpretable example by changing some input features to achieve a different output, for example, when the ML model predicts that the patient should continue with mechanical ventilation, counterfactual explanations would provide the clinician with information on which features related to this patient need to change, such as successful completion of a Spontaneous Breathing Trial (SBT), in order to change the prediction.

Generally, given an input x , an ML classifier f , and a distance metric d , a counterfactual explanation x' which produces the desired output y can be generated by solving the optimisation problem:

$$x' = \operatorname{argmin}\{y_{\text{loss}}(f(x'), y) + d(x, x')\} \quad (5.1)$$

where y_{loss} “pushes” the counterfactual x' towards a different classification than the initial input x , and the second term keeps the counterfactual x' close to the initial input x . There are four desirable properties for identifying good counterfactuals [171]. First, they should achieve the desired outcome as closely as possible, which is related to the first term in Equation 5.1. Second, the counterfactuals should be as close as possible to the original instance, which is related to the second term in Equation 5.1, i.e. the distance measure. Third, the counterfactuals should be *sparse*, i.e. an ideal counterfactual needs to change only a small number of features from the original instance. Fourth, it is desirable to have *diverse* counterfactuals, which can give the user a choice of what features to change, given the feasibility of the change. On-going research seeks to incorporate these properties in the loss function and optimisation methods. An overview of existing counterfactual explanation methods for ML is provided by Verma et al [206]. In this chapter we used the Diverse Counterfactual Examples (DiCE) method [207], which can produce diverse counterfactual examples.

Adversarial examples are typically generated by adding small, intentional perturbations to the input features to cause an ML model to make an incorrect prediction [208]. There are many techniques to create adversarial examples, e.g. by minimising the distance between the adversarial example and the input instance, which is similar to counterfactual examples. However, adversarial examples are intended to deceive the ML model instead of interpreting the model. Therefore, the changes in the inputs are often imperceptible for a human observer, which makes it more popular for use in object classification [209] [210] [211]. For example, adversarial images have been added to the training dataset to improve model robustness [212].

Influential instances are intended to identify which input instances have a strong effect on the trained model by treating the model as a function of the training data rather than fixed. Two approaches for identifying influential instances are often used – deletion diagnostics and influence functions. Deletion diagnostics is not practical for big training datasets as it needs to remove a single training instance every time to observe the effect of this instance until the effect of all of the training data has been observed. Rather than

deleting the training instance, influence functions up-weight the instance in the loss function by a very small amount in order to measure the effects of this instance on the model parameters or predictions. It is an approximation method, but more computationally efficient which is especially important when the training dataset is very large (see Section 5.4.5 for more details on the use of influence functions).

Table 5.1: Categorisation of Explainable AI Methods with examples

| Type of explanation | | Scope | Model Specific /Agnostic | Examples of explainable AI methods |
|----------------------|---------------------------------|--------|--------------------------|--|
| Interpretable Models | | Global | Specific | A model by itself interpretable for the user, e.g. linear/logistic regression, decision tree |
| Post-hoc explanation | Feature importance explanations | Local | Agnostic | LIME |
| | | Local | Agnostic | KernelSHAP |
| | | Local | Specific | TreeSHAP |
| | | Local | Specific | Gradient * Input |
| | | Local | Specific | Integrated Gradient |
| | | Local | Specific | DeepLIFT |
| | Example-based explanations | Local | Agnostic | Influential instances |
| | | Local | Agnostic | Counterfactual explanations |
| | | Local | Agnostic | Adversarial examples |

5.3 Explainability in the ML Life-cycle

In order to influence the design of ML, it is most useful to consider safety assurance in the context of ML development process. Therefore, we also explore the role of explainability in assuring safety of ML in the context of the ML development process. Different development processes have been proposed, but the essence of them are ultimately the same. For example, CRISP-DM includes six different phases, i.e. business understanding, data understanding, data preparation, modelling, evaluation and deployment [213]. Here we simplify the phases of the ML development process to include data management, ML algorithm selection, model learning and model V & V, as shown in Figure 5.2 with an explicit representation of the deployment decision. One can consider that data management is comparable to the first three phases of CRISP-DM, ML algorithm selection & model

learning are comparable to modelling in CRISP-DM, and model V & V is comparable to evaluation in CRISP-DM. Although here we show the four activities in sequence, in reality the ML development process is inevitably iterative.

Figure 5.2 also shows the relevant stakeholders who might be interested in the explanations in the different phases. Our focus here is on the development activities, but we briefly consider the potential role of explainability in operation, see Section 5.3.5; for a discussion of the wider role of explainability including incident and accident investigation see my previous publication [181]. In the rest of this section we discuss the role of explainability against each stage of the development process shown in Figure 5.2.

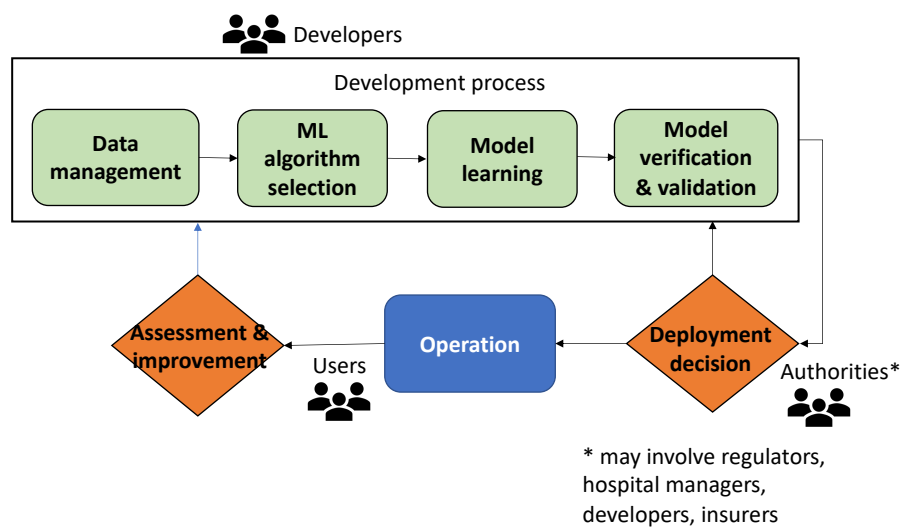


Figure 5.2: Process for development and operation of an ML System

5.3.1 Data Management

The first phase of the ML development process is data management. The Royal Society’s Policy Briefing on Explainable AI emphasises that data quality and provenance is part of the explainability pipeline, specifically saying that *“Understanding the quality and provenance of the data used in AI systems is therefore an important part of ensuring that a system is explainable”* [214]. This includes showing that the data comes from appropriate sources to address the problem concerned by the ML model. A widely accepted, harmonised framework for assessment of EHR data quality highlights conformance, completeness and accuracy [215]; we prefer accuracy to the original term plausibility because

plausibility means that the values are in the possible range but accurate means that the data is not only possible but correct. These criteria would be applicable to any ML systems developed using EHR data. In addition to these three criteria we also identify data relevance and data balance as being particularly important to the development of ML models [182] [216]. As real world data may contain biases, contain errors, and be incomplete, explaining how these five criteria are met can be at least as important as explaining the ML model itself.

For safety assurance, a safety case would need to address all five of the criteria. The evidence to ensure data quality is essentially technical, for example data conformance would include showing that data observes defined formats, e.g. correct units for weight [215]. However, demonstrating data relevance and data balance would include a judgement that the training data contained clinically relevant factors, are balanced for the problem being addressed; to do this requires clinical expertise. However, we acknowledge that often it is not possible to choose data that gives both feature balance and class balance. Instead, it might be useful to explain that some important features are balanced, e.g., gender, if the model is intended to be used for both male and female patients. In terms of class balance, this has long been an active research area in ML community [217]. It should be noted that data management is both crucial and labour intensive. Indeed, it may consume more effort than the rest of the ML life-cycle. Thus, arguments about data management will be an important part of the ML safety case.

5.3.2 ML Algorithm Selection

The second phase in the development process is ML algorithm selection (also referred to as model selection; here we use the term ML algorithm selection to avoid the confusion with model selection in the training phase where the ML algorithm is the same but hyperparameters of the model are tuned to be different). It is important to understand what kind of problem is being addressed and what kind of ML methods are suitable for the problem at hand. For example, if the problem is to identify optimal treatments in health-care, then RL might be more appropriate than others, as RL is widely used in complex decision making tasks to find an optimal policy [22]. On the other hand, if the problem is image classification then NNs might be more appropriate. In addition, another important aspect to consider at this stage is the explainability of the ML model. In Section 5.2.2 we identified that some ML models are intrinsically interpretable whereas others need to

be supplemented with *post-hoc* explainable AI methods. Guidelines on model selection, balancing model performance against explainability, have been proposed [218].

When it comes to ML algorithm selection, safety requirements are often implicitly transformed into explainability and performance requirements. Note that sometimes people make statements such as “use of deep NNs is not safe”. When they make this kind of statement, they are implicitly making the judgement that deep NNs are opaque, i.e., not explainable. This is why we argue that safety requirements are partially, but not wholly, transformed into explainability requirements. It would be ideal to have an interpretable model which can achieve performance as high as black box models. When this is not the case, a trade-off between explainability and performance would be necessary [218] and *post-hoc* explanations should be considered either in later phases of development or in operation to produce effective explanation. The rationale for the ML algorithm choice, including the performance-explainability trade-offs, needs to be documented in the safety case.

5.3.3 Model Learning

The third phase in the development process is model learning. For model learning, hyperparameter selection, loss function definition and class balance need to be considered in order to meet safety requirements. In addition, explainable AI methods can help in terms of failure class understanding and robustness. At this stage two particular explainable AI methods are relevant. One is adversarial examples and the other is influential instances, see Section 5.2.2:

- Adversarial examples are often added to training data to improve model robustness in object classification tasks. This is often referred to as adversarial training or robustness training [219] [212]. This is becoming widespread in domains such as autonomous vehicles, for example in improving performance at reading road signs under adverse conditions [220], but we believe it has wider applicability, e.g. for image classification in radiology.
- Influential instances are useful for “model debugging” as they can help to understand model behaviour and predictions by treating the model as a function of the training dataset, rather than being fixed [171]. Due to the computational cost, influential instances have not been widely used until recently with the availability of

more efficient algorithms such as influence functions which have made it possible to implement the approach on large datasets [221]. Due to these algorithmic improvements, the use of influential instances will increase, helping to determine what data to include or to exclude in the model training in order to improve model prediction and help to debug the model.

Note that these forms of explainable AI methods are of particular interest to ML model developers, but they help to ensure the soundness of the learning process and thus contribute to safety assurance.

5.3.4 Model Verification and Validation

The final phase in the development process is model V&V. We believe that explainability has a general role in validation but could also have a role in verification if there are specific explainability requirements to verify. However, such explainability requirements need to be defined in a specific situation, therefore our focus here is on validation. We derived three distinct objectives, reconciling approaches proposed by the FDA [15] and the IMDRF [14], which reflect key criteria for use of ML models in healthcare, although we note that explanations cannot guarantee that all these criteria are met [214] [218].

First is performance, which can be measured using standard ML practices, e.g. evaluation of the proportion of false positives and false negative, or AUC-ROC. This is necessary but not sufficient to assure safety of ML, see [181].

The second objective is analytical or technical validation, showing that the software for ML models is correctly constructed, and that it is accurate and reliable. Further, the ML models produce repeatable results, giving the same predictions from the same inputs. This objective can be met by employing established safety-critical software development practices including formal specifications, traceability from specification to implementation, use of test coverage criteria and static code analysis methods [130]. We do not see a role for explainable AI methods for this aspect of validation.

Third is clinical validation which measures the ability of the system to generate a clinically meaningful output associated with the intended use of the system in its operational environment. Here we define two specific sub-objectives where we believe explainable AI methods have a role in supporting clinical validation:

- Clinical association – demonstrate that the association between the system output

and the targeted clinical condition in the intended population is supported by evidence;

- Robustness – demonstrate the ability to distinguish the different classes of intended condition or recommended treatment without over-reliance on a specific input feature.

Feature importance explanations can help to demonstrate clinical association by showing that the output predictions are based on clinically meaningful and relevant factors of the input. This involves ranking input features based on their importance score or contribution score and making the rankings visible to clinicians so that they can exercise clinical judgement. In addition, this goal links back to data relevance and data balance, as data balance is directly related to the intended user population for the ML system, such as gender balance as we mentioned above. Thus clinical association is addressed from two perspectives: input features are relevant (data relevance) and outputs are based on relevant inputs (feature importance).

Example-based methods, especially counterfactual explanations, can help to assess model robustness. As mentioned in Section 5.2.2 counterfactuals are generated by minimising the distance from the original input but producing a different prediction. Therefore, the further the distance from an initial input to a counterfactual, the more robust the ML model is, i.e. the model is “harder to fool”. Thus the distance measure between the initial input and its corresponding counterfactuals can be used to define a robustness score for the ML model, see for example [222]. In an extreme case, if only one feature changed in the counterfactual examples from the original instance, this is analogous with a “single point of failure” which is a situation that needs to be avoided (the concept has origins in nuclear safety engineering [223] but is now quite widely used in critical industries). Thus, counterfactuals can also help show that this standard safety criterion is met if multiple input features have to change to produce a different classification.

The use of explainable AI methods in support of ML model V & V will contribute evidence to the safety case, complementing other activities including performance assessment and safety-critical software engineering. It should be noted that explanations should be re-generated when the ML models are updated so that they accurately reflect the state of the models.

5.3.5 Operation

As discussed in Section 5.2.1, assurance should be considered to be a “through life” activity. This would include, for example, a clinician seeking assurance about a particular prediction, especially if acting on it can have a profound impact on patient safety. Explainable AI methods can play a role here. Local feature importance explanation is relevant but counterfactual examples also have a role, for example, helping a clinician to decide whether or not a proposed change in treatment is likely to bring about the desired effect for a particular patient. Again, the role and significance of explainability in operation is examined in more detail in [181].

5.4 Clinical Case Study: Weaning from Mechanical Ventilation

This section presents our case study of weaning patients from mechanical ventilation. It describes the construction of the DSS covering each stage of the development process described in Figure 5.2, illustrating the use of explainable AI methods where appropriate. The role of the explainable AI methods in a safety argument is then shown in Section 5.4.8. The section starts by presenting the clinical context of weaning from mechanical ventilation and outlining the issues in defining a suitable DSS to support weaning decision-making.

5.4.1 Clinical Background

Mechanical ventilation via an endotracheal tube, sometimes also called invasive mechanical ventilation, is one of the most widely used interventions for patients admitted to Intensive Care Units (ICUs). Mechanical ventilation is a life-saving medical procedure used to assist or replace spontaneous breathing for patients with acute respiratory difficulties. Studies have shown that around 40% of ICU patients require invasive mechanical ventilation [224]. This consumes significant ICU resources with estimated daily costs around £1,738 in the UK [225] and \$2,300 in the US [226].

Weaning patients from mechanical ventilation covers the process of liberating the patient from mechanical support and removing the endotracheal tube (extubation). Time spent in this weaning process occupies a significant proportion of the total duration of mechanical ventilation [227]. Assessment of weaning readiness is a complex clinical task, which often includes determining whether or not the underlying disease of the patient

has been successfully treated, together with haemodynamic stability, the patient's level of consciousness, and the current values for ventilator settings. The final stage is often to conduct a series of SBTs, using either unsupported T-piece breathing or low-level Pressure Support Ventilation (PSV) over at least 30 minutes [228].

Despite advances in medical knowledge, weaning too early or too late are problematic. Delays in assessing readiness to wean are a common cause of late weaning. As a consequence, patients with prolonged ventilation might experience airway trauma, post-extubation delirium, drug dependencies, ventilator induced pneumonia, other forms of increased morbidity and even higher fatality rates [229] [230] [231]. There are also non-clinical effects including increased costs and greater strain on hospital resources, e.g. it has been reported that patients on prolonged ventilation use 37% of ICU resources [232].

On the other hand, premature extubations may lead to extubation failure, where re-intubation is required within 48-72 hours. Studies have shown that up to 25% of patients suffer extubation failure due to recurrence of respiratory insufficiency and require re-intubation [233], which can cause severe patient discomfort and result in even longer stays in the ICU with associated increases in cost and resource demands [234]. As with delayed extubation there can be increased fatality rates [235].

Considering the risks of prolonged dependence on mechanical ventilation and premature extubation, it is important to identify the ideal time point for weaning from mechanical ventilation from both a patient and healthcare provider point of view. However, there is no consensus on a standardised weaning protocol [236], even though they can be of benefit [237]. In practice protocols can vary between institutions, and may include different parameters [238]. This is mainly due to uncertainty, so an automated prediction model to indicate when extubation may be appropriate is likely to be helpful to clinicians seeking to make better-informed decisions.

5.4.2 Designing a Decision Support System

A number of projects have investigated DSS for weaning. Given the variation in weaning protocols and the clinical uncertainties it is perhaps unsurprising that a wide range of features have been considered for predicting extubation failure. These include demographic information (e.g., age, reason for intubation) [239], vital signs (e.g., heart rate, respiratory rate) [240], blood gas analysis (e.g., sodium, potassium, serum anion gap, oxygen/carbon dioxide partial pressure) [241], and respiratory parameters (e.g., duration of mechanical

ventilation, tidal volume) [239] [242]. What is striking is the variation in the factors used in different studies:

- subjects' age, reasons for intubation, duration of mechanical ventilation, Acute Physiology And Chronic Health Evaluation (APACHE II) scores, and breathing patterns obtained during a 30-minute SBT [239];
- tidal volume, minute ventilation, breathing frequency, and maximum inspiratory pressure [242];
- pre-extubation serum anion gap values and ratio of arterial oxygen partial pressure to fractional inspired oxygen (P:F ratio) [241];
- signal power of the respiratory flow obtained during the inspiratory phase [243];
- cardiorespiratory behaviour [244] and respiratory pattern parameters [245].

Generally, these studies use small numbers of features in training and prediction. In contrast, one approach employing RL uses 32 features [246]. Some work reports systematic approaches to identifying the relevant features. One study has sought to identify relevant features by comparing all combinations of three features from a total of 57, [240] eventually selecting 6 features from the best two models. Work using a Light Gradient Boosting Machine [247] initially considers 92 features and reduces them to 36 for the final model. This work [247] and the RL approach [246] analyse feature importance to help to interpret the models.

Most of the previous work, as shown above, predicts extubation outcomes, therefore we decided to build a richer model that monitors patient states every hour. This is different from, and complements the existing work and it can help clinicians choose the most appropriate action at each time step, e.g. continuing intubation or commencing extubation, including initiation of an SBT. The case study also illustrates the role of explainable AI methods, as presented in Section 5.3. In addition, it also includes explanation for data management by first presenting the rationale for data inclusion. The case study illustrates many, but not all, of the explainable AI methods for safety assurance in the context of ML development, and also briefly indicates the potential role of explainable AI methods during operation.

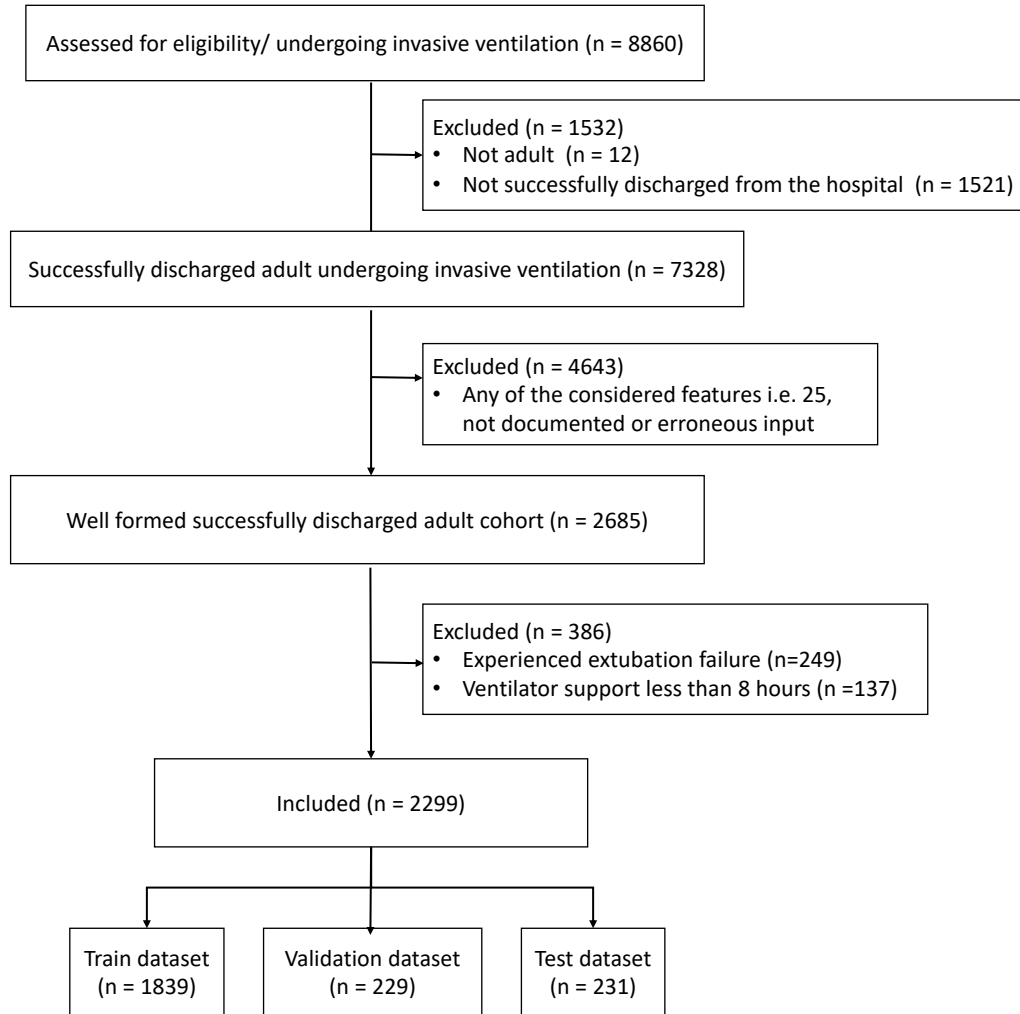


Figure 5.3: Patient inclusion diagrams in MIMIC-III

5.4.3 Data Management

Data to train the ML models was extracted from MIMIC-III [166]. We initially selected 8,860 admissions who underwent invasive mechanical ventilation from the dataset. We excluded non-adult patients and also those who died in the hospital as these fatalities can be caused by factors that are beyond the weaning process, in line with other work [246] [247]. This resulted in 7,328 adult patients who were successfully discharged following invasive ventilation, see Figure 5.3.

Based on the literature surveyed, e.g. clinical studies of “protocolized” weaning [248], and clinical judgement, we extracted 25 features including patient demographics, e.g. age, gender, ethnicity, laboratory tests, e.g. arterial pH, and vital signs, e.g. heart rate, oxygen saturation (SpO₂), and ventilator information, e.g. ventilator mode, Positive End-

Expiratory Pressure (PEEP), and mean airway pressure as shown in Table 5.5. Also, the feature correlation matrix is presented in Appendix C. We took the patient data and produced a series of records with values for the features on an hourly basis, for each patient from when the ventilator mode was recorded until the last time it was recorded. This ensured that the records covered the whole invasive ventilation period and also the non-invasive ventilation support period (NB non-invasive modes were also included). Where multiple values for a feature were available in an hour, they were averaged. Further, some features in MIMIC-III are not available for every hour so they had to be estimated. We used the previous valid value, if available, to fill in the gaps, i.e. forward propagation; if there was no valid previous value then back propagation was used. After this process, if a patient record still had missing values for some features, i.e. no values were recorded during the period considered, or was obviously erroneous then they were deleted. This processing resulted in a well-formed, successfully discharged adult cohort of 2,685 patients, see Figure 5.3. The main reason for the substantial reduction in number of patients is the absence of values of some features, so these patient records had to be discarded as deep learning will not accept missing data during training.

As a further processing stage we excluded patients who had ventilation support for less than 8 hours as they were likely to be undergoing routine ventilation following elective surgery. Post-operative extubation presents a minimal risk of adverse extubation outcomes and it was not our intention to consider such cases in this work.

There is a question about whether or not to include patients who had extubation failure which we defined as the need for re-intubation within 48 hours to be consistent with previous studies [228] [249]. Figure 5.3 shows that patients who experienced extubation failure were excluded, producing a final cohort of 2,299 patient admissions for use in our study. The rationale for this choice is presented in Section 5.4.5 where we consider the use of influential instances to understand the effects of including patients who suffered extubation failure.

Here we summarise the rationale for the data management in terms of the five criteria introduced in Section 5.3.1, based on the description above:

- Conformance – data for different patients are all processed using the same units, e.g. all of the weights are in kilograms;
- Completeness – missing feature values at each hour are established by forward and backward propagation when it is possible; if records are still incomplete, then they

are discarded;

- Accuracy – outliers are corrected using clinical knowledge when it is possible and discarded otherwise;
- Data relevance – the chosen features are based on the previous literature, see Section 5.4.2;
- Data balance – in the included cohort, 40% of the patients are female and 60% are male. The class balance for continued intubation and extubation are considered during the training, and a weighted loss function is used to guide the training.

5.4.4 ML Algorithm Selection

ML algorithm selection is strongly influenced by performance, as previously indicated. Here we use the performance metrics introduced in Section 2.1.5 to evaluate candidate ML models.

CNNs make predictions by extracting features without explicit, pre-defined knowledge of what is important in the data. CNNs have proven useful in image analysis, and their application has been explored in various other domains such as time series forecasting and data generation. The rationale for considering using a CNN for this task is that they are fast at run time and have the potential to produce accurate predictions for the type of tabular data employed here, see for example [250] [251] [252] [253].

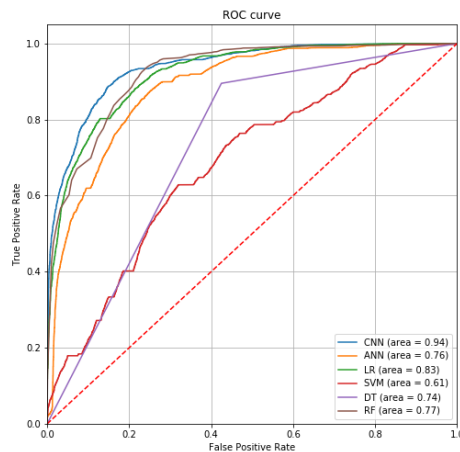


Figure 5.4: Performance of ML models

Table 5.2: Legend for Figure 5.4

For the case study, the performance of a number of ML models, including CNNs, were evaluated on the same dataset to support model selection, see Figure 5.4 and Table

Table 5.3: Performance comparison with different ML classifiers

| Methods | Accuracy | Precision | Recall | F1-Score | AUC |
|------------------------|----------|-----------|--------|----------|------|
| CNN | 86% | 82% | 86% | 84% | 0.94 |
| ANN | 85% | 84% | 76% | 79% | 0.76 |
| Logistic Regression | 82% | 78% | 84% | 79% | 0.83 |
| Support Vector Machine | 70% | 61% | 61% | 61% | 0.61 |
| Decision Tree | 81% | 76% | 74% | 74% | 0.74 |
| Random Forest Tree | 87% | 90% | 77% | 80% | 0.77 |

5.3. CNNs have the best performance and more importantly achieve better performance than intrinsically interpretable ML models such as logistic regression, but the performance difference is still considerable. As mentioned in section 5.3.2 there is a trade-off between performance and explainability. If performance over-rides the need for explainability, then CNN should be chosen. On the other hand, if intrinsic interpretability is more important, then logistic regression should be chosen. In this case study, CNNs have been chosen, and *post-hoc* explainable AI methods can be used to help to explain the model, see the rest of the section for details.

For completeness, we now give a brief overview of our CNN architecture. In CNNs, convolution computations are generally followed by non-linearities, also known as activation functions. The most commonly used activation is the Rectified Linear Unit (ReLU), given by $ReLU(x) = \max(0, x)$, where the response of a network is zeroed for negative values of the features learnt. Stacking multiple layers of convolutions and activation functions together extracts features in a CNN. These features are then passed into fully connected layers that learn to make the prediction.

The architecture of our CNN went through extensive tuning. The input features are passed through a series of 4 convolution layers with filter sizes 64, 128, 256, 256 and dropout is used in the final convolution layer. The output from this convolution layer is then flattened and passed into a fully connected layer of size 128 nodes which is then fed into the output layer, making the prediction through a sigmoid function with a threshold set at 0.5. The architecture of the CNN model is summarised in Table 5.4.

5.4.5 Model Learning

As we indicated in Section 5.3.3 there are two explainable AI methods that can be useful at this stage: adversarial examples and influential instances. Because adversarial examples

Table 5.4: CNN architecture

| |
|---|
| Conv1D with 64 filters of Kernel size 1 |
| Conv1D with 128 filters of Kernel size 1 |
| Conv1D with 256 filters of Kernel size 1 |
| Conv1D with 256 filters of Kernel size 1 |
| Dropout with probability 0.5 of leaving out units |
| Fully connected layer with 128 neurons |
| Sigmoid output |

are difficult to generate for tabular data, here we focus on the use of influential instances in their role for “debugging” ML models. This shows how they provide assurance about the appropriateness of the ML model learning process, in the context of the safety requirement.

When preparing the dataset for the case study, one issue that came up was whether or not to include the extubation failure patients. As mentioned earlier, extubation failure is defined as the need for re-intubation within 48 hours [228] [249]. Some of the literature suggests that premature extubation could cause extubation failure [254]. Therefore, the label in the dataset for extubation failure patients might not be optimal, so it might negatively influence the prediction. We can view this as a failure class as explained in Section 5.3.3. To explore this issue further, we trained two CNN models to predict the readiness for extubation in the next hour in order to observe the effect of extubation failure patients. In the first model, we excluded all of the extubation failure patients in the training dataset. In the second model, we included all of the extubation failure patents in the training dataset. The accuracy of the second model is slightly reduced by comparison with the first model. We randomly picked one of the test instances that was “interesting” in that the two models produced different predictions. For this instance, the first model predicted the patient should continue to be intubated, which is also the true (correct) label. However, the second model predicted that the patient was ready for extubation in the next hour. We used influence functions to identify the influential training instances for this test instance.

The key idea behind influence functions is to up-weight the loss of a training instance by an infinitesimally small step ϵ , which results in new model parameters:

$$\hat{\theta}_{\epsilon,z} = \operatorname{argmin} (1 - \epsilon) \frac{1}{n} \sum_{i=1}^n L(z_i, \theta) + \epsilon L(z, \theta) \quad (5.2)$$

where θ is the model parameter vector and $\hat{\theta}_{\epsilon,z}$ is the model parameter after upweighting z by ϵ . L is the loss function used for training the model. The influence of upweighting z on the parameters $\hat{\theta}$ given by Cook and Weisberg [255] is as follows:

$$I_{up,params}(z) = \left. \frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon} \right|_{\epsilon=0} = -H_{\hat{\theta}}^{-1} \nabla_{\theta} L(z, \hat{\theta}) \quad (5.3)$$

Where $H_{\hat{\theta}}$ is the Hessian matrix and $\nabla_{\theta} L(z, \hat{\theta})$ is the loss gradient with respect to the parameters for the training instance z . Next, we can apply the chain rule to calculate the influence of upweighting instance z on the loss of a test instance z_{test} :

$$\begin{aligned} I_{up,loss}(z, z_{test}) &= \left. \frac{dL(z_{test}, \hat{\theta}_{\epsilon,z})}{d\epsilon} \right|_{\epsilon=0} \\ &= \nabla_{\theta} L(z_{test}, \hat{\theta})^T \left. \frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon} \right|_{\epsilon=0} \\ &= -\nabla_{\theta} L(z_{test}, \hat{\theta})^T H_{\hat{\theta}}^{-1} \nabla_{\theta} L(z, \hat{\theta}) \end{aligned} \quad (5.4)$$

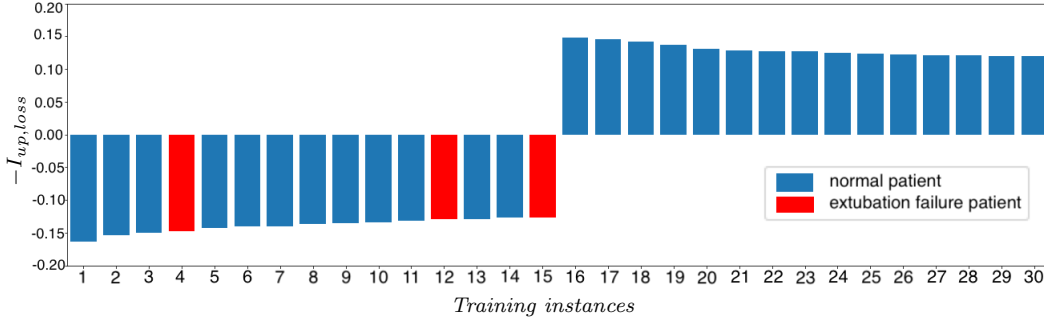


Figure 5.5: Top 30 most influential training instances

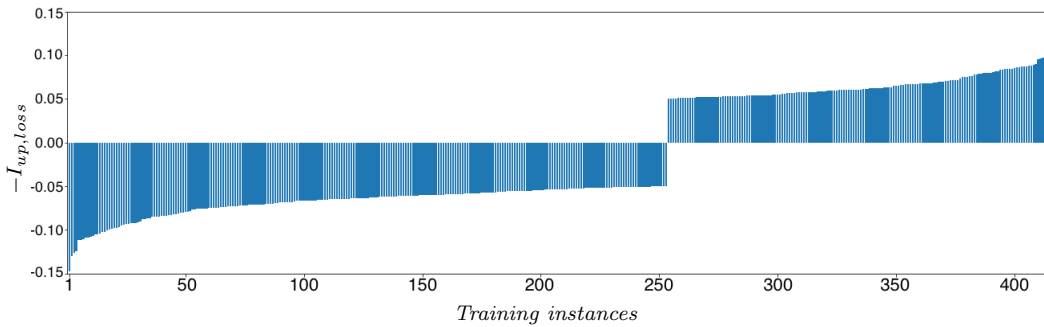


Figure 5.6: Distribution of influential instances

In this work, we use the influence functions algorithm developed by Koh and Liang [221] to calculate $-I_{up,loss}(z_i, z_{test})$ for each training instance z_i for this test instance. Figure

5.5 shows the top 15 helpful training instances (most positive $-I_{up,loss}(z_i, z_{test})$) and the top 15 harmful training instances (most negative $-I_{up,loss}(z_i, z_{test})$) for this test instance. From the figure, it shows there are three instances of patients who had extubation failure among the harmful training instances, which indicates that including the extubation failure patients made the predictions for the test instance worse. Figure 5.6 shows some of the most influential data points (magnitude of $-I_{up,loss}(z_i, z_{test})$ is large) from the extubation failure patients and that more of them have a negative influence than a positive influence. This suggests that the inclusion of extubation failure could make the prediction ready to extubate when it is not the case. Thus, we decided to exclude the extubation failure patients from the training dataset and the first CNN model was taken forward to the V&V stage. In a more general situation when prior knowledge is not available, i.e. we don't know what subset of the data could be problematic, we can still choose a test instance where the prediction is wrong and identify the influence of the training instances on this prediction. Then, further investigation could be done to understand what input features strongly impact the influence score, e.g. by perturbation [221] or by using decision trees [171]. The performance for the chosen training dataset, with the CNN architecture described in Table 5.4, is shown in Figure 5.4.

In summary, the use of influential instances has helped to show an appropriate process for meeting safety requirements and it explicitly contributes to the safety requirement “to extubate in a timely manner”.

5.4.6 Model Verification and Validation

In this section, we focus on clinical validation, as set out in Section 5.3.4, and illustrate the use of explainable AI methods for demonstrating clinical association and robustness. We do not consider analytical validation here.

5.4.6.1 Feature Importance Explanations

Here we illustrate the role of feature relevance in satisfying the clinical association safety assurance objective. This is done using DeepLIFT [202] which is a model-specific XAI method for deep NNs. When explaining deep NNs, the features are the set of inputs to the model. DeepLIFT compares the activation of each neuron to its “reference activation” and attributes to each input feature an importance score based on the difference. The “reference activation” is obtained through some user-defined reference input and in this

case, the reference sample is the minimum values of all of the input features obtained from the data set. We chose this method for two main reasons. First, it deals effectively with discontinuities in the gradient of the CNN model as it uses a difference from reference approach. Second, it avoids the problem of model saturation where using gradients would just assign zero to the features [202].

An overview of the results of using DeepLIFT is shown in Fig. 5.7; these values are averaged over the whole dataset, so this can be viewed as global feature importance. The feature ranking correlates well with clinical expectations, helping to give confidence in the model. Those features that score near zero in Fig. 5.7, e.g. ethnicity, gender and age, have little influence on the weaning decision, which is as expected. The top five features also align with clinical evidence. Patients who are undergoing invasive mechanical ventilation are often sedated to maintain physiological stability and to control pain levels. Sedation is reflected in the Richardson Agitation Scale (RAS) with negative values representing sedation and 0 meaning that they are alert and calm, thus more likely to be suitable for extubation. This is consistent with the first entry in the weaning checklist used in [256] that patients are “cooperative and pain free”. The second most important feature is “Inspired O₂ fraction” which is the third checklist entry in [256]. The third most important feature is “ventilator category”, which is the mode used for ventilation and is under direct clinician control; some modes are unsuitable for spontaneous breathing so cannot easily support weaning. The fourth and fifth most important features, peak inspiratory pressure and positive end-expiratory pressure (PEEP) set are airway pressures representing how hard the ventilator is having to work; PEEP is also the third entry in the weaning checklist in [256].

Here we have demonstrated valid clinical association through clinical evidence (relevant literature support) and expert opinion (consultation with clinicians). Overall, the benefit of the feature importance results is that they enable clinical judgement to be applied despite the opacity of the CNN model which contributes to safety assurance. Also, feature importance is of most value in making the behaviour of the ML model visible to clinicians, rather than directly to patients.

5.4.6.2 Counterfactual Explanations

The final concern in model V&V is robustness of the ML model and here we show how to use counterfactuals to demonstrate robustness. Table 5.5 shows a set of counterfactual

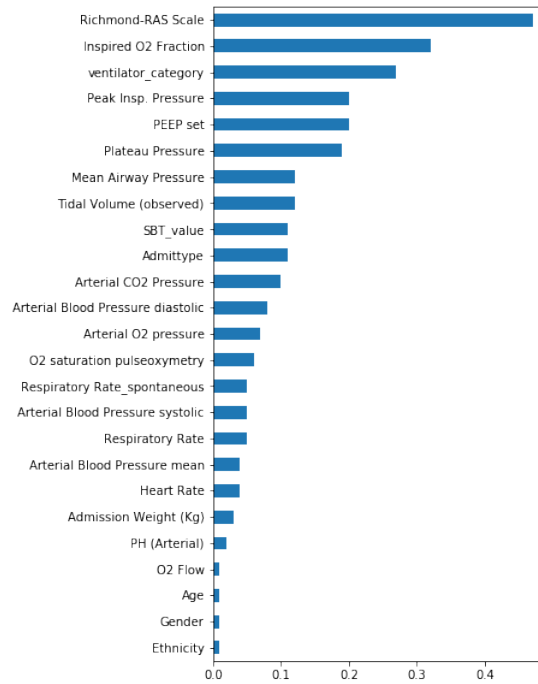


Figure 5.7: Feature Importance for the CNN Model

examples for a particular patient identifying which features need to change in order to “flip” the prediction from continued intubation to extubation. The left hand column shows the 25 features used by the model and the prediction of the ML model is included in the bottom row. The original instance is shown first, with the four rightmost column showing counterfactual examples. These counterfactual examples have been generated using DiCE [207]. Certain features cannot be varied, e.g. age and gender; the dashes in the rightmost four columns indicate no change from the original input. The change in prediction is shown in the bottom row.

Identifying counterfactual examples is undertaken by minimising the distance from the original instance to a counterfactual that produces a different prediction. Thus, given the way these counterexamples are generated, it can help to gain confidence in model robustness and the absence of single points of failure. In this case, as shown in Table 5.5, the minimum number of features that have to change to “flip” the prediction is five, showing robustness for this instance. However, one instance is not sufficient to show ML model robustness. More of the input instances in the dataset need to be investigated in order to generate a robustness score as defined in [222].

Another use of counterfactual examples to inform clinician judgement is considered in Section 5.4.7. Further comparison of counterfactuals to feature importance is presented

Table 5.5: Counterfactual examples for a given original instance

| Features | Original instance | Counterfactual Examples | | | |
|------------------------------|-------------------------|-------------------------|------------------------|------------------------|----------|
| | | 1 | 2 | 3 | 4 |
| Admit Type | Emergency | — | — | — | — |
| Ethnicity | White | — | — | — | — |
| Gender | Female | — | — | — | — |
| Age | 78.2 | — | — | — | — |
| Admission Weight | 86.5 | — | — | — | — |
| Heart Rate | 119 | — | 110 | — | — |
| Respiratory Rate | 24 | 26 | — | — | 21 |
| SpO2 | 98 | — | — | 96 | — |
| Inspired O2 Fraction | 100% | — | 40% | — | — |
| PEEP set | 10 | 5 | 5 | 5 | 0 |
| Mean Airway Pressure | 14 | — | 10 | — | 10 |
| Tidal Volume (observed) | 541 | — | — | 560 | — |
| PH (Arterial) | 7.46 | — | — | — | — |
| Respiratory Rate(Spont) | 0 | — | 24 | — | 21 |
| Richmond-RAS Scale | -1 | — | 0 | — | 1 |
| Peak Insp. Pressure | 21 | — | — | — | — |
| O2 Flow | 5 | — | — | — | 10 |
| Plateau Pressure | 19 | — | — | — | — |
| Arterial O2 pressure | 124 | 108 | 118 | — | — |
| Arterial CO2 Pressure | 33 | — | — | — | — |
| Blood Pressure (systolic) | 101 | — | — | — | — |
| Blood Pressure (diastolic) | 65 | — | — | — | — |
| Blood Pressure (mean) | 76 | — | — | — | — |
| Spontaneous breathing trials | No result | Successfully Completed | Successfully Completed | Successfully Completed | — |
| Ventilator Mode | CMV/ASSIST/ AutoFlow | PCV+ | SIMV/PSV | SIMV/PSV | CPAP/PPS |
| Predicted outcome | 0.93 | 0.44 | 0.17 | 0.36 | 0.46 |

in discussion, Section 5.5.1.

5.4.7 Operational use of the ML Model

The operation of ML models is often uncertain. Thus, there is merit in extending the notion of assurance to operation, providing support to a clinician to give confidence to act on the particular model prediction. One way of approaching this is to use local

explanations.

Figure 5.8 visualises the feature importance values for a *single* patient, i.e. local feature importance. Here, a positive feature importance score contributes to moving the output towards intubation being continued. In contrast, a negative feature importance score contributes to moving the output towards extubation. The sum of the positive contributed features are far greater than the sum of the negative contributed features, thus the prediction for this patient, for the next one hour, is to remain intubated.

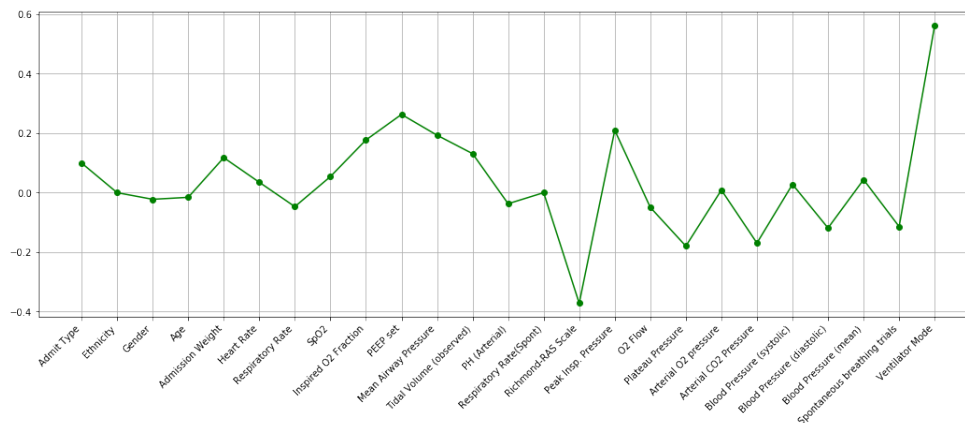


Figure 5.8: Feature Importance for a Single Patient

However, clinicians might want to find out when the patient would be ready to extubate. This brings us back to counterfactuals. The counterfactual examples shown in Table 5.5 are for the same patient shown in Figure 5.8, and could potentially help the clinician to identify actions to take so that the patient becomes ready to extubate. As shown in the table, it is beneficial to generate multiple counterfactual examples, so that the clinicians can choose one that is most practical to implement. In the counterfactual examples shown, changes in the ventilator mode and SBT successfully completed would both indicate progress towards extubation. Note our model has not been used in operation yet, so the material presented here just illustrates the possibilities.

5.4.8 Safety Arguments

As with the other case studies, the safety argument is presented using GSN. Figure 5.9 presents a *partial* safety argument for the weaning case study, emphasising the role of explainability. The top goal (G0), which states that the ML model meets its safety requirement, is set out in the context of the definition of the ML model and the associated safety requirement – that “prediction of readiness for extubation is timely”.

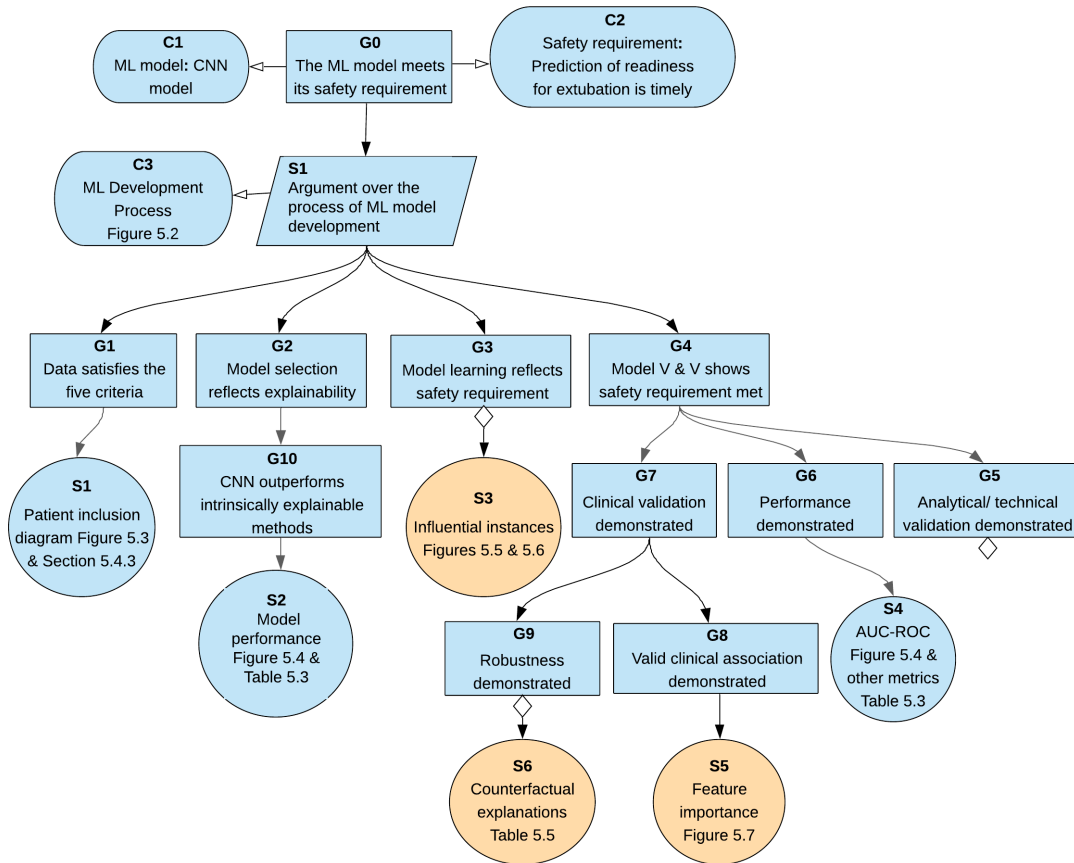


Figure 5.9: Partial Safety Argument for Weaning ML Model emphasising Explainability

The top-level argument strategy is a decomposition across the stages of the development process. We consider each of the goals in turn. The argument shows solutions for all the goals, with the exception of G5, as this is analytical/technical validation which is not covered in the case study. The amber solutions (S3, S5 and S6) reflect explainable AI methods.

G1: Data satisfies the five criteria – this is supported by the analysis in Section 5.4.3 (S1) showing that the data meet the five criteria: conformance, completeness, accuracy, data relevance and data balance introduced in Section 5.3.1.

G2: Model selection reflects explainability – this is supported by the analysis in Section 5.4.4 (S2) which shows that the CNN outperforms other available ML methods, and suitable post-hoc explainable AI methods are available.

G3: Model learning reflects safety requirement – this is directly, but *partially*, supported by the use of influential instances (S3) which show the rationale for excluding extubation failure patients. Note that other evidence is needed (hence the goal is shown

as needing development), e.g. to show appropriateness of parameter selection for training the model.

G4: Model V & V shows safety requirements met – this is broken down into G5: analytical/technical validation, which is undeveloped, G6: performance demonstrated supported by S4 and G7: clinical validation which is decomposed into two sub-goals covering the V&V criteria introduced in Section 5.3.4.

G6: Performance demonstrated – this is directly supported by the AUC-ROC performance in Figure 5.4 and other metrics shown in Table 5.3 (S4) which shows the superiority of the CNN performance to other ML models.

G8: Valid clinical association demonstrated – this is supported by the feature importance (S5) although it should be noted that clinical judgement is needed to assess the appropriateness of the feature ranking.

G9: Robustness demonstrated – this is partially supported by the counterfactuals in Table 5.5 (S6) (this is a *partial* solution to G9 as the explanations only relate to a single prediction, and there are also other ways to demonstrate robustness).

As noted above, this argument is incomplete and the evidence presented in this chapter should not be taken as sufficient to justify deployment of the CNN model described here in a clinical context. However, it is a valuable part of an overall safety case.

5.5 Discussion

Safety assurance of ML models in healthcare is an active area of research. Although explainability is often said to help in safety assurance of ML, no work so far has explored the possibilities systematically and identified precisely how explainability can help safety assurance. This chapter seeks to fill this gap. We illustrated how explainability can help in safety assurance in the context of the ML development process. Explainability used here includes explaining the data and the use of explainable AI methods to reflect the Policy Briefing on explainable AI from the Royal Society [214]. The role of the different explainable AI methods in the development and operation phases is summarised in Table 5.6 along with the interested stakeholders. We first extrapolated the safety objectives at the different stages of the ML development process. Then we used a concrete healthcare case study to demonstrate how explainable AI methods can help to meet these safety objectives, particularly in model learning and model V&V. Specifically, we have shown the value of influential instances for model learning, which is of particular interest to ML

developers. Further, we have shown the value of feature importance and counterfactuals in model V&V, which is of particular interest to ML developers, regulators and others involved in deployment decisions, see Figure 5.2. In this section, we first discuss the relationships we observed between feature importance and counterfactuals. Then, we discuss the other complementary methods that can also help safety assurance of ML.

Table 5.6: Role of Explainable AI Methods in the development and operation phases

| Phases | Activity | Explainable AI methods | Stakeholders |
|-------------|------------------------|--|--|
| Development | Data Management | N/A | ML developers Regulators Hospital managers |
| | ML Algorithm Selection | Trade-off performance & explainability | ML developers |
| | Model Learning | Adversarial examples Influential instances | ML developers |
| | Model V & V | Global feature importance Counterfactual explanations | ML developers Regulators Hospital managers Insurers |
| Operation | Decision Support | Local feature importance Counterfactual explanations | Expert users – clinicians Decision recipients – patients |

5.5.1 Relationships between Feature Importance and Counterfactuals

Although we have used the explainable AI methods to illuminate different perspectives, there is a relationship between the DeepLIFT method and counterfactual explanations. Research shows that DeepLIFT can be viewed as a variant of gradient-based methods where the gradient for the non-linearity is calculated using the ratio between the difference in output and the difference in input and the gradient for the linearity is just the weights [203]. The importance score given by DeepLIFT is equal to $(x - x')$ multiplied by the modified gradient, where x is the input features and x' is the “reference activation”, see the proof [203]. The input features for our CNN model are normalised between 0 and 1, so the minimum value of the input features is a zero-array after normalisation. Thus, in this case the importance is defined as $x \times$ the modified gradient. Where the feature has a higher score using DeepLIFT, i.e. the absolute score for a feature is greater than zero, perturbation of this feature will make a larger difference in the prediction given

that the input feature values are on a similar scale. As counterfactual explanations are proposed as a way to provide perturbations that would have changed the prediction of a model, we expect a correlation between the importance scores produced by DeepLIFT and the counterfactual explanations. This is observed in our experiments, and we consider counterfactual example 2 in Table 5.5 to illustrate this. Here, the change of ventilator mode, RAS scale, PEEP set and Inspired O₂ fraction produce a different prediction, where these features are shown to have high importance score in Figure 5.7. Therefore, changes to these features will help to “flip” the prediction.

It is also worth noting that Respiratory rate (spontaneous) has a zero score as shown in Figure 5.8. This is because the input feature value is zero. Therefore, in this case, this feature is not important but this might not be true in other cases. In the counterfactual example 2, arterial O₂ pressure also changed from 124 to 118, even though this feature only weighs 0.05 in Figure 5.8. If we investigate this counterfactual example by changing the arterial O₂ pressure back to 124, the new prediction is 0.169, which is a small change from 0.168 (this is the original prediction of the counterfactual example 2 in Table 5.5). This is consistent with our expectation as it has a small importance score in Figure 5.8.

Further, DeepLIFT can be combined with counterfactuals. Specifically, we can use DeepLIFT to assign a contribution score to each feature that changed in a counterfactual example whilst treating the original instance as the “reference activation”. This can help users to understand how much individual feature changes in the counterfactual example contribute to flipping of the prediction’s classification compared with the original instance. Where diverse counterfactual examples are available, the feature importance can help to choose between them. It can also be used to influence the generation of the counterfactual examples. For example, if there are many features in the counterfactual example that have a very low contribution score, e.g. less than 1%, then that example might be discarded or the features values not allowed to change. This facilitates the identification of sparse counterfactual examples which is particularly important when choosing between diverse counterfactuals. See my previous publication [180] for the detailed implementation.

5.5.2 Complementary Safety Assurance Methods

As indicated earlier, although the use of explainable AI methods can contribute to safety assurance, it is not enough to assure safety by itself. In this section, we will highlight some relevant complementary methods that also contribute to safety assurance.

First, any safety critical software should be developed in a quality management framework, see for example [257]. For all software, quality management includes configuration control, traceability from requirements to implementation and test, and change control. For SaMD it should also assure the quality of data used for training ML models [215]. From a safety assurance point of view, the aim of quality management is to ensure that the evidence produced to support the safety case properly reflects the build state of the system that is to be deployed. Whilst none of this is new, it may be challenging for AI/ML-based SaMD due to the highly iterative nature of the ML development process.

Second, it is important to apply established methods from safety critical software engineering, adapted as necessary for AI/ML-based SaMD. One such method is static analysis, that is analysing the code without executing it, looking for “bugs”, such as division by zero or using the wrong type of data; see [130] for an illustration of applying static analysis to ML code in healthcare. It is also standard practice to measure test coverage of the software when undertaking V&V. For conventional software, it is common to use structural coverage, e.g. ensuring that all branches in the code have been executed at least once. The obvious analogy for NNs is neuron coverage [258], although there is some debate about whether or not this is an appropriate criterion [259]. Nonetheless, coverage is significant when considering safety, as assurance is clearly undermined if there are significant parts of the ML model for which we have no test evidence. In the context of this chapter, static analysis seems most readily and immediately applicable.

Third, there are assurance methods that address the specific challenges of V&V for AI/ML-based software. We briefly consider two methods that are relevant to deep NNs. It is possible to apply formal methods (mathematical techniques of verification) to deep NNs. For example [260] applies Satisfiability Modulo Theories (SMT) solvers to find adversarial examples for NNs used in image analysis. Further, the ideas of concolic testing, which seeks to maximise code coverage, have been applied to deep NNs [261]. This work addresses structural coverage, including neuron coverage, and other properties such as Lipschitz continuity. It uses symbolic approaches to generate inputs to improve test coverage to generate a test suite for a given deep NN and also assists in finding adversarial examples.

Finally, these methods can support regulatory processes, particularly those focusing on AI/ML-based SaMD. Our aim here was not to propose alternatives to regulatory processes, but to identify where explainability could help to provide safety assurance evidence to support those processes.

5.6 Conclusion

In this chapter, we have developed a CNN model to predict readiness for extubation from mechanical ventilation, and have demonstrated how to apply explainable AI methods to the system to achieve safety assurance for the CNN model. To our knowledge, this is the first systematic attempt to explore the role of explainability in assuring safety of ML, with a particular focus on pre-deployment decision-making. We believe this will be of particular interest to regulators, as it illustrates how to use explainable AI methods to provide evidence to support relevant safety objectives, e.g. for clinical association, articulated by the FDA and IMDRF.

The case study illustrates the practical use of explainable AI methods in safety assurance. Specifically, it illustrates three different explainable AI methods:

- Influential instances – for showing how to debug the ML model, including helping to define the most appropriate training dataset;
- Feature importance – for showing valid clinical association;
- Counterfactual explanations – for showing ML model robustness and the absence of single-point failures.

From a safety assurance perspective, these uses of explainable AI methods contribute most to model learning and validation. The case study also shows how the use of these explainable AI methods feeds into a safety case, e.g. as required by healthcare standards [9]. Future work will include further exploration of explainable AI methods and development of further case studies with the aim of refining and validating the approach.

In addition, we believe it would be valuable to consider the role of explainable AI methods in accident and incident investigation for AI/ML-based SaMD. Being able to explain what happened may be crucial in order to learn from experience and to preserve confidence in a system. For example, it might be that counterfactual examples would help in understanding how an adverse event could have been avoided and thus indicate requirements for ML model retraining. This would help in achieving a TPLC approach to managing risks of AI/ML-based SaMD as proposed by the FDA [15].

Returning to the research questions, this case study provides a positive answer to question 2: *are there new opportunities for well-established safety engineering methods with the development of ML and why are they specifically good for safety in this domain?*

Opportunities have been identified for employing explainable AI methods to make contributions to safety. In model learning, influential instances help to improve safety of the learnt model. In model V&V, feature importance was used to demonstrate valid clinical association. Further, counterfactuals have been used to show robustness, especially against single failures. These two uses of explainable AI methods have a particular focus on validation. Turning to operation, the case study has demonstrated the potential to use feature importance and counterfactuals for a particular patient, providing assurance to the clinician/end user prior to them taking action.

The explainable AI methods we demonstrated here are used for supervised learning and for tabular datasets. The extension of the approach to RL and unsupervised learning is not obvious as the explainable AI methods presented in this chapter are not generally applicable to such methods. Similarly, the use of counterfactuals on image data is an ongoing research topic and how easy it is to generate such counterfactuals remains unclear.

Therefore, within the limits outlined above it seems likely that the approach developed here would generalise to other supervised ML models and application domains using tabular datasets. Many, although not all, of the explainable AI methods are model agnostic so can be applied regardless of the “base” supervised ML model being used. There is already evidence of use of explainable AI methods in other domains, e.g. autonomous vehicles. However, as with the other case studies, substantive effort would be required to validate the generalisability of the approach in other domains and such work is outside the scope of this thesis.

The code for applying various explainable AI methods is available at: https://github.com/Yanjiayork/mechanical_ventilator.

Chapter 6

Using Machine Learning to Update Safety Analysis

This chapter is based on my previous publications [262] [263] [264] and contributes to answering research question 2. Safety analysis is often viewed as subjective and uncertain; in practice, the results of safety analysis are rarely validated in operation. The use of ML presents a new opportunity to validate and update the results of well-established safety engineering methods based on the information from operation. The contribution of the chapter is summarised in Figure 6.1; as in previous chapters, the elements highlighted in red show the focus of the chapter. The case study uses well-established *Safety analysis* methods to proactively identify potential causes of medication error. As healthcare is now data rich, and a lot of data is captured in operation, it is possible to augment safety analysis with ML to discover actual causes of medication error from the data, and to identify where what was predicted in the safety analysis is inaccurate or incomplete, enabling the production of an *Updated safety analysis*, better reflecting “ground truth”. The updates feed into the *Operational safety case*.

This case study focuses on medication management for patients taking Beta-Blocker (BB) before surgery involving the thorax, e.g. oesophagectomy. Such patients are at risk of atrial fibrillation (AF) in post-operative care. The desirable treatment is to continue to give BBs after surgery to reduce the risk of developing AF. This case study combines SHARD with Bayesian network (BN) structure learning to produce the analysis results, showing the potential use of ML to update safety analysis and transforming the way that safety is managed in complex healthcare environments.

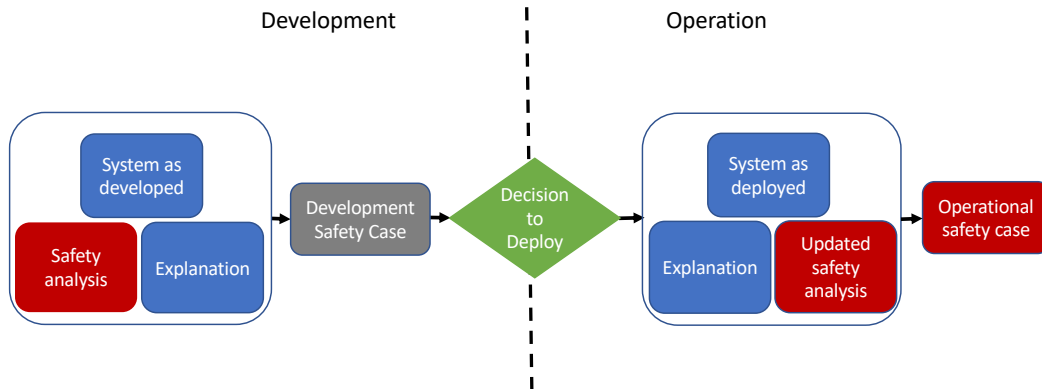


Figure 6.1: Overview for the Case Study

6.1 Introduction

Safety engineering methods are normally used predictively. They are used to identify hazards, hazard causes, and the associated risks before a system is deployed. The system is then monitored during its operation to manage risks. When systems are used in highly controlled environments and built using components with a long service history, these predictions can be accurate. However, as systems become more complex, especially in adaptive socio-technical contexts, it is more difficult to make credible predictions of safety. Indeed, recent analyses of Quantitative Risk Assessment (QRA) by Rae et al [265] show that analyses even of technical systems are rarely accurate, and the paper systematically identifies causes of deviations between prediction and reality in order to propose ways of improving QRA (and hence safety management). In general, it can be observed that safety analysis is often “open loop” and that there is little effective feedback from operation to confirm or, if necessary, refine the safety analysis to reflect ground truth.

As noted in Chapter 2, healthcare is a complex social-technical domain and includes a diverse set of activities that have to deal with its nonlinear, dynamic and unpredictable nature [47] [266], which potentially increases the gap between initial safety analysis and ground truth for medication management. Medication errors can arise in different phases of treatment and have many different potential causes ranging from clinical factors, e.g. due to comorbidities, via technical factors, e.g. due to problems with EHR, to human and organisational factors, e.g. under-staffing. These factors are variable and context-sensitive [131]. Understanding the variation and the significance of the causes of medication errors in different contexts is particularly important to support clinicians and

healthcare organisations in anticipating, monitoring and responding to medication errors to achieve medication safety [267].

Our contribution in this chapter is twofold. First, it presents a new methodology that provides a practical means of augmenting the initial safety analysis through using ML to analysis data from operation. Second, it shows that the methodology can be applied to management of medication safety, giving clinically meaningful results. This enables medication safety to be managed effectively and dynamically, using the results of ML.

The case study we present considers the complex setting of ICUs where patients may be taking multiple medications due to comorbidities, and the post-operative care is perhaps the most difficult to manage, especially when the treatment is time-critical. This work is extremely important as patients in ICUs are at high risk and medication errors can be life-threatening [268].

The rest of the chapter is structured as follows. Section 6.2 discusses the background and related work, including the use of ML in medication safety. Section 6.3 presents our methodology, showing how to use ML to update safety analysis. Section 6.4 presents the details of the case study, which concerns the management of AF in post-operative care in an ICU following thoracic surgery, to illustrate our methodology. A discussion of the methodology including some possibilities for future work is presented in Section 6.5. Section 6.6 presents conclusions.

6.2 Background and Related Work

This section considers medication safety, adding to the general discussion in Section 2.3.3, then discusses the use of ML in support of safety assurance to show the novelty of our methodology.

Medication safety is critical in healthcare and is a key factor in improving patient safety [269]. As medication errors continue to be a leading cause of avoidable harm in hospitals [270] [271], both regulatory agencies and research communities have made efforts to improve medication safety. Statistical analyses of medication errors across the whole process from prescription to administration, e.g. [272] [273], typically show that a high proportion of prescriptions in hospitals are subject to some form of error although, of course, the majority are corrected prior to administering the drugs.

As well as statistical analyses, there is work intended to establish practical and proactive means for identifying and detailing the underlying causes of errors and finding po-

tential controls for those errors, e.g. [274] [275]. The Safety Assurance of Intravenous Medication Management Systems (SAM) project [276] [277] focused on using technology to automate cross-checks to reduce certain classes of error. This is an innovative approach to the issue and is notable for considering the acceptability of the technology to patients and clinicians.

In addition, ML has been applied to different medical applications as outlined in Section 2.3.4, mainly as clinical DSSs with the aim of improving safety in healthcare, e.g. diagnosis and treatment practices. Here we present a few examples of work that has used ML to understand or support safety engineering directly. We note that, in order to find such examples, we needed to look outside healthcare.

An example from the Oil and Gas industry [278] uses Deep Neural Networks (DNNs) for risk assessment of (unintended) movements of the platform, which might ultimately lead to damage to the wellhead. The authors are cautious about their findings and note that care needs to be taken in selecting models to support safety-related decision making. Further, there have been accidents with Unmanned Air Systems (UAS), e.g. Watchkeeper, where the accident causation was very different to that predicted in the initial safety analysis [279]. Work is under way using ML to identify the causal factors that contributed to the accidents, including differences between the UAS' behaviour and the operators' perception of what was happening.

The intent of our work is to use ML to update safety analysis hence to improve medication safety. The use of ML in this way to update and enhance safety analysis is novel.

6.3 Methodology

Figure 6.2 shows the methodology we introduced for updating safety analysis using ML. It shows the development phase, the operation phase and the safety case, which is similar to the overview Figure 6.1, but with a different emphasis to illustrate how to use ML to update safety analysis. There are two elements in the development phase (Figure 6.2): *System as developed* and *Safety analysis*. Unlike the previous two clinical case studies, here the “system” is defined as clinical practice rather than ML models. Safety analysis is conducted for the clinical practice, which helps to produce the foundation of the safety case. There are two steps for the safety analysis. First, identifying hazards associated with the clinical practice of interest. Second, determining the potential causes and effects of the hazards, together with severity (degree of harm) and likelihood of the hazards, to estimate

risk associated with the hazard. There are two mirrored elements in the operation phase (Figure 6.2): *System as deployed* and *ML*. With the growing deployment of EHR and other hospital systems, it is possible to record the data from different aspects of clinical practices in healthcare. This gives an opportunity to use ML to learn from the data, e.g. to identify patterns of what goes ‘right’ and ‘wrong’ in operation. For example, using variables to represent the hazards, causes of hazards, and effects of hazards from the safety analysis can help us to understand the dependencies between the factors identified in the safety analysis or even identify other factors that we did not consider in the safety analysis in the development phase. Finally, safety analysis and also the safety case can be updated based on the ML results.

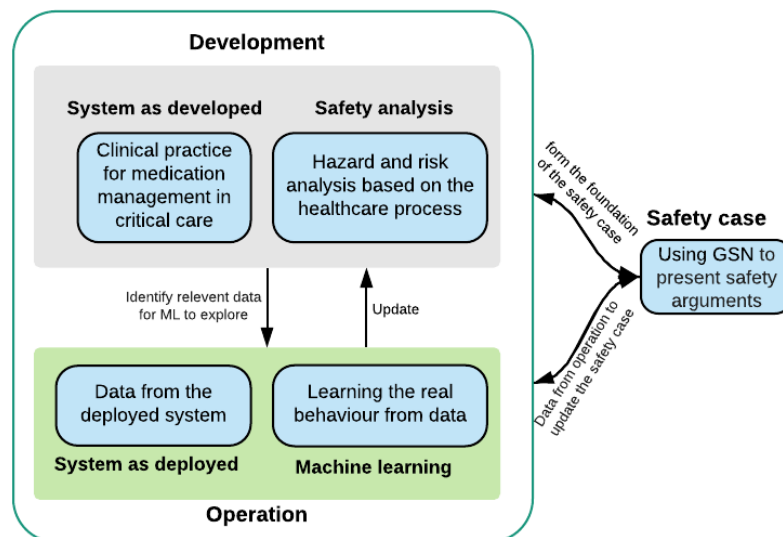


Figure 6.2: Framework for using ML to update Safety analysis

The inspiration for developing this methodology comes from Hollnagel’s characterisation of ‘Safety-I’ and ‘Safety-II’ [280]. In ‘Safety-I’, sometimes also referred to as the traditional approach to safety management, the focus is on failure – implicitly assuming that effective system design and compliance with good operating procedures will be safe. It also assumes that it is possible to analyse potential failures and to design mechanisms or procedures to control these failures and thus to assure safety. Hollnagel states that the purpose of safety management in Safety-I is to keep the number of accidents and incidents as low as possible by reacting to unacceptable events, due to system malfunction or human fallibility [280]. However, ‘Safety-I’ does not properly cater for the current socio-technical systems and will become even less effective as they become more complicated. Hollnagel

emphasises that safety management should therefore move from ensuring that ‘as few things as possible go wrong’ to ensuring that ‘as many things as possible go right’ which he characterises as ‘Safety-II’ [280]. Although our methodology is inspired by Safety-II, it is intended to go further and to provide a practical way of gaining insight from operation. More specifically, the methodology is a combination of Safety-I and Safety-II (we believe these two views are complementary and not alternatives). It uses well-established safety engineering methods to analyse clinical practice (the *System as developed* in Figure 6.2) to understand the hazards and potential causes for the hazards. This is the Safety-I point of view. Then the methodology uses ML to learn from the data which is generated from the clinical practice; this is influenced by the Safety-II point of view. Understanding what ‘goes right’ in operation can help in updating the safety analysis, for example showing that potential hazard causes do not arise in practice. We see our methodology as a way of implementing some aspects of Safety-II, which can be seen more as a broadly-stated concept than as a practical methodology.

Our methodology is intended to be ‘agnostic’ with respect to the safety analysis methods that are used, however we illustrate the methodology in the case study using a flow-based analysis method, SHARD, as this is effective in assessment of decision-making processes and helps to identify the factors that can contribute to medication errors. Similarly, our methodology is intended to be ‘agnostic’ to particular ML methods, but the case study uses BN structure learning as it is effective in identifying dependencies and can reveal fine structure in complex datasets.

Whilst the immediate focus of the methodology is on updating safety work products, it could also help relevant decision-makers to understand when it is desirable to improve clinical practices or data collection methods from EHR, if there is value in doing so, e.g., to start recording important data that was not previously available for analysis. Further, by identifying opportunities for making these improvements, we intend that the methodology can influence real-world safety management practices and ultimately improve patient safety.

6.4 Clinical Case Study: Beta-Blocker Delivery

This section presents our case study, which focuses on medication management for patients taking BBs before thoracic surgery. The case study is intended to show how to implement this methodology and to evaluate it. We start by presenting the clinical context of delivery

BBs for such patients.

6.4.1 Clinical Background

Patients undergoing thoracic surgery are at risk of disturbances of heart rhythm, typically AF in post-operative care after opening the chest [281] [282] [283]. There is debate about whether or not all patients should receive BBs following surgery but, as a minimum, treatment should continue to give BBs following thoracic surgery to reduce the risk of AF for those who were receiving BBs before surgery [284] [285].

Oesophagectomy is a thoracic operation whereby the oesophagus (food pipe) is removed, usually to treat oesophageal cancer. We use oesophagectomy as a concrete example of thoracic surgery in this case to illustrate the framework. As oesophagectomy prevents the patient from taking food or drugs orally, especially in the first week after surgery, the challenge is how to give the right form of BBs when the patient is unable to swallow. If BBs are not continued post-operatively, then there is an increased risk of developing AF which can lead to strokes that may be fatal – one analysis gives an odds ratio of death from AF of 1.5 for men and 1.9 for women [286].

There are a number of published guidelines on post-operative care for oesophagectomy. The pathway in Figure 6.3 has been synthesised from a number of publications [287] [288] [289] and focuses on the delivery of nutrition and medication. The pathway shows the differences between the presence or absence of a feeding tube (FT). Patients may be fitted with a FT during the operation and this also can be used for some forms of medicine. If there is no FT, medication has to be given by IV injection or infusion. This increases the complexity of giving BBs post-operatively, as commonly used BBs, e.g. bisoprolol, do not have an IV form. Also, the IV form of BBs, e.g. metoprolol and atenolol, may be less familiar to clinicians, and may not be immediately available on the ward. Further, the calculation of equivalent doses makes mapping between oral and IV form of BBs error-prone.

A further complicating factor is that patients may have other medications, e.g. painkillers given epidurally, and there can be an adverse interaction with BBs leading to a potentially dangerous reduction in Blood Pressure (BP). Overall, there are many potential difficulties in managing delivery of BBs in post-operative care. This makes it a rich case study to illustrate our methodology.

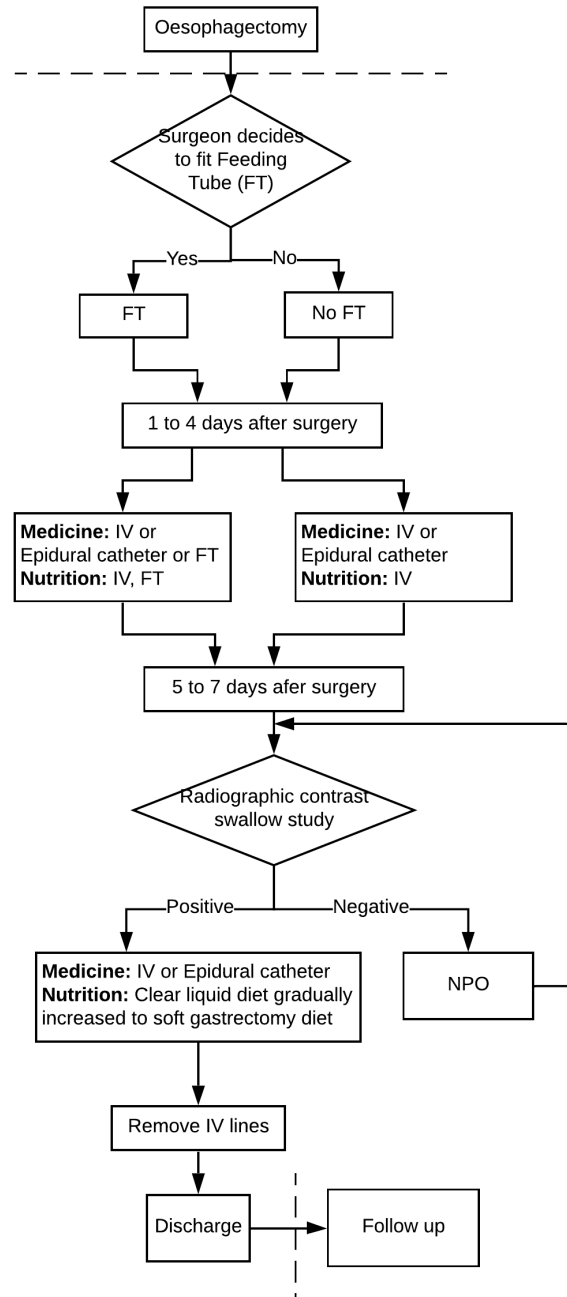


Figure 6.3: Pathway for Nutrition and Medication following oesophagectomy

6.4.2 System – the Clinical Practice

As we mentioned earlier, the system is defined as clinical practice in this case study. In order to illustrate how clinicians carry out their work in this context, we developed a decision-making model related to delivery of medication in post-operative care following an oesophagectomy.

The decision-making model used to represent the clinical practice gives the basis for

conducting the safety analysis. Section 6.4.3 presents the associated safety analysis, conducted using SHARD, summarised in a table.

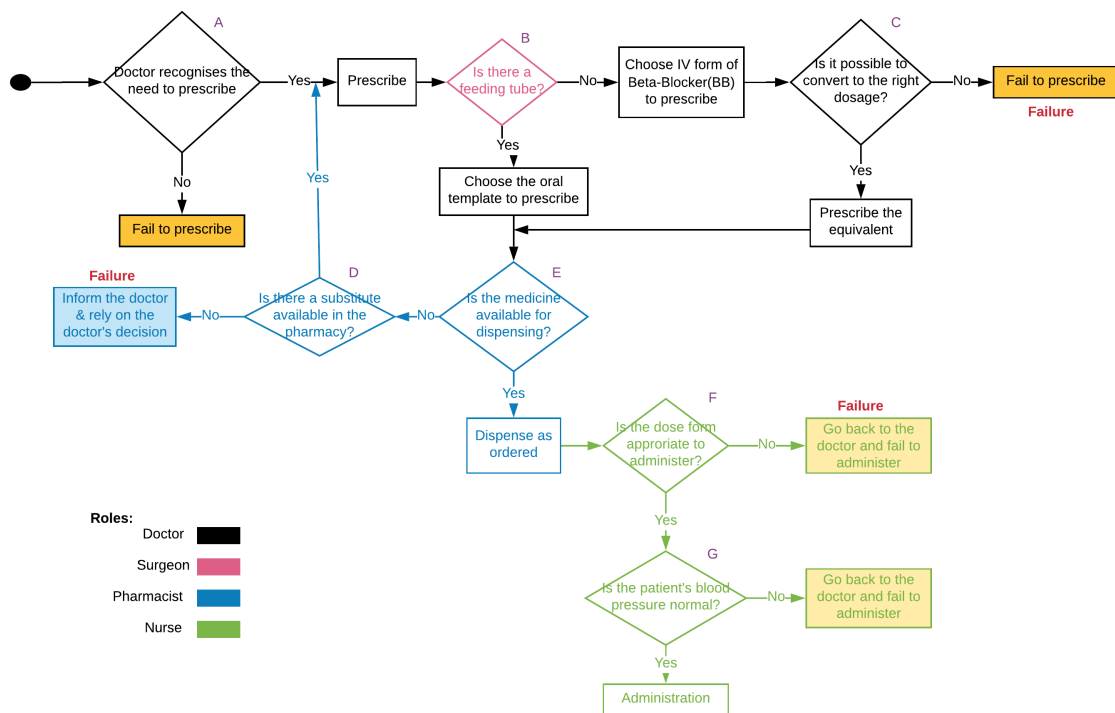


Figure 6.4: Decision-making flowchart for prescription and administration of medication

The simplified decision-making model is shown in Figure 6.4, which identifies the roles of the different professionals involved in this process. The main stakeholders in this clinical practice are the doctor, the pharmacist and the nurse with responsibilities for the prescribing, dispensing and administering phases, respectively. The surgeons are also included as they normally determine whether or not an FT is fitted, which serves as an important context for the medication decision-making. In practice, the situation is more complex, e.g. as more than one nurse will be involved in administering medications and sometimes because of understaffing, the nurses might not be able to administer the medicine as expected, but Figure 6.4 is intended to map the clinical roles rather than the work of particular individuals.

In Figure 6.4 the ‘administration’ outcome represents the success of the overall medication management process. Figure 6.4 also shows three types of failure: fail to prescribe, fail to dispense and fail to administer — the outputs of decisions C, D and F respectively. These are cases where the BB will not be administered, which might increase the risk of AF. There are also potential controls for these failures. In general, these involve the other

professionals (nurse or pharmacist) referring back to the doctor for resolution, e.g. if the medication is detected to be of an inappropriate form to administer. In addition, this decision-making model shows the complicating factors, e.g. the presence of an FT and the need to calculate equivalent doses, that we mentioned in Section 6.4.1.

6.4.3 Safety Analysis

In safety management it is common to organise risk analysis and control around the notion of a hazard – a situation which, if not controlled, could lead to harm [290], as mentioned earlier. Identifying hazards is often carried out by experts who determine specific situations which could give rise to harm, prompted by the guidewords. This case study uses SHARD guidewords (omission, commission, early, late and incorrect) to identify hazards. The identified hazards, as confirmed with the judgement of medical experts, are as follows:

- Omission – failure to administer BB to patients who took BB pre-operation (Hazard 1);
- Commission – unnecessary BB administered (Hazard 2);
- Incorrect – patient receives wrong BB (e.g. contra-indication with other medications or comorbidities) (Hazard 3);
- Incorrect – underdosage of BB (e.g. not all prescribed doses administered) (Hazard 4);
- Incorrect – overdosage of BB (e.g. repeated administration of doses) (Hazard 5).

However, it should be noted that there may be an appropriate omission in order to reduce the possibility of a different harm – worsening hypotension despite the potential increased risk of AF. Timing issues (early and late) are not separately considered as hazard categories here, as any harmful effects would be ‘caught’ by the under and over cases of the incorrect hazard category.

The clinical outcomes from these hazards could vary significantly. AF is most likely to be caused by omission (a failure to administer BBs). The effect of ‘incorrect’ depends on the medication administered; it might just cause dizziness, although the worst-case outcomes might be more severe.

Having identified hazards, it is necessary to determine potential causes of hazards as this gives a basis for defining controls and to reduce risks. Here we also use SHARD to

Table 6.1: SHARD results of the decision-making model (* assumes correct medicate)

| Guideword | Deviation | Possible Causes (Labels correspond to decisions in Fig. 3) | Potential Detection/Protection | Potential Effects |
|------------|--|--|--|---|
| Omission | No BB administered (Hazard 1) | <p>① G. Patient is suffering hypotension (may be due to epidural) and nurses decide not to administer</p> <p>② B, F. Wrong form of BB prescribed or dispensed for available route, so nurses do not administer</p> <p>③ A, C. No BB prescribed or dispensed</p> <p>④ Understaffing of wards leads to doses being missed (organisational factor)</p> <p>⑤ Complete failure of IV device or infusion pump (technical factor)</p> | <p>1. Clinicians should check BP and medications using drug chart on a daily basis</p> <p>2. Pharmacist should review prescription with clinician if suitable medication unavailable</p> <p>3. Nurse should identify the wrong form and query with clinician</p> | AF |
| Commission | Unnecessary BB administered (Hazard 2) | <p>A. Unnecessary BB prescribed</p> <p>E. Unnecessary BB dispensed</p> <p>Busy ward leads to administering medicine for wrong patient (organisational factor)</p> | <p>1. Pharmacist should review the prescriptions</p> <p>2. Nurses should check the prescriptions before administering</p> | Adverse interactions with other medication or comorbidities |
| Incorrect | Wrong BB administered (Hazard 3) | D. Incorrect substitution | 1. Pharmacist should review the prescriptions | Adverse interactions with other medication or comorbidities |
| Incorrect | Under dosage* (Hazard 4) | <p>C. Incorrect dose calculation</p> <p>G. Patient is suffering hypotension (may be due to epidural) and nurses decide not to administer</p> <p>Understaffing on wards leads to some doses being missed (organisational factor)</p> <p>Inappropriate recommendation from EPR (technical factor)</p> <p>Rate error of IV device or infusion pump (technical factor)</p> | <p>1. Order entry from the EPR might help the clinician to prescribe correct dosage</p> <p>2. Pharmacist might pick up the error</p> <p>3. Nurses might pick up the error</p> | AF |
| Incorrect | Over dosage* (Hazard 5) | <p>C. Incorrect dose calculation</p> <p>A. Doctor might prescribe both forms of BB to let the nurse choose the suitable one and both doses are given to the patient.</p> <p>Inappropriate recommendation from EPR (technical factor)</p> <p>Rate error of IV device or infusion pump (technical factor)</p> | <p>1. Order entry from the EPR might help the clinician to prescribe correct dosage</p> <p>2. Pharmacist might pick up the error</p> <p>3. Nurses might pick up the error</p> | Hypotension |

identify the causes of the hazards, by applying the guidewords to the decision-making model in Figure 6.4. The analysis results are summarised in Table 6.1 with the following columns: the guideword applied to the flow of information, the interpretation of that guideword (deviation), possible causes of the deviation (either local failures or incorrect inputs from earlier in the process), ways of detecting the deviation and protecting against it, and finally the potential effects. Using SHARD, it is common to work through the whole system description, starting at the end and working backwards. Therefore, in our case, the SHARD analysis starts at the end of the decision-making model, i.e. the administration. In addition, we also include technical and organisational factors that are not explicit in the decision-making model (the entries in blue). The entries in the deviation column correspond to the hazards identified above. The entries in the detection/protection column include use of EHR to make recommendations on order entry but are heavily dependent on the medical staff.

When assessing the risk associated with the hazard, it is common to quantify the likelihood of these causes and to use these figures to prioritise the introduction of risk controls. Where it is difficult to quantify the likelihood of hazard causes, the analysis is typically qualitative, and judgement is needed on the choice of controls to manage risk cost-effectively. In this case study, risk assessment is essentially qualitative based on clinical judgement. For example, understaffing of wards, as shown in Table 6.1, might be the most likely cause of failure to administer BBs to patients following a thoracic operation. This information can then be used to prioritise the introduction of new controls. However, even this qualitative approach can be difficult in a medical setting because of the many shaping factors such as the patient's general health, comorbidities, etc. This is a further motivation for using ML to complement the safety analysis. We show in Section 6.4.5.2 that hazard 1 – failure to administer BB to patients following a thoracic operation – would cause a 11% increase in the likelihood of AF post-operation, a result which no qualitative analysis could produce.

6.4.4 Data Generated from Clinical Practice

Following our safety analysis, we have identified that factors such as the presence of hypotension (maybe due to epidural) can influence the decision to administer BBs (the first entry in Possible Causes Column in Table 6.1). To understand whether or not these factors really are significant requires analysis of real data generated from the clinical practice.

Whilst this could be seen as simply confirming expert opinion, experts' views can vary, so being able to base the results on extensive datasets helps to resolve inconsistent opinions. That is, the safety analysis has given us a hypothetical view of the causes and effects reflecting Safety-I. However, the actual risk still needs to be evaluated and validated based on real clinical data supporting Safety-II.

For brevity, we use Hazard 1 – no BB administered – as an illustration in this case study, i.e. the 1st row in Table 6.1. Given the clinical context, there are three primary variables for us to consider, Pre_beta, Surgery, Post_beta (see Table 6.2). The reason why we categorise Surgery based on whether it involves the thorax or not is that, 1) any major thoracic surgery (not only oesophagectomy) carries the risk of AF in post-operative care especially for patients taking BBs before the surgery, see the research [283]; 2) in order to get more data to carry out ML, it is useful to consider all thoracic surgery rather than just oesophagectomy. Thus, the data analysis should identify these patients, along with the potential causes and effects of medication error, i.e. the 3rd and 5th columns in the 1st row in Table 6.1, which means that we should consider: hypotension, epidural, busy ward, understaffing of wards, failure or error rate of IV device or infusion pump, from the causes column and AF from the effect column.

In this case study, we only focus on a subset of the potential causes of Hazard 1 – Hypotension and Epidural – because of the scope of the MIMIC-III dataset. We use it to provide the 'real data' that generated from the clinical practice. All the SQL queries used to extract the data are available online at <https://github.com/Yanjiayork/papers>. Six of the variables identified above can be found in the MIMIC-III dataset and are described in Table 6.2.

We used the Current Procedural Terminology (CPT) codes [291] to determine whether patients definitely had thoracic surgery, definitely did not, or possibly did (where there are alternative ways of doing the operation) in MIMIC-III. For example, CPT code 43415 is defined as 'suture of an oesophageal wound or injury; transthoracic or transabdominal approach' which clearly can be conducted via the chest or the abdomen, hence we give it the value 1. On the other hand, CPT code 31760 is 'tracheoplasty; intrathoracic' which is definitely thoracic surgery, hence is given the value 2, and, of course, excision procedures on the oesophagus (oesophagectomy) are all given value 2.

The MIMIC-III records are time-stamped, and the records are analysed to identify patients taking BBs at any time before an operation whilst in hospital (Pre_beta), and

Table 6.2: Variables extracted from MIMIC-III Dataset

| Variables | Variable Code | Variable Values |
|--------------------------------|---------------|---|
| Surgery | Surgery | Value = 0, 'not thoracic' Value = 1, 'might be thoracic' Value = 2, 'definitely thoracic' |
| Receiving BB before surgery | Pre_beta | Value = 0, 'not receiving BB' Value = 1, 'receiving BB' |
| Receiving BB after surgery | Post_beta | Value = 0, 'not receiving BB' Value = 1, 'receiving BB' |
| Hypotension | Hypotension | Value = 0, 'no Hypotension' Value = 1, 'has Hypotension' |
| Epidural catheter placed | Epidural | Value = 0, 'no Epidural' Value = 1, 'has Epidural' |
| Having AF during the encounter | AF | Value = 0, 'no AF' Value = 1, 'has AF' |

within 24 hours after surgery (Post_beta), as this is the most critical time. The value for Hypotension is based on the first reading after 6am on the day following surgery (less than 100mmHg is viewed as Hypotension and given the value 1). This time is chosen as it is when patients would normally have their BBs, so this BP reading is the one that is most likely to affect the nurse's decision about whether or not to administer the BB (this is entry ① for hazard 1 in Table 6.1). Information about AF is inferred from the International Classification of Diseases, Ninth Revision (ICD 9) code, beginning with 427.

After the data preparation, 7,202 encounters were identified as relevant to this study, and had associated 'flags' indicating whether or not the patients suffered from AF, etc. A potential limitation for the data extraction is that when inferring the development of AF, we used the diagnosis table in MIMIC-III to determine which patient has AF after surgery, but because we do not have the medication history information for the patients (maybe due to our limitations in understanding the database), we might include chronic AF in the dataset and not be aware of it. It would be ideal if we could identify and exclude any such patients.

6.4.5 Applying ML to the Generated Data

After extracting the data generated from the clinical practice, we use ML to learn the real behaviour based on these data. It is important to note that many ML methods could be used to explore the data. However, in order to validate the safety analysis results, it is preferable to choose a ML method which can present the result so that it is understandable for humans. In other words, intrinsically interpretable ML models should be preferred. As we introduced in Section 5.2.2, logistic regression and Bayesian models are all viewed as interpretable model. Although, logistic regression can also highlight the most significant correlations between pairs of variables, a learned Bayesian network model can reveal a much finer structure by distinguishing between direct and indirect dependencies [27], which is particularly helpful in this case as it can be directly used to compare to the resulting structure from the safety analysis. For the reasons above, we decided to use BN structuring learning. Here we first use BN structure learning to understand the relationships between the different factors we extracted from the safety analysis in Section 6.4.3 that might compromise medication safety. Then, we use parameter learning to quantify the dependencies among these factors, which gives valuable clinical findings. Finally, the result of applying ML is used to update and enhance the safety analysis.

6.4.5.1 Learning Bayesian Structure from Data

A BN is a directed graph of nodes and edges connecting those nodes. Each node represents a random variable, while the edge between the nodes represents probabilistic dependencies among the corresponding variables. Associated with each node is a conditional probability table which specifies the probability of each node state given every combination of states of parent nodes. We use BNs as they have the potential to reveal a much finer structure by distinguishing between direct and indirect dependencies, by comparison with other statistical methods, such as logistic regression which focus on the most significant correlations.

BNs have a two-phase lifecycle. First, they are constructed, either by hand based on domain knowledge or by ‘structure learning’ from observational data [27]. In this case, the structure of a BN was learnt automatically using ML. To do this requires that we define a hypothesis space of possible structures for searching as well as a score function to measure each structure by a defined searching algorithm, such as greedy search [292]. In this case study, we use a greedy search-and-score methodology to learn the BN structure.

BN structure learning provides a means to make sense of the complex correlations in clinical data that have hampered other approaches. Secondly, it is necessary to determine the probability distribution of each node in order to fully specify BNs.

The structure of the BN was learnt from the six variables shown in Table 6.2 using BDeu (Bayesian Dirichlet equivalence uniform) score [293] based on the data preparation described in Section 6.4.4. BDeu is a widely-used scoring metric for learning BN structures for discrete data. Figure 6.5 presents the results. Note that arrow directions in the structure learnt should not be interpreted as showing causality, only a statistical correlation.

The model in Figure 6.5 generally reflects the safety analysis in Table 6.1, but there is one point of interest. It shows no direct dependency between Epidural and Hypotension, despite the fact that it shows Post_beta has an individual direct dependency with both Hypotension and Epidural. This is a very interesting finding, as our safety analysis shows that Hypotension should be the reason for patients not receiving Post_beta rather than Epidural. Epidural might affect Post_beta because it has the potential to cause Hypotension, but itself should not influence whether or not to administer BB. So, we expect there to be a direct dependency between Epidural and Hypotension, especially when Epidural has a direct dependency with Post_beta as shown in Figure 6.5.

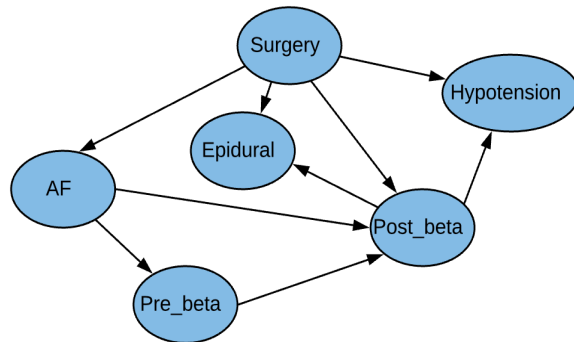


Figure 6.5: Learnt Bayesian Network Structure based on Safety analysis

We initially thought that it might be to do with confounding factors, e.g. normal BP readings will be obtained if they are measured whilst vasopressor medications such as phenylephrine or nor-epinephrine are being given by infusion. However, even when we ‘corrected’ the value of the variable by considering Hypotension to be present although the BP reading is normal, whilst such infusions were being given did not alter the learnt

structure. On reflection, this might best be explained as follows.

Epidural can cause Hypotension, but it might just cause a slight drop of the BP, not sufficiently severe to count as Hypotension [294]. Also, the means of giving Epidural, e.g. infusion or bolus will influence the result. If Epidural is given by bolus, it is more likely to cause Hypotension [295]. In addition, different patients will react differently and often the patients can bring up their BP in a short time (then keep it up) with their own bodily mechanisms. As we chose one time to read BP, this might not be immediately after epidural. Thus, it is not surprising that there is no direct dependency between Epidural and Hypotension.

Alternatively, Figure 6.5 might suggest a new pattern of how nurses carry out their work in the real world. If a nurse is aware that the patient has an epidural (and it should be quite obvious), a decision might be made not to give the BB, even when their BP is normal as they know the potential effect of epidural and BBs on BP. The combination of these two events is capable of causing a severe drop in BP in some situations [295]. Thus, a direct dependency is learnt between Epidural and Post_beta, even when there is no direct dependency between Epidural and Hypotension.

To confirm what really happened, it will be useful to refer back to clinicians by asking them focused questions based on the ML analysis results or observing their behaviours. This shows that ML can not only help to update and enhance the safety analysis, but also refine the questions that need to be asked in order to understand the real world.

6.4.5.2 Learning Parameters from Data

Based on the structure learnt in Figure 6.5, we used Bayesian estimation to learn the parameters for the network. Once the parameters of the BN have been specified, it allows exploration of the impact of decisions as the context evolves, i.e. probabilistic inference. As specific information about the context is known (e.g. the patient who had Pre_beta underwent thoracic Surgery), we instantiate the variables corresponding to the context in the network (i.e. Pre_beta = 1 Surgery = 2), which revises the probability for other variables (e.g. Post_beta or AF) in the BN to the posterior probability conditioned on the known context. 80% of the dataset was used to estimate the parameters, and 20% was used to test them by predicting the development of AF given the values for the remainder of the variables. Table 6.3 compares the BN and logistic regression methods for predicting AF. It shows that BN had slightly better prediction accuracy than logistic regression.

Table 6.3: Predictive accuracy of estimation methods

| Methods | Accuracy | Recall | Specificity |
|---------|----------|--------|-------------|
| BN | 72% | 6% | 98% |
| LR | 70% | 5% | 96% |

Table 6.4: Effects of post_beta on AF for patients with pre_beta and undergoing thoracic surgery

| Development of AF | Post_beta = 0 | Post_beta = 1 |
|-------------------|---------------|---------------|
| AF = 0 | 40% | 51% |
| AF = 1 | 60% | 49% |

In order to understand the extent to which Post_beta affects the development of AF, when patients underwent thoracic Surgery and had Pre_beta (i.e. the effect of Hazard 1) we assessed the posterior probability of developing AF conditioned on Surgery = 2, Pre_beta = 1 and Post_beta. The results are given in Table 6.4 and show that giving BBs after surgery reduces the probability of developing AF from 60% to 49%. In medical terms this is referred to as an 11% absolute risk reduction, or it may be expressed as the number needed to treat of 9 which is good for a medical intervention [296] [297]. This is an important finding as it not only confirms our SHARD analysis, but also is clinically significant. This confirms that giving Post_beta to patients who have Pre_beta and have had thoracic Surgery is beneficial in controlling AF.

Further, in order to assess how significant an influence Hypotension has on patients not receiving BB (first row of Table 6.1), we determined the posterior probability of Post_beta conditioned on Surgery = 2, Pre_beta = 1 and Hypotension, see Table 6.5. This shows that presenting Hypotension decreases the probability of getting Post_beta from 45% to 24%. Again, this is an important clinical finding.

Table 6.5: Effects of hypotension on post_beta for patients with pre_beta and undergoing thoracic surgery

| Post_beta | Hypotension = 0 | Hypotension = 1 |
|---------------|-----------------|-----------------|
| Post_beta = 0 | 55% | 76% |
| Post_beta = 1 | 45% | 24% |

6.4.6 Updating Safety Analysis and the Safety Argument

The result from applying the ML to the generated data can be used to update the safety analysis. Based on the learnt BN structure and Tables 6.4 & 6.5, it can be confirmed that the hazard cause ① in Table 6.1, i.e. patient is suffering hypotension and nurses decided not to administer, is valid, but the hypotension is not due to epidural. This also needs to be reflected in the safety case. Traditionally, safety cases strongly reflect the safety analysis of the system, but our methodology presents a way to update the safety analysis based on the ML results. As earlier, we use GSN for the safety argument.

A partial argument for control of the risks associated with AF is presented in Figure 6.6. The top goal is ‘Prevention of AF’ – amplified to say control of the risk of AF through use of BBs. The context includes patient characteristics and the hospital setting assumed to be an ICU, etc.

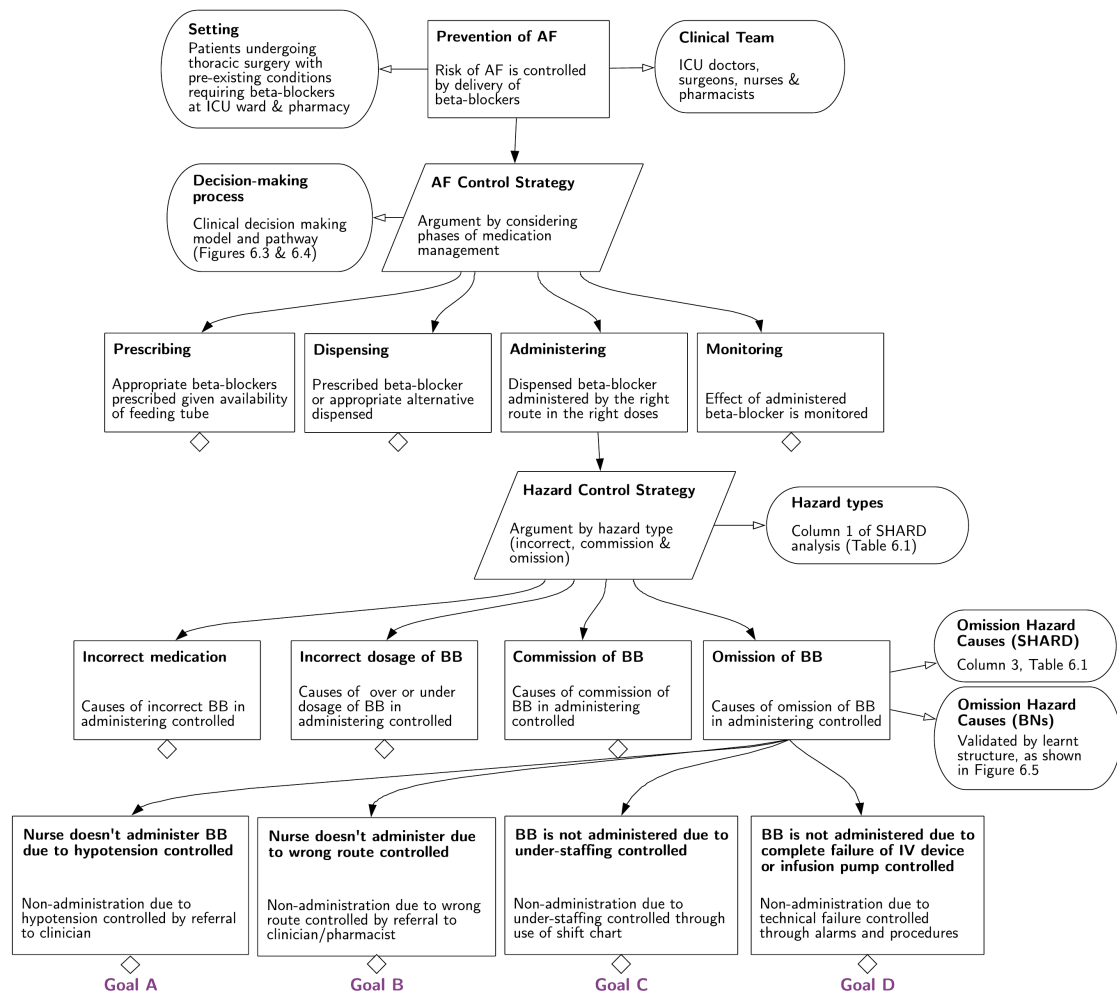


Figure 6.6: Safety Argument for Prevention of AF (with Emphasis on Omission of BBs)

The argument is broken down in several layers reflecting the decision-making model (Section 6.4.2) and the safety analysis (Section 6.4.3). The top strategy is a breakdown across the phases of medication management, reflecting the structure of Figure 6.4. Below this, the structure is organised first by types of hazard with under and over dosage grouped together in the incorrect dosage goal, and then by types of controls over the hazard causes. For brevity, Figure 6.6, only provides detail on Hazard 1 to illustrate the concept, leaving the rest of the argument undeveloped. The safety analysis and the learnt BN structure together provide the context for the ‘Omission of BB’ goal, which relates to Hazard 1. The goal is further decomposed to show that the causes associated with this hazard are controlled. Thus the sub-goals correlate strongly with the possible causes column in the first row in the safety analysis (Table 6.1), except for leaf Goal A that nurse doesn’t administer BB due to hypotension controlled. This goal doesn’t refer to epidural, to reflect the update to the safety analysis based on the ML results. Cause ③ in row 1 of Table 6.1 arises from prescribing and dispensing, therefore this cause is not dealt with under the administering branch.

Furthermore, the BNs also alert us to the level of risk related to Hypotension and show that the presence of Hypotension will reduce the chance of getting Post_beta by 21%, as shown in Table 6.5. This may be a key point to introduce stronger controls. After the introduction of new controls, we can use BNs to continue to learn from the new data, which should show that the effect of presence of Hypotension on getting Post_beta is reduced, if the control is effective.

6.5 Discussion

This chapter presented a new methodology, shown in Figure 6.2, recognising the distinctions between Safety-I and Safety-II, and providing a way of reconciling the two views of safety. The methodology proposed is novel, and has many potential benefits.

First, the methodology provides a practicable means of using ML to update safety analysis. This is intended to be a generic methodology, reflecting the fact that many systems are now data rich and ML can be used on available datasets to provide feedback from operation to the safety analysis. Our clinical case study used to illustrate the methodology demonstrated that safety engineering methods and ML are mutually supportive. Conducting safety analysis can help us proactively identify the relevant variables to explore in ML and to understand what kind of knowledge we expect to derive, rather than just believing

what the ML is telling us (treating it as a ‘black box’). The ML is used to update and enhance the safety analysis to show whether or not it is sound and reflecting ground truth rather than being merely hypothetical or subjective or seen as a mere paper-based exercise with little relevance to actual clinical practice.

Second, the ‘culture of paper safety’ has been a major threat to the validity of safety cases for complex systems in socio-technical environments [298] [299] [300]. Our methodology can help to avoid the problem of ‘paper safety’ and to focus on ‘real safety’. In our case study, although we can’t go back to influence the practices in the hospitals from which the MIMIC-III data was collected, the methodology has the potential to do so if it is used in a ‘live’ setting with access to clinicians.

Third, it would be valuable to identify ways to automatically update the safety analysis and safety case based on the ML results. One possible approach would be to use dynamic safety cases as part of the feedback mechanism. A dynamic safety case [301] has the evidence, and potentially the arguments, updated as the system operates. In the context of our methodology, this would mean using ML to continue learning from data generated from the clinical practice to update the safety case. To be useful, such updates also need to be associated with criteria for alerting clinicians, e.g. if there is a trend in the data that suggests that planned interventions are ceasing to be effective. This would further assist in focusing on ‘real safety’.

Whilst ML is very powerful, it is also possible to end up with misleading results. Here, we focus on the pragmatic issues that affected the production of the BNs used in this clinical case study, specifically the process for structure learning. Our approach to structure learning for BNs uses the BDeu scoring function, for which we must determine a single hyper-parameter, the equivalent sample size α . We found that the learnt structure is quite sensitive to the hyper-parameter α . There is a general trend that with the increase of α , more arcs are added in the structure, but not necessarily for every increase of α . Currently, there is no generally accepted rule for determining the right value of α , although there is some ongoing research into rules to set the value of α [302] [303].

Finally, it is possible to combine other methods to increase understanding of, and confidence in, the learnt structures. We have previously shown how the BN structure learning can be augmented with process mining [262]. BN structure learning can show statistical correlations between different factors, no matter whether they occur or not, e.g. the presence of absence of hypotension, and it can also evaluate the effects of one factor

on another when the context is known, see Section 6.4.5.2. Process mining can show what has really happened, and cannot explicitly identify when activities do not happen as there is no event and no timestamp [304]. Thus, it can discover temporal relationships between different activities which the BNs cannot. In our case study, process mining can help to give further insights into the results of the BN structure learning but is not a substitute for it, see my previous publication for the details [262].

6.6 Conclusion

This chapter has introduced a new methodology for updating and enhancing safety analysis carried out in the development phase by applying ML to the data generated from clinical practice in the operation phase. This gives insights that can then be used to help to ensure medication safety by updating the initial safety analysis. We believe that the combination of safety analysis methods and ML methods to enable reduction of the gap between development and operation is a unique approach, especially valuable when systems become more data rich and goes some way to implementing Hollnagel's notion of Safety-II without losing the merits of Safety-I.

We employed a case study focusing on post-operative care following thoracic surgery both to illustrate and to validate the methodology. In the case study we were able to identify clinically meaningful results relating to thoracic surgery. It also shows the unanticipated mis-alignment between the development phase and operation phase reflected through the initial safety analysis and the ML results. Although, for this case study, only clinical factors were considered in the ML analysis, it would be ideal to cover the range of factors that can cause and control hazards, including technical and organisational, as well as clinical factors. It would be necessary to obtain data on each factor to enable accurate analysis of the influences on medication error, and undesirable outcomes, e.g. AF. To provide this dataset requires much more extensive preparation, e.g. drawing on hospital administrative data, as well as clinical data such as is available in MIMIC-III. This would help to maximise the potential value of this methodology. In addition, this methodology might be able to support the FDA's proposed TPLC approach by using ML to update safety analysis and the safety case in order to assure the safety of SaMD through life.

This chapter complements the second case study and provides a further answer to question 2, *are there new opportunities for well-established safety engineering methods with the development of ML and why are they specifically good for safety in this domain?*

Opportunities have been identified and illustrated for improving safety analysis using ML. The work of the chapter has shown how ML can be used for validating the SHARD safety analysis showing that the majority of predicted hazard causes do arise. It has also shown that it is possible to refine the safety analysis based on insights from operational data by identifying that a predicted hazard cause does not arise and could be discounted. Note that safety analysis is often “open loop”, i.e. there is a lack of systematic means to validate safety analysis, so this helps to get feedback for improving the application of safety engineering methods.

To investigate generalisability of this methodology, it is worth mentioning that results from this case study can't be assumed for other hospital settings, even when the same clinical practices are of concern. But the methodology would still hold, which means that the safety analysis and data generated should all reflect the local clinical practice. Further, the work presented here has only used BN structure learning for tabular datasets, and how easy it is to use other ML methods remains an open question. Because of the nature of the methodology we need to find variables to represent the hazards, causes of hazards, effects of hazards, so this approach would most likely only work for tabular datasets.

Chapter 7

Conclusions

In concluding, we return to the two questions raised in the introduction, provide a summary of the contributions in the thesis, consider generalisability of the contributions, then outline possible future work.

7.1 Research Question 1

The first question that this thesis addresses is *are well-established safety engineering methods still appropriate and effective in assuring the safety of ML in some representative healthcare scenarios?*

7.1.1 Response to Research Question 1

Healthcare is very complex and clinicians may encounter situations that are unprecedented. However, clinical pathways are defined for commonly occurring situations – which we can therefore characterise as representative. Further, the clinical pathway provides the foundation for applying established safety engineering methods. Thus, the use of RL in support of sepsis treatment in Chapter 4 provides a positive answer to research question 1. It shows that *well-established safety engineering methods are still appropriate and effective in assuring the safety of ML* by providing links from hazard analysis conducted against the clinical pathway, via DSRs to concrete changes to the RL model development.

As well as answering question 1, undertaking this work has identified some insights from the use of SHARD and some limitations of the method in this context. Key insights which we believe would apply to other uses of SHARD in healthcare are:

- It is critical to define the clinical pathway before applying SHARD (it is a precondition);
- The clinical pathway needs to be defined at a suitable level to include the ML-based SaMD, i.e. not too abstract, so it is helpful in identifying interface hazards between the human and the SaMD, and not too detailed, so the analysis can be completed with a reasonable level of effort;
- The identified *interface hazards* can further be used to determine DSRs for the ML-based SaMD;
- Identifying the potential causes of the deviations is the most laborious part of the method;
- Clinical knowledge is essential for determining the effects and credible causes of hazards.

The importance of the clinical pathway and clinical knowledge for the effective use of SHARD is also seen in the case study in Chapter 6.

SHARD also has some limitations in the context of analysing ML-based SaMD in healthcare, and we use NNs to illustrate the problems. First, when SHARD is used for a conventional software system it is applied to data flows between major functional blocks. There will typically be a fairly modest number of blocks, each block will have a fairly discrete function, and the blocks will not be highly interconnected. This means that analysing each flow and working back from the output to the input is fairly straightforward, even if it is laborious. However, NNs have very complex data flows and layers in an NN are fully interconnected making following a “flow” through the network practically impossible for anything but the simplest NN.

Second, the individual neurons in an NN are quite primitive so it is hard to determine appropriate guidewords and to make sound judgements about potential deviations at such a low level of detail. Also, there may be thousands of neurons in a large NN, making manual analysis impossible. Trying to analyse at the level of layers in the NN would also pose problems as it will be hard to identify the function performed by the layer (by comparison with blocks in conventional software).

Thus, SHARD and similar methods can support analysis of ML models in context, e.g. integrated in a clinical pathway, showing the potential impact of hypothetical failure

modes of the ML model in the wider context, but it is very difficult if not impossible to use SHARD to analyse the failure behaviour of the ML model itself, especially when it is complex.

All the case studies used GSN which is another well-established safety-engineering method, although the uses in each case study are rather different. The usage in Chapter 4 is perhaps the most typical example of GSN, showing that the risks associated with an ML-based DSS for sepsis treatment are controlled, and supporting this with evidence from design and development of the ML model. Chapter 5 highlights the role of explainability in showing that an ML Model meets its safety requirements and in validating those safety requirements. Finally, Chapter 6 shows how the safety argument can be updated based on analysis of real-world data; this is a relatively unusual usage of GSN but may contribute to development of dynamic safety cases [301].

GSN is a very flexible notation and this has been of value in constructing arguments for the three case studies; no difficulties have been found in applying the notation, despite the very different nature of the arguments. In some contexts, safety argument patterns have been identified for commonly-arising situations, e.g. arguing over all hazards associated with a system. It may be that, in time, argument patterns can be produced for situations exemplified by the case studies in this thesis, but more experience would be needed to have confidence that the approaches followed here are general enough to encode as patterns. It should be noted that there are argument patterns for dealing with assurance of ML, see AMLAS [182], but these are domain-independent patterns and there remains a “gap” between these generic patterns and what needs to be done in a healthcare context to demonstrate regulatory compliance as well as safety – although, as we show below, the work presented here goes some way towards meeting regulatory requirements.

7.1.2 Regulatory Context

Here, we further consider this question in the context of the regulatory approaches in the USA, Europe and the UK presented in Section 2.4, drawing mainly on the work presented in Chapter 4.

The regulatory frameworks in the three jurisdictions cover, and the associated standards identify, a number of requirements for applying safety engineering methods to ML-based SaMD. Some of the more significant requirements identified in Chapter 2 are set out here together with a discussion of how they have been addressed in this thesis:

- ISO 14971 [30]: it is necessary to identify hazards in normal operation, under fault conditions and arising from human error – the SHARD analysis of the clinical pathway (see Figure 4.4 and Table 4.1) fulfils this role, where “fault conditions” are taken to include potentially unsafe behaviour of the RL model;
- ISO 24971 [32] (1): results of hazard and risk analysis should be used to inform design – the SHARD analysis helps to identify DSRs (see Table 4.2), which specifically influence the RL learning process, e.g. shaping the cost function, hence they directly inform design of the ML to reduce risk;
- ISO 24971 [32] (2): ensure that design changes do not introduce new hazards – the iterative nature of the methodology (see Figure 4.2) ensures that changes are scrutinised to see whether or not they introduce new hazards, although in our case study, only one major hazard is considered;
- Benefit vs risk (underlying many standards/regulations): there is a need to weigh benefits against risks – this can be seen in Chapter 4 especially in Tables 4.4 and 4.6 which show that the original learnt policy has a better performance when using off-policy evaluation, but it raises a new safety concern as there are many sudden changes in vasopressor dose for one patient. In contrast the modified policy presents a lower performance when compared to the original learnt policy, but it reduces the rate of sudden changes in vasopressor dose, which is considerably safer based on clinical knowledge. This reflects the balance between risk and benefit;
- CQC/MHRA sandbox [108] and CONSORT-AI guidelines [122]: provide clarity on the intended use of ML devices within clinical pathways to ensure high-quality care – this is done by incorporating the ML model in the clinical pathway in Figure 4.4.

The standards and regulatory documents state *requirements* for the safety assurance of ML, as shown above. As well as showing that the established safety engineering methods are appropriate and effective applied to representative ML-based systems in healthcare, we have also shown above that they help to support regulatory compliance.

7.2 Research Question 2

It is easy to think of ML as a “problem”, rather than as a “solution”. The second question takes the “solution” point of view seeking to understand *are there new opportunities for*

well-established safety engineering methods with the development of ML and why are they specifically good for safety in this domain?

7.2.1 Response to Research Question 2

The work presented in Chapters 5 and 6 show quite different opportunities for safety engineering methods with the development of ML. We briefly summarise each but focus mainly on the “why” part of the question.

Chapter 5 shows the opportunity for explainable AI methods to make contributions to safety. In development, the methods help to improve safety of the learnt model and to demonstrate clinical association and robustness of the ML model as a whole. In operation the case study has demonstrated how to use explainable AI methods to provide assurance to the clinician/end user relating to treatment of a single patient.

Why are these methods specifically good? Established safety engineering practices work, in part, by “flowing down” DSRs to system components. The adaptation of this approach to ML models was shown in Chapter 4 with the DSRs constraining the state space and cost function in the RL model (rather than being more direct requirements on the system itself as is typically the case [163]). In addition, in Chapter 5 we show how to use of influential instances in support of “design for safety”. So in this case, the “why” is because the use of explainable AI methods allows the *intent* of the established safety engineering methods to be preserved, even with complex ML models. Similarly, the use of counterfactuals gives a way of implementing another established safety engineering principle, to show that no single point of failure can give rise to a hazard – by exploring the details of the learnt model in a way that would be impossible using established methods such as FMEAs. Finally, the use of feature importance supports clinical validation of the ML model as learnt – in safety engineering terms this is showing that the nominal behaviour is safe. This would traditionally be done by expert review or simulation – the use of feature importance enables clinicians to undertake this expert review which they would otherwise be unable to do. In summary “why are they specifically good for safety in this domain?”; it is that they use ML methods to enable established safety engineering *principles* to be applied to ML models, where this cannot be done using the established safety engineering *methods*. Thus in “applying ML to ML” we are able to bring complex ML models within the scope of established safety engineering principles.

Chapter 6 has identified and demonstrated opportunities for improving safety analysis

using ML. In particular it has shown how to validate and refine safety and hazard analysis based on insights gained by using BNs to identify safety-relevant correlations in the operational data and comparing them with the predictions from the SHARD analysis.

Why is this approach specifically good? In *principle*, safety engineering is undertaken “through life”. In *practice*, safety analysis is often “open loop” unless an accident or incident occurs; in such cases the investigation may identify desirable improvements to safety analysis as well as to the system itself. Thus, the use of ML to analyse operational data gives the opportunity to improve established safety engineering practices, by allowing the “loop to be closed” during ongoing operations, not only after an incident or accident when it is arguably too late. In summary “why are they specifically good for safety in this domain?”; it is that they have the potential to fill a gap in current safety engineering *practices* (as opposed to *principles*) and one that will become ever more important as systems become more complex. However, we must acknowledge that we haven’t “applied ML to ML” in this case and it remains to be seen how effective the approach explored in Chapter 6 would be if applied to ML-based SaMD.

7.2.2 Regulatory Context

The notion of using ML as part of the “solution” is not explicit in regulations, but it is implicit in some of the exploratory work in the wider community. The IMDRF has done work on Quality Management Systems (QMSs) for SaMD [14] and on Clinical Evaluation of SaMD [257]. The IMDRF proposals indicate challenges for evaluation and validation of ML-based SaMD – which also represent opportunities. A regulatory sandbox conducted by the CQC and MHRA [118] also identifies opportunities for ML methods. We outline how the work presented in this thesis can help to realise these opportunities.

- IMDRF Clinical Evaluation [14]: is there a valid clinical association between your SaMD output, *based on the inputs and algorithms selected*, and your SaMD’s targeted clinical condition? – the global feature importance (see Figure 5.7) gives confidence that the model is considering relevant clinical factors, e.g. ventilator mode and peak inspiratory pressure;
- CQC and MHRA sandbox [118] (1): those interacting with the tool have a good understanding of what it does and doesn’t do – the explanations of the data (see Figure 5.3) show that the tool only deals with adult patients (what it *does* do) and

doesn't deal with those who undergo elective surgery and need breathing support (what it *doesn't* do);

- CQC and MHRA sandbox [118] (2): clinicians who interact directly with ML systems understand how they work and how to use them – the counterfactual explanations (see Table 5.5) focus on a specific patient, helping clinicians to see what actions to take (*how to use* the outputs from the system) to increase the chances of a successful extubation.

The regulators, especially the FDA, emphasise the need to manage safety through life although they are not specific about how this might be done or managed. However, there are some ideas about how to do this in the research literature. We present an assessment of how these ideas can be realised using the work presented in this thesis.

- Work-as-imagined vs work-as-done [305]: in Hollnagel's terms, safety analysis is undertaken on work-as-imagined, i.e. as modelled in development, not work-as-done, i.e. in operation. He says that there is a need to understand work-as-done, including 'what is done right', to ensure safety. Our approach has taken this abstract concept and shown how to make it useful and actionable by employing ML to analyse data from work-as-done. Thus, our work could help to achieve sound adoption of Hollnagel's ideas and the associated concepts of Safety-I and Safety-II [280] which are quite influential in healthcare;
- Updating safety cases [301]: it is necessary to update the reasoning about the safety of ongoing operations – the role of the BN in updating the reasoning in the safety case is explicit in Chapter 6, specifically in the context for the goal "Omission of BB" (see Figure 6.6). The FDA ML-based SaMD action plan [106] identified the need to collect and monitor real-world data to implement TPLC, which is proposed in order to manage ML models that continue learning in operation. Therefore, updating safety cases will be critical in order to reflect the changes in risk in operation;
- Dynamic safety cases [301]: these are an extension of the ideas of updating safety cases to enable the argument and evidence to be updated automatically. Some researchers [43] have suggested using dynamic safety cases to support the TPLC approach proposed by FDA to allow the ML model to continue to learn in operation. Using ML to analyse operational data seems inevitable to enable such concepts to be implemented, and we see our work as an initial step in this direction.

This work shows the role of ML in supporting through life safety management which is clearly important in healthcare, whether or not ML-based SaMD are used. It also sheds some light on the ideas of dynamic safety cases although it is far from realising the overall concept as identified in the literature, e.g. [301]. The work also gives some insight into possible ways of providing through-life support for systems, potentially contributing to realising the FDA’s TPLC concept although more would need to be done to understand how to “apply ML to ML” as noted above.

7.3 Summary and Directions for Future Work

This thesis has explored issues at the intersection of ML, safety engineering and healthcare. Whilst significant progress has been made there are many more opportunities to explore. We close by summarising the key contributions, presenting some general findings then identifying some broad themes for future work.

7.3.1 Summary

This thesis has made three major contributions:

1. Showed how well-established safety engineering methods can be applied to ML-based systems to assure safety, both supporting “design for safety” and producing evidence to demonstrate safety (see Chapter 4);
2. Demonstrated the role of explainability in helping to provide safety assurance for ML-based SaMD (see Chapter 5);
3. Demonstrated how to use ML to enhance well-established safety engineering methods including updating safety analysis from operational data (see Chapter 6).

The three case studies do not illustrate all the possible synergies between ML and safety engineering but they do provide substantive and complementary contributions. Further these contributions address some of the requirements for demonstrating safety of ML-based SaMD identified by regulators (see Sections 7.1.2 and 7.2.2).

7.3.2 General Findings

Two broad themes have emerged from this work.

First, it is very important to model the pathways in which ML-based SaMD are used so as to conduct effective safety engineering. The experience in conducting the research presented in this thesis is that clinical pathways are often weakly defined with a lot of implicit assumptions. Further, there can be many quite different published pathways for the same clinical issue. Hence, we believe that real benefit can accrue from up-front effort in defining clinical pathways precisely enough, including identifying the role of ML-based SaMD in the pathway, so that they can be subjected to hazard and safety analysis and validating these pathways with clinicians prior to their use.

Second, we have presented a number of methodologies in this thesis, and it might be expected that future work would be to try to unify these into one overarching methodology. However, we do not take this view. Healthcare is an extremely complex socio-technical enterprise and trying to produce one comprehensive methodology is likely to be very difficult – and, more importantly, not very useful. We believe that the methodologies we have produced are appropriate for their purpose, and that Figure 3.1 serves to show how the different elements of this work fit together.

7.3.3 Generalisability

There is a question of generalisability of the work undertaken in this thesis. The scope of this thesis is in healthcare. Therefore, we do not have evidence that the methodologies and techniques we used apply in other domains, but there may be merit in exploring them in other domains. In addition, all of the case studies use tabular data (from MIMIC-III) so we would have more confidence in getting satisfactory results if the methodologies are applied to other tabular data in healthcare. Also, the work presented in this thesis uses RL and supervised learning (CNNs and BN structure learning). We do not have evidence that the methodologies would work well with unsupervised learning.

The first methodology should generalise well to other healthcare applications using deep RL. The second methodology should generalise well to other healthcare applications using supervised learning as many of the explainable AI methods are model agnostic, i.e. can be applied independent of the “base” supervised ML model used. The third methodology should generalise well to other healthcare applications that have good data from operation. We expect this to include cases which incorporate ML-based SaMD within clinical practice.

It is tempting to seek to generalise the results of the research as widely as possible,

however we would suggest caution. In approaching other case studies it will be instructive to see if their characteristics are similar to those we have presented here. If they are similar, this suggests that the methodologies developed should be evaluated for use on the new case study and they may well be usable with minimal adaptation.

7.3.4 Future Work

In Chapters 4 to 6, we considered future work for the individual methodology, e.g. by applying the methodology in new healthcare settings, or seeking to progress the work using different ML methods. Here, we consider three broad perspectives on future work rather than focus on the individual methodologies.

First, all our analysis drew on the MIMIC-III dataset. Whilst this is a very valuable data source, it would facilitate ML and safety research in healthcare further if there were more open source healthcare datasets. In addition, MIMIC-III only includes clinical information. As noted above, healthcare is a socio-technical system, so it would be valuable to have a dataset that also includes organisational information, e.g. on shift patterns or stress factors on a ward, that might contribute to risk and lead to undesired outcomes. To take forward some of the work described here, particularly that in Chapter 6, would need access to a richer dataset that included these additional factors. Whilst there might be concerns from healthcare staff about such data collection, without such data it will be difficult to achieve through-life learning to improve patient safety, addressing all the potentially relevant factors. Thus work to identify and gather such datasets will be an important element of any future work in this area.

Second, development and assessment of ML-based SaMD are highly iterative. The case study in Chapter 4 was greatly facilitated by using AdvoCATE and it was easier to track the changes in models, DSRs, etc. with the tool than without it. However, the tool was also limited in the sense that it embeds concepts that are more relevant to aerospace which, to an extent, get in the way when considering healthcare situations. For example, the tool will automatically generate safety arguments from the BTDs (using built-in patterns) but we had to bypass this mechanism in our work. Developing effective tools is difficult, but there would be merit in exploring how to develop tools like AdvoCATE but better adapted to the healthcare domain, e.g. supporting definitions of clinical pathways, and so on.

Finally, the work carried out in this thesis is multi-disciplinary. Several of the papers which inspired the individual case studies, e.g. on sepsis, emphasised the need to take a

multi-disciplinary approach. However, to our knowledge, we are the first to fully embrace safety engineering within the multi-disciplinary team. It seems essential that future work in this area takes a multi-disciplinary approach – and ideally expands the team to include human factors specialists, statisticians, and so on. If this happens, then our work on embracing ML in safety assurance in healthcare might be seen as a key step towards developing a fully inclusive approach to addressing the enormous challenges facing the global healthcare system.

Appendix A

SHARD analysis for the clinical workflow in case study 1

This appendix presents the full SHARD analysis for the clinical workflow shown in Figure 4.4, using the approach described in Section 4.4.2 in case study 1 presented in Chapter 4. The detailed description of SHARD can be found in [8]. Our analysis focuses on vasopressor administration; provision of oxygen and administration of antibiotics are not analysed here.

Tables A.1 and A.2 present the summary of the SHARD analysis results and is the basis for Table 4.1. Tables A.3 to A.7 below present analyses of five key activities in the workflow, working back from nurse administration (Table A.3) to input of patient feature data (Table A.7). Working back (rather than forward) means that hazards are identified at the outset where we focus on the final output of the workflow, and the later analyses can focus on finding possible causes for the identified hazards. Tables A.8 and A.9 are a summary analysis for the RL recommendation in the workflow that enable us to identify *interface hazards* – what can go wrong with the RL model and present problems that the clinicians must manage.

Severity is only included for Tables A.1 and A.2 (as they show the final effects on the patients). The classification: Minor, Significant, Considerable, Major, Catastrophic is used based on NHS Digital standard DCB160 [9].

In all the tables the leftmost column is the prompt (SHARD guideword) and the next column is the deviation or deviations that can arise. This is followed by possible causes of the deviation(s). The remaining columns differ between the summary (Tables A.1 and A.2) and the more detailed tables.

Table A.1: Overall Administration of Vasopressors, Part 1

| Guide word | Deviation | Possible Causes | Effects | Severity | Justification or Design recommendations |
|------------|---|--|---|--------------------|---|
| Omission | No vasopressor administered | 1. Nurse fails to administer the vasopressor (e.g. due to workload) 2. Doctor does not produce final decision (e.g. due to workload) 3. Initial recommendation by doctor is not produced (not considered) 4. Failure to output results due to software fault in RL agent (e.g. errors in the display code) 5. Algorithm or hardware does not have the capability to process the input data (e.g. inadequate memory allocation) 6. Data corruption (e.g. invalid or wrong data produced by over-writing patient's features) 7. Some features are not entered (e.g. due to staff workload) 8. Some test results are missing or not available 9. Wrong form of patient features entered | Continuous hypotension | Considerable | |
| Commission | Unnecessary vasopressor administered | 1. Nurse administers to wrong patient 2. Doctor produces unnecessary final decision (e.g. due to miscommunication at handover) | Potential danger to the patient, e.g. Cardiac Arrhythmia | Considerable | |
| Incorrect | 1. Wrong vasopressor administered | 1.1 Allergy is not recorded, or wrong information provided 1.2 Doctor fails to check allergies and thus wrong vasopressor decided 1.3 Patient allergic to initial vasopressor recommended by doctor (not considered) 1.4 Patient allergic to vasopressor recommended by RL agent (not credible) | 1. Adverse drug effect | Major/considerable | |
| | 2. Wrong dose administered (this hazard concerns single dose) | 2.1 Wrong initial dose recommended by Doctor (not considered) 2.2 RL recommends wrong equivalent dose and doctor accepts the advice, e.g. due to automation bias 2.3 Failure to update display with new recommendation 2.4 Algorithm or hardware does not have the capability to process the input data correctly (e.g. software fault) | 2. Potential danger to the patient | Considerable | Careful interface design required for using RL agent, especially causes (2&3&4) 6 to 9. (2&3&4) 10 to 11 might also need to be considered when designing the interface for using the RL model, but it might also arise from errors in the lab. Careful RL model design required, especially causes (2&3&4) 1 to 5 |
| | 3. Sudden drop of vasopressor dose administered | 3.1 Kink of line 3.2 The pump fails, e.g. due to electrical problem or bag/syringe not installed correctly 3.3 The delivery line might not be connected to patient's central line, e.g. due to the patient pulling out the central line 3.4 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) 3.5 Initial recommendation by doctor has a sharp drop in dose and doctor carried through the recommendation (not considered) 3.6 RL agent recommends sharp drop in dose and doctor accepts the advice, e.g. due to automation bias | 3.1 Acute Hypotension 3.2 Strokes, 3.3 Renal failure 3.4 Heart attack | Major/considerable | |
| | 4. Sudden increase of vasopressor dose administered (3&4 hazards concern two consecutive doses) | 4.1 Initial recommendation by doctor has a sharp rise in dose and doctor carried through the recommendation (not considered) | 4.1 Hypertension 4.2 Cardiac Arrhythmia 4.3 Strokes 4.4 Raised intracranial pressure 4.5 Pulmonary oedema | Major/considerable | |

Table A.2: Overall Administration of Vasopressors, Part 2

| | | | | | |
|-------|------------------------------------|---|---|--------------|--|
| | | <p>4.2 RL agent recommends sharp rise in dose and doctor accepts the advice, e.g. due to automation bias</p> <p>(3&4) 1. Inappropriate titration of dose by nurse (3&4) 2. Doctor fails to check current dose</p> <p>(2&3&4)1. Features in state space of the RL model are not sufficient to represent the patient conditions for sepsis decision making (2&3&4)2. Reward function used for RL model is coarse (2&3&4)3. Cost function used for RL model development is not appropriate (2&3&4)4. Hyperparameters used for RL model development are not optimised (2&3&4)5. Training data for RL model development is not appropriate</p> <p>(2&3&4) 6. Nurse prepared wrong dose (e.g. due to calculation error) (2&3&4) 7. Data corruption (e.g. invalid or wrong data produced by over-writing patient's features) (2&3&4) 8. Features for wrong patient entered (2&3&4) 9. Wrong patient feature values entered (e.g. due to unit difference) (2&3&4) 10. Test results for wrong patient received (2&3&4) 11. Incorrect test results received</p> | | | |
| Late | Delay of administering vasopressor | <p>1. Late to get central line access 2. Late to get the vasopressor 4. Delay in administration (e.g. due to nurse workload) 5. Delay of the initial recommendation by the Doctor (not considered) 6. Algorithm or hardware does not have the capability to process the input data in a timely fashion (e.g. inefficient algorithm or infinite loop or inadequate memory allocation or hardware limitations) 7. Some patient features are not entered at all or not entered on time (e.g. due to staff workload) 8. Some test results are delivered late, missing or not available 9. Failure to output results due to software fault in RL agent (e.g. errors in the display code) 10. Data corruption (e.g. invalid or wrong data produced by over-writing patient's features) 11. Wrong form of patient features entered</p> | Continuous Hypotension or increased mortality | Considerable | |
| Early | N/A | N/A | N/A | N/A | There is ongoing clinical research about whether to deliver vasopressor earlier to increase MAP for sepsis treatment, thus it is not discussed further here. |

In Tables A.1 and A.2, the final three columns show the effect on the patient, followed by its severity and a justification or design recommendations. Design recommendations are included if it is judged that the risk can be, or needs to be, better controlled. In Tables A.3 to A.7, the final two columns are the output deviation – what “problem” can be passed on to the next activity in the workflow, and a justification or design recommendations.

Several of the rows in Tables A.1 and A.2 are simple. For example, with *omission* there is a single deviation and a simple list of possible causes. Cause 3 is in blue indicating it is not analysed further; this colour coding is also used elsewhere in the tables with the same meaning.

The row for *incorrect* in Tables A.1 and A.2 is more complex. There are four possible deviations, i.e. four possible ways in which the administered vasopressor could be incorrect. The first sub-row (wrong vasopressor) is simple, like *omission*. Deviations 2-4 are grouped into one sub-row as they have many common possible causes. The label $x.y$, e.g. 2.1, is the y th possible cause of deviation x . Those labelled (3&4) or (2&3&4) show that they can cause two or all three of the deviations. The effect column is sub-divided again to reflect the different clinical effects of the three different sub-cases of *incorrect*.

Early is not analysed in Table A.2 or any of the subsequent tables as the clinical effects are a matter of debate.

Table 4.1 is derived from Tables A.1 and A.2. It covers deviations 3 and 4 for *incorrect*, but grouped together as “sudden change” rather than separating out sudden increase or sudden decrease. Combining entries 3.5 & 4.1 and entries 3.6 & 4.2 reduces the 21 possible causes in Tables A.1 and A.2 to the 19 presented in Table 4.1. Entries (2&3&4)1-5 are the RL-related causes and appear as possible causes 10-14 in Table 4.1.

Table A.3 considers the final activity — administration in the workflow. The output deviations here are the deviations in Tables A.1 and A.2; for example, compare the entries for *incorrect*. Propagation can also be seen between the activities in the workflow, e.g. the input deviations in Table A.3 are the output deviations in Table A.4. The numbering of the input and output deviations in one table shows internal “flows”, specifically the number following the input deviation shows that it contributes to the corresponding output deviation.

Table A.4 shows the penultimate activity — final decision by the doctor in the workflow. All input deviations labelled X can contribute to different classes of deviation, e.g. *omission* to *late* in this case, which are highlighted in red. This colour code applies to

Table A.3: Nurses Administer Vasopressors as Advised by Doctor

| Guide word | Input Deviation | Internal Deviations | Output Deviation | Justification or Design recommendations |
|------------|---|---|--|--|
| Omission | Final decision is not provided | Nurse fails to administer the vasopressor (e.g. due to workload) | No vasopressor administered | |
| Commission | Unnecessary final decision is provided | Nurse administers to wrong patient | Unnecessary vasopressor administered | |
| Incorrect | Wrong vasopressor is decided (1) Wrong dose is decided (2) A sharp drop in dose is decided (3) A sharp rise in dose is decided (4) | 1.1 Allergy is not recorded, or wrong information provided 2.&3&4 Nurse prepared wrong dose (e.g. due to calculation error) 3.1. Kink of line 3.2 The pump fails, e.g. due to electrical problem or bag/syringe not installed correctly 3.3 The delivery line might not be connected to patient's central line, e.g. due to the patient pulling out the central line 3.4 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) 3&4. Inappropriate titration of dose | 1. Wrong vasopressor administered 2. Wrong dose administered (this hazard concerns single dose) 3. Sudden drop of vasopressor dose administered 4. Sudden increase of vasopressor dose administered (3 & 4 concern two consecutive doses) | |
| Late | Delay in final decision | 1. Late to get central line access 2. Late to get the vasopressor 3. Delay in administration (e.g. due to nurse workload) | Delay of administering vasopressor | |
| Early | N/A | N/A | N/A | There is ongoing clinical research about whether to deliver vasopressor earlier to increase MAP for sepsis treatment, thus it is not discussed further here. |

other tables as well. They are labelled with O meaning *omission*, C meaning *commission*, etc. The red deviation in this case is an *omission* from the RL agent.

Tables A.8 and A.9 assess the RL recommendation in the workflow as a whole, i.e. the dotted green box in Figure 4.4. The protections are ways of controlling the possible causes. Some of these reflect the practice in developing the initial model; the design requirements are those changes that were identified to produce the modified model. They correspond to safety requirements R1-R5 in Table 4.2 and provide the detail behind the recommendation “Careful RL model design required, especially causes (2&3&4) 1 to 5” in Tables A.1 and A.2.

Table A.4: Final Decision/Final Dose Decided by Doctor

| Guide word | Input Deviation | Internal Deviations | Output Deviation | Justification or Design recommendations |
|------------|--|---|---|---|
| Omission | Initial recommendation by Doctor is not produced (not considered) XO. RL agent does not produce recommendation | Doctor does not produce final decision (e.g. due to workload) | Final decision is not provided | |
| Commission | N/A (RL agent produces unnecessary recommendation is covered by Incorrect) | Doctor produces unnecessary final decision (e.g. due to miscommunication at handover) | Unnecessary final decision is provided | |
| Incorrect | Patient allergic to initial vasopressor recommended by doctor (not considered) Patient allergic to vasopressor recommended by RL agent (not credible) Wrong equivalent dose recommended by the RL agent (2) Wrong initial dose recommended by doctor (not considered) Initial recommendation by doctor has a sharp drop in dose (not considered) Recommendation by RL agent has a sharp drop in dose (3) Initial recommendation by doctor has a sharp rise in dose (not considered) Recommendation by RL agent has a sharp rise in dose (4) | 1.1 Doctor fails to check allergies and thus wrong vasopressor decided 1.2 Allergy information is incorrect 2.1 RL recommends wrong equivalent dose and doctor accepts the advice, e.g. due to automation bias 3.1. RL agent recommends sharp drop in dose and doctor accepts the advice, e.g. due to automation bias 4.1. RL agent recommends sharp rise in dose and doctor accepts the advice, e.g. due to automation bias 3&4. Doctor fails to check current dose | 1. Wrong vasopressor is decided 2. Wrong dose is decided 3. A sharp drop in dose is decided 4. A sharp rise in dose is decided | Patient allergic to vasopressor recommended by RL agent is not credible as the RL agent only recommends noradrenaline-equivalent dose The input deviation which are marked as not considered come from recommendation 1 from the doctors and this is out of the scope of the paper |
| Late | Delay of the recommendation from the RL agent Delay of the initial recommendation by the doctor (not considered) XO. RL agent does not produce recommendation | | Delay in final decision | Delay of the initial recommendation by the doctor is from recommendation 1, so is not considered further |
| Early | N/A | N/A | N/A | |

Table A.5: Recommendation by RL Agent

| Guide word | Input Deviation | Internal Deviations | Output Deviation | Justification or Design recommendations |
|------------|--|---|---|---|
| Omission | Processing of patient data does not produce a result (crash) | Failure to output results due to software fault in RL agent (e.g. errors in the display code) | XO. RL agent does not produce recommendation | |
| Commission | N/A | N/A | N/A (RL agent produces unnecessary recommendation is covered by Incorrect) | |
| Incorrect | Processing of the patient data is incorrect (2&3&4) | 2. Failure to update display with new recommendation (2&3&4) 1. Features in state space of the RL model are not sufficient to represent the patient conditions for sepsis decision making (2&3&4) 2. Reward function used for RL model is coarse (2&3&4) 3. Cost function used for RL model development is not appropriate (2&3&4) 4. Hyperparameters used for RL model development are not optimised (2&3&4) 5. Training data for RL model development is not appropriate | 1. Patient allergic to vasopressor recommended by RL agent (not credible) 2. Wrong equivalent dose recommended by RL agent 3. Recommendation by RL agent has a sharp drop in dose 4. Recommendation by RL agent has a sharp rise in dose | |
| Late | Processing of the patient data is slow | N/A | Delay of the recommendation from the RL agent | |
| Early | N/A | N/A | N/A | |

Table A.6: RL Agent processes the patient data

| Guide word | Input Deviation | Internal Deviations | Output Deviation | Justification or Design recommendations |
|------------|---|--|--|---|
| Omission | Some features are missing XI. Patient features are in wrong form, e.g. string when numeric expected | Algorithm or hardware does not have the capability to process the input data (e.g. inadequate memory allocation) | Processing of patient data does not produce a result (crash) | Use good systems engineering practice, e.g. resource analysis |
| Commission | N/A (covered by Incorrect) | N/A | N/A | |
| Incorrect | Wrong patient data provided Patient data provided incorrectly | Algorithm or hardware does not have the capability to process the input data correctly (e.g. software fault) | Processing of the patient data is incorrect | Use good software engineering practice, e.g. static analysis |
| Late | Patient features are provided late | Algorithm or hardware does not have the capability to process the input data in a timely fashion (e.g. inefficient algorithm or infinite loop or hardware limitations) | Processing of the patient data is slow | Use good software engineering practice, e.g. algorithm complexity analysis, timing analysis |
| Early | N/A | N/A | N/A | |

Table A.7: Input Patient Features

| Guide word | Input Deviation | Internal Deviations | Output Deviation | Justification or Design recommendations |
|------------|--|---|---|--|
| Omission | Some test results are missing or not available | 1. Data corruption (e.g. invalid or wrong data produced by over-writing patient's features) 2. Some features are not entered (e.g. due to staff workload) | Some features are missing | |
| Commission | Duplicate test results received (not hazardous) | N/A (covered by Incorrect) | N/A (covered by Incorrect) | |
| Incorrect | Test results for wrong patient received (2) Incorrect test results received (3) | 1&2&3. Data corruption (e.g. invalid or wrong data produced by over-writing patient's features) 1.1 Wrong form of patient features entered 2.1 Features for wrong patient entered 3.1 Wrong patient feature values entered (e.g. due to unit difference) | XI. Patient features are in wrong form, e.g. string when numeric expected 2. Wrong patient data provided 3. Patient data provided incorrectly | |
| Late | Some test results are delivered late | Some patient features are not entered on time (e.g. due to staff workload) | Patient features are provided late | Timing errors in algorithms are unlikely to be significant given response time of system |
| Early | N/A | N/A | N/A | |

Table A.8: Interface between RL agent and Clinical care, Part 1

| Guide word | Deviation | Possible Causes | Linked hazard | Protections | Design requirements |
|------------|--|---|--|---|--|
| Omission | RL agent does not produce recommendation | <ol style="list-style-type: none"> 1. Failure to output results due to software fault in RL agent (e.g. errors in the display code) 2. Algorithm or hardware does not have the capability to process the input data (e.g. inadequate memory allocation) 3. Data corruption (e.g. invalid or wrong data produced by over-writing patient's features) 4. Some features are not entered (e.g. due to staff workload) 5. Some test results are missing or not available 6. Wrong form of patient features entered | No vasopressor administered | | Apply good practice from systems engineering, software engineering and user interface design (details out of the scope of the paper). |
| Commission | N/A (RL agent produces unnecessary recommendation is covered by incorrect) | N/A | N/A | | |
| Incorrect | 1. Patient allergic to vasopressor recommended by RL agent (not credible) | | 1. Wrong vasopressor administered | | |
| | <ol style="list-style-type: none"> 2. Wrong equivalent dose recommended by RL agent 3. RL agent recommends a sudden drop in dose 4. RL agent recommends a sudden increase in dose | <ol style="list-style-type: none"> 2.1 Failure to update display with new recommendation 2.2 Algorithm or hardware does not have the capability to process the input data correctly (e.g. software fault) <p>(2&3&4) 1. Features in state space of the RL model are not sufficient to represent the patient conditions for sepsis decision making (2&3&4) 2. Reward function used for RL model is coarse (2&3&4) 3. Cost function used for RL model development is not appropriate (2&3&4) 4. Hyperparameters used for RL model development are not optimised (2&3&4) 5. Training data for RL model development is not appropriate</p> <p>(2&3&4) 6. Data corruption (e.g. invalid or wrong data produced by over-writing patient's features) (2&3&4) 7. Features for wrong patient entered (2&3&4) 8. Wrong patient feature values entered (e.g. due to unit difference)</p> | <ol style="list-style-type: none"> 2. Wrong dose administered 3. Sudden drop of vasopressor dose administered 4. Sudden increase of vasopressor dose administered | <ol style="list-style-type: none"> 1.1 The 47 included features, e.g. demographics, lab values, were chosen to represent the most important parameters that clinicians would consider when deciding treatment for patients. 1.2 Add one extra feature to enable the RL agent to take account of the relative vasopressor dose change. 2. The terminal reward is based on patients' 90-day-mortality and an intermediate reward is also added based on SOFA score and Arterial Lactate level. 3.1 Use a standard double DQN loss function plus one regularisation term in cost function to penalise output Q-values when it is outside of the allowed thresholds. 3.2 Add a second regularization term to the cost function to penalise sudden dosage change while learning the optimal policy. 4. Multiple values are tried and the best value is chosen based on the validation error. | <ol style="list-style-type: none"> 1. Feature representation in the state space shall be sufficient to allow the control of sudden changes in recommended dose 2. An appropriate reward function shall be defined to allow the recognition of desired clinical outcome 3. An appropriate cost function shall be defined to penalise hazardous behaviours 4. Hyperparameters shall be optimised based on the validation dataset 5. Patient cohort shall be defined using recognised criteria, i.e. sepsis-3 <p>For causes (2&3&4) 6-10 Apply good practice from systems engineering, software engineering and user interface design (details out of the scope of the paper).</p> |

Table A.9: Interface between RL agent and Clinical care, Part 2

| | | | | | |
|-------|---|--|------------------------------------|--|--|
| | | (2&3&4) 9. Test results for wrong patient received (2&3&4) 10. Incorrect test results received | | 5.1 Real world data (MIMIC III) is used and sepsis 3 definition is used to determine patient cohort 5.2 Outliers are corrected and missing data is addressed, e.g. using sample-and-hold approach. 1 to 5: Hold-out test data was used to evaluate the appropriateness of the learnt policy in comparison with the clinician policy. | |
| Late | Delay of the recommendation from the RL agent | 1. Algorithm or hardware does not have the capability to process the input data or in a timely fashion (e.g. inefficient algorithm or infinite loop or hardware limitations) 2. Some patient features are not entered at all or not entered on time (e.g. due to staff workload) 3. Some test results are delivered late, missing or not available 4. Failure to output results due to software fault in RL agent (e.g. errors in the display code) 5. Data corruption (e.g. invalid or wrong data produced by over-writing patient's features) 6. Wrong form of patient features entered | Delay of administering vasopressor | | |
| Early | N/A | N/A | N/A | | |

Appendix B

Feature correlation matrix for the RL model in case study 1

This appendix presents the 47 features used to represent the state space in the RL model in case study 1 presented in Chapter 4, see Table B.1. A feature correlation matrix is also presented. As the matrix is big, it is split into three parts in Figures B.1 to B.3.

Table B.1: List of features used in the RL model

| Feature Abbreviation | Feature Description |
|-----------------------------|-------------------------------|
| Gender | Gender |
| Age | Age |
| re_admission | Readmission to intensive care |
| mechvent | Mechanical ventilation |
| Weight_kg | Weight |
| GCS | Glasgow coma scale |
| SysBP | Systolic blood pressure |
| MeanBP | Mean blood pressure |
| DiaBP | Diastolic blood pressure |
| HR | Heart rate |
| RR | Respiratory rate |
| Temp_C | Temperature |
| FiO2_1 | FiO2 |
| Potassium | Potassium |

Appendix B: Feature correlation matrix for the RL model in case study 1

| | |
|-------------------|--|
| Sodium | sodium |
| Chloride | chloride |
| Glucose | Glucose |
| Magnesium | Magnesium |
| Calcium | Calcium |
| Hb | Hemoglobin |
| WBC_count' | White blood cells count |
| Platelets_count | Platelets count |
| PTT | PTT |
| PT | PT |
| Arterial_pH | Arterial PH |
| paO2 | PaO2 |
| paCO2 | PaCO2 |
| Arterial_BE | Arterial base excess |
| HCO3 | Bicarbonate |
| Arterial_lactate | Arterial lactate |
| SOFA | SOFA |
| SIRS | SIRS |
| Shock_Index | Shock Index |
| PaO2_FiO2 | PaO2/FiO2 ratio |
| cumulated_balance | Cumulated fluid balance since admission (includes preadmission data when available) |
| SpO2 | SpO2 |
| BUN | BUN |
| Creatinine | Creatinine |
| SGOT | SGOT |
| SGPT | SGPT |
| Total_bili | total bilirubin |
| INR | INR |
| max_dose_vaso | Maximum dose of vasopressor over 4h |
| input_total | total input since hospital (when pre-ICU data available) or ICU admission |

| | |
|----------------|--|
| input_4hourly | Current IV fluid intake over 4h |
| output_total | total input since hospital (when pre-ICU data available) or ICU admission |
| output_4hourly | Urine output over 4h |

| | gender | mechvent | re_admission | age | Weight_kg | GCS | HR | SysBP | MeanBP | DiaBP | RR | Temp_C | FIO2_1 | Potassium | Sodium | Chloride | Glucose |
|-------------------|--------|----------|--------------|-------|-----------|-------|-------|-------|--------|-------|-------|--------|--------|-----------|--------|----------|---------|
| gender | 1.00 | -0.04 | 0.00 | 0.08 | -0.16 | 0.04 | 0.02 | 0.00 | -0.04 | -0.08 | 0.01 | -0.02 | -0.04 | -0.07 | 0.01 | -0.01 | 0.00 |
| mechvent | -0.04 | 1.00 | -0.06 | -0.06 | 0.07 | -0.48 | 0.02 | -0.03 | 0.01 | 0.00 | -0.04 | 0.07 | 0.17 | -0.00 | 0.10 | 0.11 | 0.02 |
| re_admission | 0.00 | -0.06 | 1.00 | 0.02 | -0.03 | 0.06 | 0.00 | -0.05 | -0.06 | -0.04 | -0.00 | -0.04 | -0.02 | 0.05 | -0.02 | -0.08 | 0.03 |
| age | 0.08 | -0.06 | 0.02 | 1.00 | -0.11 | 0.04 | -0.19 | -0.01 | -0.19 | -0.25 | 0.03 | -0.06 | 0.01 | 0.05 | 0.05 | 0.02 | 0.04 |
| Weight_kg | -0.16 | 0.07 | -0.03 | -0.11 | 1.00 | -0.04 | 0.01 | 0.02 | 0.02 | 0.02 | -0.00 | 0.03 | 0.04 | 0.07 | -0.01 | -0.05 | 0.07 |
| GCS | 0.04 | -0.48 | 0.06 | 0.04 | -0.04 | 1.00 | -0.04 | 0.01 | -0.01 | -0.02 | -0.01 | -0.05 | -0.17 | -0.01 | -0.07 | -0.09 | -0.03 |
| HR | 0.02 | 0.02 | 0.00 | -0.19 | 0.01 | -0.04 | 1.00 | -0.03 | 0.12 | 0.19 | 0.28 | 0.09 | 0.08 | -0.02 | -0.02 | -0.01 | 0.02 |
| SysBP | 0.00 | -0.03 | -0.05 | -0.01 | 0.02 | 0.01 | -0.03 | 1.00 | 0.69 | 0.49 | 0.04 | 0.02 | -0.06 | -0.06 | 0.10 | 0.01 | 0.08 |
| MeanBP | -0.04 | 0.01 | -0.06 | -0.19 | 0.02 | -0.01 | 0.12 | 0.69 | 1.00 | 0.75 | 0.04 | 0.02 | -0.05 | -0.08 | 0.10 | 0.04 | 0.04 |
| DiaBP | -0.08 | 0.00 | -0.04 | -0.25 | 0.02 | -0.02 | 0.19 | 0.49 | 0.75 | 1.00 | 0.04 | 0.01 | -0.05 | -0.07 | 0.05 | 0.02 | -0.00 |
| RR | 0.01 | -0.04 | -0.00 | 0.03 | -0.00 | -0.01 | 0.28 | 0.04 | 0.04 | 0.04 | 1.00 | 0.06 | 0.09 | -0.04 | 0.04 | 0.00 | 0.04 |
| Temp_C | -0.02 | 0.07 | -0.04 | -0.06 | 0.03 | -0.05 | 0.09 | 0.02 | 0.02 | 0.01 | 0.06 | 1.00 | 0.00 | -0.03 | 0.03 | 0.01 | -0.01 |
| FIO2_1 | -0.04 | 0.17 | -0.02 | 0.01 | 0.04 | -0.17 | 0.08 | -0.06 | -0.05 | -0.05 | 0.09 | 0.00 | 1.00 | 0.08 | -0.02 | -0.02 | 0.05 |
| Potassium | -0.07 | -0.00 | 0.05 | 0.05 | 0.07 | -0.01 | -0.02 | -0.06 | -0.08 | -0.07 | -0.04 | -0.03 | 0.08 | 1.00 | -0.20 | -0.11 | 0.08 |
| Sodium | 0.01 | 0.10 | -0.02 | 0.05 | -0.01 | -0.07 | -0.02 | 0.10 | 0.10 | 0.05 | 0.04 | 0.03 | -0.02 | -0.20 | 1.00 | 0.66 | -0.02 |
| Chloride | -0.01 | 0.11 | -0.08 | 0.02 | -0.05 | -0.09 | -0.01 | 0.01 | 0.04 | 0.02 | 0.00 | 0.01 | -0.02 | -0.11 | 0.66 | 1.00 | -0.08 |
| Glucose | 0.00 | 0.02 | 0.03 | 0.04 | 0.07 | -0.03 | 0.02 | 0.08 | 0.04 | -0.00 | 0.04 | -0.01 | 0.05 | 0.08 | -0.02 | -0.08 | 1.00 |
| Magnesium | -0.04 | 0.06 | -0.02 | 0.09 | 0.05 | -0.04 | -0.08 | -0.00 | -0.04 | -0.04 | 0.00 | -0.02 | 0.04 | 0.19 | 0.06 | -0.00 | 0.03 |
| Calcium | 0.03 | -0.05 | 0.03 | -0.01 | 0.02 | 0.03 | -0.06 | 0.10 | 0.07 | 0.05 | -0.03 | -0.02 | -0.00 | 0.12 | 0.05 | -0.18 | 0.03 |
| Hb | -0.07 | -0.06 | -0.12 | -0.02 | 0.02 | 0.01 | -0.01 | 0.04 | 0.11 | 0.12 | -0.01 | -0.00 | 0.03 | 0.02 | 0.01 | -0.07 | 0.02 |
| WBC_count | 0.02 | 0.07 | -0.03 | 0.03 | 0.01 | -0.06 | 0.08 | -0.03 | -0.04 | -0.04 | 0.07 | 0.00 | 0.06 | 0.05 | -0.02 | -0.01 | 0.04 |
| Platelets_count | 0.06 | -0.02 | 0.04 | -0.01 | -0.03 | 0.02 | 0.05 | 0.05 | 0.03 | 0.01 | 0.07 | 0.02 | 0.00 | 0.05 | -0.01 | -0.12 | 0.01 |
| PTT | -0.00 | 0.02 | 0.03 | 0.05 | -0.00 | -0.03 | 0.01 | -0.08 | -0.08 | -0.05 | 0.03 | -0.02 | 0.03 | 0.02 | -0.05 | -0.03 | 0.02 |
| PT | -0.02 | -0.02 | 0.08 | 0.04 | 0.01 | -0.00 | 0.03 | -0.08 | -0.08 | -0.04 | 0.03 | -0.03 | 0.03 | 0.03 | -0.02 | -0.04 | 0.01 |
| Arterial_ph | -0.01 | -0.02 | -0.04 | -0.01 | -0.03 | 0.04 | -0.02 | 0.08 | 0.09 | 0.06 | 0.03 | 0.04 | -0.09 | -0.29 | 0.06 | -0.03 | -0.09 |
| paO2 | -0.00 | 0.02 | -0.03 | -0.00 | -0.03 | -0.05 | -0.03 | 0.01 | 0.02 | 0.01 | -0.08 | -0.01 | 0.02 | 0.04 | -0.02 | 0.03 | 0.01 |
| paCO2 | 0.01 | 0.02 | 0.04 | 0.02 | 0.05 | 0.02 | -0.02 | 0.01 | -0.02 | -0.03 | -0.04 | -0.01 | 0.03 | 0.11 | 0.03 | -0.15 | 0.04 |
| Arterial_BE | 0.00 | 0.01 | 0.00 | 0.01 | 0.02 | 0.05 | -0.04 | 0.08 | 0.07 | 0.04 | -0.01 | 0.02 | -0.06 | -0.18 | 0.08 | -0.16 | -0.05 |
| HCO3 | -0.00 | -0.00 | 0.03 | 0.03 | 0.05 | 0.06 | -0.06 | 0.07 | 0.04 | 0.01 | -0.03 | 0.02 | -0.01 | -0.11 | 0.12 | -0.37 | -0.03 |
| Arterial_lactate | -0.00 | 0.02 | 0.00 | -0.01 | 0.01 | -0.09 | 0.06 | -0.04 | -0.04 | -0.01 | 0.04 | -0.02 | 0.08 | 0.08 | -0.02 | -0.04 | 0.13 |
| SOFA | -0.04 | 0.25 | 0.08 | 0.02 | 0.05 | -0.47 | 0.03 | -0.21 | -0.24 | -0.16 | 0.02 | -0.00 | 0.30 | 0.12 | -0.06 | -0.04 | 0.02 |
| SIRS | 0.01 | 0.04 | -0.01 | -0.08 | 0.00 | -0.08 | 0.55 | -0.02 | 0.04 | 0.08 | 0.49 | 0.05 | 0.09 | -0.01 | -0.00 | 0.01 | 0.06 |
| Shock_Index | 0.02 | 0.04 | 0.04 | -0.13 | -0.01 | -0.04 | 0.76 | -0.64 | -0.34 | -0.18 | 0.18 | 0.06 | 0.10 | 0.03 | -0.08 | -0.02 | -0.03 |
| PaO2_FIO2 | 0.02 | -0.14 | -0.01 | -0.01 | -0.05 | 0.09 | -0.05 | 0.03 | 0.03 | 0.04 | -0.09 | -0.01 | -0.44 | -0.01 | -0.02 | 0.02 | -0.02 |
| cumulated_balance | -0.01 | 0.07 | 0.00 | -0.03 | 0.02 | -0.03 | 0.04 | -0.06 | -0.03 | -0.04 | -0.07 | -0.01 | 0.04 | 0.02 | -0.04 | 0.07 | -0.02 |
| SpO2 | 0.01 | 0.17 | 0.01 | -0.04 | -0.06 | -0.10 | -0.09 | 0.05 | 0.07 | 0.04 | -0.15 | -0.00 | -0.12 | -0.04 | 0.02 | 0.10 | -0.03 |
| BUN | -0.10 | 0.06 | 0.13 | 0.27 | 0.07 | -0.06 | -0.09 | 0.01 | -0.08 | -0.11 | 0.03 | -0.05 | 0.04 | 0.27 | 0.08 | -0.02 | 0.14 |
| Creatinine | -0.13 | -0.03 | 0.12 | 0.08 | 0.09 | -0.00 | -0.07 | -0.00 | -0.07 | -0.07 | -0.03 | -0.04 | 0.01 | 0.27 | -0.02 | -0.10 | 0.07 |
| SGOT | -0.03 | 0.07 | -0.04 | -0.05 | 0.02 | -0.08 | 0.03 | -0.01 | 0.02 | 0.03 | 0.01 | -0.00 | 0.04 | 0.04 | -0.02 | -0.02 | 0.02 |
| SGPT | -0.03 | 0.05 | -0.04 | -0.05 | 0.03 | -0.06 | 0.02 | 0.02 | 0.05 | 0.05 | 0.01 | -0.00 | 0.02 | 0.01 | -0.00 | -0.02 | 0.04 |
| Total_bili | -0.04 | 0.04 | -0.00 | -0.07 | 0.03 | -0.05 | 0.03 | -0.02 | -0.00 | 0.01 | 0.01 | -0.00 | 0.01 | -0.04 | -0.03 | -0.02 | -0.04 |
| INR | -0.03 | -0.01 | 0.09 | 0.04 | 0.02 | -0.02 | 0.04 | -0.10 | -0.10 | -0.05 | 0.04 | -0.02 | 0.03 | 0.02 | -0.02 | -0.03 | 0.01 |
| input_total | -0.05 | 0.25 | -0.13 | -0.09 | 0.03 | -0.11 | 0.04 | -0.01 | 0.02 | -0.00 | -0.01 | 0.04 | -0.01 | -0.11 | 0.06 | 0.18 | -0.04 |
| input_4hourly | -0.02 | 0.34 | -0.05 | -0.08 | 0.01 | -0.15 | 0.07 | -0.04 | -0.01 | -0.01 | -0.05 | 0.03 | 0.08 | -0.07 | 0.03 | 0.15 | 0.00 |
| output_total | -0.02 | 0.18 | -0.19 | -0.04 | 0.04 | -0.04 | -0.00 | 0.07 | 0.08 | 0.04 | 0.05 | 0.04 | -0.03 | -0.17 | 0.11 | 0.14 | -0.01 |
| output_4hourly | -0.02 | 0.20 | -0.14 | -0.02 | 0.04 | -0.03 | 0.01 | 0.06 | 0.08 | 0.04 | 0.02 | 0.03 | 0.04 | -0.12 | 0.09 | 0.16 | 0.01 |

Figure B.1: Feature correlation matrix for case study concerning sepsis treatment, Part 1

Appendix B: Feature correlation matrix for the RL model in case study 1

| | Magnesium | Calcium | Hb | WBC_count | Platelets_count | PTT | PT | Arterial_pH | paO2 | paCO2 | Arterial_BE | HCO3 | Arterial_lactate | SOFA | SIRS | Shock_Index |
|-------------------|-----------|---------|-------|-----------|-----------------|-------|-------|-------------|-------|-------|-------------|-------|------------------|-------|-------|-------------|
| gender | -0.04 | 0.03 | -0.07 | 0.02 | 0.06 | -0.00 | -0.02 | -0.01 | -0.00 | 0.01 | 0.00 | -0.00 | -0.04 | 0.01 | 0.02 | 0.02 |
| mechvent | 0.06 | -0.05 | -0.06 | 0.07 | -0.02 | 0.02 | -0.02 | -0.02 | 0.02 | 0.02 | 0.01 | -0.00 | 0.02 | 0.25 | 0.04 | 0.04 |
| re_admission | -0.02 | 0.03 | -0.12 | -0.03 | 0.04 | 0.03 | 0.08 | -0.04 | -0.03 | 0.04 | 0.00 | 0.03 | 0.00 | 0.08 | -0.01 | 0.04 |
| age | 0.09 | -0.01 | -0.02 | 0.03 | -0.01 | 0.05 | 0.04 | -0.01 | -0.00 | 0.02 | 0.01 | 0.03 | -0.01 | 0.02 | -0.08 | -0.13 |
| Weight_kg | 0.05 | 0.02 | 0.02 | 0.01 | -0.03 | -0.00 | 0.01 | -0.03 | -0.03 | 0.05 | 0.02 | 0.05 | 0.01 | 0.05 | 0.00 | -0.01 |
| GCS | -0.04 | 0.03 | 0.01 | -0.06 | 0.02 | -0.03 | -0.00 | 0.04 | -0.05 | 0.02 | 0.05 | 0.06 | -0.09 | -0.47 | -0.08 | -0.04 |
| HR | -0.08 | -0.06 | -0.01 | 0.08 | 0.05 | 0.01 | 0.03 | -0.02 | -0.03 | -0.02 | -0.04 | -0.06 | 0.06 | 0.03 | 0.55 | 0.76 |
| SysBP | -0.00 | 0.10 | 0.04 | -0.03 | 0.05 | -0.08 | -0.08 | 0.08 | 0.01 | 0.01 | 0.08 | 0.07 | -0.04 | -0.21 | -0.02 | -0.64 |
| MeanBP | -0.04 | 0.07 | 0.11 | -0.04 | 0.03 | -0.08 | -0.08 | 0.09 | 0.02 | -0.02 | 0.07 | 0.04 | -0.04 | -0.24 | 0.04 | -0.34 |
| DiaBP | -0.04 | 0.05 | 0.12 | -0.04 | 0.01 | -0.05 | -0.04 | 0.06 | 0.01 | -0.03 | 0.04 | 0.01 | -0.01 | -0.16 | 0.08 | -0.18 |
| RR | 0.00 | -0.03 | -0.01 | 0.07 | 0.07 | 0.03 | 0.03 | 0.03 | -0.08 | -0.04 | -0.01 | -0.03 | 0.04 | 0.02 | 0.49 | 0.18 |
| Temp_C | -0.02 | -0.02 | -0.00 | 0.00 | 0.02 | -0.02 | -0.03 | 0.04 | -0.01 | 0.02 | 0.02 | 0.02 | -0.02 | -0.00 | 0.05 | 0.06 |
| FiO2_1 | 0.04 | -0.00 | 0.03 | 0.06 | 0.00 | 0.03 | 0.03 | -0.09 | 0.02 | 0.03 | -0.06 | -0.01 | 0.08 | 0.30 | 0.09 | 0.10 |
| Potassium | 0.19 | 0.12 | 0.02 | 0.05 | 0.05 | 0.02 | 0.03 | -0.29 | 0.04 | 0.11 | -0.18 | -0.11 | 0.08 | 0.12 | -0.01 | 0.03 |
| Sodium | 0.06 | 0.05 | 0.01 | -0.02 | -0.01 | -0.05 | -0.02 | 0.06 | -0.02 | 0.03 | 0.08 | 0.12 | -0.02 | -0.06 | 0.00 | -0.08 |
| Chloride | -0.00 | -0.18 | -0.07 | -0.01 | -0.12 | -0.03 | -0.04 | -0.03 | 0.03 | -0.15 | -0.16 | -0.37 | -0.04 | -0.04 | 0.01 | -0.02 |
| Glucose | 0.03 | 0.03 | 0.02 | 0.04 | 0.01 | 0.02 | 0.01 | -0.09 | 0.01 | 0.04 | -0.05 | -0.03 | 0.13 | 0.02 | 0.06 | -0.03 |
| Magnesium | 1.00 | 0.14 | -0.00 | 0.05 | 0.00 | 0.04 | 0.03 | -0.03 | -0.01 | 0.01 | -0.02 | 0.01 | 0.02 | 0.08 | -0.01 | -0.06 |
| Calcium | 0.14 | 1.00 | 0.13 | -0.02 | 0.07 | -0.01 | -0.00 | 0.03 | 0.02 | 0.07 | 0.08 | 0.18 | 0.03 | -0.01 | -0.06 | -0.11 |
| Hb | -0.00 | 0.13 | 1.00 | 0.03 | -0.03 | -0.04 | -0.04 | -0.04 | 0.02 | 0.02 | -0.02 | 0.03 | 0.03 | -0.08 | 0.00 | -0.03 |
| WBC_count | 0.05 | -0.02 | 0.03 | 1.00 | 0.22 | 0.02 | 0.03 | -0.04 | -0.02 | -0.02 | -0.05 | -0.08 | 0.03 | 0.06 | 0.29 | 0.08 |
| Platelets_count | 0.00 | 0.07 | -0.03 | 0.22 | 1.00 | -0.04 | -0.04 | 0.02 | -0.01 | 0.05 | 0.06 | 0.11 | -0.06 | -0.23 | 0.11 | 0.00 |
| PTT | 0.04 | -0.01 | -0.04 | 0.02 | -0.04 | 1.00 | 0.25 | -0.04 | -0.00 | -0.03 | -0.06 | -0.07 | 0.09 | 0.12 | 0.03 | 0.06 |
| PT | 0.03 | -0.00 | -0.04 | 0.03 | -0.04 | 0.25 | 1.00 | -0.04 | -0.02 | -0.02 | -0.06 | -0.06 | 0.10 | 0.15 | 0.05 | 0.08 |
| Arterial_pH | -0.03 | 0.03 | -0.04 | -0.04 | 0.02 | -0.04 | -0.04 | 1.00 | 0.02 | -0.36 | 0.57 | 0.18 | -0.19 | -0.15 | 0.00 | -0.07 |
| paO2 | -0.01 | 0.02 | 0.02 | -0.02 | -0.01 | -0.00 | -0.02 | 0.02 | 1.00 | -0.09 | -0.04 | -0.03 | 0.02 | -0.21 | -0.05 | -0.03 |
| paCO2 | 0.01 | 0.07 | 0.02 | -0.02 | 0.05 | -0.03 | -0.02 | -0.36 | -0.09 | 1.00 | 0.48 | 0.31 | -0.09 | -0.01 | -0.12 | -0.02 |
| Arterial_BE | -0.02 | 0.08 | -0.02 | -0.05 | 0.06 | -0.06 | -0.06 | 0.57 | -0.04 | 0.48 | 1.00 | 0.43 | -0.24 | -0.15 | -0.10 | -0.09 |
| HCO3 | 0.01 | 0.18 | 0.03 | -0.08 | 0.11 | -0.07 | -0.06 | 0.18 | -0.03 | 0.31 | 0.43 | 1.00 | -0.14 | -0.17 | -0.12 | -0.09 |
| Arterial_lactate | 0.02 | 0.03 | 0.03 | 0.03 | -0.06 | 0.09 | 0.10 | -0.19 | 0.02 | -0.09 | -0.24 | -0.14 | 1.00 | 0.17 | 0.07 | 0.08 |
| SOFA | 0.08 | -0.01 | -0.08 | 0.06 | -0.23 | 0.12 | 0.15 | -0.15 | -0.21 | -0.01 | -0.15 | -0.17 | 0.17 | 1.00 | 0.09 | 0.18 |
| SIRS | -0.01 | -0.06 | 0.00 | 0.29 | 0.11 | 0.03 | 0.05 | 0.00 | -0.05 | -0.12 | -0.10 | -0.12 | 0.07 | 0.09 | 1.00 | 0.42 |
| Shock_Index | -0.06 | -0.11 | -0.03 | 0.08 | 0.00 | 0.06 | 0.08 | -0.07 | -0.03 | -0.02 | -0.09 | -0.09 | 0.08 | 0.18 | 0.42 | 1.00 |
| PaO2_FiO2 | -0.03 | 0.02 | 0.02 | -0.05 | -0.01 | -0.01 | -0.03 | 0.04 | 0.79 | -0.09 | -0.01 | -0.03 | -0.00 | -0.31 | -0.08 | -0.06 |
| cumulated_balance | -0.02 | -0.07 | -0.01 | 0.03 | -0.15 | 0.02 | 0.05 | -0.06 | 0.02 | -0.06 | -0.10 | -0.14 | 0.06 | 0.17 | 0.01 | 0.07 |
| SpO2 | -0.03 | -0.01 | -0.10 | -0.02 | 0.01 | -0.02 | -0.04 | 0.07 | 0.12 | -0.07 | 0.01 | -0.05 | -0.07 | -0.03 | -0.09 | -0.11 |
| BUN | 0.29 | 0.08 | -0.12 | 0.09 | -0.10 | 0.08 | 0.12 | -0.12 | -0.03 | 0.01 | -0.10 | -0.14 | 0.06 | 0.30 | 0.04 | -0.05 |
| Creatinine | 0.19 | 0.08 | -0.07 | 0.04 | -0.13 | 0.09 | 0.12 | -0.16 | -0.02 | -0.03 | -0.17 | -0.22 | 0.07 | 0.33 | -0.01 | -0.04 |
| SGOT | 0.04 | 0.01 | 0.03 | 0.02 | -0.10 | 0.05 | 0.08 | -0.03 | -0.01 | -0.04 | -0.06 | -0.07 | 0.16 | 0.22 | 0.04 | 0.03 |
| SGPT | 0.03 | 0.01 | 0.05 | 0.01 | -0.07 | 0.02 | 0.04 | -0.01 | -0.01 | -0.02 | -0.03 | -0.03 | 0.12 | 0.14 | 0.03 | 0.00 |
| Total_bili | 0.04 | 0.05 | -0.01 | 0.01 | -0.15 | 0.06 | 0.12 | 0.03 | -0.02 | -0.07 | -0.03 | -0.06 | 0.11 | 0.41 | 0.04 | 0.04 |
| INR | 0.03 | -0.01 | -0.06 | 0.05 | -0.06 | 0.30 | 0.89 | -0.04 | -0.03 | -0.04 | -0.07 | -0.09 | 0.11 | 0.19 | 0.07 | 0.10 |
| input_total | 0.01 | -0.12 | -0.11 | 0.04 | -0.05 | -0.01 | -0.03 | 0.05 | -0.02 | -0.08 | -0.02 | -0.08 | -0.00 | 0.00 | 0.03 | 0.04 |
| input_4hourly | -0.04 | -0.14 | -0.07 | 0.06 | -0.05 | -0.01 | 0.01 | -0.01 | -0.03 | -0.05 | -0.05 | -0.12 | 0.02 | 0.04 | 0.06 | 0.09 |
| output_total | 0.03 | -0.06 | -0.07 | -0.00 | 0.01 | -0.03 | -0.07 | 0.11 | -0.10 | -0.01 | 0.08 | 0.07 | -0.06 | -0.25 | 0.01 | -0.05 |
| output_4hourly | 0.01 | -0.09 | -0.02 | 0.01 | 0.00 | -0.05 | -0.06 | 0.08 | -0.05 | -0.01 | 0.06 | 0.03 | -0.07 | -0.38 | 0.01 | -0.04 |

Figure B.2: Feature correlation matrix for case study concerning sepsis treatment, Part 2

| | PaO2_FiO2 | cumulated_balance | SpO2 | BUN | Creatinine | SGOT | SGPT | Total_bili | INR | input_total | input_4hourly | output_total | output_4hourly |
|-------------------|-----------|-------------------|-------|-------|------------|-------|-------|------------|-------|-------------|---------------|--------------|----------------|
| gender | 0.02 | -0.01 | 0.01 | -0.10 | -0.13 | -0.03 | -0.03 | -0.04 | -0.03 | -0.05 | -0.02 | -0.02 | -0.02 |
| mechvent | -0.14 | 0.07 | 0.17 | 0.06 | -0.03 | 0.07 | 0.05 | 0.04 | -0.01 | 0.25 | 0.34 | 0.18 | 0.20 |
| re_admission | -0.01 | 0.00 | 0.01 | 0.13 | 0.12 | -0.04 | -0.04 | -0.00 | 0.09 | -0.13 | -0.05 | -0.19 | -0.14 |
| age | -0.01 | -0.03 | -0.04 | 0.27 | 0.08 | -0.05 | -0.05 | -0.07 | 0.04 | -0.09 | -0.08 | -0.04 | -0.02 |
| Weight_kg | -0.05 | 0.02 | -0.06 | 0.07 | 0.09 | 0.02 | 0.03 | 0.03 | 0.02 | 0.03 | 0.01 | 0.04 | 0.04 |
| GCS | 0.09 | -0.03 | -0.10 | -0.06 | -0.00 | -0.08 | -0.06 | -0.05 | -0.02 | -0.11 | -0.15 | -0.04 | -0.03 |
| HR | -0.05 | 0.04 | -0.09 | -0.09 | -0.07 | 0.03 | 0.02 | 0.03 | 0.04 | 0.04 | 0.07 | -0.00 | 0.01 |
| SysBP | 0.03 | -0.06 | 0.05 | 0.01 | -0.00 | -0.01 | 0.02 | -0.02 | -0.10 | -0.01 | -0.04 | 0.07 | 0.06 |
| MeanBP | 0.03 | -0.03 | 0.07 | -0.08 | -0.07 | 0.02 | 0.05 | -0.00 | -0.10 | 0.02 | -0.01 | 0.08 | 0.08 |
| DiaBP | 0.04 | -0.04 | 0.04 | -0.11 | -0.07 | 0.03 | 0.05 | 0.01 | -0.05 | -0.00 | -0.01 | 0.04 | 0.04 |
| RR | -0.09 | -0.07 | -0.15 | 0.03 | -0.03 | 0.01 | 0.01 | 0.01 | 0.04 | -0.01 | -0.05 | 0.05 | 0.02 |
| Temp_C | -0.01 | -0.01 | -0.00 | -0.05 | -0.04 | -0.00 | -0.00 | -0.00 | -0.02 | 0.04 | 0.03 | 0.04 | 0.03 |
| FiO2_1 | -0.44 | 0.04 | -0.12 | 0.04 | 0.01 | 0.04 | 0.02 | 0.01 | 0.03 | -0.01 | 0.08 | -0.03 | 0.04 |
| Potassium | -0.01 | 0.02 | -0.04 | 0.27 | 0.27 | 0.04 | 0.01 | -0.04 | 0.02 | -0.11 | -0.07 | -0.17 | -0.12 |
| Sodium | -0.02 | -0.04 | 0.02 | 0.08 | -0.02 | -0.02 | -0.00 | -0.03 | -0.02 | 0.06 | 0.03 | 0.11 | 0.09 |
| Chloride | 0.02 | 0.07 | 0.10 | -0.02 | -0.10 | -0.02 | -0.02 | -0.02 | -0.03 | 0.18 | 0.15 | 0.14 | 0.16 |
| Glucose | -0.02 | -0.02 | -0.03 | 0.14 | 0.07 | 0.02 | 0.04 | -0.04 | 0.01 | -0.04 | 0.00 | -0.01 | 0.01 |
| Magnesium | -0.03 | -0.02 | -0.03 | 0.29 | 0.19 | 0.04 | 0.03 | 0.04 | 0.03 | 0.01 | -0.04 | 0.03 | 0.01 |
| Calcium | 0.02 | -0.07 | -0.01 | 0.08 | 0.08 | 0.01 | 0.01 | 0.05 | -0.01 | -0.12 | -0.14 | -0.06 | -0.09 |
| Hb | 0.02 | -0.01 | -0.10 | -0.12 | -0.07 | 0.03 | 0.05 | -0.01 | -0.06 | -0.11 | -0.07 | -0.07 | -0.02 |
| WBC_count | -0.05 | 0.03 | -0.02 | 0.09 | 0.04 | 0.02 | 0.01 | 0.01 | 0.05 | 0.04 | 0.06 | -0.00 | 0.01 |
| Platelets_count | -0.01 | -0.15 | 0.01 | -0.10 | -0.13 | -0.10 | -0.07 | -0.15 | -0.06 | -0.05 | -0.05 | 0.01 | 0.00 |
| PTT | -0.01 | 0.02 | -0.02 | 0.08 | 0.09 | 0.05 | 0.02 | 0.06 | 0.30 | -0.01 | -0.01 | -0.03 | -0.05 |
| PT | -0.03 | 0.05 | -0.04 | 0.12 | 0.12 | 0.08 | 0.04 | 0.12 | 0.89 | -0.03 | 0.01 | -0.07 | -0.06 |
| Arterial_pH | 0.04 | -0.06 | 0.07 | -0.12 | -0.16 | -0.03 | -0.01 | 0.03 | -0.04 | 0.05 | -0.01 | 0.11 | 0.08 |
| paO2 | 0.79 | 0.02 | 0.12 | -0.03 | -0.02 | -0.01 | -0.01 | -0.02 | -0.03 | -0.02 | -0.03 | -0.10 | -0.05 |
| paCO2 | -0.09 | -0.06 | -0.07 | 0.01 | -0.03 | -0.04 | -0.02 | -0.07 | -0.04 | -0.08 | -0.05 | -0.01 | -0.01 |
| Arterial_BE | -0.01 | -0.10 | 0.01 | -0.10 | -0.17 | -0.06 | -0.03 | -0.03 | -0.07 | -0.02 | -0.05 | 0.08 | 0.06 |
| HCO3 | -0.03 | -0.14 | -0.05 | -0.14 | -0.22 | -0.07 | -0.03 | -0.06 | -0.09 | -0.08 | -0.12 | 0.07 | 0.03 |
| Arterial_lactate | -0.00 | 0.06 | -0.07 | 0.06 | 0.07 | 0.16 | 0.12 | 0.11 | 0.11 | -0.00 | 0.02 | -0.06 | -0.07 |
| SOFA | -0.31 | 0.17 | -0.03 | 0.30 | 0.33 | 0.22 | 0.14 | 0.41 | 0.19 | 0.00 | 0.04 | -0.25 | -0.38 |
| SIRS | -0.08 | 0.01 | -0.09 | 0.04 | -0.01 | 0.04 | 0.03 | 0.04 | 0.07 | 0.03 | 0.06 | 0.01 | 0.01 |
| Shock_Index | -0.06 | 0.07 | -0.11 | -0.05 | -0.04 | 0.03 | 0.00 | 0.04 | 0.10 | 0.04 | 0.09 | -0.05 | -0.04 |
| PaO2_FiO2 | 1.00 | -0.01 | 0.10 | -0.05 | -0.01 | -0.03 | -0.02 | -0.02 | -0.03 | -0.04 | -0.09 | -0.08 | -0.09 |
| cumulated_balance | -0.01 | 1.00 | 0.01 | -0.00 | 0.08 | 0.06 | 0.03 | 0.09 | 0.05 | 0.15 | 0.16 | -0.14 | -0.10 |
| SpO2 | 0.10 | 0.01 | 1.00 | -0.02 | -0.02 | -0.02 | -0.02 | -0.00 | -0.04 | 0.07 | 0.08 | 0.00 | 0.03 |
| BUN | -0.05 | -0.00 | -0.02 | 1.00 | 0.70 | 0.06 | 0.05 | 0.08 | 0.15 | -0.06 | -0.05 | -0.10 | -0.10 |
| Creatinine | -0.01 | 0.08 | -0.02 | 0.70 | 1.00 | 0.05 | 0.02 | 0.04 | 0.15 | -0.11 | -0.06 | -0.24 | -0.22 |
| SGOT | -0.03 | 0.06 | -0.02 | 0.06 | 0.05 | 1.00 | 0.85 | 0.41 | 0.10 | 0.06 | 0.05 | 0.02 | 0.00 |
| SGPT | -0.02 | 0.03 | -0.02 | 0.05 | 0.02 | 0.85 | 1.00 | 0.33 | 0.06 | 0.05 | 0.03 | 0.04 | 0.02 |
| Total_bili | -0.02 | 0.09 | -0.00 | 0.08 | 0.04 | 0.41 | 0.33 | 1.00 | 0.17 | 0.08 | 0.05 | 0.02 | -0.02 |
| INR | -0.03 | 0.05 | -0.04 | 0.15 | 0.15 | 0.10 | 0.06 | 0.17 | 1.00 | -0.03 | 0.01 | -0.07 | -0.06 |
| input_total | -0.04 | 0.15 | 0.07 | -0.06 | -0.11 | 0.06 | 0.05 | 0.08 | -0.03 | 1.00 | 0.44 | 0.48 | 0.31 |
| input_4hourly | -0.09 | 0.16 | 0.08 | -0.05 | -0.06 | 0.05 | 0.03 | 0.05 | 0.01 | 0.44 | 1.00 | 0.27 | 0.39 |
| output_total | -0.08 | -0.14 | 0.00 | -0.10 | -0.24 | 0.02 | 0.04 | 0.02 | -0.07 | 0.48 | 0.27 | 1.00 | 0.69 |
| output_4hourly | -0.09 | -0.10 | 0.03 | -0.10 | -0.22 | 0.00 | 0.02 | -0.02 | -0.06 | 0.31 | 0.39 | 0.69 | 1.00 |

Figure B.3: Feature correlation matrix for case study concerning sepsis treatment, Part 3

Appendix C

Feature correlation matrix for the weaning model in case study 2

Figure C.1 shows the feature correlation matrix for the weaning model in case study 2 presented in Chapter 5.

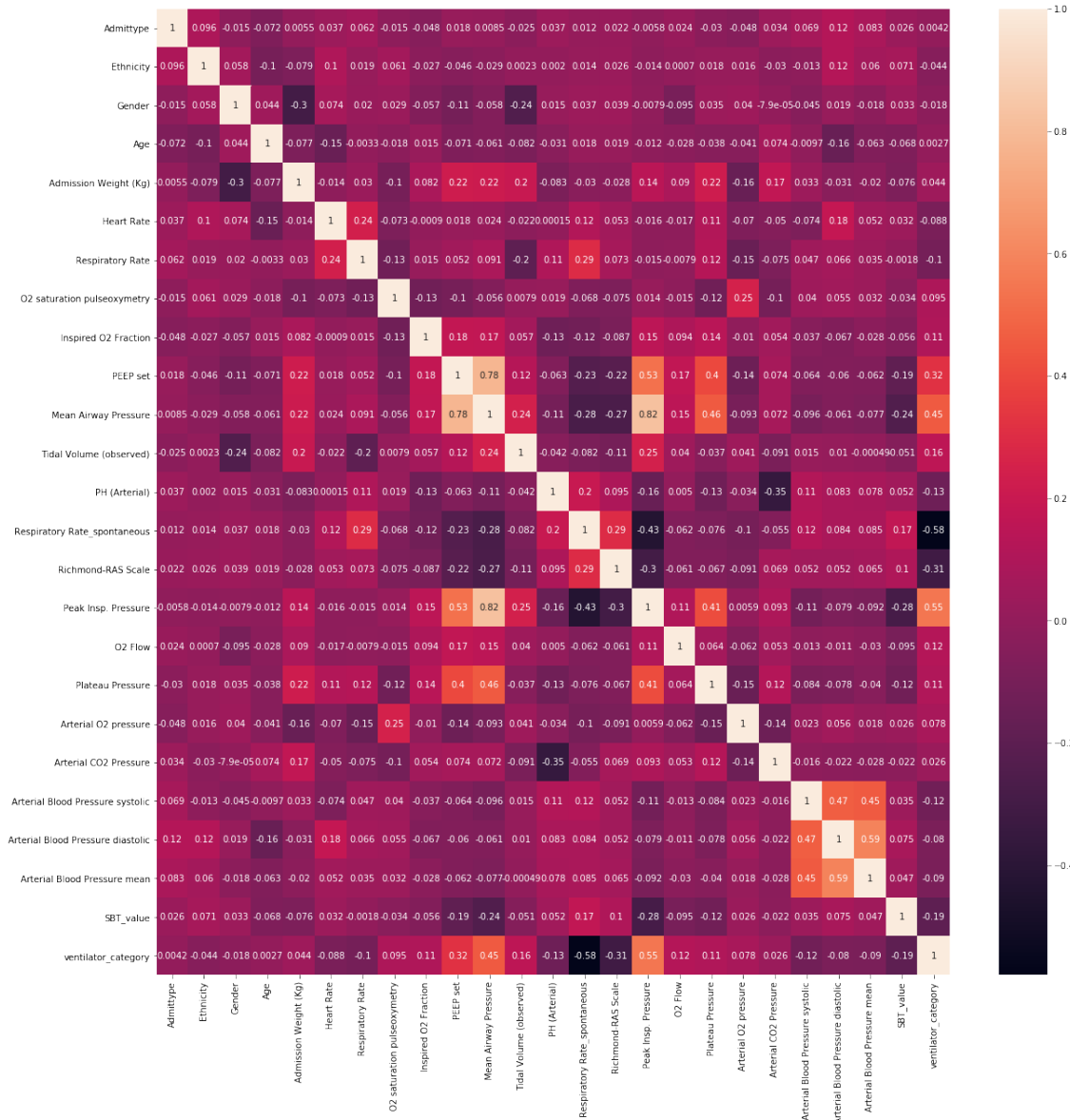


Figure C.1: Feature correlation matrix for case study concerning ventilator weaning

Abbreviations

AF Atrial Fibrillation

AI Artificial Intelligence

AMLAS Assurance of Machine Learning for Autonomous Systems

AUC Area Under the Curve

AUC-ROC Area Under The Receiver Operating Characteristics Curve

BB Beta-Blocker

BN Bayesian Network

BP Blood Pressure

BSI British Standards Institution

BTD Bow Tie Diagram

CAE Claims Argument Evidence

CE Conformité Européenne

CNN Convolutional Neural Network

CQC Care Quality Commission

DiCE Diverse Counterfactual Examples

DNN Deep Neural Network

DSR Derived Safety Requirement

DSS Decision Support System

EHR Electronic Health Record

EMA European Medicines Agency

ETA Event Tree Analysis

EWS Early Warning Score

FDA Federal Drug Administration

FMEA Failure Modes and Effects Analysis

FTA Fault-Tree Analysis

GDPR General Data Protection Regulations

GSN Goal Structuring Notation

HAZOP Hazard and Operability Analysis

HIT Healthcare IT

ICU Intensive Care Unit

IEEE Institute of Electrical and Electronics Engineers

IMDRF International Medical Device Regulators Forum

IoM Institute of Medicine

ISMP Canada The Institute for Safe Medication Practices Canada

LIME Local Interpretable Model-Agnostic Explanations

MAP Mean Arterial Pressure

MHRA Medicines & Healthcare products Regulatory Agency

ML Machine Learning

NN Neural Network

PEEP Positive End-Expiratory Pressure

PGM Probabilistic Graphical Model

PSV Pressure Support Ventilation

QMS Quality Management System

RF Random Forest

RL Reinforcement Learning

ROC Receiver Operating Characteristic Curve

SAM Safety Argument Manager

SaMD Software as a Medical Device

SBT Spontaneous Breathing Trial

SHAP SHapley Additive exPlanations

SHARD Software Hazard Analysis and Resolution in Design

SOFA Sequential Organ Failure Assessment

SVM Support Vector Machine

TPLC Total Product Life-Cycle

UKCA UK Conformity Assessed

WHO World Health Organisation

References

- [1] 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.
- [2] E. Alpaydin, “Introduction to machine learning, (adaptive computation and machine learning),” 2010.
- [3] T. M. Mitchell, *The discipline of machine learning*, vol. 9. Carnegie Mellon University, School of Computer Science, Machine Learning . . . , 2006.
- [4] W. E. Vesely, F. F. Goldberg, N. H. Roberts, and D. F. Haasl, “Fault tree handbook,” tech. rep., Nuclear Regulatory Commission Washington dc, 1981.
- [5] Society of Automotive Engineers, “Recommended Practices for Non-Automobile Applications, ARP 5580,” 2001.
- [6] R. Kenarangui, “Event-tree analysis by fuzzy probability,” *IEEE Transactions on Reliability*, vol. 40, no. 1, pp. 120–124, 1991.
- [7] T. A. Kletz, *HAZOP and HAZAN: identifying and assessing process industry hazards*. IChemE, 1999.
- [8] D. J. Pumfrey, *The principled design of computer system safety analyses*. PhD thesis, University of York, 1999.
- [9] NHS Digital, “DCB0160: Clinical risk management: its Application in the Deployment and Use of health IT Systems,” 2018.
- [10] US Food and Drug Administration, “Infusion pumps total product life cycle: Guidance for industry and FDA staff,” *Food and Drug Administration Std*, pp. 0910–0766, 2014.

- [11] R. Redfern, “A regional examination of surgery and fracture treatment in iron age and roman britain,” *International Journal of Osteoarchaeology*, vol. 20, no. 4, pp. 443–471, 2010.
- [12] World Health Organisation (WHO), “What is patient safety?.” <http://www.who.int/patientsafety/about/en/>. Accessed 2021-05-20.
- [13] R. M. Gallagher and D. Melnyk, “The national coordinating council for medication error reporting and prevention: 25 years of building medication safety.” <https://www.nccmerp.org/sites/default/files/nccmerp-25-year-report.pdf>, 2020. Accessed 2021-05-20.
- [14] IMDRF SaMD Working Group, “Software as a Medical Device(SaMD): Clinical Evaluation – Guidance for Industry and Food and Drug Administration Staff,” International Medical Device Regulators Forum, 2017.
- [15] 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.
- [16] N. M. Murray, M. Unberath, G. D. Hager, and F. K. Hui, “Artificial intelligence to diagnose ischemic stroke and identify large vessel occlusions: a systematic review,” *Journal of neurointerventional surgery*, vol. 12, no. 2, pp. 156–164, 2020.
- [17] Y. Jia, T. Lawton, S. White, and I. Habli, “Developing a safety case for electronic prescribing,” in *Studies in Health Technology and Informatics: MEDINFO2019*, vol. 264, pp. 629–633, Aug. 2019.
- [18] I. Habli, Y. Jia, S. White, G. Gabriel, T. Lawton, M. Sujjan, and C. Tomsett, “Development and piloting of a software tool to facilitate proactive hazard and risk analysis of health information technology,” *Health informatics journal*, vol. 26, no. 1, pp. 683–702, 2020.
- [19] W. Chao, “Machine learning tutorial,” *National Taiwan University*, 2011.
- [20] R. Xu and D. Wunsch, “Survey of clustering algorithms,” *IEEE Transactions on neural networks*, vol. 16, no. 3, pp. 645–678, 2005.
- [21] I. K. Fodor, “A survey of dimension reduction techniques,” *Center for Applied Scientific Computing, Lawrence Livermore National Laboratory*, vol. 9, pp. 1–18, 2002.

-
- [22] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
- [23] P. Domingos, “A few useful things to know about machine learning,” *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.
- [24] A. Abraham, “Artificial neural networks,” *handbook of measuring system design*, 2005.
- [25] W. Koehrsen, “10 facts on patient safety.” <https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d>, December 2017. Accessed 2021-05-20.
- [26] S. Patel, “Chapter 2 : SVM (Support Vector Machine)-Theory.” <https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72>, May 2017. Accessed 2021-05-20.
- [27] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
- [28] T. Knode, D. Schonacher, N. Ritchie, *et al.*, “Wellsite risk management improvement including human factors,” in *SPE International Conference and Exhibition on Health, Safety, Security, Environment, and Social Responsibility*, Society of Petroleum Engineers, 2018.
- [29] P. Clarkson, D. Bogle, J. Dean, M. Tooley, J. Trewby, L. Vaughan, E. Adams, P. Dudgeon, N. Platt, and P. Shelton, “Engineering better care: a systems approach to health and care design and continuous improvement,” 2017.
- [30] ISO, “ISO 14971: Medical devices — Application of risk management to medical devices.” <https://www.iso.org/standard/72704.html>. Accessed 2021-05-20.
- [31] US Food and Drug Administration, “Factors to Consider Regarding Benefit-Risk in Medical Device Product Availability, Compliance, and Enforcement Decisions,” 2016.
- [32] ISO, “ISO/TR 24971: Medical devices — Guidance on the application of ISO 14971.” <https://www.iso.org/standard/74437.html>, 2020. Accessed 2021-05-20.

- [33] J. Forder, C. Higgins, J. McDermid, and G. Storrs, “Sam—a tool to support the construction, review and evolution of safety arguments,” in *Directions in safety-critical systems*, pp. 195–216, Springer, 1993.
- [34] T. P. Kelly, *Arguing safety: a systematic approach to managing safety cases*. PhD thesis, University of York, 1999.
- [35] Adelard, “Claims, Arguments and Evidence (CAE).” <https://www.adelard.com/asce/choosing-asce/cae.html>. Accessed 2021-05-20.
- [36] S. Toulmin, *The Uses of Argument*. Cambridge University Press, 1958.
- [37] IEC 62304, “Medical device software – Software life cycle processes.” <https://www.iso.org/standard/38421.html>, May 2006. Accessed 2021-05-20.
- [38] J. McDermid and D. Pumfrey, “Software safety: Why is there no consensus?,” in *proceedings of the international system safety conference*, Citeseer, 2001.
- [39] A. Wassyng, T. Maibaum, M. Lawford, and H. Bherer, “Software certification: Is there a case against safety cases?,” in *Foundations of Computer Software. Modeling, Development, and Verification of Adaptive Systems* (R. Calinescu and E. Jackson, eds.), (Berlin, Heidelberg), pp. 206–227, Springer Berlin Heidelberg, 2011.
- [40] N. G. Leveson, “The use of safety cases in certification and regulation,” https://www.csb.gov/assets/1/7/leveson_paper.pdf, 2011.
- [41] US Food and Drug Administration, “Examples of Reported Infusion Pump Problems.” <https://www.fda.gov/medical-devices/infusion-pumps/examples-reported-infusion-pump-problems>. Accessed 2021-05-20.
- [42] M. Sujan and I. Habli, “Safety cases for digital health innovations: can they work?,” *BMJ Quality & Safety*, 2021.
- [43] C. Petersen, J. Smith, R. R. Freimuth, K. W. Goodman, G. P. Jackson, J. Kannry, H. Liu, S. Madhavan, D. F. Sittig, and A. Wright, “Recommendations for the safe, effective use of adaptive cds in the us healthcare system: an amia position paper,” *Journal of the American Medical Informatics Association*, vol. 28, no. 4, pp. 677–684, 2021.

- [44] K. Luxford, “‘first, do no harm’: shifting the paradigm towards a culture of health,” *Patient Experience Journal*, vol. 3, no. 2, pp. 5–8, 2016.
- [45] R. E. Herzlinger, “Why Innovation in Health Care Is So Hard.” <https://hbr.org/2006/05/why-innovation-in-health-care-is-so-hard>, May 2006. Accessed 2021-05-20.
- [46] WRHA Infection Prevention & Control, “Routine Practices.” <http://www.wrha.mb.ca/extranet/ipc/files/routine-practices/InfoSheet-Education.pdf>. Accessed 2021-05-20.
- [47] C. Vincent, *Patient safety*. John Wiley & Sons, 2011.
- [48] N. Kapur, A. Parand, T. Soukup, T. Reader, and N. Sevdalis, “Aviation and health-care: a comparative review with implications for patient safety,” *JRSM open*, vol. 7, no. 1, p. 2054270415616548, 2015.
- [49] R. Levaggi and P. C. Smith, “Decentralization in health care: lessons from public economics,” *Health Policy and Economics: Opportunities and Challenges (Maidenhead: Open University Press, 2005)*, pp. 223–247, 2003.
- [50] Department of Health, “Keeping the NHS local - a new direction of travel.” http://webarchive.nationalarchives.gov.uk/20120214222221/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_4085947.pdf, 2003. Accessed 2021-05-20.
- [51] J. Carlson, “Two additional patient deaths linked to Medtronic infusion pump.” <http://www.startribune.com/2-additional-patient-deaths-linked-to-medtronic-infusion-pump/399576491/>, November 2016. Accessed 2021-05-20.
- [52] B. M. Association *et al.*, “Caring, supportive, collaborative? doctors’ views on working in the nhs,” *London: BMA*, 2018.
- [53] G. Livio and A. Padula, “Defensive medicine in europe: a ‘full circle’?,” *The European Journal of Health Economics: HEPAC*, vol. 21, no. 2, pp. 165–170, 2020.
- [54] L. Emanuel, D. Berwick, J. Conway, J. Combes, M. Hatlie, L. Leape, J. Reason, P. Schyve, C. Vincent, and M. Walton, “What exactly is patient safety,” *Advances in patient safety: new directions and alternative approaches*, vol. 1, pp. 1–17, 2008.

- [55] S. M. Erickson, J. Wolcott, J. M. Corrigan, P. Aspden, *et al.*, *Patient safety: achieving a new standard for care*. National Academies Press, 2003.
- [56] World Health Organisation (WHO), “Conceptual Framework for the International Classification for Patient Safety.” https://www.who.int/patientsafety/taxonomy/icps_full_report.pdf. Accessed 2021-05-20.
- [57] E. Newhook, “What You Should Know About Preventable Harm.” <https://mha.gwu.edu/blog-preventable-harm/>, August 2015. Accessed 2021-05-20.
- [58] M. Nabhan, T. Elraiyah, D. R. Brown, J. Dilling, A. LeBlanc, V. M. Montori, T. Morgenthaler, J. Naessens, L. Prokop, V. Roger, *et al.*, “What is preventable harm in healthcare? a systematic review of definitions,” *BMC health services research*, vol. 12, no. 1, p. 128, 2012.
- [59] S. M. Mark, J. D. Little, S. Geller, and R. J. Weber, *Principles and Practices of Medication Safety*, ch. 5. New York, NY: The McGraw-Hill Companies, 2011.
- [60] M. Grissinger, “Tall man letters are gaining wide acceptance,” *Pharmacy and Therapeutics*, vol. 37, no. 3, p. 132, 2012.
- [61] ISMP Canada, “Near Miss Identification and Reporting.” <http://www.ismp-canada.org/download/safetyBulletins/ISMPCSB2007-07NearMiss.pdf>, December 2007. Accessed 2021-05-20.
- [62] K. N. Barker, E. A. Flynn, G. A. Pepper, D. W. Bates, and R. L. Mikeal, “Medication errors observed in 36 health care facilities,” *Archives of internal medicine*, vol. 162, no. 16, pp. 1897–1903, 2002.
- [63] NHS England, “Patient safety alert stage three: Directive improving medication error incident reporting and learning.” <https://www.england.nhs.uk/wp-content/uploads/2014/03/psa-sup-info-med-error.pdf>, March 2014. Accessed 2021-05-20.
- [64] M. Lisby, L. P. Nielsen, B. Brock, and J. Mainz, “How are medication errors defined? a systematic literature review of definitions and characteristics,” *International Journal for Quality in Health Care*, vol. 22, no. 6, pp. 507–518, 2010.

- [65] National Coordinating Council for Medication Error Reporting and Prevention, “What is a Medication Error?.” <https://www.nccmerp.org/about-medication-errors>. Accessed 2021-05-20.
- [66] D. W. Bates, J. M. Teich, J. Lee, D. Seger, G. J. Kuperman, N. Ma’Luf, D. Boyle, and L. Leape, “The impact of computerized physician order entry on medication error prevention,” *Journal of the American Medical Informatics Association*, vol. 6, no. 4, pp. 313–321, 1999.
- [67] J. K. Aronson, “Medication errors: definitions and classification,” *British journal of clinical pharmacology*, vol. 67, no. 6, pp. 599–604, 2009.
- [68] B. Dean, N. Barber, and M. Schachter, “What is a prescribing error?,” *BMJ Quality & Safety*, vol. 9, no. 4, pp. 232–237, 2000.
- [69] Mersey Care NHS Trust, “Guidelines for the Management of Medicines Errors within Mersey Care NHS Trust.” <https://www.merseycare.nhs.uk/media/2445/guidelines-for-the-management-of-medicines-errors-mm09.pdf>, May 2015. Accessed 2021-05-20.
- [70] J. K. Aronson, “Medication errors: what they are, how they happen, and how to avoid them,” *QJM: An International Journal of Medicine*, vol. 102, no. 8, pp. 513–521, 2009.
- [71] G. P. Velo and P. Minuz, “Medication errors: prescribing faults and prescription errors,” *British journal of clinical pharmacology*, vol. 67, no. 6, pp. 624–628, 2009.
- [72] J. P. Marcin, M. Dharmar, M. Cho, L. L. Seifert, J. L. Cook, S. L. Cole, F. Nasrollahzadeh, and P. S. Romano, “Medication errors among acutely ill and injured children treated in rural emergency departments,” *Annals of emergency medicine*, vol. 50, no. 4, pp. 361–367, 2007.
- [73] M. Dharmar, N. Kuppermann, P. S. Romano, N. H. Yang, T. S. Nesbitt, J. Phan, C. Nguyen, K. Parsapour, and J. P. Marcin, “Telemedicine consultations and medication errors in rural emergency departments,” *Pediatrics*, pp. 1090–1097, 2013.
- [74] R. E. Ferner and J. K. Aronson, “Clarification of terminology in medication errors,” *Drug safety*, vol. 29, no. 11, pp. 1011–1022, 2006.

- [75] World Health Organisation (WHO), “Medication Errors: Technical Series on Safer Primary Care.” <http://apps.who.int/iris/bitstream/handle/10665/252274/9789241511643-eng.pdf?sequence=1>, 2016. Accessed 2021-05-20.
- [76] “Council Directive 93/42/EEC of 14 June 1993 concerning medical devices.” <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A31993L0042>, 1993.
- [77] MEDDEV 2.1/6, “Medical Devices: Guidance document - Qualification and Classification of stand alone software,” 2016.
- [78] Medicines & Healthcare product Regulatory Agency (MHRA), “Guidance: Medical device stand-alone software including apps (including IVDMDs).” https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/717865/Software_flow_chart_Ed_1-05.pdf. Accessed 2021-05-20.
- [79] R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards, and K. Tsaneva-Atanasova, “Artificial intelligence, bias and clinical safety,” *BMJ Quality & Safety*, vol. 28, no. 3, pp. 231–237, 2019.
- [80] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, “Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.
- [81] E. J. Hwang, S. Park, K.-N. Jin, J. I. Kim, S. Y. Choi, J. H. Lee, J. M. Goo, J. Aum, J.-J. Yim, C. M. Park, *et al.*, “Development and validation of a deep learning-based automatic detection algorithm for active pulmonary tuberculosis on chest radiographs,” *Clinical Infectious Diseases*, vol. 69, no. 5, pp. 739–747, 2019.
- [82] R. Singh, M. K. Kalra, C. Nitiwarangkul, J. A. Patti, F. Homayounieh, A. Padole, P. Rao, P. Putha, V. V. Muse, A. Sharma, *et al.*, “Deep learning in chest radiography: detection of findings and presence of change,” *PloS one*, vol. 13, no. 10, p. e0204155, 2018.

-
- [83] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” *nature*, vol. 542, no. 7639, pp. 115–118, 2017.
- [84] S. S. Han, M. S. Kim, W. Lim, G. H. Park, I. Park, and S. E. Chang, “Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm,” *Journal of Investigative Dermatology*, vol. 138, no. 7, pp. 1529–1538, 2018.
- [85] J. Zhang, S. Gajjala, P. Agrawal, G. H. Tison, L. A. Hallock, L. Beussink-Nelson, M. H. Lassen, E. Fan, M. A. Aras, C. Jordan, *et al.*, “Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy,” *Circulation*, vol. 138, no. 16, pp. 1623–1635, 2018.
- [86] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O’Donoghue, D. Visentin, *et al.*, “Clinically applicable deep learning for diagnosis and referral in retinal disease,” *Nature medicine*, vol. 24, no. 9, pp. 1342–1350, 2018.
- [87] R. F. Thompson, G. Valdes, C. D. Fuller, C. M. Carpenter, O. Morin, S. Aneja, W. D. Lindsay, H. J. Aerts, B. Agrimson, C. Deville Jr, *et al.*, “Artificial intelligence in radiation oncology: a specialty-wide disruptive transformation?,” *Radiotherapy and Oncology*, vol. 129, no. 3, pp. 421–426, 2018.
- [88] 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.
- [89] K. Orphanou, A. Stassopoulou, and E. Keravnou, “Risk assessment for primary coronary heart disease event using dynamic bayesian networks,” in *Conference on Artificial Intelligence in Medicine in Europe*, pp. 161–165, Springer, 2015.
- [90] C. G. Walsh, J. D. Ribeiro, and J. C. Franklin, “Predicting risk of suicide attempts over time through machine learning,” *Clinical Psychological Science*, vol. 5, no. 3, pp. 457–469, 2017.
- [91] Babylon Health, “Babylon health services.” <https://www.babylonhealth.com/product>. Accessed 2021-05-20.

- [92] K. Middleton, M. Butt, N. Hammerla, S. Hamblin, K. Mehta, and A. Parsa, "Sorting out symptoms: design and evaluation of the 'babylon check' automated triage system," *arXiv preprint arXiv:1606.02041*, 2016.
- [93] M. Burgess and N. Kobie, "The messy, cautionary tale of how Babylon disrupted the NHS." <https://www.wired.co.uk/article/babylon-health-nhs>, 2019. Accessed 2021-05-20.
- [94] MEDDEV 2.7/1 revision 4, "Clinical evaluation: A guide for manufacturers and notified bodies under directives 93/42/EEC and 90/385/EEC," 2016.
- [95] European Commission, "Guidance document - Market surveillance - Guidelines on a Medical Devices Vigilance System - MEDDEV 2.12/1 rev.8." <https://ec.europa.eu/docsroom/documents/32305/attachments/1/translations>, 2013. Accessed 2021-05-20.
- [96] US Food and Drug Administration, "Classify Your Medical Device." <https://www.fda.gov/medical-devices/overview-device-regulation/classify-your-medical-device>, 2020. Accessed 2021-05-20.
- [97] US Food and Drug Administration, "Premarket Approval (PMA)." <https://www.fda.gov/medical-devices/premarket-submissions/premarket-approval-pma>, 2019. Accessed 2021-05-20.
- [98] US Food and Drug Administration, "510(k) Clearances." <https://www.fda.gov/medical-devices/device-approvals-denials-and-clearances/510k-clearances>, 2018. Accessed 2021-05-20.
- [99] US Food and Drug Administration, "De Novo Classification Request." <https://www.fda.gov/medical-devices/premarket-submissions/de-novo-classification-request>, 2019. Accessed 2021-05-20.
- [100] US Food and Drug Administration, "Humanitarian Device Exemption." <https://www.fda.gov/medical-devices/premarket-submissions/humanitarian-device-exemption>, 2019. Accessed 2021-05-20.
- [101] U. J. Muehlematter, P. Daniore, and K. N. Vokinger, "Approval of artificial intelligence and machine learning-based medical devices in the usa and europe (2015–20): a comparative analysis," *The Lancet Digital Health*, 2021.

-
- [102] CHPSO, “FDA Medical Device Clearance Process Criticized.” <https://www.chpso.org/newsletter/fda-medical-device-clearance-process-criticized>. Accessed 2021-05-20.
- [103] US Food and Drug Administration, “Global Approach to Software as a Medical Device.” <https://www.fda.gov/medical-devices/software-medical-device-samd/global-approach-software-medical-device>, 2017. Accessed 2021-05-20.
- [104] US Food and Drug Administration, “Deciding When to Submit a 510(k) for a Software Change to an Existing Device.” <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/deciding-when-submit-510k-software-change-existing-device>, 2018. Accessed 2021-05-20.
- [105] US Food and Drug Administration, “Digital Health Software Precertification (Pre-Cert) Program.” <https://www.fda.gov/medical-devices/digital-health-center-excellence/digital-health-software-precertification-pre-cert-program>, 2020. Accessed 2021-05-20.
- [106] US Food and Drug Administration, “Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan,” 2021.
- [107] European Commission , “Medical Devices - Sector: Overview.” https://ec.europa.eu/health/md_sector/overview_en. Accessed 2021-05-20.
- [108] European Union, “Regulation (EU) 2017/745 of the european parliament and of the council of 5 april 2017.” <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02017R0745-20170505>. Accessed 2021-05-20.
- [109] European Union, “Council Directive 93/42/EEC of 14 June 1993.” <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:01993L0042-20071011>. Accessed 2021-05-20.
- [110] MEDDEV 2.7/4, “Guidelines on clinical investigation: A guide for manufacturers and notified bodies,” 2010.
- [111] HMA and EMA, “HMA-EMA Joint Big Data Taskforce.” https://www.ema.europa.eu/en/documents/minutes/hma/ema-joint-task-force-big-data-summary-report_en.pdf, 2019. Accessed 2021-05-20.

- [112] European AI Alliance, “Artificial Intelligence and Machine Learning in Software as a Medical Device: discussion Paper and Request for Feedback.” <https://ec.europa.eu/futurium/en/european-ai-alliance/artificial-intelligence-and-machine-learning-software-medical-device-discussion>. Accessed 2021-05-20.
- [113] European Commission, “White Paper On Artificial Intelligence – A European approach to excellence and trust, Brussels, 19.2.2020 COM(2020) 65 final.” https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf. Accessed 2021-05-20.
- [114] European Commission, “Commission Report on safety and liability implications of AI, the Internet of Things and Robotics.” https://ec.europa.eu/info/publications/commission-report-safety-and-liability-implications-ai-internet-things-and-robotics-0_en, 2020. Accessed 2021-05-20.
- [115] T. Minssen, S. Gerke, M. Aboy, N. Price, and G. Cohen, “Regulatory responses to medical machine learning,” *Journal of Law and the Biosciences*, 2020.
- [116] “Using the UKCA marking.” <https://www.gov.uk/guidance/using-the-ukca-marking>, Dec 2020. Accessed 2021-05-20.
- [117] “Regulating medical devices in the UK.” <https://www.gov.uk/guidance/regulating-medical-devices-in-the-uk>, Dec 31 2020. Accessed 2021-05-20.
- [118] CQC and MHRA, “Using machine learning in diagnostic services: A report with recommendations from CQC’s regulatory sandbox,” 2020.
- [119] M. Gould, “Regulating AI in health and care.” <https://healthtech.blog.gov.uk/2020/02/12/regulating-ai-in-health-and-care/>, Feb 2020. Accessed 2021-05-20.
- [120] REFORM, “Data-driven healthcare: Regulation & regulators.” <https://reform.uk/research/data-driven-healthcare-regulation-regulators>. Accessed 2021-05-20.
- [121] T. J. Hwang, A. S. Kesselheim, and K. N. Vokinger, “Lifecycle regulation of artificial intelligence–and machine learning–based software devices in medicine,” *Jama*, vol. 322, no. 23, pp. 2285–2286, 2019.

-
- [122] X. Liu, S. C. Rivera, D. Moher, M. J. Calvert, and A. K. Denniston, “Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the consort-ai extension,” *bmj*, vol. 370, 2020.
- [123] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, “Concrete problems in ai safety,” *arXiv preprint arXiv:1606.06565*, 2016.
- [124] P. M. Bossuyt and J. B. Reitsma, “The STARD initiative,” *The Lancet*, vol. 361, no. 9351, p. 71, 2003.
- [125] NHS, “The Topol Review Preparing the healthcare workforce to deliver the digital future,” 2019.
- [126] S. Benjamens, P. Dhunoo, and B. Meskó, “The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database,” *NPJ digital medicine*, vol. 3, no. 1, pp. 1–8, 2020.
- [127] S. Gerke, B. Babic, T. Evgeniou, and I. G. Cohen, “The need for a system view to regulate artificial intelligence/machine learning-based software as medical device,” *NPJ digital medicine*, vol. 3, no. 1, pp. 1–4, 2020.
- [128] Y. Jia, J. Burden, T. Lawton, and I. Habli, “Safe reinforcement learning for sepsis treatment,” in *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 1–7, 2020.
- [129] Y. Jia, T. Lawton, J. Burden, J. McDermid, and I. Habli, “Safety-driven design of machine learning for sepsis treatment,” *Journal of Biomedical Informatics*, vol. 117, p. 103762, 2021.
- [130] J. McDermid and Y. Jia, “Safety of artificial intelligence: A collaborative model,” in *AISafety@IJCAI*, 2020.
- [131] I. Habli, S. White, M. Sujjan, 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.
- [132] M. Hutson *et al.*, “Even artificial intelligence can acquire biases against race and gender,” *Science Magazine, Science AAAS*, vol. 13, 2017.

- [133] 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.
- [134] 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.
- [135] 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.
- [136] 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.
- [137] C. Yu, J. Liu, and S. Nemati, “Reinforcement learning in healthcare: a survey,” *arXiv preprint arXiv:1908.08796*, 2019.
- [138] 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.
- [139] 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.
- [140] 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.
- [141] 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.
- [142] 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.
- [143] 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.
- [144] D. C. Angus, “Randomized Clinical Trials of Artificial Intelligence,” *JAMA*, vol. 323, pp. 1043–1045, 03 2020.
- [145] 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.
- [146] C. Picardi, R. Hawkins, C. Paterson, and I. Habli, “A pattern for arguing the assurance of machine learning in medical diagnosis systems,” in *International Conference on Computer Safety, Reliability, and Security*, pp. 165–179, Springer, 2019.
- [147] 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.
- [148] 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.
- [149] J. Gallagher, “‘Alarming’ one in five deaths due to sepsis.” <https://www.bbc.co.uk/news/health-51138859>, 2020. Accessed 2021-05-20.
- [150] 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.
- [151] 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 2021-05-20.
- [152] 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.
- [153] K. L. Mosier and L. J. Skitka, “Human decision makers and automated decision aids: Made for each other?,” in *Automation and human performance: Theory and applications*, pp. 201–220, CRC Press, 2018.
- [154] NHS Improvement, “Sepsis is a medical emergency!.” https://improvement.nhs.uk/documents/652/Sepsis_Ae_Easy_Guide.pdf. Accessed 2021-05-20.
- [155] Royal College of Physicians, “National early warning score.” <https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2>. Accessed 2021-05-20.
- [156] 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 2021-05-20.
- [157] 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.
- [158] 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.
- [159] 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.
- [160] Hospira UK Ltd, “Noradrenaline (Norepinephrine) 1 mg/ml Concentrate for Solution for Infusion.” <https://www.medicines.org.uk/emc/product/4115/smpc>, 2018. Accessed 2021-05-20.
- [161] J. M. Allen, “Understanding vasoactive medications: focus on pharmacology and effective titration,” *Journal of Infusion Nursing*, vol. 37, no. 2, pp. 82–86, 2014.

- [162] 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.
- [163] P. Fenelon, J. McDermid, M. Nicolson, and D. Pumfrey, “Towards integrated safety analysis and design,” *ACM SIGAPP Applied Computing Review*, vol. 2, no. 1, pp. 21–32, 1994.
- [164] 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.
- [165] C. J. C. H. Watkins, *Learning from delayed rewards*. PhD thesis, King’s College, University of Cambridge, 1989.
- [166] 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.
- [167] “An end-to-end open source machine learning platform.” <https://www.tensorflow.org>. Accessed 2021-05-20.
- [168] 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.
- [169] N. K. Jong and P. Stone, “Model-based function approximation in reinforcement learning,” in *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, pp. 1–8, 2007.
- [170] N. Jiang and L. Li, “Doubly robust off-policy value evaluation for reinforcement learning,” in *International Conference on Machine Learning*, pp. 652–661, PMLR, 2016.
- [171] C. Molnar, *Interpretable Machine Learning*. Lulu.com, 2020.
- [172] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, pp. 5–32, 2001.
- [173] 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.

- [174] 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.
- [175] E. Denney and G. Pai, “Tool support for assurance case development,” *Automated Software Engineering*, vol. 25, no. 3, pp. 435–499, 2018.
- [176] M. Komorowski, “Clinical management of sepsis can be improved by artificial intelligence: yes,” 2019.
- [177] 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.
- [178] Y. Jia, C. Kaul, T. Lawton, R. Murray-Smith, and I. Habli, “Prediction of weaning from mechanical ventilation using convolutional neural networks,” *Artificial Intelligence in Medicine*, vol. 117, p. 102087, 2021.
- [179] Y. Jia, J. McDermid, T. Lawton, and I. Habli, “The role of explainability in assuring safety of machine learning in healthcare,” *Journal of Biomedical Informatics (Submitted)*, 2021.
- [180] Y. Jia, J. McDermid, and I. Habli, “Enhancing the value of counterfactual explanations for deep learning,” in *AIME 2021: Artificial Intelligence in Medicine in Europe*, Porto, 2021.
- [181] J. McDermid, Y. Jia, Z. Porter, and I. Habli, “AI Explainability: The Technical and Ethical Dimensions,” in *Phil. Trans. R. Soc. A. 379: 20200363*, PMLR, 2021.
- [182] R. Hawkins, C. Paterson, C. Picardi, Y. Jia, R. Calinescu, and I. Habli, “Guidance on the assurance of machine learning in autonomous systems (amlas),” 2021.
- [183] D. S. Watson, J. Krutzinna, I. N. Bruce, C. E. Griffiths, I. B. McInnes, M. R. Barnes, and L. Floridi, “Clinical applications of machine learning algorithms: beyond the black box,” *Bmj*, vol. 364, 2019.
- [184] FDA, “Transparency of Artificial Intelligence/ Machine Learning (AI/ML)-Enabled Medical Devices: FDA Virtual Public Workshop,” 2021.
- [185] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, “Explaining explanations: An overview of interpretability of machine learning,” in *2018 IEEE 5th*

-
- International Conference on data science and advanced analytics (DSAA)*, pp. 80–89, IEEE, 2018.
- [186] B. Goodman and S. Flaxman, “European union regulations on algorithmic decision-making and a “right to explanation”,” *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017.
- [187] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” *arXiv preprint arXiv:1702.08608*, 2017.
- [188] Z. C. Lipton, “The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery,” *Queue*, vol. 16, no. 3, pp. 31–57, 2018.
- [189] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in neural information processing systems*, pp. 4765–4774, 2017.
- [190] J. Montano and A. Palmer, “Numeric sensitivity analysis applied to feedforward neural networks,” *Neural Computing & Applications*, vol. 12, no. 2, pp. 119–125, 2003.
- [191] B. Liang, H. Li, M. Su, P. Bian, X. Li, and W. Shi, “Deep text classification can be fooled,” *arXiv preprint arXiv:1704.08006*, 2017.
- [192] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *European conference on computer vision*, pp. 818–833, Springer, 2014.
- [193] M. T. Ribeiro, S. Singh, and C. Guestrin, ““ why should i trust you?” explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.
- [194] L. S. Shapley, “A value for n-person games,” *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307–317, 1953.
- [195] U. Bhatt, P. Ravikumar, and J. M. Moura, “Towards aggregating weighted feature attributions,” *arXiv preprint arXiv:1901.10040*, 2019.
- [196] E. Štrumbelj and I. Kononenko, “Explaining prediction models and individual predictions with feature contributions,” *Knowledge and information systems*, vol. 41, no. 3, pp. 647–665, 2014.

- [197] J. Chen, L. Song, M. J. Wainwright, and M. I. Jordan, “L-shapley and c-shapley: Efficient model interpretation for structured data,” *arXiv preprint arXiv:1808.02610*, 2018.
- [198] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, “From local explanations to global understanding with explainable ai for trees,” *Nature Machine Intelligence*, vol. 2, no. 1, pp. 2522–5839, 2020.
- [199] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” *arXiv preprint arXiv:1312.6034*, 2013.
- [200] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, “Not just a black box: Learning important features through propagating activation differences,” *arXiv preprint arXiv:1605.01713*, 2016.
- [201] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” *arXiv preprint arXiv:1703.01365*, 2017.
- [202] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning important features through propagating activation differences,” in *International Conference on Machine Learning*, pp. 3145–3153, PMLR, 2017.
- [203] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, “Towards better understanding of gradient-based attribution methods for deep neural networks,” *arXiv preprint arXiv:1711.06104*, 2017.
- [204] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box: Automated decisions and the gdpr,” *Harv. JL & Tech.*, vol. 31, p. 841, 2017.
- [205] D. Kahneman and A. Tversky, “The Simulation Heuristic,” tech. rep., Stanford Univ Ca Dept Of Psychology, 1981.
- [206] S. Verma, J. Dickerson, and K. Hines, “Counterfactual explanations for machine learning: A review,” *arXiv preprint arXiv:2010.10596*, 2020.

-
- [207] R. K. Mothilal, A. Sharma, and C. Tan, “Explaining machine learning classifiers through diverse counterfactual explanations,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 607–617, 2020.
- [208] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” *arXiv preprint arXiv:1312.6199*, 2013.
- [209] C. Xie, M. Tan, B. Gong, J. Wang, A. L. Yuille, and Q. V. Le, “Adversarial examples improve image recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 819–828, 2020.
- [210] R. Jia and P. Liang, “Adversarial examples for evaluating reading comprehension systems,” *arXiv preprint arXiv:1707.07328*, 2017.
- [211] M. Sato, J. Suzuki, H. Shindo, and Y. Matsumoto, “Interpretable adversarial perturbation in input embedding space for text,” *arXiv preprint arXiv:1805.02917*, 2018.
- [212] E. W. Ayers, F. Eiras, M. Hawasly, and I. Whiteside, “Parot: a practical framework for robust deep neural network training,” in *NASA Formal Methods Symposium*, pp. 63–84, Springer, 2020.
- [213] R. Wirth and J. Hipp, “Crisp-dm: Towards a standard process model for data mining,” in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, vol. 1, pp. 29–40, Manchester, 2000.
- [214] The Royal Society, “Explainable AI: the basics POLICY BRIEFING.” https://ec.europa.eu/futurium/en/system/files/ged/ai-and-interpretability-policy-briefing_creative_commons.pdf, 2019.
- [215] M. G. Kahn, T. J. Callahan, J. Barnard, A. E. Bauck, J. Brown, B. N. Davidson, H. Estiri, C. Goerg, E. Holve, S. G. Johnson, *et al.*, “A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data,” *Egems*, vol. 4, no. 1, 2016.
- [216] C. Paterson, R. Calinescu, and R. Ashmore, “Assuring the machine learning lifecycle: Desiderata, methods, and challenges,” *ACM Computing Surveys*, 2021.

- [217] G. E. Batista, R. C. Prati, and M. C. Monard, “A study of the behavior of several methods for balancing machine learning training data,” *ACM SIGKDD explorations newsletter*, vol. 6, no. 1, pp. 20–29, 2004.
- [218] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, “The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies,” *Journal of Biomedical Informatics*, p. 103655, 2020.
- [219] R. Huang, B. Xu, D. Schuurmans, and C. Szepesvári, “Learning with a strong adversary,” *arXiv preprint arXiv:1511.03034*, 2015.
- [220] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, “Robust physical-world attacks on deep learning visual classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1625–1634, 2018.
- [221] P. W. Koh and P. Liang, “Understanding black-box predictions via influence functions,” in *International Conference on Machine Learning*, pp. 1885–1894, PMLR, 2017.
- [222] S. Sharma, J. Henderson, and J. Ghosh, “Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models,” *arXiv preprint arXiv:1905.07857*, 2019.
- [223] IEEE, “IEEE Standard Application of the Single-Failure Criterion to Nuclear Power Generating Station Safety Systems,” *IEEE Std 379-1994*, pp. 1–16, 1994.
- [224] H. Wunsch, J. Wagner, M. Herlim, D. Chong, A. Kramer, and S. D. Halpern, “ICU occupancy and mechanical ventilator use in the united states,” *Critical care medicine*, vol. 41, no. 12, 2013.
- [225] J. Marti, P. Hall, P. Hamilton, S. Lamb, C. McCabe, R. Lall, J. Darbyshire, D. Young, and C. Hulme, “One-year resource utilisation, costs and quality of life in patients with acute respiratory distress syndrome (ards): secondary analysis of a randomised controlled trial,” *Journal of intensive care*, vol. 4, no. 1, p. 56, 2016.

-
- [226] L. M. Cooper and W. T. Linde-Zwirble, “Medicare intensive care unit use: analysis of incidence, cost, and payment,” *Read Online: Critical Care Medicine— Society of Critical Care Medicine*, vol. 32, no. 11, pp. 2247–2253, 2004.
- [227] T. Chockalingam, “Weaning and extubation,” *J Lung Pulm Respir Res*, vol. 2, no. 3, p. 00043, 2015.
- [228] J.-M. Boles, J. Bion, A. Connors, M. Herridge, B. Marsh, C. Melot, R. Pearl, H. Silverman, M. Stanchina, A. Vieillard-Baron, *et al.*, “Weaning from mechanical ventilation,” *European Respiratory Journal*, vol. 29, no. 5, pp. 1033–1056, 2007.
- [229] L. M. Bigatello, H. T. Stelfox, L. Berra, U. Schmidt, and E. M. Gettings, “Outcome of patients undergoing prolonged mechanical ventilation after critical illness,” *Critical care medicine*, vol. 35, no. 11, pp. 2491–2497, 2007.
- [230] C. G. Hughes, S. McGrane, and P. P. Pandharipande, “Sedation in the intensive care setting,” *Clinical pharmacology: advances and applications*, vol. 4, p. 53, 2012.
- [231] A. Esteban, A. Anzueto, F. Frutos, I. Alía, L. Brochard, T. E. Stewart, S. Benito, S. K. Epstein, C. Apezteguía, P. Nightingale, *et al.*, “Characteristics and outcomes in adult patients receiving mechanical ventilation: a 28-day international study,” *Jama*, vol. 287, no. 3, pp. 345–355, 2002.
- [232] D. Wagner, “Economics of prolonged mechanical ventilation,” *American Journal of Respiratory and Critical Care Medicine*, vol. 140, 1989.
- [233] M. J. Tobin, “Advances in mechanical ventilation,” *New England Journal of Medicine*, vol. 344, no. 26, pp. 1986–1996, 2001.
- [234] J. S. Krinsley, P. K. Reddy, and A. Iqbal, “What is the optimal rate of failed extubation?,” *Critical Care*, vol. 16, no. 1, pp. 1–5, 2012.
- [235] J. Whiting, J. Gowardman, D. Huntington, *et al.*, “The effect of extubation failure on outcome in a multidisciplinary australian intensive care unit,” *Critical Care and Resuscitation*, vol. 8, no. 4, p. 328, 2006.
- [236] G. Conti, J. Mantz, D. Longrois, and P. Tonner, “Sedation and weaning from mechanical ventilation: time for ‘best practice’ to catch up with new realities?,” *Multidisciplinary respiratory medicine*, vol. 9, no. 1, p. 45, 2014.

- [237] H. M. Horst, D. Mouro, R. A. Hall-Jenssens, and N. Pamukov, "Decrease in ventilation time with a standardized weaning process," *Archives of Surgery*, vol. 133, no. 5, pp. 483–489, 1998.
- [238] H. Al Mandhari, W. Shalish, E. Dempsey, M. Keszler, and P. Davis, "Po-0726 international survey on peri-extubation practices in extremely premature infants," 2014.
- [239] H.-J. Kuo, H.-W. Chiu, C.-N. Lee, T.-T. Chen, C.-C. Chang, and M.-Y. Bien, "Improvement in the prediction of ventilator weaning outcomes by an artificial neural network in a medical icu," *Respiratory care*, vol. 60, no. 11, pp. 1560–1569, 2015.
- [240] A. Mikhno and C. M. Ennett, "Prediction of extubation failure for neonates with respiratory distress syndrome using the mimic-ii clinical database," in *2012 Annual international conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5094–5097, IEEE, 2012.
- [241] B. Saugel, P. Raketle, A. Hapfelmeier, C. Schultheiss, V. Phillip, P. Thies, M. Treiber, H. Einwächter, A. von Werder, R. Pfab, *et al.*, "Prediction of extubation failure in medical intensive care unit patients," *Journal of critical care*, vol. 27, no. 6, pp. 571–577, 2012.
- [242] A. Gottschalk, M. C. Hyzer, and R. T. Geer, "A comparison of human and machine-based predictions of successful weaning from mechanical ventilation," *Medical Decision Making*, vol. 20, no. 2, pp. 160–169, 2000.
- [243] J. A. Chaparro and B. F. Giraldo, "Power index of the inspiratory flow signal as a predictor of weaning in intensive care units," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 78–81, IEEE, 2014.
- [244] L. J. Kanbar, C. C. Onu, W. Shalish, K. A. Brown, G. M. Sant'Anna, D. Precup, and R. E. Kearney, "Undersampling and bagging of decision trees in the analysis of cardiorespiratory behavior for the prediction of extubation readiness in extremely preterm infants," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4940–4944, IEEE, 2018.

-
- [245] J. A. Chaparro, B. F. Giraldo, P. Caminal, and S. Benito, “Performance of respiratory pattern parameters in classifiers for predict weaning process,” in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4349–4352, IEEE, 2012.
- [246] N. Prasad, L.-F. Cheng, C. Chivers, M. Draugelis, and B. E. Engelhardt, “A reinforcement learning approach to weaning of mechanical ventilation in intensive care units,” *arXiv preprint arXiv:1704.06300*, 2017.
- [247] T. Chen, J. Xu, H. Ying, X. Chen, R. Feng, X. Fang, H. Gao, and J. Wu, “Prediction of extubation failure for intensive care unit patients using light gradient boosting machine,” *IEEE Access*, vol. 7, pp. 150960–150968, 2019.
- [248] G. D. Perkins, D. Mistry, S. Gates, F. Gao, C. Snelson, N. Hart, L. Camporota, J. Varley, C. Carle, E. Paramasivam, *et al.*, “Effect of protocolized weaning with early extubation to noninvasive ventilation vs invasive weaning on time to liberation from mechanical ventilation among patients with respiratory failure: the breathe randomized clinical trial,” *Jama*, vol. 320, no. 18, pp. 1881–1888, 2018.
- [249] A. Esteban, F. Frutos, M. J. Tobin, I. Alía, J. F. Solsona, V. Valverdu, R. Fernández, M. A. de la Cal, S. Benito, R. Tomás, *et al.*, “A comparison of four methods of weaning patients from mechanical ventilation,” *New England Journal of Medicine*, vol. 332, no. 6, pp. 345–350, 1995.
- [250] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.
- [251] C. Kaul, N. Pears, and S. Manandhar, “Sawnet: A spatially aware deep neural network for 3d point cloud processing,” *arXiv preprint arXiv:1905.07650*, 2019.
- [252] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, “Capture, learning, and synthesis of 3d speaking styles,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10101–10111, 2019.
- [253] C. Kaul, N. Pears, and S. Manandhar, “Fatnet: A feature-attentive network for 3d point cloud processing,” *arXiv preprint arXiv:2104.03427*, 2021.

- [254] N. Ambrosino and L. Gabbrielli, “The difficult-to-wean patient,” *Expert review of respiratory medicine*, vol. 4, no. 5, pp. 685–692, 2010.
- [255] R. D. Cook and S. Weisberg, *Residuals and influence in regression*. New York: Chapman and Hall, 1982.
- [256] T. Walsh, S. Dodds, and F. McArdle, “Evaluation of simple criteria to predict successful weaning from mechanical ventilation in intensive care patients,” *British journal of anaesthesia*, vol. 92, no. 6, pp. 793–799, 2004.
- [257] IMDRF SaMD Working Group, “Software as a Medical Device(SaMD): Application of Quality Management System,” International Medical Device Regulators Forum, 2015.
- [258] Y. Sun, X. Huang, D. Kroening, J. Sharp, M. Hill, and R. Ashmore, “Structural test coverage criteria for deep neural networks,” *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 18, no. 5s, pp. 1–23, 2019.
- [259] Z. Li, X. Ma, C. Xu, and C. Cao, “Structural coverage criteria for neural networks could be misleading,” in *2019 IEEE/ACM 41st International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)*, pp. 89–92, IEEE, 2019.
- [260] X. Huang, M. Kwiatkowska, S. Wang, and M. Wu, “Safety verification of deep neural networks,” in *International conference on computer aided verification*, pp. 3–29, Springer, 2017.
- [261] Y. Sun, X. Huang, D. Kroening, J. Sharp, M. Hill, and R. Ashmore, “Deepconcolic: Testing and debugging deep neural networks,” in *2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, pp. 111–114, IEEE, 2019.
- [262] Y. Jia, T. Lawton, J. McDermid, E. Rojas, and I. Habli, “A framework for assurance of medication safety using machine learning,” *arXiv preprint arXiv:2101.05620*, 2021.
- [263] Y. Jia, “Improving medication safety using machine learning,” in *AIME Doctoral Consortium: Artificial Intelligence in Medicine in Europe*, 2019.

-
- [264] J. 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.
- [265] A. Rae, R. Alexander, and J. McDermid, “Fixing the cracks in the crystal ball: A maturity model for quantitative risk assessment,” *Reliability Engineering & System Safety*, vol. 125, pp. 67–81, 2014.
- [266] L. A. Lipsitz, “Understanding health care as a complex system: the foundation for unintended consequences,” *Jama*, vol. 308, no. 3, pp. 243–244, 2012.
- [267] C. Vincent and R. Amalberti, *Safer healthcare: strategies for the real world*. Springer Nature, 2016.
- [268] E. Tissot, C. Cornette, P. Demoly, M. Jacquet, F. Barale, and G. Capellier, “Medication errors at the administration stage in an intensive care unit,” *Intensive care medicine*, vol. 25, no. 4, pp. 353–359, 1999.
- [269] NHS, “The medicines safety improvement programme.” <https://www.england.nhs.uk/patient-safety/national-medicines-safety-programme/>. Accessed 2021-05-20.
- [270] World Health Organization (WHO), “Patient safety: making health care safer,” tech. rep., World Health Organization, 2017.
- [271] D. W. Bates, E. B. Miller, D. J. Cullen, L. Burdick, L. Williams, N. Laird, L. A. Petersen, S. D. Small, B. J. Sweitzer, M. Vander Vliet, *et al.*, “Patient risk factors for adverse drug events in hospitalized patients,” *Archives of internal medicine*, vol. 159, no. 21, pp. 2553–2560, 1999.
- [272] I. Lyons, D. Furniss, A. Blandford, G. Chumbley, I. Iacovides, L. Wei, A. Cox, A. Mayer, J. Vos, G. H. Galal-Edeen, *et al.*, “Errors and discrepancies in the administration of intravenous infusions: a mixed methods multihospital observational study,” *BMJ quality & safety*, vol. 27, no. 11, pp. 892–901, 2018.
- [273] R. Kaushal, D. W. Bates, C. Landrigan, K. J. McKenna, M. D. Clapp, F. Federico, and D. A. Goldmann, “Medication errors and adverse drug events in pediatric inpatients,” *Jama*, vol. 285, no. 16, pp. 2114–2120, 2001.

- [274] D. Furniss, B. Dean Franklin, and A. Blandford, “The devil is in the detail: how a closed-loop documentation system for iv infusion administration contributes to and compromises patient safety,” *Health informatics journal*, vol. 26, no. 1, pp. 576–591, 2020.
- [275] A. J. Ross, T. Murrells, T. Kirby, P. Jaye, and J. E. Anderson, “An integrated statistical model of emergency department length of stay informed by resilient health care principles,” *Safety Science*, vol. 120, pp. 129–136, 2019.
- [276] M. Suján, D. Furniss, D. Embrey, M. Elliott, D. Nelson, S. White, I. Habli, and N. Reynolds, “Critical barriers to safety assurance and regulation of autonomous medical systems,” in *Proceedings of the 29th European safety and reliability conference (ESREL 2019)*, 2019.
- [277] M. Suján, “Safety Assurance of Autonomous Intravenous Medication Management Systems (SAM).” <https://www.york.ac.uk/assuring-autonomy/projects/sam/>. Accessed 2021-05-20.
- [278] N. Paltrinieri, L. Comfort, and G. Reniers, “Learning about risk: Machine learning for risk assessment,” *Safety science*, vol. 118, pp. 475–486, 2019.
- [279] Defence Safety Authority, “Service inquiry, loss of watchkeeper (wk043) unmanned air vehicle over cardigan bay in west wales 24 mar 17,” tech. rep., DSA/DAIB/17/006, 2019.
- [280] E. Hollnagel, *Safety-I and safety-II: the past and future of safety management*. CRC press, 2018.
- [281] L.-T. Chen and C.-Y. Jiang, “Impact of atrial arrhythmias after esophagectomy on recovery: a meta-analysis,” *Medicine*, vol. 97, no. 23, 2018.
- [282] T. Ojima, M. Nakamori, M. Nakamura, M. Katsuda, K. Hayata, T. Kato, J. Kitadani, H. Tabata, A. Takeuchi, and H. Yamaue, “Randomized clinical trial of landomolol hydrochloride for the prevention of atrial fibrillation and postoperative complications after oesophagectomy for cancer,” *Journal of British Surgery*, vol. 104, no. 8, pp. 1003–1009, 2017.
- [283] S. Dixit, “Atrial fibrillation after major thoracic surgery,” *Journal of the American College of Cardiology*, vol. 54, no. 22, pp. 2049–2051, 2009.

- [284] S. P. Stawicki, M. P. Prosciak, A. T. Gerlach, M. Bloomston, H. T. Davido, D. E. Lindsey, M. E. Dillhoff, D. C. Evans, S. M. Steinberg, and C. H. Cook, “Atrial fibrillation after esophagectomy: an indicator of postoperative morbidity,” *General thoracic and cardiovascular surgery*, vol. 59, no. 6, pp. 399–405, 2011.
- [285] J.-H. Chin, Y.-J. Moon, J.-Y. Jo, Y. A. Han, H. R. Kim, E.-H. Lee, and I.-C. Choi, “Association between postoperatively developed atrial fibrillation and long-term mortality after esophagectomy in esophageal cancer patients: an observational study,” *PLoS One*, vol. 11, no. 5, p. e0154931, 2016.
- [286] E. J. Benjamin, P. A. Wolf, R. B. D’Agostino, H. Silbershatz, W. B. Kannel, and D. Levy, “Impact of atrial fibrillation on the risk of death: the framingham heart study,” *Circulation*, vol. 98, no. 10, pp. 946–952, 1998.
- [287] Memorial Sloan Kettering Cancer Center, “Esophagectomy Pathway.” https://www.mskcc.org/sites/default/files/node/20707/document/esophagectomy-pathway_080715.pdf. Accessed 2021-05-20.
- [288] S. C. Tomaszek, S. D. Cassivi, M. S. Allen, K. R. Shen, F. C. Nichols III, C. Deschamps, and D. A. Wigle, “An alternative postoperative pathway reduces length of hospitalisation following oesophagectomy,” *European journal of cardio-thoracic surgery*, vol. 37, no. 4, pp. 807–813, 2010.
- [289] G. H. Berkelmans, B. J. Wilts, E. A. Kouwenhoven, K. Kumagai, M. Nilsson, T. J. Weijs, G. A. Nieuwenhuijzen, M. J. van Det, and M. D. Luyer, “Nutritional route in oesophageal resection trial ii (nutrient ii): study protocol for a multicentre open-label randomised controlled trial,” *BMJ open*, vol. 6, no. 8, 2016.
- [290] N. G. Leveson and J. P. Thomas, “STPA handbook,” *Cambridge, MA, USA*, 2018.
- [291] T. Leslie-Mazwi, J. Bello, R. Tu, G. Nicola, W. Donovan, R. Barr, and J. Hirsch, “Current procedural terminology: history, structure, and relationship to valuation for the neuroradiologist,” *American Journal of Neuroradiology*, vol. 37, no. 11, pp. 1972–1976, 2016.
- [292] N. Friedman and D. Koller, “Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks,” *Machine learning*, vol. 50, no. 1, pp. 95–125, 2003.

- [293] W. Buntine, “Theory refinement on bayesian networks,” in *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, UAI’91, (San Francisco, CA, USA), p. 52–60, Morgan Kaufmann Publishers Inc., 1991.
- [294] A. Clemente and F. Carli, “The physiological effects of thoracic epidural anesthesia and analgesia on the cardiovascular, respiratory and gastrointestinal systems,” *Minerva Anestesiol*, vol. 74, no. 10, pp. 549–63, 2008.
- [295] M. A. Gerhardt, V. B. Gunka, and R. J. Miller, “Hemodynamic stability during labor and delivery with continuous epidural infusion,” *The Journal of the American Osteopathic Association*, vol. 106, no. 12, pp. 692–698, 2006.
- [296] A. Laupacis, D. L. Sackett, and R. S. Roberts, “An assessment of clinically useful measures of the consequences of treatment,” *New England journal of medicine*, vol. 318, no. 26, pp. 1728–1733, 1988.
- [297] C. A. Chong, G. Tomlinson, L. Chodirker, N. Figdor, M. Uster, G. Naglie, and M. D. Krahn, “An unadjusted nnt was a moderately good predictor of health benefit,” *Journal of clinical epidemiology*, vol. 59, no. 3, pp. 224–233, 2006.
- [298] C. Cave, “An independent review into the broader issues surrounding the loss of the raf nimrod mr2 aircraft xv230 in afghanistan in 2006,” *The Stationary Office, Tech. Rep*, 2006.
- [299] T. Kelly, “Are safety cases working,” *Safety Critical Systems Club Newsletter*, vol. 17, no. 2, pp. 31–33, 2008.
- [300] I. Habli and T. Kelly, “Safety case depictions vs. safety cases—would the real safety case please stand up?,” in *2nd Institution of Engineering and Technology International Conference on System Safety*, pp. 245–248, IET, 2007.
- [301] 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.
- [302] H. Steck, “Learning the bayesian network structure: Dirichlet prior versus data,” *arXiv preprint arXiv:1206.3287*, 2012.

- [303] T. Silander, P. Kontkanen, and P. Myllymaki, “On sensitivity of the map bayesian network structure to the equivalent sample size parameter,” *arXiv preprint arXiv:1206.5293*, 2012.
- [304] E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro, “Process mining in healthcare: A literature review,” *Journal of biomedical informatics*, vol. 61, pp. 224–236, 2016.
- [305] E. Hollnagel, “Why is work-as-imagined different from work-as-done?,” in *Resilient health care, Volume 2*, pp. 249–264, CRC Press, 2015.