Automation of Patient Trajectory Management

A deep-learning system for critical care outreach

GEORGINA KENNEDY

B.Eng (Hons), M.Eng, M.Res

Supervisor: Assoc. Prof. Blanca Gallego Luxan
Associate Supervisor: Assoc. Prof. Mark Dras

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Faculty of Medicine — Centre for Big Data Research in Health
The University of New South Wales
Sydney, Australia

23 June 2021

Mandatory Statements

THESIS TITLE & ABSTRACT

Thesis Title

Automation of Patient Trajectory Management: A deep-learning system for critical care outreach

Thesis Abstract

The application of machine learning models to big data has become ubiquitous, however their successful translation into clinical practice is currently mostly limited to th e field of imaging. Despite much interest and promise, there are many complex and interrelated barriers that exist in clinical settings, which must be addressed systemati cally in advance of wide-spread adoption of these technologies. There is limited evidence of comprehensive efforts to consider not only their raw performance metrics, b ut also their effective deployment, particularly in terms of the ways in which they are perceived, used and accepted by clinicians.

The critical care outreach team at St Vincent's Public Hospital want to automatically prioritise their workload by predicting in-patient deterioration risk, presented as a wa tch-list application. This work proposes that the proactive management of in-patients at risk of serious deterioration provides a comprehensive case-study in which to un derstand clinician readiness to adopt deep-learning technology due to the significant known limitations of existing manual processes.

Herein is described the development of a proof of concept application uses as its input the subset of real-time clinical data available in the EMR. This data set has the no teworthy challenge of not including any electronically recorded vital signs data. Despite this, the system meets or exceeds similar benchmark models for predicting in-pa tient death and unplanned ICU admission, using a recurrent neural network architecture, extended with a novel data-augmentation strategy.

This augmentation method has been re-implemented in the public MIMIC-III data set to confirm its generalisability. The method is notable for its applicability to discrete t ime-series data. Furthermore, it is rooted in knowledge of how data entry is performed within the clinical record and is therefore not restricted in applicability to a single c linical domain, instead having the potential for wide-ranging impact.

The system was presented to likely end-users to understand their readiness to adopt it into their workflow, using the Technology Adoption Model. In addition to confirmin g feasibility of predicting risk from this limited data set, this study investigates clinician readiness to adopt artificial intelligence in the critical care setting. This is done wit h a two-pronged strategy, addressing technical and clinically-focused research questions in parallel.

ORIGINALITY, COPYRIGHT AND AUTHENTICITY STATEMENTS

ORIGINALITY STATEMENT

☑ I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

COPYRIGHT STATEMENT

☑ I hereby grant the University of New South Wales or its agents a non-exclusive licence to archive and to make available (including to members of the public) my thesis or dissertation in whole or part in the University libraries in all forms of media, now or here after known. I acknowledge that I retain all intellectual property rights which subsist in my thesis or dissertation, such as copyright and patent rights, subject to applicable law. I also retain the right to use all or part of my thesis or dissertation in future works (such as articles or books).

For any substantial portions of copyright material used in this thesis, written permission for use has been obtained, or the copyright material is removed from the final public version of the thesis.

AUTHENTICITY STATEMENT

☑ I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis.

INCLUSION OF PUBLICATION STATEMENTS

☑ The candidate has declared that **their thesis has publications - either published or submitted for publication - incorporated into it in lieu of a Chapter/s. Details of these publications are provided below.**.

Publication Details #1

Full Title:	Clinical prediction rules: a systematic review of healthcare provider opinions and preferences.
Authors:	Kennedy, G., Gallego, B.
Journal or Book Name:	International journal of medical informatics
Volume/Page Numbers:	123, 1–10
Date Accepted/Published:	Accepted 11 December 2018, Available online 12 December 2018.
Status:	published
The Candidate's Contribution to the Work:	GK: Study design, search piloting, screening, data extraction, thematic analysis, manuscript preparation BG: Screening, thematic analysis iteration, project supervision
Location of the work in the thesis and/or how the work is incorporated in the thesis:	This publication has been incorporated as chapter 2 of the thesis (Systematic Review)

Publication Details #2

Full Title:	Developing a deep learning system to drive the work of the critical care outreach team
Authors:	Kennedy, G., Rihari-Thomas, J., Dras, M., Gallego, B.,
Journal or Book Name:	BMC Medical Informatics and Decision Making
Volume/Page Numbers:	Submitted - under review
Date Accepted/Published:	Submitted - under review
Status:	submitted
The Candidate's Contribution to the Work:	GK: Data analysis, model development, model architecture conception, manuscript preparation JRT: Clinical guidance and review MD: Technical guidance and review BG: Overall guidance and direction of model development, project supervision
Location of the work in the thesis and/or how the work is incorporated in the thesis:	This manuscript has been incorporated in the thesis as chapter 4. A preprint has been made available on medrxiv.

Publication Details #3

Full Title:	Augmentation of electronic medical record data for deep learning
Authors:	Kennedy, G., Dras, M., Gallego, B.,
Journal or Book Name:	MEDINFO conference proceedings
Volume/Page Numbers:	Submitted - under review
Date Accepted/Published:	Submitted - under review
Status:	submitted
The Candidate's Contribution to the Work:	GK: Data analysis, model development, study design, manuscript preparation MD: Technical guidance and review BG: Overall guidance, project supervision
Location of the work in the thesis and/or how the work is incorporated in the thesis:	This manuscript has been incorporated in the thesis as chapter 5. A preprint has been made available on medrxiv. Note that MEDINFO makes full papers available as part of their conference proceedings.

Publication Details #4

Full Title:	Clinician readiness to adopt A.I. for critical care prioritisation
Authors:	Kennedy, G., Gallego, B.
Journal or Book Name:	HIC conference proceedings
Volume/Page Numbers:	Call for papers is yet to open for 2021, but pre-print has been made available that has been formatted for submission to this venue.
Date Accepted/Published:	
Status:	submitted
The Candidate's Contribution to the Work:	GK: Study design, data capture, data analysis, manuscript preparation BG: Overall guidance, project supervision
Location of the work in the thesis and/or how the work is incorporated in the thesis:	This manuscript has been incorporated in the thesis as chapter 7. A preprint has been made available on medrxiv. Note that HIC makes full papers available as part of their conference proceedings.

Candidate's Declaration

I confirm that where I have used a publication in lieu of a chapter, the listed publication(s) above meet(s) the requirements to be included in the thesis. I also declare that I have complied with the Thesis Examination Procedure.

Acknowledgements

To my dear Andrew, I quite literally could not have done this without you. I know that you know how I feel about your support, and how grateful I am, but I hope that everyone else realises it too.

To my girls, I definitely could have done this without you — probably in half the time — but it wouldn't have been half as fun, nor would I have cared anywhere near as much.

To Blanca and Mark, I am grateful for all your supervision and guidance. You gave me sufficient freedom to allow me to find my own feet, but coupled this with strong guidance when it was required — I greatly appreciate all the myriad ways in which you shared your wisdom and shaped both my work and the learnings that I can take away from this process.

To my wonderful parents, sisters and in-laws, your practical support was of course a key factor in my ability to get this done, but more than that, your genuine pleasure in every small success and heartbreak with each setback experienced along the way kept me going.

To the many colleagues and friends at St Vincent's Hospital, Macquarie University and UNSW, your guidance, encouragement and collegiality has been invaluable. For their technical guidance, I would like to thank in particular Tim Churches, Richard Green and Juan Quiroz. I am also grateful to the clinicians who helped ground my study in their real practice — especially Rosemary Kennedy, Lauren McFarlane-Bentley and John Rihari-Thomas. For your enthusiasm and encouragement, Ly Tong and Liliana Laranjo, you were invaluable.

I also want to acknowledge the privilege it has been live in a society where I can even approach this piece of work, with the infrastructure and financial support that allowed me to do so. I do not take this for granted in the slightest, and hope that I can contribute back many-fold.

Executive Summary

In recent years, the application of machine learning models to big data has become ubiquitous, however their successful translation into clinical practice is currently limited to the field of imaging for the most part. Despite much interest and promise, there are many complex and interrelated barriers that exist in clinical settings, which must be addressed systematically in advance of any wide-spread adoption of these technologies. There is limited evidence in the literature of any comprehensive efforts to approach clinical prediction problems in a way that considers not only their raw performance metrics, but also their effective deployment, particularly in terms of the ways in which they are perceived, used and accepted by clinicians.

The genesis of this body of work came from the critical care outreach team at St Vincent's Public Hospital, Sydney. They want to be able to automatically prioritise their workload through the prediction of in-patient deterioration risk, presented in the form of a watch-list application. This work proposes that the proactive management of in-patients at risk of serious deterioration provides a comprehensive case-study in which to understand clinician readiness to adopt deep-learning technology due to the significant known limitations of existing manual processes.

Herein is described the development of a proof of concept application that meets the requirements as stated by critical care stakeholders. This system is based around models that use as their input the subset of real-time clinical data that is available in the electronic medical record at the target institution. This dataset has the noteworthy challenge of not including any electronically recorded vital signs data. Despite this, the system meets or exceeds similar benchmark models for predicting in-patient death and unplanned ICU admission, using a recurrent neural network architecture that has been extended with a novel data-augmentation strategy.

This data-augmentation method has been re-implemented for demonstration in the publicly available MIMIC-III dataset in order to establish both its specific effect and

generalisability of the technique. This novel augmentation method is notable for its applicability to discrete time-series data. Furthermore, it is rooted in knowledge of how data entry is performed within the clinical record and is therefore not restricted in applicability to a single clinical domain, instead having the potential for wide-ranging impact.

This system was then presented to likely end-users in order to understand their readiness to adopt this technology into their workflow, with the use of the Technology Adoption Model.

In addition to the confirmation of feasibility of generating suitable predictions of risk from this limited dataset, this study presents an investigation of clinician readiness to adopt artificial intelligence in the critical care setting, specifically for the purpose of supporting the critical care outreach role. This is done with a two-pronged strategy, addressing technical and clinically-focused research questions in parallel. The overarching aim is to deliver a proposed system that is demonstrated not only to meet a technical benchmark for statistical performance, but also one which meets a real clinical need, and has been designed in such a way that it is ready for acceptance by clinical end-users.

List of Abbreviations

AI Artificial Intelligence

ATT Attitude towards use construct from the Technology Adoption Model

AUROC Area Under the Receiver Operating Characteristic curve

AWS Amazon Web Services

BI Behavioural Intention to use construct from the Technology Adoption Model

CCO Critical Care Outreach

CCOM Critical Care Outreach Medical Officer

CCON Critical Care Outreach Nurse

CCU Critical Care Unit

CERS Clinical Emergency Response System

CFI Comparative Fit Index

CPR Clinical Prediction Rule

CPU Central Processing Unit

DIA Dialysis unit

ED Emergency Department

EM Emergency Medicine

EMR Electronic Medical Record

EOL End of Life

EWS Early Warning Score

FHIR Fast Healthcare Interoperability Resources

GAN Generative Adversarial Network

GP General Practitioner

GPU Graphical Processing Unit

GRU Gated Recurrent Unit

HL7 Health Level 7 messaging standard

ICD International Classification of Diseases

ICU Intensive Care Unit

LIME Local Interpretable Model-Agnostic Explanations

LRAR Low Risk Ankle Rule

LSTM Long Short Term Memory unit

MBS Medicare Benefits Schedule

MET Medical Emergency Team

NASSS Framework to explain non-adoption, abandonment, and failure to scale, spread and reach sustainability

NEWS National Early Warning Score

NLP Natural Language Processing

OAR Ottawa Ankle Rule

OBR HL7 Observation Request segment

OBX HL7 Observation Result segment

ORM Object Relational Model

PEOU Perceived Ease of Use construct from the Technology Adoption Model

PU Perceived Utility construct from the Technology Adoption Model

RMSEA Root Mean Square Error of Approximation

RNN Recurrent Neural Network

SHAP SHapley Additive exPlanations

SMOTE Synthetic Minority Over-Sampling Technique

SQL Structured Query Language

SVH St Vincent's Public Hospital, Sydney

TAM Technology Adoption Model

TTE Time to Event

WDR Workup to Detection Ratio

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Background

For a patient in an acute care setting, there are many complex and interrelated factors that affect their likely trajectory toward either recovery or deterioration. Prior to significant deterioration events, there are observable patterns in clinical features that indicate this change in acuity [1–4]. These warning signs may be present as much as 48 hours prior to the adverse outcome [1], however they are often overlooked.

In addition, there is evidence that sub-optimal care (including delayed or missed interventions) in general hospital wards is a key contributing factor to both unplanned ICU admissions and preventable inpatient mortality [5, 6].

These factors have combined to drive the modern desire for tools and processes that can faithfully highlight patients at risk of deterioration on the general wards such that interventions can be deployed sooner, improving both patient outcomes and resource utilisation.

1.1.1 Context

1.1.1.1 Clinical Emergency Response to the Deteriorating Patient

A Clinical Emergency Response System (CERS) is defined by the Clinical Excellence Commission as the established procedures for escalation of care for the deteriorating

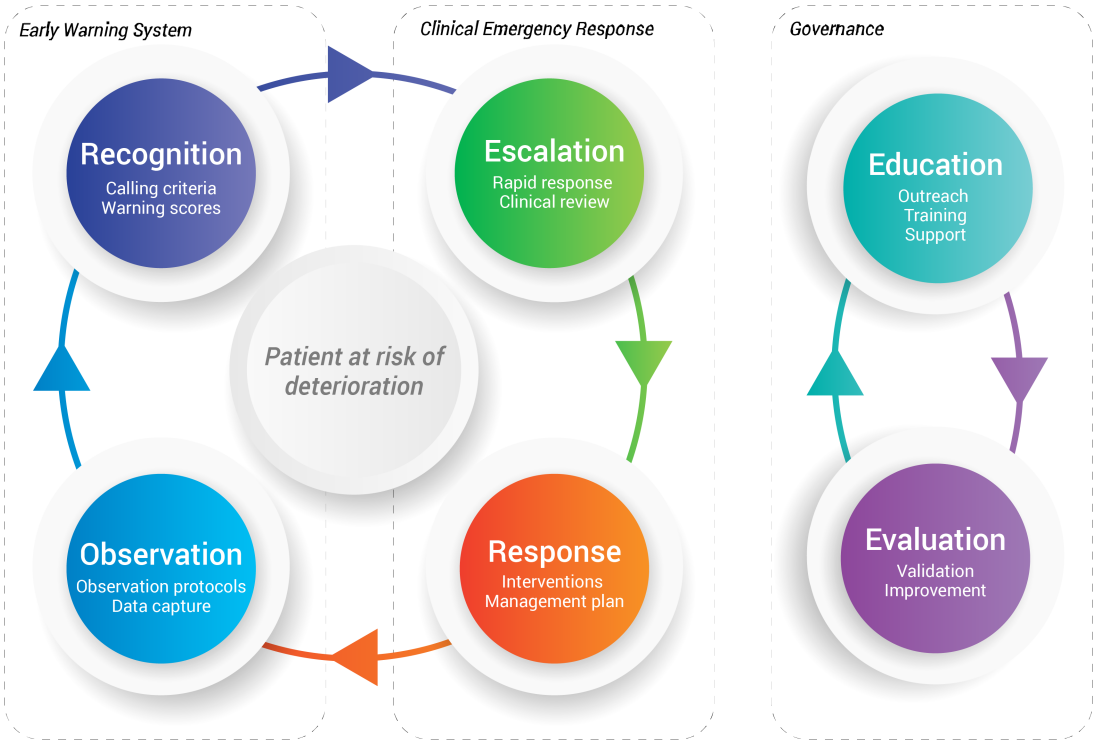


FIGURE 1.1. Recognising and managing the deteriorating patient

patient, based on standard calling criteria [7]. This system is defined locally, according to general principles and standards that are mandated at the state level.

Such systems were developed in the early 1990s as an expansion of the existing practice of dedicated resuscitation teams [8], and have become common globally. Their precise terminology varies somewhat (e.g. resuscitation team vs. medical emergency team vs. cardiac response team), despite this common conceptual grounding. Note that where there is inconsistency, this work defers to definitions that are applicable in the domain of public hospitals in New South Wales, Australia.

More generally, there are three key components to a comprehensive patient deterioration management system, as seen in Figure 1.1 - the early warning system (EWS), the clinical emergency response protocols, and the associated governance structures required to enact them.

TABLE 1.1. CERS classifications at SVH [11]

Level	Target	Triggers	Response	Responders
Code Blue	Immediately life threatening	Cardiac arrest Airway obstruction Unresponsive patient	Immediate	MET
Rapid Response	Urgent review required	Vital sign observations in 'red zone' Vital sign observations in 'yellow zone' that reflect patient deterioration	30 min	Rapid response medical officer
Clinical Review	Patient review required	Vital sign observations in 'yellow zone'	30 min	Attending medical team or designated responder

An EWS observes patients according to a defined protocol, allowing evaluation of their condition against set criteria in order to trigger an emergency response. Together, the EWS and CERS define the interactions between general ward staff and specialist emergency responders in order to affect patient stasis in a cycle of recognition and response [9].

The medical emergency team (MET) is made up of nursing and medical staff who have received specialist resuscitation training [10] who are required to respond in a medical emergency. The calling protocol in the target institution consists of three levels of clinical urgency (see Table 1.1). Additionally, all levels of emergency response may be also triggered by staff, patient or family concern. This tiered protocol was introduced in 2009.

1.1.1.2 Pre-emergency Management of Deterioration Risk

Beyond this definition of emergency response protocols, which has become widespread, many institutions now implement an outreach functionality that draws on resources within the critical care unit to proactively manage patients at risk of deterioration on the ward.

The purpose of this role is to integrate critical care skills into the general care wards. This is done by providing resources to follow those discharged from intensive care unit (ICU)

beds to support recovery and anticipating deterioration that could potentially be averted in order to reduce unplanned ICU admissions [12, 13].

The task of preempting patient deterioration across the entire hospital is one that requires awareness of vastly more patients and events than is typically required of clinical staff. For this role to be effectively prioritised and directed, an existing EWS must be enriched to include patients prior to the point at which deterioration triggers an emergency response [14].

1.1.1.3 Setting

St Vincent's Public Hospital (SVH) is a major, government funded, quaternary care hospital in urban Sydney, Australia. As with many hospitals, they have developed strategies to reduce preventable ICU admissions. Two critical care outreach (CCO) roles (one nursing and one medical) have been created to bridge the gap between the ICU and the wards. This is a role in addition to the MET and is designed not only to respond to emergency situations, but also to manage care proactively on the wards and anticipate the needs of patients relative to the available ICU resources.

To support these roles, a 'watch-list' has been developed that generates a list of patients who are deemed to be at risk of deterioration, with the intention that these patients can form the basis of discussion between ward and CCO staff at handover time, and also to drive the work priorities of the CCO staff. This watch-list is not intended to predict emergency situations appropriate for MET responses, rather highlight patients who require additional monitoring and care coordination. Currently, this watch-list is generated heuristically based on rapid response calls and patient movements in the preceding 8-24 hours. There has been limited uptake of this watch-list, due in part to its perceived lack of relevance stemming from the inclusion of patients clearly not at risk. For example, the inclusion of all patients transferred into the hospital from the Emergency Department is of insufficient granularity to helpfully identify deteriorating patients.

1.2 Objective

This project was initiated to automate the generation of a new watch-list that is able to include a richer set of input factors, in order to make the risk estimation more relevant to the goal of of detecting and responding to patient deterioration. The dataset used for this work comprises hospital administrative data plus medication and pathology records. This represents the subset of patient data that is available in real-time at this institution, and therefore can be reasonably used for a prediction model.

This work will describe the design and development of a prototype for a fully automated system that can augment the existing manual clinical emergency response system at St Vincent's Public Hospital.

1.2.1 Research Questions

This thesis has a dual focus on both the clinical and technological domains, with the goal of presenting a system that has credible capacity for both translation and clinical utility.

1.2.1.1 Technical

T.1 Determine an appropriate modeling architecture that can, in principle, identify patients at risk of deterioration in the short term from the clinical record, in real-time and without access to vital signs data.

T.2 Measure how well such an architecture can generalise within the target institution.

1.2.1.2 Clinical

C.1 Understand the qualities of predictive models that are most valued by clinical end-users.

C.2 Apply these qualities to the delivered model as a prototype, and measure the success of this application as perceived by likely stakeholders.

1.3 Approach

In order to formulate responses to the proposed research questions, the research approach is structured as follows:

Chapter 1 - Introduction: Provides an overview of the setting, motivations and high-level results of the project.

Chapter 2 - Systematic Review (C.1): In order to understand the relationship of end users with predictive models in a hospital setting, a systematic review of qualitative research is described. This provides a clinical use-case and foundation for the overall design and implementation of the technical solution presented in later chapters.

Chapter 3 - Technical Architecture (T.1): An in-depth description of the data flow, analytic methods and processing architecture that has been implemented.

Chapter 4 - Model Results (T.1, T.2): This chapter is a stand-alone work describing the delivered model, within the context of its use-case.

Chapter 5 - Data Augmentation for Discrete Time-Series Data (T.1, T.2): This chapter presents a novel strategy for data augmentation. It also acts as a validation of the primary model, as the core novel elements of the data processing pipeline and model development are re-implemented in a publicly available dataset.

Chapter 6 - Watch-list User Interface (C.2): This chapter proposes a concrete implementation interface for the models described in Chapter 4.

Chapter 7 - Face-validity Study (C.2): In which the technical output of this work is presented to likely end-users of such an application, for the purpose of determining its

perceived value within their clinical workflow. This captured via responses to a structured web questionnaire that is based on the Technology Acceptance Model. [15].

Chapter 8 - Discussion: Ties together individual elements of the project as they relate back to the proposed research questions and provides analysis of the strengths and limitations of this work. Summarises the findings of this project, presenting a case for the translation of this model into practice.

1.4 Publication notes

Chapters 2, 4, 5 and 7 have been published, submitted for publication, or prepared for submission as follows in Table 1.2.

Ch.	Status	Reference	Author contributions
2	Published	**Kennedy, G.**, Gallego, B., Clinical prediction rules: a systematic review of healthcare provider opinions and preferences. *International journal of medical informatics* **123**, 1–10 (2018). [16]	GK: Study design, search piloting, screening, data extraction, thematic analysis, manuscript preparation BG: Screening, thematic analysis iteration, project supervision
4	Submitted to BMC Medical Informatics and Decision Making	**Kennedy, G.**, Rihari-Thomas, J., Dras, M., Gallego, B., Developing a deep learning system to drive the work of the critical care outreach team *medRxiv*, (2020). [17]	GK: Data analysis, model development, model architecture conception, manuscript preparation JRT: Clinical guidance and review MD: Technical guidance and review BG: Overall guidance and direction of model development, project supervision
5	Submitted to MEDINFO	**Kennedy, G.**, Dras, M., Gallego, B., Augmentation of electronic medical record data for deep learning *medRxiv*, (2021) [18]	GK: Data analysis, model development, study design, manuscript preparation MD: Technical guidance and review BG: Overall guidance, project supervision
7	Preprint available - to be submitted to HIC 2021 upon opening of call for submissions	**Kennedy, G.**, Gallego, B., Clinician readiness to adopt A.I. for critical care prioritisation *medRxiv*, (2021) [19]	GK: Study design, data capture, data analysis, manuscript preparation BG: Overall guidance, project supervision

TABLE 1.2. Publication Notes

A preamble has been added to each of these chapters in order to provide necessary context within the body of work. There are minor modifications in formatting and cross-referencing in order to preserve the coherence of the entire thesis, but these chapters have been included here with their content otherwise exactly as published / submitted.

CHAPTER 2

Clinical prediction rules: Systematic review of health-care provider opinions & preferences

---

2.1 Preamble

This publication is reproduced exactly as published in [16], with the exception of this preamble.

In the face of limited evidence of successful translation of comprehensive prediction models incorporating the full breadth of the clinical record, the purpose of this chapter is to seek to understand the more mature technology of the clinical prediction rule, and from this to extrapolate the requirements of clinical end-users of the watch-list technology.

2.2 Abstract

Objective. The act of predicting clinical endpoints and patient trajectories based on past and current states is on the precipice of a technological revolution. This systematic review summarises the available evidence describing healthcare provider opinions and preferences with respect to the use of clinical prediction rules. The primary goal of this work is to inform the design and implementation of future systems, and secondarily to identify gaps for the development of clinician education programs.

Methods. Five databases were systematically searched in May 2016 for studies collecting empirical opinions of healthcare providers regarding clinical prediction rule usage.

Reference lists were scanned for additional eligible materials and an update search was made in August 2017. Data was extracted on high-level study features, before in-depth thematic analysis was performed.

Conclusions. Some of the objections and preferences stated by healthcare providers are inherent to the nature of the clinical problem addressed, which may or may not be within the designer's capacity to change; however, others (in particular — actionability, validation, integration and provision of high quality education materials) should be considered by prediction rule designers and implementation teams, in order to increase user acceptance and improve uptake of these tools. We summarise these findings across the clinical prediction rule lifecycle and pose questions for the rule developers, in order to produce tools that are more likely to successfully translate into clinical practice.

2.3 Introduction

Two patients present to the same hospital, scheduled to undergo identical procedures at the hands of equally skillful and qualified surgeons. One recovers speedily, while the other struggles with major complications requiring complex interventions. The benefit of a reliable method to describe in advance the likelihood of each of these trajectories for a particular patient is clear. Credible foreknowledge of expected outcomes and individual response to treatment can inform decision making of both clinicians and patients, allow for responsive resource allocation to eliminate waste and improve outcomes for high-risk patients, and more accurately benchmark performance of facilities based on their risk-adjusted case-mix than has been possible in the past.

The current technical and infrastructural capacity for predictive analytics seen routinely in other fields exceeds what is implemented in typical clinical practice [20, 21], although significant progress is anticipated [22]. Even in institutions with advanced systems, healthcare data are plagued by technical and procedural limitations that inhibit successful big data analysis. This follows a familiar story in clinical information technology projects, which have typically been shaped by slow uptake, reluctant user acceptance, organisational and training issues, decentralised implementation and a piecemeal design approach [23–

25]. On the assumption that the data exists and is available in a timely fashion, however, there is evidence for the ability to predict patient outcomes with high accuracy [26].

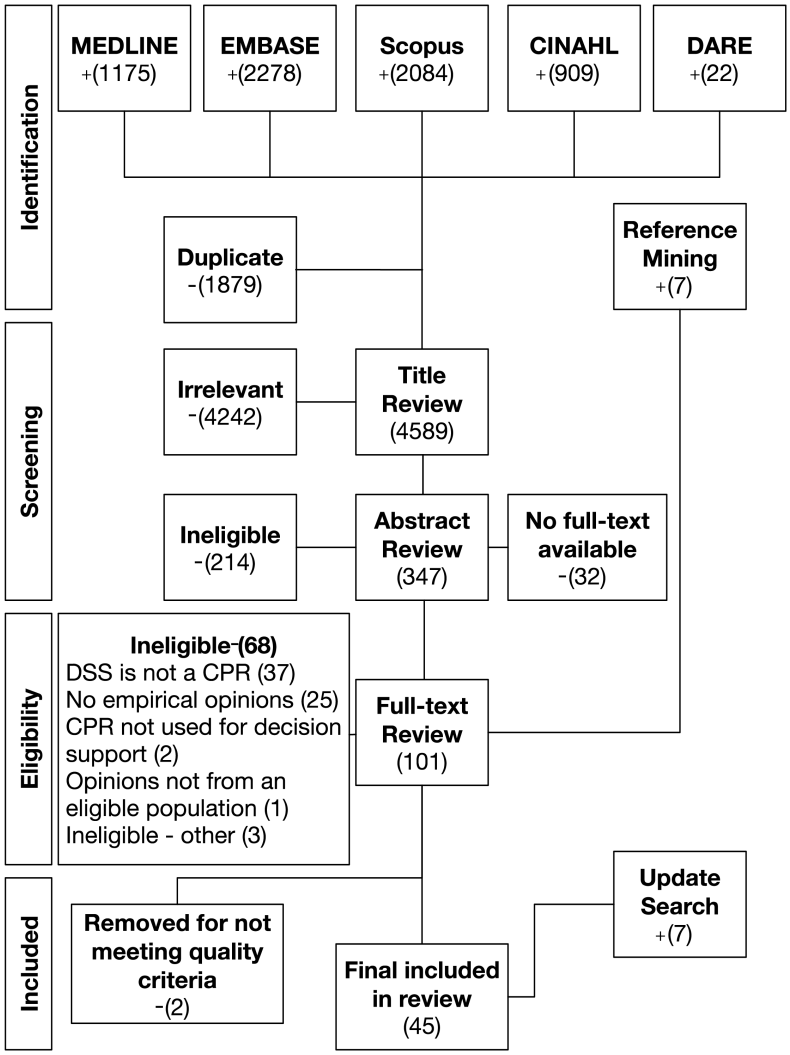
Informal prediction forms the foundation of clinical practice — patients are continually compared to a physician's experience and available knowledge base. Likewise, the practice of evidence-based medicine is definitionally predictive in nature [27] — interventions are applied based on their likelihood for success, established through prior observed outcomes within patient populations. Although more advanced systems have been proposed [21, 28], a common way in which prediction is formally applied in a typical clinical setting is through the use of clinical prediction rules (also known as decision models or risk scores — see definition in Table 2.1) [29]. These rules help clinicians synthesise clinical characteristics with the evidence base and produce a likely diagnosis, risk profile or recommendation for intervention for their patient.

We propose that it is a valuable and timely enterprise to understand the current and future role of CPR in clinical practice, as understood by clinicians. Investment in larger scale predictive analytics projects may be wasteful unless this comparatively simple relationship can be navigated successfully through to translated outcomes. To this end, this paper presents a systematic review of the perspectives of healthcare personnel on CPR. We will use this review to identify characteristics of successfully implemented and broadly used CPRs.

2.3.1 Outline

The remainder of this article is organised as follows. Section 2.4 describes the search strategy, data extraction and data synthesis that was performed. Section 2.5 contains details of specific characteristics of the included papers and their subject CPRs. Section 2.6 then details the extracted themes and describes the context of these themes within the included papers. Finally Sections 2.7 and 2.8 summarise this work and provide conclusions and recommendations.

FIGURE 2.1. PRISMA flow diagram.



2.4 Methods

The protocol for this review was developed in advance, and has been registered as PROSPERO ID 42016039098.

2.4.1 Search strategy

A systematic search of the literature regarding predictive models is challenging, due to lack of standard terms. Therefore, Ingui & Rogers [30] developed and tested a search strategy to retrieve studies of CPR from MEDLINE (since updated [31]).

After testing these searches, however, it became clear that a high proportion of known eligible papers were missed. This was because primarily papers describing the development of new CPR were returned, rather than the qualitative studies that are the target of this review.

Search terms were defined to capture studies which jointly address two high-level concepts (1) clinician attitudes, preferences and practices and (2) prediction of risk via a model or decision rule. See Table 2.1 for the PICOS strategy and Appendix 8.3.2 for final searches used after extensive piloting.

The literature was searched in May 2016 and an update search performed in August 2017. After screening for duplicates and eligibility, and adding references obtained from reference mining, 45 eligible papers were included in the final review (Figure 2.1).

Studies were screened independently by authors GK and BG, with discrepancies resolved via consensus. The inter-rater agreement was excellent, with a Cohen's kappa statistic of $\kappa = 0.84$ and $\kappa = 0.88$ for abstract and full-text phases respectively.

2.4.2 Data extraction

A data extraction form was developed a priori to capture high-level study characteristics. The results of this extraction can be seen in Tables 2.2, 2.3 and 2.4.

2.4.3 Data synthesis

Thematic coding was performed using the NVivo qualitative research software (version 11.3) by author GK. This coding was reviewed iteratively with feedback by BG until saturation was achieved, and a meaningful hierarchy of codes was developed.

TABLE 2.1. PICOS Search Strategy

Component	**Details**
Population	Healthcare providers - including physicians, non-physician clinicians & health-care administrators.
Intervention	Use or knowledge of a clinical prediction rule in current clinical practice. This includes hypothetical prediction rules (such as studies gathering types of rules that healthcare personnel wish to see developed or potential perceived barriers).
Comparator	Existing accepted best practice
Outcome	Empirically gathered healthcare provider opinions
Setting	All healthcare settings
Study Design	Qualitative studies, including surveys, interviews, focus groups and usability testing.
Other Eligibility Criteria	Patient opinions are excluded, except as reported secondarily by healthcare personnel. Studies that describe only the validation of the rule itself and not its design or implementation are excluded.
Data Sources	MEDLINE, EMBASE, Scopus, CINAHL and DARE databases were searched.
Reference Mining	Reference lists were scanned for all papers included in the full-text review. Potentially relevant papers were then also reviewed for eligibility.
Definition	For these purposes, we define CPR broadly - as a proceduralised effort (automated or otherwise) that assesses the current or historical characteristics of a patient in order to derive either an estimate of the future risk of target outcomes (prognostic), the likelihood of a current specified disease state (diagnostic), or likely response to treatment (therapeutic).
Dates	No date restrictions were applied to searches

2.5 Results

2.5.1 Characteristics of included studies

TABLE 2.2. Study setting

Country*		Recruitment Setting*		Study Methods*	
Australia	10	Academic hospital(s)	4	Focus Group	9
Canada	6	Academic primary care network	2	Semi-structured interview	14
France	2	Educational institution	2	Survey	28
Germany	1	Non-academic hospital(s)	3	Usability testing	4
Netherlands	2	Non-academic primary care network	14		
Spain	2	Professional body membership	14		
Switzerland	1	Targeted approach	2		
UK	11	Enrolment in an existing RCT	2		
USA	17	Other†	4		
		Study Participants*			
		Hospital or specialised practice nurses	5	Infectious disease specialists	1
		Anaesthetists	1	Neurologists	1
		Cancer specialists	1	Physiotherapists	4
		Cardiologists	1	Primary care nurses	2
		Diabetes specialists	1	General practitioners	21
		Dental care providers	1	Support staff & leadership	2
		Emergency physicians	10	Surgeons	4
		Hospitalists (medical)	3		

† Conference attendees, snowball, dental care network or users of a GP website

TABLE 2.3. CPR characteristics

Rule Type*		Rule Domain*		Clinical Specialty*	
Diagnostic	22	All-cause mortality	1	Breast surgery	1
Prognostic	30	Back pain	2	Dentistry	1
Therapeutic	4	Cancer	6	Diabetes Management	1
		Cardiovascular disease	12	Emergency	11
Rule Output*		Caries/peridontal disease	1	General Practice	20
Diagnosis (& likelihood)	4	Diabetic foot disease	1	General Surgery	2
Intervention guidance	7	Infection	2	Hospital (medical)	3
Patient risk/risk category	26	Multiple	6	Intensive Care	1
Unclear or unspecified	9	Post-op adverse events	1	Neurology	1
		Post-op vomiting/nausea	1	Oncology	1
		Response to treatment	1	Physiotherapy	4
		Shoulder pain	1		
		TIA & stroke	4		
		Trauma	7		

* Totals do not add to 45 due to mixed-methods and mixed group studies, or studies referring to hypothetical or multiple (listed or unrestricted) CPRs

2.5.1.1 All included studies

TABLE 2.4. CPRs included in review

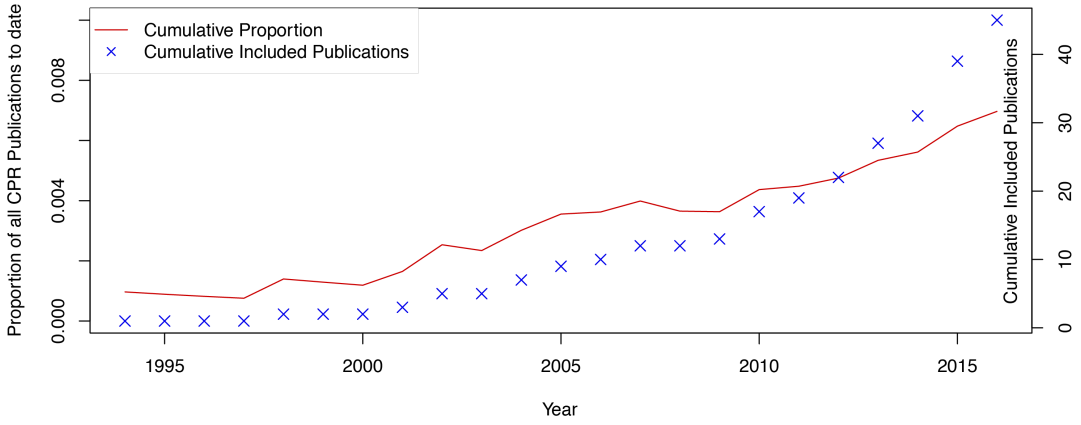
Rule Name	Year	Ref(s)	Description
ABCD/ABCD2	2010 2016	[32] [33]	Estimate stroke risk within 7 days of diagnosis of first TIA
ACI-TIPI	2006	[34]	Support diagnosis of acute cardiac ischemia in emergency patients
Adjuvant! Online	2015	[35]	Estimate breast cancer patient's survival and treatment benefit likelihood
APACHE III	2007	[36]	Predict mortality and morbidity of critical care patients
CAMBRA, PEMBRA	2015	[37]	Calculate risk of developing caries or peridontal disease (respectively) based on patient characteristics
Canadian CT Head Rule	2016	[38]	Guide use of computerised tomography for minor head injuries
CAPER Cancer RATs	2012 2015	[39] [40]	Identify and quantify risk of cancer in symptomatic primary care patients
DS3	2015	[41]	Provides preoperative patient-level risk estimates for postoperative adverse events
Framingham (or Framingham based)	2002 2009 2011 2013	[42] [43] [44] [45] [46]	Calculates future cardiovascular risk using patient characteristics and medical history. A number of the referenced articles adapted the existing algorithms for specific applications or populations.
GRAIDS	2002	[47]	Assess risk of cancer based on family history in primary care
HeartDecision	2012	[48]	Calculates risk of cardiac event in the next 10 years
In-house unnamed rules	1994 2014 2016	[49] [50] [51]	Describe a rule that has been developed locally but does not have an unambiguous name by which it is referred e.g. *a rule for calculation of [domain] risk was developed...*
Keele Stroke Model	2004	[52]	Estimate individual risk and benefit of prophylactic therapy in stroke patients
Low Risk Ankle Rule	2010	[53]	Inform management of blunt ankle trauma in children prescribing whether ankle radiographs indicated

Multiple unrestricted rules	2005 2007 2010 2013 2014 2015 2016	[54] [55] [56] [57][58] [59][60] [61][62] [63][64]	These studies ask clinicians about multiple (>2) unrelated CPR or CPR in general e.g. *which [domain] CPR are you aware of?* or *do you use any of the following list of CPR in your practice?*
Ottawa Ankle Rule	1998 2001 2005	[65] [66] [67]	Inform management of ankle or foot injury by providing information as to whether radiography is indicated
Ottawa Knee Rule	1998 2001	[65] [66]	Inform management of knee injury by providing information as to whether radiography is indicated
PECARN TBI Rules	2013	[68]	Inform management of traumatic brain injury in children by providing information as to whether CCT is indicated
PRISM	2016	[69]	Predict risk of emergency admission for patients with chronic illness
QCancer	2015	[70]	Assesses symptoms and provides risk of cancer diagnosis by type
QRISK	2013	[46]	Estimate lifetime risk of cardiovascular disease
SCI-DC foot assessment	2010	[71]	Stratifies the diabetic population based on characteristics and calculated likelihood of developing diabetic foot ulcers
Theoretical proposed rule	2004 2011 2014 2015	[72] [73] [74] [75]	These studies ask clinicians about the perceived potential benefit of a new CPR in a given clinical work-flow, and in some instances elicit specific requirements.
Walsh rule, Heckerling rule	2012	[76]	Predict diagnosis of streptococcal pharyngitis and pneumonia respectively

2.5.1.2 Publication dates

As seen in Figure 2.2, the rate at which studies meeting the inclusion criteria have been published has increased over time. Using the updated MEDLINE search string as a baseline [31], the rate that CPRs are studied for acceptability or usability is found to slightly outpace

FIGURE 2.2. Included studies over time



the rate of general CPR publications. This proportion remains extremely low, but does indicate a relative increase in interest in the qualitative analysis of CPRs over time.

2.5.1.3 Methods

A majority of included studies (28, 62%) employed a survey, with a significant minority (14, 31%) performing semi-structured interviews. Usability tests and focus groups were performed more rarely (4, 9%) due to the higher resources required. 9 studies used mixed methods, typically an initial survey with follow-up interviews and/or focus groups.

The most common recruitment strategy was through primary care network membership (16, 36%) followed by contacting members of a professional body (14, 31%). Response rate was provided in 23 studies, with an average of 51.3% (s.d. 22.5%)[1]. For recruitment directly via medical practice, network or hospital(s), there were more studies targeting health-care providers within non-academic than academic institutions (16 and 8 respectively).

[1]Where more than one response rate reported, the most general was used for aggregation

2.5.1.4 Uptake of CPRs

More than half of the included studies (24, 53%) do not report usage or uptake of the CPRs in question. Amongst the remaining 21 studies, only two report observed uptake in an experimental setting [50, 76]. One reports uptake (acceptance when triggered) by encounter and the other by clinician. The rest describe the self-reported use. Direct comparisons are not possible due to heterogeneity in the reporting and quantification of CPR use.

The only measure that could be directly compared was clinicians' awareness of a specific named rule — this was reported in 6 studies, covering 5 named rules, and one group of domain-related rules — see Figure 2.3. Note in particular that the Low Risk Ankle Rule (LRAR) is far less familiar than the Ottawa Ankle Rule (OAR), which is likely due to the overwhelming popularity of the OAR, despite the LRAR's higher sensitivity and specificity.

2.5.1.5 Health provider perspectives

Overall, three high-level categories emerged in the themes of included studies — *Utility*, *Credibility* and *Usability* — which reflect the three distinct phases in the lifecycle of a CPR — *Development*, *Validation* and *Implementation* — respectively.

These findings have been summarised in Figure 2.4, along with questions that reflect the thematic analysis results, which CPR designers can use to interrogate the design and architecture of new tools in order to bring them inline with the stated health provider perspectives.

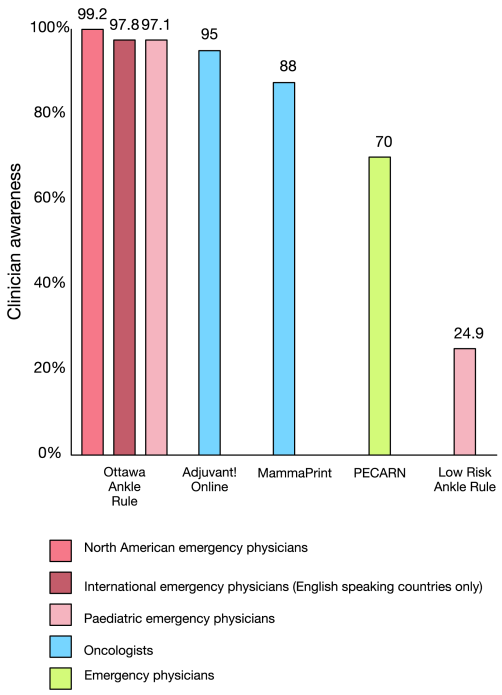


FIGURE 2.3. Physician awareness of included CPR: Ottawa Ankle Rule [53, 66, 67]; Adjuvant! Online & MammaPrint [35]; PECARN [68]; LRAR[53]

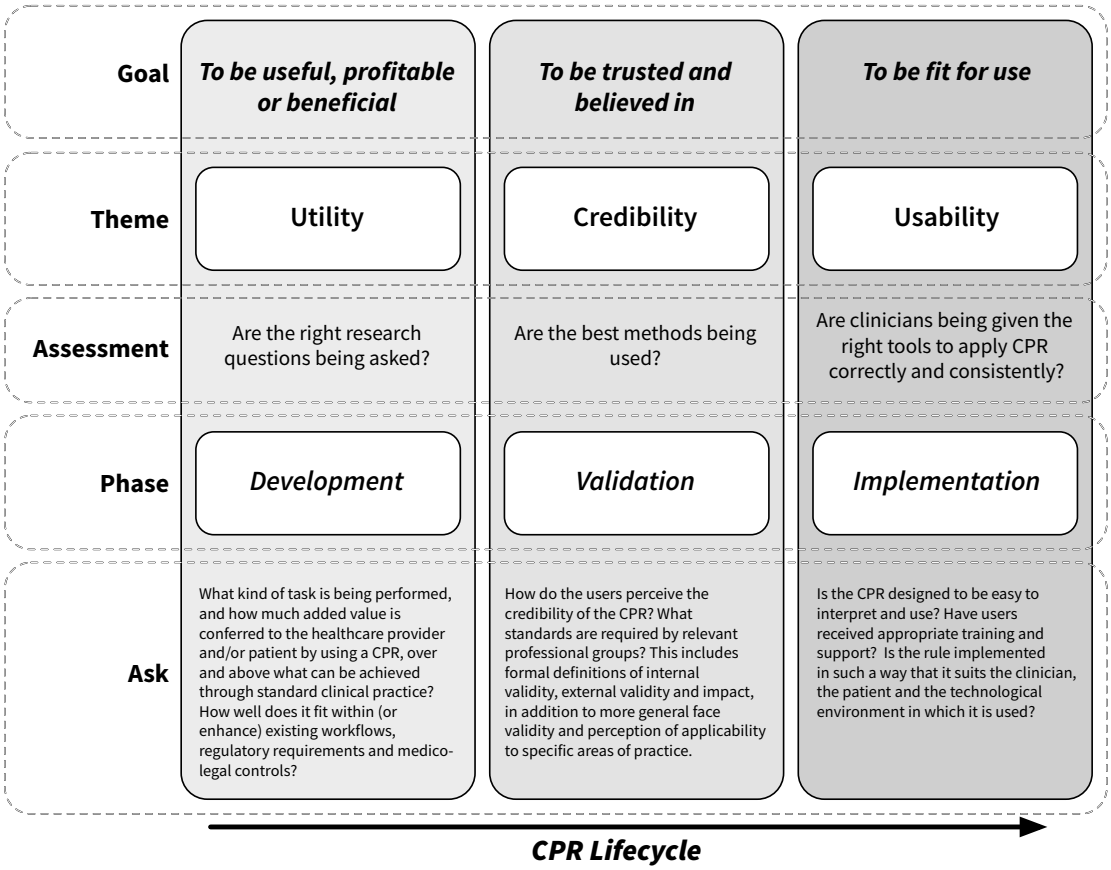
2.6 Results in context

2.6.1 Utility

2.6.1.1 Specialty

General practice and emergency medicine are by far the most CPR-served clinical specialties, represented in 31 of 45 included studies. This is consistent with the most commonly observed rule domains — cardiovascular disease (CVD) (12, 27%), cancer (6, 13%) and trauma (7, 16%) — and study participants — GPs and emergency physicians (21 and 10 respectively). These specialties require practitioners to be generalists; to recognise, support and treat a variety of conditions, and it is not practicable to expect consistent knowledge across a broad domain without flexible and accessible decision aids. Additionally, general

FIGURE 2.4. Identified themes by development phase



practice is the specialty where physicians are likely to have an ongoing relationship with patients, following up progress over time. As such, a focus on helping patients to understand their personal risks, suggested behaviour changes and treatment pathways is valued [44, 63]. In hospital settings, emergency physicians also rate CPRs as more aligned with their workflow and thought processes than internists do [62].

Cancer and CVD were both highly represented domains, however it is noteworthy that there is no correspondingly high prevalence of studies targeting cancer specialists or cardiologists (1 study reporting on each). Most of these rules were instead defined in the general practice setting, targeting early diagnosis and management of patient risk over time.

2.6.1.2 Audience

Numerous papers report higher acceptability and knowledge of CPR amongst clinicians with fewer years of experience [49, 59, 63, 67, 68, 70]. In these instances, it is assumed that CPRs function as a substitute for clinical experience and reassure younger clinicians of their judgment, and may even help teach clinical reasoning [61].

The most commonly cited reason for this is an inverse correlation between clinician confidence and the utility of a CPR, for example this verbatim quote from an experienced family physician in [63]:

> It depends how confident you are, in your decision making...like the PHQ-9 I am confident enough taking a mental health history and a depression history...I don't feel that that score replaces my own clinical judgement but there would be some scores where you know I would feel that if the score told me something that I wasn't sure of I would rely on the score more than my own because I don't feel my own clinical acumen is good enough in that area to replace the score.

This implies that in an area of true clinical equipoise, where even experienced clinicians express a lack of confidence, is one that is most likely to benefit from the development of novel CPRs.

The perception that a CPR is really the formalism of existing best practice or traditional reasoning [59] can explain some of the perceived utility for inexperienced clinicians, however it may also extend to the point that CPRs are seen as a 'crutch' (contributing to negative views) [34]. This may also be impacted by a reluctance of clinicians to change long-held beliefs, leaving younger participants more open to statistical tools [70], and a lower overall comfort level with technology and/or evidence-based medicine in older respondents [37].

The most effective CPR users were clinicians who worked full-time and reported using rules frequently [67] — underscoring the benefit seen with frequent usage.

Only one of the included studies reports actual performance differences between groups who use the studied CPR and those who do not. In [68], clinicians were presented with a vignette where an imaging test is not indicated according to current best-practice guidelines. Prior to CPR decision support intervention, more experienced clinicians were significantly more likely to adhere to the guidelines. Clinicians who were shown CPR-based decision support were significantly more likely to change their assessment in line with best practice than those who were not, in particular if they reported some preexisting use of the PECARN rules in their practice, although no further breakdown was provided as to the characteristics of clinicians most likely to respond to CPR recommendations.

2.6.1.3 Added value

Multiple studies reported that the way in which cancer [40, 74] and CVD [58] CPRs most added value to primary care was by distinguishing patients who had a slightly elevated risk, as opposed to identifying those with a greatly increased risk (who should be readily identifiable). A side effect of this observation is that outside controlled validation study settings, clinicians may not apply CPRs uniformly [57, 63]. This is important, since a rule that is in practice used only in low-certainty cases will necessarily underperform its theoretical accuracy.

Critical care is identified by advanced practice nurses as a specialty with a large degree of uncertainty [36] and thus should have a proliferation of CPR, due to the high potential to add value through confidence in treatment decisions. This is not borne out by this review, which includes only one such study, although study participants report a reduction in anxiety at end of life when the decision to remove life-support systems is supported by objective data [36].

This desire for reassurance and objectivity is also seen in EM, where there was a positive correlation between likelihood to use a rule and severity of outcome [41, 49]. A rule to identify patients for prophylactic treatment against post operative nausea and vomiting is one of the most poorly received CPRs in the review [51], which is attributed by anaesthetists to the low burden that this issue has on patients, compared to the side effects of available treatments.

CPR for common complaints are perceived to be beneficial [37, 68], likely due to the increase in memorability and a decrease in the time to apply a given rule with more frequent use.

2.6.1.4 Actionability

Healthcare providers consistently prefer CPR where results are actionable or directive, as opposed to strictly numerical [51, 63, 66, 68] due to challenges interpreting risk-based recommendations in isolation. In this context, an actionable outcome refers to a recommendation that states a specific course of action, such as whether or not to prescribe prophylactic treatment for PONV [51], or guidance as to whether or not an imaging study is indicated for a particular patient [66, 68].

This requirement for actionability is also seen with respect to rules that attempt to diagnose patients or classify them into sub-types. If treatment decisions are not different between groups, further detail is not important in a clinical (i.e. not research) context [75].

Other actionable outcomes favourably viewed by clinicians avoid time-consuming, invasive or costly procedures [68], identify risk factors that will have the highest impact on patient outcomes [42], prioritise tests or referrals in the face of non-specific symptoms [70] and systematically assess combinations of symptoms instead of in isolation [40].

2.6.1.5 Medico-legal and regulatory environment

There is evidence that US physicians are more likely to believe CPRs increase the risk of being sued [34, 53, 66], whereas the converse is generally true in other English-speaking countries, where they are viewed as protective against such suits [66, 68] by providing documented evidence of a rationale to forming certain decisions. A corollary to this, however, is the concern that in the instance that a CPR disagrees with the clinical judgment of a physician, they may order procedures they believe are unnecessary if there is a paper trail indicating that they were prompted that this was necessary [39].

Physician autonomy is an important factor for participants in many of the included studies, although the level of concern varied greatly. One study reported a global pattern which closely reflected the pattern by which physicians from each country viewed their increased risk of being sued when using CPRs [66] — a higher perception of loss of physician control correlating with a greater fear of litigation.

2.6.1.6 Psychosocial factors

Due to this desire for autonomy, multiple studies found that clinicians would augment or replace the CPRs with other factors [49, 50, 67, 70], or apply in a manner that was contrary to the tool design such as restricting the applicable patient population [36, 57, 63]. This explains discrepancies between results in validation and impact studies.

A perceived result of loss of autonomy is the imposition of external control or reduction in services for patients predicted not to respond to treatment [36, 51, 59] if insurers have access to CPR results. Physicians feel threatened by this as an artifice of a rigid framework that is not suitable for all patients [36], and that it may result in 'intellectual sloppiness', as providers become dependent on these tools and cannot form judgments independently [59, 61].

It is possible to argue that some loss of autonomy is a positive outcome of CPR usage, with authorities able to benchmark performance of multiple clinicians and ensure standardised care is provided to all patients. There is evidence, however, that this is perceived as a threatening or overbearing action by some clinicians [69], which may affect their successful implementation.

2.6.1.7 Patient/clinician interaction

Consultation time is a finite and valuable resource. CPR implementations typically interrupt the flow of the consultation in some way — whether explicitly such as a pop-up window or in a clinician-initiated fashion — prompting concerns that CPR use may cause longer consultations, or cause other important issues to be deemphasised [39, 40, 43, 44, 48, 52,

70, 76]. This is particularly problematic in instances requiring double data entry or context switching [48, 71], where computer usage gets in the way of more relational interactions [52]. Some papers suggest that carefully selected trigger points, responding to and better reflecting the true nature of the consultation, would address this issue [44, 76].

Emergency physicians report less concern regarding time taken to apply CPRs, and are more likely to feel that they are time saving devices [53, 65]. This may be due to the high representation of the Ottawa Ankle and Knee Rules, which are not only simple to apply, requiring no data entry, but also have demonstrably decreased unnecessary tests, which directly affects the emergency physician workflow — unlike specialist and primary care contexts, where tests fall outside consultation time so any time saved is decoupled from the main patient care phase.

Some physicians are reluctant to initiate CPRs in front of a patient if they cannot anticipate results, especially if these results are potentially confronting, as in the case of a high-risk cancer prognosis [70] or if they feel that patients will have trouble contextualising the risk [52].

2.6.2 Credibility

2.6.2.1 Face validity

This review found that rules that do not have clear face validity are rarely acceptable to clinicians. Studies report that tools with conspicuously missing risk factors will be viewed skeptically [34–36, 43, 55, 74], whereas those that reflect current best practice clinical reasoning have good credibility [40, 50, 59, 61]. This skepticism holds, largely regardless of the verified performance of the rule in practice.

This disadvantages machine learning models with complex feature engineering and hidden layers, at least partially explaining the gap between what is technically feasible and that which is actually observed. This call for biological plausibility is reported directly by clinicians [59], and also observable in use. Physicians feel that weight bearing ability

is relevant to diagnosis of ankle fractures, and manually include it in their calculations, despite the 100% sensitivity of the LRAR. The outcome of this is a reduction in specificity, with no additional instances of fracture diagnosed.

This is linked to the desire for actionability — with a premium on time and resources, there is a preference to only use tools that will have an impact on patient care. For prognostic models, this requires that identifiable risk factors can be discerned and modified. Diagnostically and therapeutically this makes less sense, however clinicians do not seem to discriminate, requiring face validity in all circumstances. Similarly, without a discernible causative pathway some clinicians do not view the results as truly personalised, crediting only applicability at the group level [51].

2.6.2.2 Validation study design and availability

Practitioners must be confident in the evidence supporting a tool before they are willing to overcome other barriers to implementation, however there are few included studies that spell out what validation is required to meet this bar. The OAR is cited as an example of rigorous development standards [65, 67], but only in a general sense, not detailing by which factors this is defined.

Out of 29 studies that look at named included rules (not including theoretical, in-house or unrestricted sets of rules discussed in the abstract), 23 refer to contemporaneous validation studies in external populations.

The endorsement and/or mandate by professional bodies is a valuable strategy in dissemination and successful uptake of a new CPR [63, 65, 74], and is given great weight by clinicians.

2.6.3 Usability & Implementation

2.6.3.1 Usage & Usability

In a number of papers, usability is equated with memorability [67, 75], with some clinicians only willing to use tools that can be applied without referring to the computer at all. Similarly, high value is placed model simplicity and low variable count [50], which should be easily measurable and available without elaborate equipment and testing. Benefits of a simple tool are realised by both clinicians and patients, as patients can be involved in the decision making process [75], and the benefits of CPR usage are readily understood [70].

Simplicity in both variable selection and interface extends into the learnability and accurate application of the tool. Even tools that have eventually high usability demonstrate a period of familiarisation where error rates are higher while clinicians become accustomed to the system [42].

Few papers (4) apply any usability testing methods directly; however even in a low-resource and low-experience setting, this is shown to be an effective tool in increasing satisfaction [42], and simple tests can expose usability flaws [75, 76].

2.6.3.2 Information technology & integration

40% of CPRs as implemented in the studies of this review were fully or partially integrated (patient data populated directly from the EMR), 20% offline, 11% online but not integrated (requiring double data entry) and 29% unclear or hypothetical implementations only. For CPR requiring data input, the clear preference is for the system to be integrated with patient data [41, 43, 44, 48, 70, 71, 74, 76], with the importance of this requirement increasing in recent years. This preference for integration puts additional value on CPR designs that rely on relatively few commonly available data points, ideally in their most generic form, to allow effective integration across diverse systems. Integration into the health record can also provide additional context to filter CPRs to only those applicable to the patient under review.

Physicians in the included studies do not demonstrate a sophisticated concern about data security or persistence [77] however they are sensitive to any technological failures or outages, and perceived delays [36, 39, 71].

2.6.3.3 Results presentation & visualisation

Interpretation of risk-based outputs presents a challenge to clinicians, particularly differences between absolute and relative risks [52], and even when output can be correctly interpreted, this doesn't necessarily translate into consistent treatment [42].

Visual representation is beneficial in accurate interpretation of risks [43, 77], particularly consistent use of traffic-lights for risk-based information, which evoke an emotional response [39, 43, 70], helping clinicians understand when to treat, and patients understand physician recommendations. Visual representation is also found to speedup review of results, allowing rapid interpretation whilst remaining patient-focused [39] and avoiding presenting too much information, which may be confrontational [43].

Other presentation factors preferred by clinicians are the option to print out results and tailored supporting documentation for patients to take home and digest in their own time [48], comparative displays to show the impact of modifying risk factors [48] and multiple formats to improve understanding [47, 74].

2.6.3.4 Education & dissemination

One way in which clinicians express frustration with CPRs is with the lack of training and support for them to apply rules accurately and consistently [36, 39, 52]. The lack of simple and accessible training materials is a highly impactful barrier to implementation [36, 40].

An effective roll-out will be situationally dependent, however the following desirable characteristics were identified by users included in this review: integration into the patient flow at the appropriate time [68], convincing evidence of effectiveness above current accepted practice [53], materials that address correct usage, including how to combine the

tool with clinical acumen and characteristics that may appear to be erroneously omitted [53, 67], patient communication strategies [53], quick-reference materials and memory aids [67], publication in high-profile journals [65], endorsement by professional organisations [65], continued support and interaction post roll-out [40], information on CPR development methodology and standards of evidence [46, 65], and short and well-planned training sessions or materials with key information clearly highlighted [36, 39].

2.7 Discussion

2.7.1 Limitations

It is clear that few CPRs are studied for usability or utility - consequently there is little to no evidence of how they are used in practice. Our search was intentionally restricted to qualitative studies, which significantly limited the result set.

There are no commonly accepted variables available for aggregation, which limits reviews to qualitative synthesis. Physician awareness of CPRs and intended versus actual uptake would be useful data points in future work. Without summary statistics, it is not possible to precisely track trends over time, and the overall conclusions of the review are more susceptible to bias.

Due to the unavailability of directly comparable and aggregatable variables, it is difficult to make clear distinctions of trends over time. It is possible to see an increase in demand for integration into health records, with all studies stating this this as a high-priority item published since 2009, however few of the resultant themes show such a distinct trend.

2.7.1.1 Definition

There is no precise definition of CPRs as a distinct subset of decision support. For the purposes of this review, we selected studies that either self-identify as CPR, or where the decision support described within clearly takes current or past measurable patient

characteristics into account to derive a relevant clinical likelihood. It is possible that CPR form a basis of many other decision support systems that were not included due to a lack of clear definition.

The lack of consistent terminology for CPRs made an exhaustive search of literature challenging. For the final search implemented in this review, relatively generic terms were favoured over a curated list of known prediction rules in order to avoid biasing the results by domain. Consistency of observed results across specialties demonstrates that this is nonetheless likely to be a representative sample of available qualitative studies and therefore the results can be expected to be generalisable.

2.7.1.2 Big-data extrapolation

None of the included studies present a clear path forward from static predictions to a living big-data system. These systems are only available in a very limited scope currently, with much research, but very few implemented in real healthcare settings [22, 78], thus it is not possible to collect clinician opinions of the usage in daily practice.

Despite this, we chose to use the acceptance of CPR in clinical practice as the closest available proxy for this next step in maturation of clinical prediction. This is additionally confounded by the fact that a number of the included CPR are designed to be used as manually-applied decision aids (for example, the OAR) with no technological component. The results of this review are, however, consistent between these manual interventions and the more obviously comparable computerised tools, particularly with respect to utility and credibility.

Given the nascent application of big-data tools in clinical practice, it is unlikely that a more directly representative review could be made for a number of years, and therefore we believe this work provides an important foundational block, as steps toward truly personalised EMR-based predictions are taken.

2.7.1.3 Correlation of opinions to real-world experience

Physician opinions are an important predictor of the success of a new CPR in clinical practice, but they are of course not the only factor required for well-received implementations. This review addresses the perceptions of CPRs across a diverse set of paradigms, covering the nature of clinical practice alongside human factors and implementation strategies, which will, when followed, increase the likelihood of successful implementation and uptake. This does not address the question of impact and true, measurable benefit to patient populations.

2.8 Conclusions & Recommendations

Holding constant factors that relate only to the nature of the clinical problem addressed, for broad acceptance of a new CPR, developers should prioritise the utility, credibility and usability of their models. These goals are reflected primarily in the rule's actionability, face validity and simplicity (respectively).

Figure 2.4 presents a summary of the findings of this review that allow CPR developers to interrogate their methods and goals in order to produce a model that is highly translatable and will be viewed favourably by clinicians.

2.8.1 Utility

The most commonly observed utility of CPRs relates to assisting GPs with diagnosis and risk management of cancer and CVD patients. This is followed by providing emergency physicians with rapid reassurance in the face of uncertainty.

The utility of CPRs decrease if the predicted outcome does not have significant impact to the patient (severe condition or serious potential side effects), or if the outcome is numerical and not clearly actionable. CPRs should include directive outcomes and causal

pathways, with attention paid to the discriminative performance and calibration of patients in the 'yellow' zone (slightly elevated risk).

2.8.2 Credibility

Clinician's perception can be more important for translatability than proven performance. GPs in particular do not generally report knowledge of CPR performance, while some emergency physicians and specialists pay attention to the discriminative power (in particular, sensitivity) of CPRs.

Face validity can be improved by:

- Performing feature engineering steps to ensure biological plausibility
- Complexity reduction
- Clear direction on modifiable risk factors
- Where the above are not possible, directly addressing any potential concerns in educational materials

Additionally, the roll-out phase should include steps to educate professional organisations and comply with their requirements for validation and training. Ideally a validation phase should include an impact assessment, which will address potential inconsistency in application across patient groups.

2.8.3 Usability

Simplicity must be preserved to improve both the technical implementation and integratability. This facilitates a smooth fit within the clinician/patient interaction by limiting data entry, disruption in communication and unnecessary context switching. In order to ensure that CPRs meet these goals, they must be tested for usability more frequently than has been demonstrated in these results. Where resources are limited, the authors advocate a 'think aloud' protocol as demonstrated in [79].

Clinicians report misuse of CPRs when attempting to add features or inconsistently apply rules. Where risk factors are modifiable, or any output is actionable, the CPR and associated materials should communicate this clearly.

CHAPTER 3

Processing Pipeline Implementation

3.1 Source Data

3.1.1 Data Ethics

This study was approved as ethics application HREC/15/SVH/403 under the St Vincent's Hospital Human Research Ethics Committee (SVH reference SVH/259). This approval was granted on the 10th of August, 2016 and is valid until the 2nd of August, 2021.

3.1.2 Content

The data made available for this study by the St Vincent's Hospital data custodian represent realistic input that would be available for real-time analysis. Pending delivery and acceptance of a model developed from this dataset, it is thus theoretically possible to integrate it into the live clinical information management system. Any data points that would be unavailable at each relevant prediction time were carefully masked (i.e. a delay of 3 days inserted between discharge and the availability of coded discharge diagnoses in the clinical record; 28 day readmission flags inserted only after the appropriate delay; length of stay, discharge code, leave days, separation mode, transfer to other hospital and total ICU hours only available for prior admissions).

3.1.2.1 Data Dictionary

The requested data fall into two categories - admission-centric and patient-centric (Figure 3.1). This distinction is due to the fact that it is valid for a patient to have a pathology result, medication administration or alert event as an outpatient, where other clinical events such as ward movements or surgical procedures are only experienced within admission episodes.

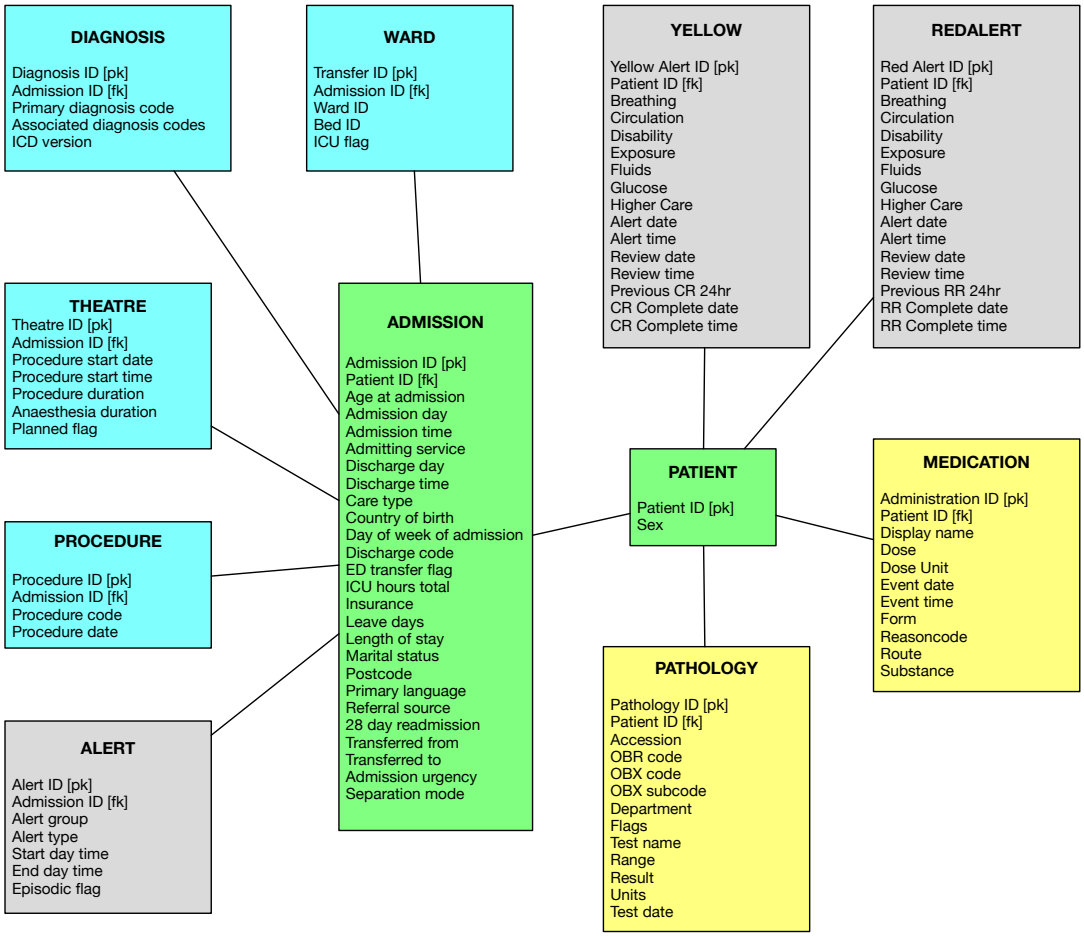


FIGURE 3.1. Source data relationship diagram: blue - admission-centric; yellow - patient-centric; grey - dropped for data quality issues

3.1.3 Data Quality

There were a number of serious quality issues observed when processing the data. These issues are noted here to illustrate challenges experienced when applying statistical techniques to naturalistically collected data, where assumptions of uniformity may not hold. In

particular it is worth observing that the last two issues were not evident under any general statistical summarisations and were exposed only under detailed time-series analysis.

Firstly, there was a process change during the study period in how rapid response calls (red alerts), clinical reviews (yellow alerts) and general alerts were recorded in the EMR. As a consequence, for each of these tables a non-representative sample of records was available — covering only a small proportion of the time window of interest. This meant that the clinically relevant target endpoint of predicting clinical emergencies (other than death) was not available for many admissions, and it was not possible to integrate these emergency calls into the final models, thus all alert tables were dropped.

In addition, the transfer of data to a new Patient Administration System was incorrectly mapped, which meant that ward transfer times were incorrect in the source system for the earliest 10% of admissions. As ICU admissions (identified by a ward transfer from general to intensive-care wards) are a key target endpoint, it was therefore necessary to discard all admissions prior to this data migration date.

The final large-scale data error was fortunately systematic and therefore recoverable, as the date-obfuscation process (part of the project-specific patient de-identification research ethics requirements) was applied differently to admission-centric and patient-centric tables. Once this issue was identified, it was possible to correct simply by applying the required offset to the medication and pathology tables.

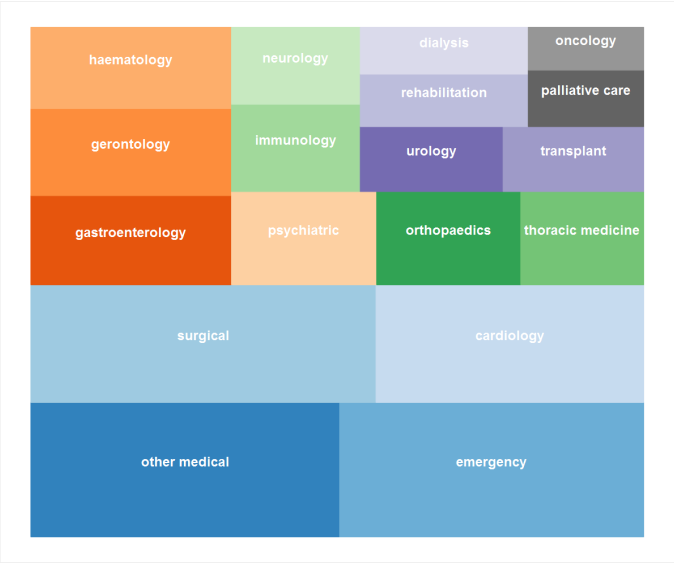
3.2 Summary Statistics

After the correction of the above issues, 192,883 admissions remained, shared by some 92,802 patients. See Table 3.1 and Figures 3.2 and 3.3 for detailed summary statistics. It is important to notice the extreme skew in the distribution of some data elements across patient records, in particular pathology and medication records, where the mean number of records per patient outstrips the mode by several orders of magnitude. This lack of uniformity in the richness of input data is a key challenge for any predictive task.

TABLE 3.1. Population Statistics

Item	Count	Mode	Mean	St.Dev.
Patients	92,802			
Male	51,844 (55.9%)			
Female	40,958 (44.1%)			
Age		71	59.96	20.34
Admissions	192,883			
Admissions per patient		1	2.07	3.92
All diagnoses	892,629			
Primary diagnoses	192,863			
Comorbid diagnoses	699,766			
Diagnoses per patient		1	9.62	17.73
Diagnoses per admission		1	4.63	4.08
Distinct diagnoses per patient		1	7.12	8.61
Medication administration events	12,524,922			
Medication events per patient		0	134.96	506.50
Per patient with $\geq$ 1 event		1	252.25	670.74
Pathology results	41,871,520			
Pathology results per patient		0	451.19	1221.09
Per patient with $\geq$ 1 result		2	479.85	1253.80
Surgical procedures	117,658			
Surgical procedures per admission		0	0.61	1.36
Surgical procedures per patient		0	1.27	2.68
Per admission with $\geq$ 1 procedure		1	1.92	1.82
Per patient with $\geq$ 1 procedure		1	2.85	3.42
Ward movements	676,193			
Ward movements per admission		1	3.51	3.26
Ward movements per patient		2	7.29	11.12
Time between admissions (days)		2	105	238
Length of stay (hours)		3	98	239
Length of stay for stays $\geq$ 24h		27	196	316

FIGURE 3.2. Admissions by admitting service



3.3 Technical Architecture

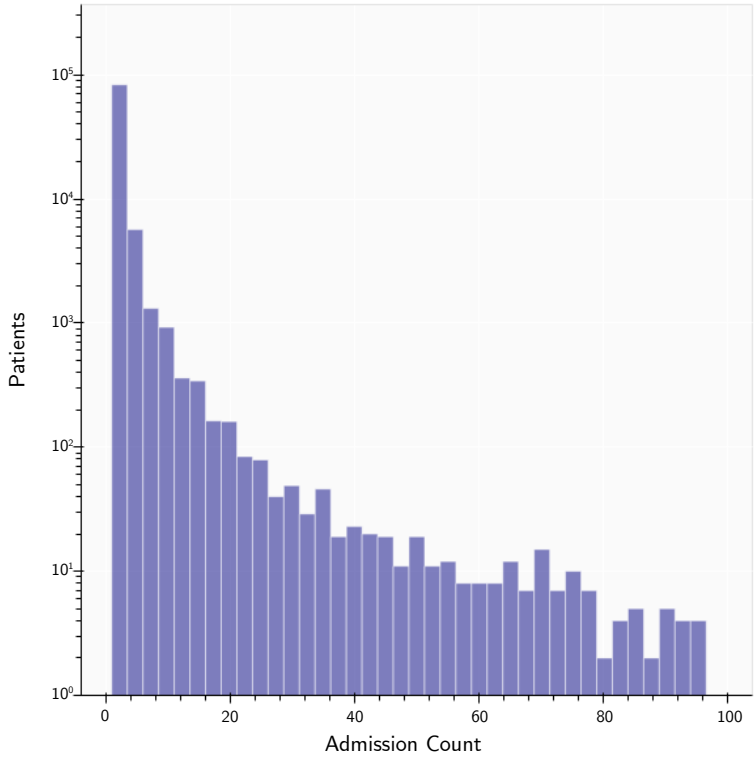
This prototype system was developed in a secure Amazon Web Services Elastic Compute Cloud (AWS EC2) environment. Data processing was performed in Python 3.7, notably leveraging the SQLAlchemy, NumPy, Pandas, Matplotlib, Scikit-Learn and TensorFlow libraries.

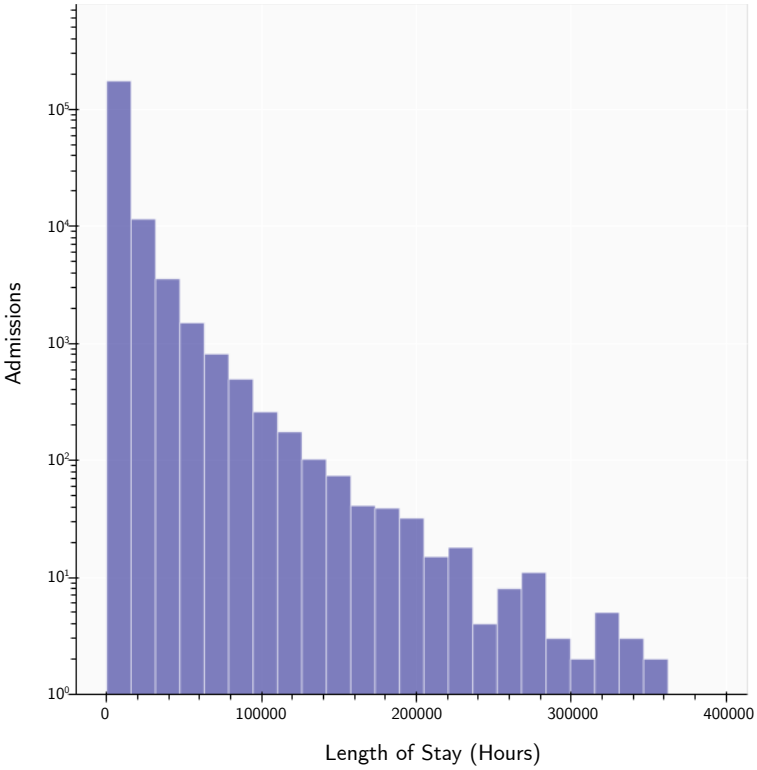
By implementing a SQLAlchemy ORM abstraction layer, the code is transferable to numerous database back-ends according to the target institution requirements.

3.3.1 Data Processing Pipeline

The data processing pipeline can be broken into 4 high-level steps — trajectory generation, model input curation, model training and results evaluation. A summary of the main logic is provided for each of these steps below.

FIGURE 3.3. Distributions: admissions per patient, length of stay





3.3.1.1 Trajectory Generation

Input. Raw data as described in Section 3.1.2.1, queried from a PostgreSQL database via the SQLAlchemy ORM abstraction layer.

Output. For each admission, the following items were generated:

(1) A set of ordered, discrete tokens, which represent the full patient history up until the prediction time.
(2) Endpoints (binary outcome and time to event) for each target event (death and unplanned ICU admission), calculated from prediction time.

Process. This output was generated by the following steps:

(1) Discarding admissions shorter than 24 hours, define the time of prediction as 24 hours after admission
(2) Select all events available for this patient (including both the current and any historical admissions) which occurred prior to this prediction time
(3) Mask any items in this set of events that occurred earlier than the prediction time, but would not have reasonably been available in the clinical record in real time (i.e. administrative data that is either entered or coded post-hoc)
(4) Tokenize discrete events (e.g. surgical procedures, ward movements) by defining their categorical label
(5) Tokenize continuous events (e.g. numerical pathology results, surgery durations) by replacing with their decile and translating to their corresponding categorical label
(6) Interleave all tokenized events in the order of their occurrence to form the patient trajectory until the time of prediction
(7) Label each trajectory with the target outcomes and time-to-event

Upon completion of the trajectory generation, seven token types were retained for analysis. See Table 3.2 for a summary and examples. In this way, all historical data tokens were

TABLE 3.2. Vocabulary Statistics

Token type	Description	Details/Examples	Vocabulary Size
Admission	Admission-related data and demography.	Admitting speciality Marriage status Urgency of admission Discharge code (prior admissions only)	3503
Diagnoses	Discharge diagnoses (primary and comorbid - prior admissions only). ICD-10-AM encoding.	Z49.1, I10, N18.9, Z72.0, E78.0	8034
Pathology	Institution-specific encoding. Includes panel name, test name and results as observed in the HL7 OBR and OBX segments.	Urea (mmol/L), eGFR (ml/mn/1.73m^2), PO2 (mmHg), Lipase (U/L), HCT, INR, FiO2 (%)	7759 distinct tests; 11,784 when combined with result decile information
Medication	Medication administration events (substance, dose, route, form, time of admission).	Paracetamol (500mg) tablet: 1g Oral, Oxycodone (5mg) tablet: 5mg Oral, Heparin Sodium (5000units/0.2mL) Injection: 5000 units Subcutaneous	9368
Ward	Ward and bed assignments across the admission.	ED, ICU, CCU, DIA	809
Procedures	Medicare Benefits Schedule (MBS) encoding of surgical procedures.	Oesophagoscopy, Central vein catheterisation, Selective coronary angiography, Repair of soft tissue wound	1964
Theatre	Theatre movements and details of anaesthesia.	Procedure and anaesthetic duration as per-procedure-type decile; Elective/emergency	99,711

available as potential input for each patient, although the most rare tokens (appearing in fewer than 2% of trajectories) were discarded.

3.3.1.2 Model Input Curation

Input. Tokenized patient trajectories and endpoint labels as defined above

Output. TensorFlow record files (`TFRecord`) containing all necessary endpoints and input data for rapid model training

Process. Train, validation and test record files were created by serialising input data into a format that is optimised for training TensorFlow models. This step ensures that the models can batch load and prepare data on the CPU cores whilst training the prior batch on the GPU. Such parallelisation greatly increases the model training efficiency, although it comes with a trade-off of unwieldy data structures that are not well suited to more traditional error-checking and model-introspection processes.

3.3.1.3 Model Training & Results Evaluation

The description of model training and results evaluation forms the core of Chapter 4.

CHAPTER 4

Developing a deep learning system to drive the work of the critical care outreach team

---

4.1 Preamble

This chapter is a stand-alone work that has been submitted for publication as [17], reproduced exactly as submitted with the exception of this preamble.

In the context of the work of this thesis, this chapter serves to provide the technological proof of concept that the proposed watch-list technology can meet acceptable performance benchmarks, in particular with respect to the lack of vital signs data in the source system.

All overarching functional requirements were sourced from the critical care team at the requesting institution. This set of functional requirements are the key motivation behind this project, and thus define the final form of model endpoints and overall workflow.

4.2 Abstract

Care of patients at risk of deterioration on acute medical and surgical wards requires timely identification, increased monitoring and robust escalation procedures. The critical care outreach role brings specialist-trained critical care nurses and physicians into acute wards to facilitate these processes. Performing this role is challenging, as the breadth of information synthesis required is both high and rapidly updating.

We propose a novel automated 'watch-list' to identify patients at high risk of deterioration, to help prioritise the work of the outreach team.

This system takes data from the electronic medical record in real-time and creates a discrete tokenized trajectory, which is fed into a recurrent neural network model. These models achieve an AUROC of 0.928 for inpatient death and 0.778 for unplanned ICU admission (within 24 hours), which compares favourably with existing early warning scores and is comparable with proof of concept deep learning systems requiring significantly more input data.

4.3 Background

4.3.1 Clinical Setting

For a patient in an acute care setting, there are many complex and interrelated factors that affect their likely trajectory toward either recovery or deterioration. Prior to significant deterioration events, there are observable patterns in clinical features that indicate this change in acuity [1–4]. These warning signs may be present as much as 48 hours prior to the adverse outcome [1], however they are often overlooked.

In addition, there is evidence that sub-optimal care (including delayed or missed interventions) in general hospital wards is a key contributing factor to both unplanned ICU admissions and preventable inpatient mortality [5, 6].

These factors have combined to drive the modern desire for tools and processes that can accurately highlight patients at risk of deterioration on the general wards such that interventions can be deployed sooner, improving both patient outcomes and resource utilisation. This commonly takes the form of an early warning score such as the National Early Warning Score (NEWS) [80], which tracks physiological variables and raises an alert when they fall outside of acceptable limits.

It may also include the establishment of a critical care outreach team whose purpose is to integrate critical care skills of advanced assessment into the general care wards [12, 13]. This is a challenging role, requiring a rapidly updating awareness of events and patients across the whole hospital. In order to effectively prioritise their distributed workload, critical care outreach nurses and medical officers (CCON & CCOM) must synthesise information on a broader scale than is required of typical ward staff.

A physiological early warning score such as NEWS is intended to provide a trigger for emergency response, however the remit of the outreach role is broader than this — including the goal of identifying potential future deterioration in order to allow intervention prior to emergency onset. Risk models used to prioritise this work may therefore benefit from the inclusion of alternative risk factors such as pathology results or complex comorbidities. In addition to this, the reliance of existing models on vital signs indicators alone limits their capacity for automation in settings where these observations are not captured electronically.

4.3.2 Technological Setting

There has been much interest in the development of deep learning models derived from electronic medical record (EMR) data. Deep-learning techniques are robust to heterogeneous, sparse and messy data, which are defining characteristics of the EMR. EMR data also fit naturally into recurrent neural network (RNN) architectures due to the discrete, episodic, time-series nature of the patient trajectory, which draws robust analogies to models of language. These language models have recently been expanded to account for the variable time intervals present in the patient record [81–83] by incorporating time-modulation gates or weightings for elapsed time.

Importantly, deep-learning techniques based on sequential tokens have the capacity to learn from rare events that would have insufficient predictive power in traditional models. Contextual embeddings such as the skipgram algorithm [84] transform high-dimensional one-hot encoded concepts into a lower-dimensional vector representation that can describe not only the exact event type, but also where the event type fits within a conceptual ‘neighbourhood’ [85]. This is done by learning a representation of events as they relate to

adjacent events in the clinical trajectory — inferring that events that consistently appear in the same context will often contribute similarly to the patient's overall risk profile.

Recurrent models have been developed from EMR data with high accuracy for diagnostic, phenotyping and prognostic purposes in diverse clinical domains. In particular, such systems have been demonstrated to perform well when used to predict inpatient mortality and ICU admission [81, 86, 87], which are the most important end-points for understanding short-term risk in a general patient population.

4.3.3 Aim

The primary aim of this project is to investigate the feasibility of an automatically generated watch-list that provides outreach staff with an ordered list of patients most at risk of short-term deterioration. By analysing all available data in the medical record as it is generated, this list can supplement the clinical judgement of the CCON & CCOM and help them to proactively identify patients in need of early intervention to improve outcomes, avoid unnecessary or ineffective ICU admissions and reduce the risk of unexpected death.

The watch-list does not attempt to form a specific diagnosis nor prognosis but rather produces a priority list that can sit alongside clinical judgment. Users are therefore less tied to strictly explainable inference, requiring only a meaningfully calibrated relative risk. As such, we propose that it is a good candidate for piloting a real deep learning system in the clinical workflow. Preliminary user discussions suggest an openness to augment their workflow in this way, and a lower barrier for requiring exhaustive model scrutability due to the fact that the existing mental model for this role is so burdensome.

A significant limitation in this setting is the lack of any electronically recorded vital signs in the source data. All identified comparison deterioration models (both traditional [80, 88–92] and deep-learning [81, 86, 87]) rely on patient vital signs and physiological observations as key predictors. We are therefore also aiming to establish the viability of an alternative for predicting short-term patient deterioration where vital signs observations are not available. A study found that in settings where vital signs data are routinely

documented using a mix of paper and electronic records, there are high levels of invalid and incomplete data [93], meaning that this limitation is sufficiently wide-spread that the automation of existing deterioration models would not be universally possible, and such an alternative is worth seeking.

4.4 Methods

4.4.1 Data

For this work, we used a dataset of hospital admissions from a metropolitan quaternary-care hospital in Sydney, Australia. The data were gathered retrospectively and approved for use by the target institution's Human Research Ethics Committee.

All historical entries in the EMR were converted to discrete token values, based on their event type (admission/discharge, historical diagnosis, pathology results, medication administration, ward movement, surgical procedure or demography). These tokenized events were then concatenated to form a list of discrete values describing the patient's historical trajectory that could be fed into the prediction model.

4.4.1.1 Example

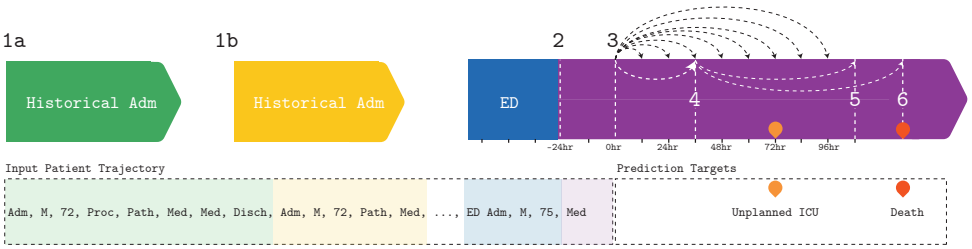


FIGURE 4.1. Example patient trajectory

Figure 4.1 shows an example of the inputs and prediction targets used to develop these predictive models. This example patient has two historical admissions (1a, 1b) prior to the current admission. Both historical admissions were for planned procedures, and include a mix of demographic and clinical tokens.

In the target admission, the patient was admitted via the Emergency Department. Admission time (2) is the time that the patient was transferred to the medical wards. Prediction time (3) is set to 24 hours after admission time (t=0hr). All demographic and clinical tokens up to prediction time are included in the input data. Thus the input trajectory is an ordered list of all tokens occurring in any available historical admission(s), the patient's ED stay, and the first 24 hours of the target admission.

4.4.2 Targets

Events of interest are defined as in-hospital death and unplanned ICU admission, as a reduction in these events is the core premise supporting the establishment the critical care outreach team. There is no distinction made as to whether a death occurs in general wards, theatre or in the ICU.

No predictions are generated for patients in the ICU at the time of prediction, as they are already under the care of the core ICU team.

An ICU admission is classified as planned if it follows immediately from a surgical procedure, as there is no data available that specifically captures ICU admission intention. In the case that an admission to ICU direct from surgical theatres is indeed unplanned (i.e. due to unexpected in-theatre deterioration or adverse event), there is no intervention required from outreach staff, therefore the inability of the model to identify such cases is unlikely to be impactful.

Patients admitted directly to ICU are excluded from these models (363 admissions). In order to allow all states to be mutually exclusive and thus avoid the additional imbalance that would be introduced under a multiclass classifier accounting for death/ICU admission/both/neither, we train separate models for ICU admission and death risk.

Prediction time ($t = 0hr$) is set to 24 hours after a patient is admitted to general medical wards, either directly or via transfer from the emergency department. Prediction endpoints are measured at 12 hourly intervals, up to 4 days into the future ($t = 0 + [12, 24, 36...96]hours$).

4.4.3 Data Preparation

In order to take advantage of the contextual embeddings that were initially developed for natural language processing (NLP) tasks, and as per prior deep learning work with EMR data [94], we converted each entry in the clinical database into token(s) of one of the following types: admission, discharge, pathology result, medication administration, ward movement, surgical procedure.

Pathology results and surgical procedure details contain continuous data types (numerical results, duration respectively), which cannot be handled by a straightforward contextual embedding model. These numerical values are therefore converted to decile results for each test or procedure type respectively. These tokens are then concatenated for each patient, with their associated time-delta since time of index admission, in order to describe their care trajectory, such as in Figure 4.1.

All data are inserted into the care trajectory at the time that they become available in the EMR. Ward movements, medication admission, pathology result, procedure and theatre movements are incorporated into the EMR in real-time. Some demography data are available at triage time, whilst some variables are input only at discharge. Coded diagnoses are not available in the EMR until some time after the time of discharge due to manual coding procedures. We therefore mask diagnosis codes associated with the target admission and only include historical diagnoses that end at least 3 days prior. Any variables containing multiple or inconsistent time-stamps were only inserted in the trajectory at the latest associated time stamp. Similarly, we take a pessimistic view of time to data entry for details pertaining to historical admissions, also inserting a delay of 3 days for discharge-related information such as discharge code, separation mode and total ICU hours.

4.4.4 Time-sensitive Concept Embedding

Before feeding such tokens into a deep-learning model, we must represent them numerically so that they may be used in the matrix algebra that forms the basis of the learning algorithm. An integer label for each distinct token in the vocabulary is insufficient for this purpose, as it implies an ordinality that does not exist and thus performs poorly. A one-hot encoding is possible, where each token is represented as an n-dimensional vector with a single '1' corresponding to the specific term being described, and n is the number of distinct terms in the vocabulary. Such a representation typically leads to intractable calculations where n is a non-trivial number of available terms, and importantly does not take advantage of semantic similarity between terms (in this instance, perhaps a condition and its treatment are found to co-occur with sufficient frequency and particularity such that they may be treated similarly). These tokens were therefore transformed into a lower dimensional embedding space using a modification of the skipgram algorithm [84], which is a commonly used technique for assigning tokens a semantically-meaningful spatial representation.

Temporal and relational knowledge was encoded within the embedding by using a sampling function that was weighted inversely proportionally to both the time-delta between two events, and also whether or not the event occurred in the same admission. In equation 4.1, *w* is the weighted likelihood of selecting a particular pair of events as input to the skipgram algorithm, *s* is the distance between the two events by admission (for events in the same admission, *s=0*, for events in the admission immediately prior, *s=1* etc.), and *t* is the time interval between the two events in hours.

$$w = \frac{1}{(s+1)(t+1e^{-3})^{\frac{1}{100}}} \tag{4.1}$$

This weighting was then used to distribute the likelihood of sampling token pairs for inclusion in the embedding model. This is important because it allows the use of wide context windows in order to capture relationships between events occurring in rapid succession, as we want to preserve the strong relationship between temporally linked events (e.g. pathology results, where full test panels may return many results simultaneously) without introducing extraneous relationships between more loosely associated concepts captured

within the same broad context window only incidentally due to the fact that there were no interposing events. This is a known challenge when learning low-dimensional embedding representations of clinical events [95, 96] without allowing for the time dimension. The effect of this decay factor is conceptually similar to the time-based dynamic windowing techniques in [97].

4.4.5 Data Balancing

The targets of this model have a highly imbalanced distribution, which represents a significant challenge in the development of a useful model [98], with imbalances as skewed as 1 event in 160 for unplanned ICU admission and 1 in 180 for death within the shortest time-frames. At such significant levels of imbalance, it was found that resampling alone was insufficient to produce a model of adequate performance, as the models rapidly overtrained on the (numerically and proportionally few) samples from the minority class. We therefore use a data augmentation strategy that allows the models to weight the loss functions appropriately and learn a more accurate representation of both the majority and minority classes.

Data augmentation is common in the domain of image processing tasks, where deep-learning has the longest history. It is typical to flip, rotate, skew, scale and mask portions of the input image in order to create multiple synthetic samples that retain the same class as the source, but allow a network to learn a more robust set of features that are less likely to over-learn idiosyncrasies related strictly to scale and positioning rather than the content of the image itself. Similarly, [99] applies window slicing and window warping strategies to provide synthetic samples from time-series data.

Following from these techniques, we implemented a data augmentation algorithm that can be applied to discrete time-series events such as those present in the EMR.

After copying trajectories and then randomly truncating the copies to 20-100% of their original length (by dropping the oldest events), time-series events were bucketed into 1 hour windows. 1 hour windows were chosen given the likelihood of meaningless time

distinctions at any higher resolution based on an assumption of primarily manual data entry processes. Events within each of these 1-hour windows were then randomly shuffled and/or masked to create modulated patient trajectories which could be used to augment the input data. In the training set, each trajectory not including the target event was randomly augmented 4 times. Trajectories that included the target were augmented at a rate that was inversely proportional to the time to event (thus emphasising indicators of proximal deterioration), producing a balanced dataset. In the validation and test datasets, all trajectories were augmented 30 times, regardless of outcome.

4.4.6 Final Models

The final model architecture was made up of three sub models that were trained jointly (Figure 4.2).

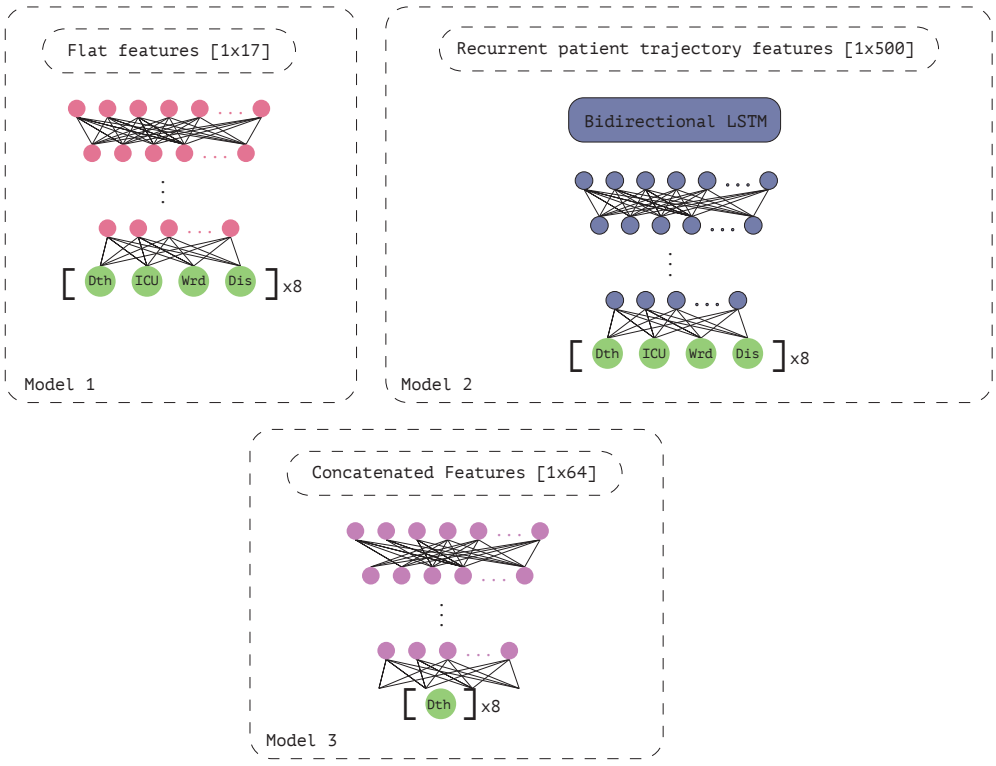


FIGURE 4.2. Model Architecture

Model 1: A flat set of features was created for each admission (see Table 4.1). These flat features were fed into a dense feed-forward network with a 4 dimensional output branch

Feature	Range	Available (%)	Most common
Age (yrs)	18-114	100	71
Marital Status	8 distinct	96.2	Married/partner
Aboriginal or Torres Strait Islander Ethnicity	7 distinct	98.7	Neither
Insurance Status	94 distinct	100	Medicare - overnight
Postcode of residence	1497 distinct	100	Postcode of hospital
Country of birth	220 distinct	98.8	Australia
Relative admission day	0-2000	100	1703
Admission hour	1-23	100	7 (07:00hr)
Admission day of week	0-7	100	3 (Tuesday)
Admission speciality	47	98.6	Emergency
Last discharge code	9 distinct	99.8	9 (discharge alive)
Days since last stay	0-2000	62	2
Last length of stay (LOS)	0-475	62	0
Historical total LOS	0-592	62	0
Historical average LOS	0-237	62	0
Historical ICU hours total	0-2733	62	0
Historical ICU hours mean	0-1665	62	0

TABLE 4.1. Flat demography and historical summary features for each admission

(Death, ICU, Discharge, Ward) for each of 8 time points (12, 24, 36, 48, 60, 72, 84 & 96 hours in the future). Terminal layer activation was set to Softmax, all prior layers had a LeakyReLU activation.

Model 2: The most recent 500 tokens in the patient trajectory were fed into a bi-directional LSTM layer, which then connected to a densely connected network, trained with the same 8 output branches as Model 1. Activations were also set as per Model 1.

Model 3: The 64 output variables from models 1 and 2 were concatenated into a single vector and used to train a densely connected network, with binary outcomes (i.e., death/~death or ICU/~ICU) at each of the target times.

4.4.6.1 Training Process

These models were trained jointly, meaning a single training batch was fed into models 1 and 2, with the resulting gradients back propagated, and then the output of this same batch was fed into model 3 and back propagated before moving onto the next training batch.

The models were trained on all 8 output times (12 to 96 hour forecasts), and then the loss function was modified to attend to the first 4 output times only and trained further in order to prioritise detection of imminent deterioration, whilst still allowing the model to learn from the more plentiful short to medium term deterioration end-points.

A 10% test set was held out with no processing applied until both ICU and Death model training was completed, with the remaining 90% used in a 5-fold cross validation process. At each fold, the training set was split into 80% training, 5% calibration and 15% validation sets. Although it is arguable that the cross validation procedure alone would be sufficient to demonstrate generalisability, due to the many iterative stages of model piloting, a more cautious approach was taken with the holding out of an explicit test set.

4.4.7 Calibration

A reference distribution of risks and uncertainty were produced by generating 300 predictions for each patient in the calibration set as per the validation data. We extend upon the binned calibration methods in [100] to transform the model output into a clinically-meaningful probability of deterioration.

For such short-term deterioration, it is a reasonable expectation that the proportion of patients deemed at low risk will far outweigh those at high risk. As such, instead of the fixed bin-widths in [100], we follow the argument in [101] for the use of unevenly spaced bins to generate measures of calibration quality to its logical conclusion and use these unevenly spaced bins to form the basis of the recalibration function itself.

This distribution was bucketed using a stick-breaking process at the quantiles $[0, 1 - \frac{1}{2}^{0+\alpha}, 1 - \frac{1}{2}^{1+\alpha}, ..., 1 - \frac{1}{2}^{10+\alpha}]$ to generate scoring thresholds that appropriately reflected the far higher proportion of patients in low risk categories. A different α was selected for each category (correct, correct+72 hours, correct in admission) to reflect the different target distributions in the calibration set.

The risk score between 0 and 10 was then generated by comparing the predicted risk for each patient in the test set against these cutoff thresholds.

4.5 Results

4.5.1 Summary Statistics

Input data for these models included 192,883 hospitalisations, belonging to 92,802 adult patients (44.05% female), undergoing 117,658 surgical procedures over the period from June 2008 to June 2016. Patients had between one and 899 visits in the time period. Patients with 100 or more admissions (129 patients - all receiving regularly scheduled dialysis or rehabilitation treatments) were removed from the dataset so that they did not overwhelm the models, leaving a range of 1-99 admissions per patient (mean 2.08 $\pm$ 3.92).

Patients had an average of 3864 $\pm$ 7221 included clinical tokens at admission time. For admissions lasting more than 24 hours, 65 $\pm$ 40 additional events were captured within the first day.

Admissions had one primary diagnosis and up to 44 associated comorbidities (mean 4.63 $\pm$ 4.08). Every admission included by definition at least one ward movement (the ward to which the patient was initially admitted). Detailed summary statistics of the data can be found in Chapter 3.

4.5.2 Endpoint Rates

Data imbalance is a well known challenge in the development of machine learning models. This is particularly relevant when the minority class is the class of interest, which is frequently the case in models that predict mortality, specific diagnoses or other important clinical end-points.

In the source admissions, there was an overall inpatient death rate of 1.53% and unplanned ICU admission rate of 3.22%. These rates change over the course of admission time, however, and drop drastically as the time windows become shorter (see Figure 4.3). At 24 hours after admission, the rate of death in the next 24 hours is 0.35% and for unplanned ICU admission it is 0.61%.

Unplanned ICU admission rates peak in the first day of admission and remain steady after that. Once an admission lasts more than 12 hours, the death rate becomes much higher. This is likely to represent the low death incidence within day-surgery admissions. From 12 hours onwards, the rate rises more gradually as the less severely ill patients are discharged. As death rates rise, unplanned ICU rates fall, which is indicative of an overall increase in acuity over time despite a decrease in instability.

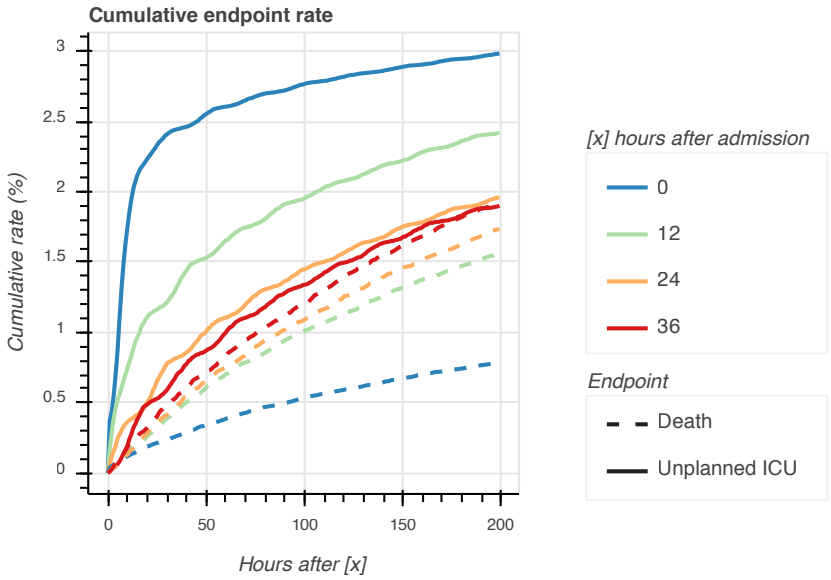


FIGURE 4.3. Endpoint rates in source data, relative to the number of patients still admitted at the given prediction point.

4.5.3 Reported Metrics

We report here metrics that test the output predictions against three measures:

(1) A strictly correct forecast (model predicts endpoint within t hours, and this reflects accurately the presence of this endpoint within t hours).
(2) A forecast that is correct with a clinically relevant tolerance. This tolerance is set to 72 hours (model predicts endpoint within t hours, and this reflects accurately the presence of this endpoint within $t + 72$ hours), to account for patients where similar response from outreach staff may be appropriate, given the desire for early intervention.
(3) A forecast that is correct within the target admission (endpoint is predicted within t hours, and this is not necessarily accurate, however the endpoint of interest does occur prior to discharge). This gives a better sense of the true burden of false positives and false negatives on both patients and outreach staff.

In the example from figure 4.1, there is an unplanned ICU admission at t=72hr, and the patient dies outside of the prediction window, but within this admission. At t=36hr (4), neither endpoint has occurred, so a prediction of false is strictly correct. Unplanned ICU admission does occur within 36+72 hours however (5), and therefore a prediction of ICU=true would be correct within the tolerance window and a prediction of death would be correct within the target admission.

For prediction use-cases with such high degrees of imbalance as those targeted by these models, with far more negative cases than cases of interest, reporting the area under the receiver operator curve (AUROC) alone can be highly misleading [102]. Despite this, it remains the most commonly reported statistic of model quality.

For this reason, we also report here the sensitivity and workup to detection ratio (WDR) for every prediction target. Model sensitivity is calculated as true positive predictions divided by all positive cases, or $\frac{TP}{TP+FN}$. WDR is the inverse of the model positive predictive value, and provides the ratio of all positive predictions to all true positive predictions i.e. $\frac{1}{PPV}$, or $\frac{TP+FP}{TP}$.

Sensitivity is the key outcome measure from the perspective of at-risk patients. This is because a false negative corresponds to potential missed interventions and directly impacts their outcomes. WDR is the key metric for outreach staff however, as an increase in the burden of false positives will heavily reduce the usefulness of any predictive model, and may draw clinicians away from truly deteriorating patients. If balanced appropriately, these measures will result in the predictive model with the highest clinical utility.

[x]	12	24	36	48	60	72	84	96
Death								
Correct forecast	0.918	0.928	0.921	0.915	0.906	0.911	0.902	0.902
Correct forecast with tolerance	0.921	0.917	0.911	0.917	0.903	0.902	0.904	0.901
Target within admission	0.901	0.902	0.903	0.902	0.890	0.890	0.891	0.890
Unplanned ICU admission								
Correct forecast	0.747	0.778	0.777	0.776	0.782	0.776	0.789	0.781
Correct forecast with tolerance	0.754	0.783	0.779	0.774	0.781	0.779	0.789	0.786
Target within admission	0.725	0.757	0.743	0.750	0.757	0.753	0.768	0.767

TABLE 4.2. Area under the receiver operating curve for prediction within [x] hours, using data available 24 hours after admission time.

Note that the AUROC frequently decreases as the tolerance increases, which is somewhat counter-intuitive, since a more permissive calculation could be expected to necessarily improve model performance. This is due to the fact that the tolerance does not only increase for the model predictions, but also for the model targets. Thus as the target event frequency increases the sensitivity calculation changes in both a positive and negative fashion, as more targets are correctly specified but more again are missed. This illustrates further the issue with reporting AUROC as the sole metric of model performance.

4.5.4 Mortality Prediction

At 24 hours after admission, death within the following 24 hours was predicted with an AUROC of 0.928 (see Table 4.2 for all time points). This is higher than the baseline score NEWS [80] (0.89), however as outlined above, this measure alone is unlikely to tell the whole story of model utility. Note also that the NEWS baseline could not be replicated in the source data due to the unavailability of patient vital signs so is compared only to the AUROC as reported in the cited study.

Figure 4.4 demonstrates the discriminative value of this model, i.e. that the output does indeed correspond to prediction of clinically meaningful risk. Although the sensitivity is poor at the earliest time point (due to the enormous class imbalance) later forecasts can be expected to correctly predict between a quarter and a third of patients who will deteriorate rapidly. Sensitivity drops as the tolerance increases to 72 hours, as there is now a higher proportion of target events. The workup to detection ratio decreases much more rapidly, however, demonstrating that the clinical burden of a false positive in this model is low, and that responding to a patient with even moderate risk is likely to be worthwhile.

There are a number of reasons that can explain the observed plateau of risk for most predictions from 48 hours onward seen in Figure 4.4. In referring back to Figure 4.3, we note that as duration of each admission extends beyond 24 hours, the rates of each endpoint observed in the data relative to the patients still admitted at that point becomes steadier, indicating that the cohort that is still admitted at this time becomes more stable, despite its increased overall average severity of illness. A patient's overall general risk increases with time as lower severity patients are discharged, but their risk within a given time window levels off. This is also an artifact of the training methodology whereby the models were initially trained on all 8 output times, before further training was done attending to the first four time-points only. This was done to balance the priority of predicting patients at risk of imminent deterioration with the lack of available data for these shortest time-points, however, it does somewhat mute the discrimination between medium and longer-term risk predictions. This is deemed to be an acceptable trade-off due to the original remit of this work, which had a strong preference for predicting short-term deterioration.

4.5.5 Unplanned ICU admission

There is a significant difference between the AUROC of the mortality prediction models and the corresponding unplanned ICU admission models. This is likely to be due to the fact that ICU admission criteria are strongly coupled to vital sign triggers, and therefore a prediction model that does not include this data will under-perform.

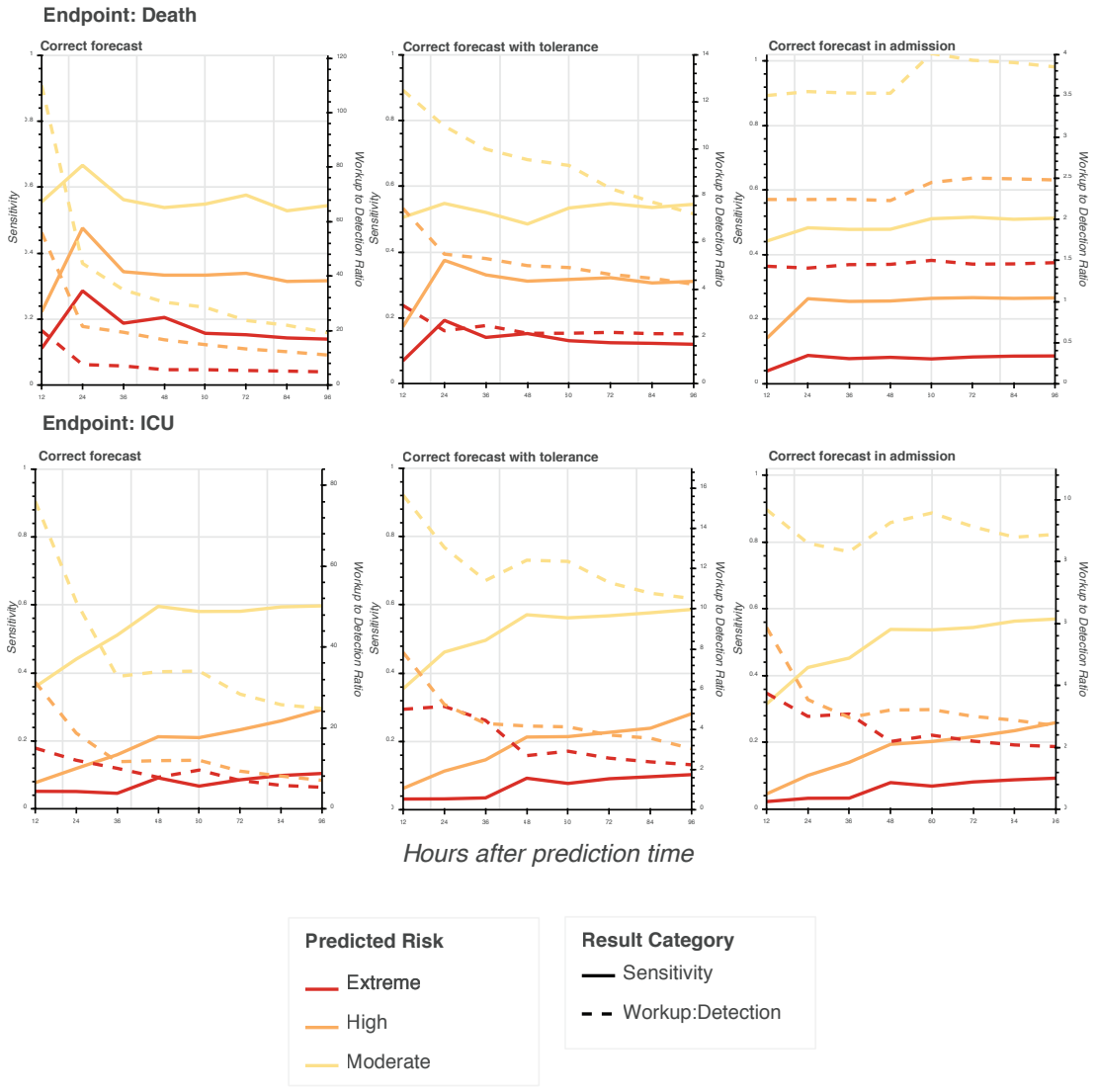


FIGURE 4.4. Mortality and unplanned ICU prediction — sensitivity and WDR of death prediction at future time points using data available at 24 hours after admission. For the purposes of risk stratification, extreme risk is here defined as a calibrated risk score of 6 or more, high risk as a score of 4 or 5, and moderate risk as a score of 2 or 3.

Despite this, from Figure 4.4, it remains possible to predict unplanned ICU admissions within the following 48 hours with a sensitivity of around 20% of all cases, and a corresponding WDR of 1 in 17. When allowing a 72 hour window of tolerance, a WDR of 1 in 12 gives up to 60% sensitivity, and therefore still represents a tool with meaningful clinical applications.

4.5.6 Model Calibration

The raw results produced by this model had poor calibration, despite their good discriminative power, meaning that the probabilities output by the models could not be directly interpreted as the actual probability of the event occurring. This is typical of neural net techniques [101], which tend to be overconfident, or 'sharp' in their predictions.

There was a very low positive class count (not only proportionally, but also numerically) in the small calibration set. This meant that typical recalibration methods of isotonic regression [103] and Platt scaling [104] were ineffective (see Figure 4.5), and it also put techniques such as [105] out of reach.

We find that the highest probability that we can assign to precise death forecasts is 40%, deaths within 72 hours of their forecast time have a maximal confidence of 80% and in-admission death has a maximum confidence of 90%. This matches the expectation that clinical trajectories are non-deterministic, particularly over the short term, but as the precise prediction time expands, confidence increases.

4.5.7 Implementation in External Dataset

In order to demonstrate an external validation of this model, we have re-implemented the full pipeline in the MIMIC-IV dataset [106]. Due to the lack of electronically-recorded vital signs at the source institution it was not possible to implement NEWS directly as a baseline, so the implementation in this additional dataset also provides the advantage of being able to compare against this commonly-used benchmark score.

4.5.7.1 NEWS Baseline in External Dataset

The MIMIC-IV dataset includes 523,740 admissions from 256,878 patients. We calculated a NEWS score for every admission that had at least one full vital-signs set recorded in the first 24 hours of their admission. Where more than one recorded value exists within the

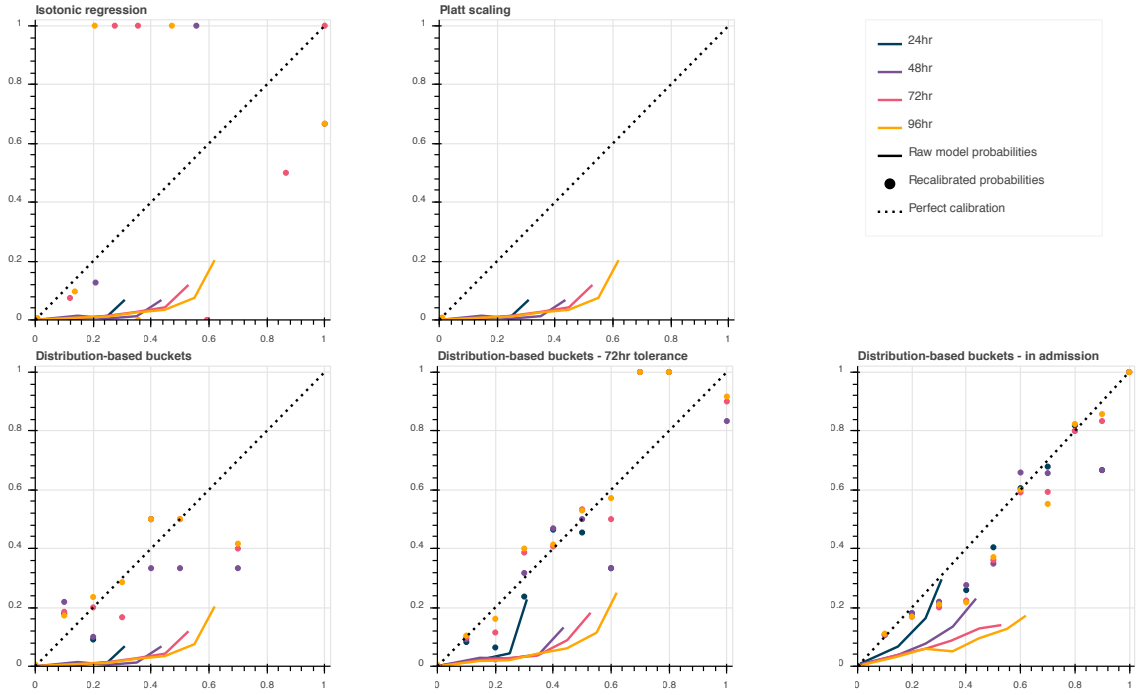


FIGURE 4.5. Recalibration techniques for death model predicted at 24 hours after admission. Note that Platt scaling reduces all probabilities to a single point close to the origin.

first 24 hours the most recent was used. It was possible to calculate a full NEWS score for 53,528, or around 10% of all admissions. Of these, 1209 admissions were shorter than 24 hours in duration and 854 patients died within the first 24 hours, leaving 51,465 admissions for which a comparison score could validly be calculated at the prediction time. In 4474 of these admissions (8.69%) the patient died before being discharged from hospital and 684 (1.33%) died within the next 24 hours. Applying the standard NEWS calculation, the resultant AUROC for in-hospital death and death within 24-hours were found to be 0.76 and 0.86 respectively. Death within 24 hours was similar to the benchmark reported in the original NEWS development study [80] (0.89), with the small difference likely attributable to the different composition of the cohort (ICU patients instead of all medical admissions) and the relatively low proportion of admissions having a full set of vital-signs within the first 24 hours.

4.5.7.2 External Validation

In order to perform an external validation, tokens were similarly collected from the MIMIC IV dataset that most closely resembled the tokens used in the source data. This meant that vital sign values were not utilised, despite the fact that in this case, they are indeed available. The tokens used for validation included medication administration events, procedures, historical diagnoses, ward movements, pathology results (tokenised by quantile) and historical demographic data.

The flat portion of the model was also calculated similarly, including age, marital status, insurance status, and numeric features describing the duration and recency of a patient's historical admissions. The results for this model can be found in Table 4.3.

The NEWS baseline, with its access to vital signs data, clearly outperforms the simple ML models for death within 24 hours, although this does not carry across to in-admission death. This matches expectations for the importance of vital signs for imminent deterioration, whilst simultaneously boosting the case for using administrative data to predict longer term patient trajectories.

The application of the same strategy to a new dataset outperforms all other candidate models in this data, in many cases by a significant margin, proving its viability in new settings.

4.5.7.3 Comparison to Non-recurrent Models

It is a key hypothesis of the design of this solution that there is value in the use of a recurrent model, over and above simpler and potentially more scrutable ML architectures, due to the time-series nature of the patient trajectory. As a point of comparison, we therefore trained models using 6 different supervised algorithms (logistic regression, XGBoost, simple feed-forward deep-learning network, random forest, AdaBoost and a Bayesian network). In order to do this, features were selected that were most highly correlated with in-hospital death and least strongly correlated with each other. The candidate features for

the algorithm were those that most strongly resembled the features used in the recurrent model, i.e. demography, historical admissions, historical diagnoses, procedures, pathology results and medication administration events. Resultant AUROC, WDR and sensitivity values are given in Table 4.3.

Model	Death within 24 hours			Death within hospitalisation		
	AUROC	WDR	Sens.	AUROC	WDR	Sens.
Logistic Regression	0.57	42.4	0.5	0.78	5.6	0.49
XGBoost	0.85	6.3	0.5	0.81	2.7	0.5
Feed-forward deep-learning	0.83	9.7	0.5	0.73	4.2	0.5
Random Forest	0.80	7.3	0.5	0.81	2.9	0.5
AdaBoost	0.84	7.1	0.5	0.81	2.9	0.5
Bayesian Network	0.82	10.8	0.5	0.79	4.6	0.5
NEWS	0.86	12.9	0.55	0.76	4.9	0.56
CCO watch-list (ext. validation)	0.89	4.5	0.5	0.88	2.1	0.5

TABLE 4.3. Area under the receiver operating curve, workup to detection ratio at sensitivity cutoff as close to 0.5 as available in the distribution, using data available 24 hours after admission time in the MIMIC IV dataset. Note that a threshold of 0.5 for sensitivity is not available for NEWS, given its discrete scoring.

4.6 Discussion

4.6.1 Source Data Limitations

Scores or tools that target imminent patient deterioration typically aim to detect derangement of physiological signs and symptoms. This is based on the observation of predictable patterns of changes in patient vital signs prior to each of the relevant deterioration endpoints cardiopulmonary arrest, unplanned ICU admissions and death [1–3, 107, 108].

Although a physiological early warning score (EWS) is used as a manual trigger of emergency response at the target institution [109], due to a lack of availability of vital sign data within the EMR, it is not currently possible to use such a score as the basis for a fully automated watch-list.

This, along with variable importance analyses in logistic regression models such as [90], serve to highlight the importance of vital sign data as the key element underpinning the vast

majority of current best practice for prediction of inpatient deterioration. The limitation seen in our data is a realistic one, however, that should be considered for implementation of a fully automated system. It is characteristic of many EMR systems to serve the purposes of hospital administration first, and support clinically relevant data only where this aligns with the requisite billing and logistical goals, and/or where the clinical utility is high enough to justify the additional documentation burden above what can be provided with paper charts. Thus, it is unsurprising to observe in this data set that all theatre-based procedures are fully available in the clinical record, as they are not only billable, but also require the booking of resources from a central pool, compared with typical bedside procedures and nursing observations that go unrecorded for the inverse reasons.

This limitation in the breadth of input data is significant, however encourages a model that is built primarily around administrative data points, which are likely to be more reliably and consistently available in the EMR.

4.6.2 Error Analysis

In order to understand the limitations of this model in these contexts, and to inspect the model for evidence of causal leakage, we ran the false positive samples with highest predicted risk (predicted death within 36 hours with a probability of 0.6 or higher but discharged alive) and the false negative samples with lowest predicted risk (died within 24 hours but death probability at 96 hours was lower than 0.2) through the LIME Text Explainer module [110]. LIME is an algorithm that provides insights into a 'black-box' model by learning a locally interpretable model that can explain which input data was most relevant to a given prediction. In contrast to SHAP [111], the LIME methods are model-agnostic, and therefore possible to apply to a nested set of models such as those developed here.

There was no evidence of causal leakage, with no highly weighted tokens that reflected the target endpoints directly.

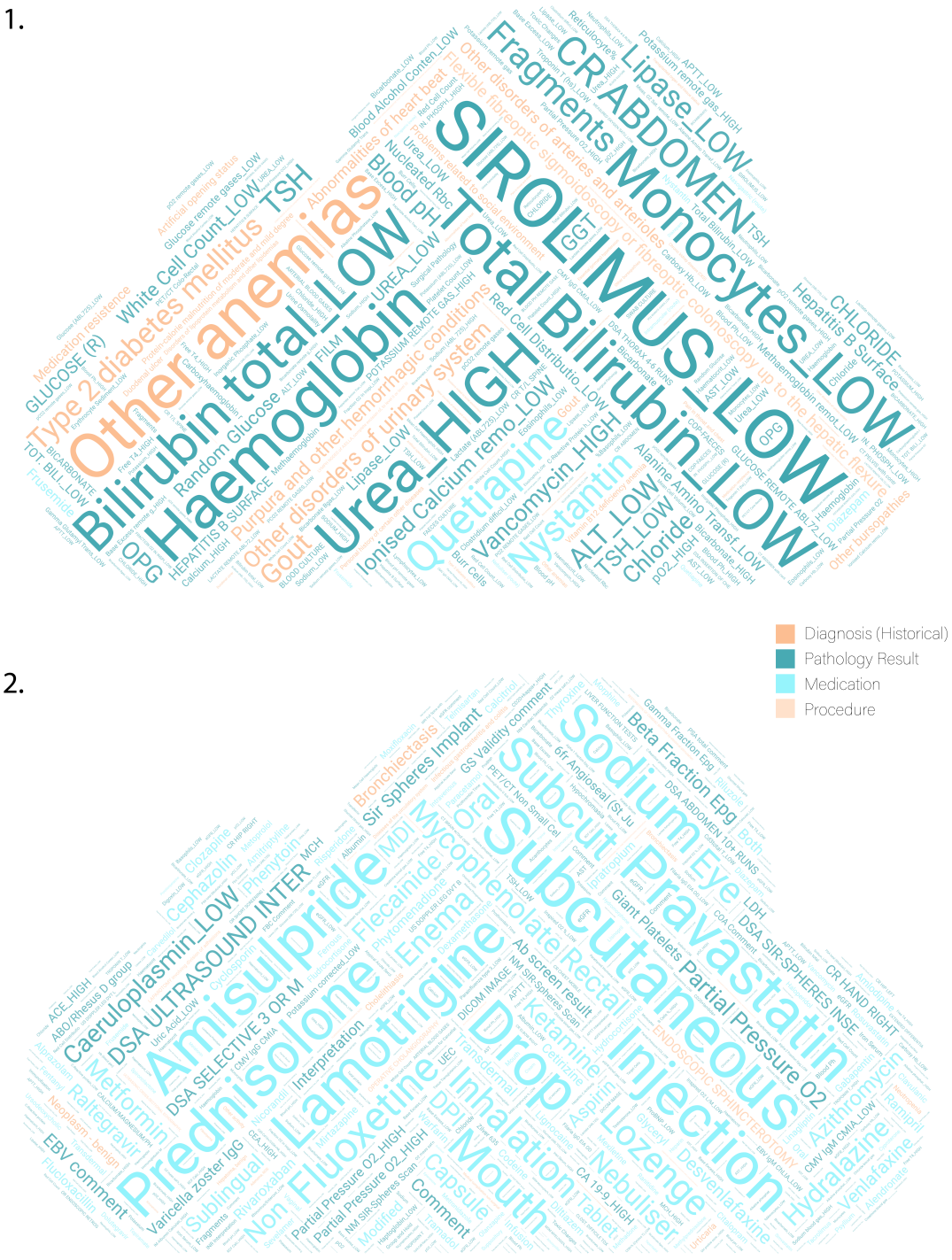


FIGURE 4.6. Word clouds demonstrating the most highly weighted terms for (1) false positive predictions and (2) false negative predictions.

In the word clouds in Figure 4.6, the size of a word corresponds to its weighted frequency as associated with each error type (false positives and false negatives). There is a clear pattern between the factors that contribute strongly to a prediction of high risk versus those contributing strongly to a low risk prediction. Lab results are generally indicative

of a risk increase, where medications and medication-related tokens dominate lower risk predictions.

For false negatives, most of these drug terms represent the highest-frequency tokens in the corpus. Their interpretation therefore is limited to the fact that they are evidence of a sort of regression to the mean, where these patients simply do not have enough distinctive data at the point of prediction to make an accurate risk assessment. Overall, despite having a comparable number of unique tokens, the medication terms each individually tend to have higher frequency than other token types. This holds true even when accounting for the repeated administration of medications, as these tokens on average each appear in more distinct patient trajectories than other event token types (excluding ward movement tokens).

In the list of terms contributing to false positives, there are numerous terms that may indicate that the patient has a complex history or is in a high-risk category, e.g. low white cell count, high blood urea, medication resistance, artificial opening status, sirolemus testing, low lipase. There are also, however, terms that either don't have a sensible interpretation with respect to deterioration risk, e.g. low bilirubin, low blood alcohol content, Nystatin administration, or that are not sufficiently specific to make an informed interpretation of risk e.g. anaemia, sigmoidoscopy procedure, abdominal x-ray. This system is therefore insufficient to provide directed actions or interventions and its use must be limited to the prioritisation of attention.

4.6.3 Congruence with Current Clinical Practice

The use of rapid response systems is intended to act as a safety net for deteriorating patients via the monitoring of a standardised subset of patient vital signs. It has, however, been argued that this drives nursing practice towards the detection of deterioration that is already well underway, as opposed to highlighting at-risk patients who are yet to go downhill [112]. By removing the reliance on vital signs, this model affords the capacity to move away from detection and into the realm of prediction.

Studies have also found that workloads and hospital work culture affect the likelihood of staff triggering rapid response calls according to the prescribed protocols [113]. Although calling criteria are nominally specified to allow triggering of the rapid response protocol based on clinical intuition alone (even when vital signs based criteria are not yet met) nursing staff who wish to act upon early signs of deterioration report themselves to be reluctant to do so in the face of potential criticism. This is true despite the fact that nursing intuition can preempt deterioration identified by vital signs alone [114]. A system that is able to provide contextualisation of such minor changes in patient state is therefore well placed to augment existing escalation protocols.

4.6.4 Comparison Models

As a baseline, we present in Table 4.4 a selection of models that have been developed with the goal of detecting the early stages of short-term patient deterioration in a general ward population. Not all of these baselines can be compared directly to the models presented in this work due to the variability of endpoints and prediction times, giving instead an overview of the general targets and performances in existing models.

Note that it is only possible to compare WDR to baselines reported in different populations if a fixed incidence rate is chosen in order to standardise this measure. Where it was possible to make this calculation, the fixed rate was set to 0.35%, which is the death rate within 24 hours in this population, per section 4.5.2.

The traditional models were identified from a recent review paper that is closely aligned with the target use-case [115] in addition to the NEWS model [80], which is a highly cited and widely implemented early warning score that forms the basis for comparison for many similar works.

In order to capture potential deep learning baselines, the reference list of two systematic reviews [116, 117] were filtered to identify EMR-based patient deterioration prediction models. General deterioration endpoints not applicable to the CCON/CCOM role were excluded, e.g. readmission, death other than short-term, or studies only applicable to

patients already within the ICU. Notably, many deep learning models do not fit our use-case as they either predict only inpatient or longer-term mortality e.g. [81, 118, 119], target a specific morbidity such as congestive heart failure or sepsis e.g. [120, 121] or are developed using data for patients already admitted to the ICU e.g. [122, 123] (largely due to the wide utilisation of the freely-available MIMIC-III database [124]). [81] was retained as the deep learning baseline, as it is closest to meeting the target use-case. Interestingly, this reference uses the NEWS model as a mortality baseline, despite the fact that NEWS was developed to detect 24-hour mortality where the deep learning model predicts inpatient mortality.

This summary of baselines exposes a number of issues with the comparison of such predictive systems. In particular, the precise definition of endpoints is inconsistent. We also note that all mortality endpoints reported here are for in-hospital mortality only, i.e. they are unable to report full mortality as an endpoint due to the lack of data linkage and potential loss to followup. Only Kipnis et al [91] have access to network-level data linkage, but this is not utilised as a primary endpoint. Rajkomar et al [81] go further by redefining readmission to include only readmission to the same institution. The availability of linked data as per [125] would provide additional insight and allow expansion of these models to include identification of patients at the end of life.

TABLE 4.4. Comparison to baseline models

Model	Target Endpoint	Incl/Excl Criteria	Prediction Time	AUROC	Sens.	Spec.	Standardised WDR
		Traditional models					
NEWS [80]	In-hospital death (within 24hr)	Ex: Discharged before midnight of admission day; admitted directly to ICU	Time observations taken in medical assessment unit	0.89	-	-	-
	Unplanned ICU (within 24hr)			0.86	-	-	-
	Cardiac arrest (within 24hr)			0.72	-	-	-
	Combined 24hr deterioration			0.87	-	-	31.5
Alvarez et al [88]	Resuscitation events and death	Inc: Adult patients admitted to internal medicine ward or ICU. Ex: admitted directly to surgery; DNR order at admission; obstetrics admission; events on first day of admission	Daily prediction	0.85	0.52	0.94	35.6
Churpek et al (a) [89]	Cardiac arrest (in admission)	Inc: Adult patients with documented vital signs	Every 8 hours	0.88	-	-	-
	Unplanned ICU (in admission)			0.77	0.54	0.90	55.6
	Cardiac arrest (within 24hr)			0.88	0.65	0.93	33.2
	Unplanned ICU (within 24hr)			0.76	-	-	-
Churpek et al (b) [90]	Combined 8hr deterioration	Inc: Adult patients with documented vital signs	Every 8 hours	0.80	0.50	0.93	42.8
Kipnis et al [91]	Combined 12hr deterioration	Inc: Adult patients. Ex: out of network transfers; childbirth admissions, 'comfort care only' orders.	Hourly	0.82	0.49	0.92	49.5
Green et al [92]	Combined 24hr deterioration	Inc: All admissions.	At time of vital sign observation	0.80	0.50	0.90	59.9
		Deep learning models					
Rajkomar et al [81]	In-admission death	Inc: Length of stay $>$ 24hr; adult patients	24hr after admission	0.95	-	-	-
CCO watch-list (this work)	Unplanned ICU (within 48hr)	Inc: Length of stay $>$ 24hr; adult patients, fewer than 100 visits, not admitted directly to ICU	24hr after admission	0.77	0.50	0.88	71.2
	In-hospital death (within 24hr)			0.93	0.47	0.97	21.3

Notes: Where more than one result available for same end-point, result with highest AUROC is reported.

Where more than one prediction time is available, most clinically relevant prediction time for that end-point is reported.

Where multiple cutoff points are available, sensitivity and specificity are reported as per review paper [115].

Workup to detection ratio is only reported where it is possible to standardise this measure to a fixed reference prevalence rate.

4.6.5 Data Processing

Many clinical prediction scores rely on highly regulated data collection that may not reflect existing clinical processes, thus requiring additional data entry or hand calculation. Our noisy dataset reflects true practice and availability, with pre-processing limited to routines that can be performed with no human input. Within this pre-processing of data, we do not attempt to normalise the labelling of medications and pathology — e.g. different spellings are present for the same test across different panels — instead, allowing contextual embeddings to handle this noise.

Because we rely on the naturalistic data ecosystem, rather than one requiring abstraction, we assume that we are reducing errors caused by hand calculations or operational error, and robust to errors preexisting within the EMR. The trade-off with this strategy is that we cannot expect these models to achieve generalisation in a new setting without re-training to accommodate local vocabularies and idiosyncrasies of data entry. An external validation study will therefore require translation of the entire model pipeline, rather than transfer and mapping of only the model inputs themselves.

To this end, the full breadth of the clinical record that was available for this project was incorporated in the input data. The tokenisation procedure included a lower-frequency bound whereby tokens appearing in fewer than 0.5% of patient trajectories were replaced with a placeholder 'RARE' token, but beyond this, there was intentionally no attempt to manually remove low-information features. We do this on the assumption that the more hands off we are in data preparation, the more robust the results will be to changing practice and the lower effort required by both implementers, and the end-users. We also do not make any effort to handle multiple recordings at the same time, or detect outliers for this same reason. This is similar to the data preparation strategy in [81], which is promoted by those authors as a scalable approach, creating a model that has real-world productionalisable qualities.

Implicit in this strategy of hands-off data preparation and delivery of a pipeline, rather than a model that would be translated without retraining, is that any and all manual curation of patient sub-types is out of scope. This follows existing all-cause mortality and deterioration

models such as [80, 81, 125], and is supported by the capacity of the contextual embeddings to sub-type patient classes as a by-product of the training process itself.

4.6.6 Calibration Measure

It is not feasible to calculate the Hosmer-Lemeshow statistic of calibration for this model due to the large sample size and excessive degrees of freedom [126, 127]. Alternative calibration statistics were reviewed for their applicability such as [128], however were found to be unsuitable due to their focus on density. This makes sense for many use-cases, where it is valuable to prioritise areas of the calibration curve that represent the majority of samples, however in this situation it is not suitable, as the differences between probabilities at the low end of the risk scale are not clinically meaningful. Instead, the differences in the most sparse regions must be prioritised — outreach staff may be expected to treat patients at 80% risk quite differently to those at 90% risk, despite there being very few patients in those risk categories, where their response will differ very little for patients at 10% risk vs. 20% risk.

This knowledge-based interpretation of the utility of a model's calibration cannot be quantified without some parameters set by target users a priori.

4.7 Conclusion

Based on these results, we can conclude that it is technically feasible to build a set of predictive models that meet the needs of the critical care outreach role, based on a limited set of real-time clinical data. These models compare favourably with the current practice of using physiological early warning scores to highlight deteriorating patients when compared numerically in terms of accuracy, AUROC and workup to detection ratio, although there remains a significant amount of work to successfully implement them in practice.

CHAPTER 5

Augmentation of Electronic Medical Record Data for Deep Learning

---

5.1 Preamble

This chapter has been submitted for publication as [18], and is reproduced as submitted, with the exception of this preamble and minor updates to cross-referencing for thesis cohesion.

In this chapter, we demonstrate the result of applying the novel time-series data augmentation strategy described in Chapter 4. Although this technique is mature in the domain of image processing, it is uncommon as applied to discrete time-series data. The confirmation of generalisability in a publicly available dataset serves as the basis of a validation study in external data, and quantifies the effect of this domain-specific data processing technique.

We also provide more in-depth implementation details in the form of published code, for the purposes of research reproducibility.

Beyond the obvious benefit of being able to make code available that is useful in a public dataset, the other reason for demonstrating the specific effect of this technique in the MIMICIII dataset instead of the core dataset of this project was due to the significantly higher costs of the secure infrastructure required to process this private data compared to standard cloud compute resources.

5.2 Abstract

Data imbalance is a well-known challenge in the development of machine learning models. This is particularly relevant when the minority class is the class of interest, which is frequently the case in models that predict mortality, specific diagnoses or other important clinical end-points.

Typical methods of dealing with this include over- or under-sampling training data, or weighting the loss function in order to boost the signal from the minority class. Data augmentation is another method that is employed frequently — particularly for models that use images as input data. In the case of discrete time-series data, however, there is no consensus method of data augmentation.

We propose a simple data augmentation strategy that can be applied to discrete time-series data from the EMR. This strategy is then demonstrated using a publicly available data-set, in order to provide proof of concept for the work undertaken in Chapter 4, where data is unable to be made open.

5.3 Background

5.3.1 Premise

Clinical prediction models frequently target rare endpoints such as mortality within a specific time-frame or other adverse events. This is a known challenge when developing machine learning models [98], as it is easy to over-train to the majority set, producing a classifier of high accuracy, but low utility.

In machine learning by gradient descent, the weights of a model are updated based on the overall distance of the model output from the target state using the gradients of a predefined differentiable function. This function can act as either a penalty to be minimised (the 'cost' of each error), or a target to be maximised (the 'reward' for each correct output).

For simplicity, we will refer to cost functions and minimisation for the remainder of this work, however can be simply extrapolated by reversing the target class definition ($true = 1$ becoming $true = 0$).

If each training data point contributes equally to this cost function, in a data set with a large imbalance between the majority and minority class the calculation quickly favours accuracy in the majority and will err on the side of under-classifying the class of interest. Under extreme levels of imbalance, this is true even in the instance where there is a strong signal from the minority class.

Formally, if the model M takes inputs $\bar{x}$ and produces predictions $\hat{y}$, we calculate the loss $\mathcal{L}$ as the sum of the cost C across a batch of size n, where C is some predefined distance metric between $\hat{y}$ and the target labeled output y.

$$\hat{y} = M(\bar{x})$$
$$\mathcal{L} = \sum_{i=1}^{n} C(y_i - \hat{y}_i)$$

If the classes are imbalanced by some factor imb, we can separate samples $\bar{x}$ and labels y belonging to the majority and minority classes into $(y_{maj},\ y_{min})$ and $(x_{maj},\ x_{min})$ respectively such that $x = x_{maj} \cup x_{min}$ and $x_{maj} \cap x_{min} = \emptyset$, with $imb * n$ samples in (y_{maj}, x_{maj}) and $(1 - imb) * n$ samples in (y_{min}, x_{min}). In the case where $imb = 0.5$ (a balanced data set), the loss for each batch is equally dependent on costs from each class. As $imb \to 1$, $\mathcal{L} \to \mathcal{L}_{maj}$, increasing the likelihood of simply learning a majority class classifier, which predicts the majority class for all input samples.

If applying an under- or over- sampling strategy, the loss function is artificially balanced by masking a portion of (y_{maj}, x_{maj}) or repeating a portion of (y_{min}, x_{min}) respectively, until the signal of each class is able to affect the final model weights at a level that is appropriate for the particular use-case.

5.3.2 Data Augmentation

Data augmentation is an alternative to oversampling, where instead of repeating the same samples exactly, synthetic samples are created and used to expand the dataset more richly than repetition alone.

Data augmentation in this context has two goals. If samples belonging to each class are augmented at a rate that is inversely proportional to their imbalance, this has an effect equivalent to an oversampling strategy as described above. In addition to this, it is possible to introduce an element of spatial or temporal invariance that improves the ability of the model to recognise patterns in unseen samples [129]. In an image classification task for instance, one would not want the model to rely on the precise orientation or positioning of the input to be able to detect the presence of the target class. Thus, by repeating each input image with random rotation, scale and skew factors, the model becomes robust in the face of input images that were captured in different contexts.

More recently, data augmentation strategies using generative adversarial networks (GANs) have been applied to data from the electronic medical record (EMR) with some success [130], although this brings with it some additional challenges due to the complexity of the implementation and cost of significant additional model training. A GAN uses two models with opposing (adversarial) goals to produce realistic data samples — a generator network that creates synthetic data and a discriminator network that tries to differentiate these synthetic samples from the real data. As the discriminator becomes unable to differentiate between real and generated data samples, these samples are deemed sufficiently realistic, and treated as though they were part of the original dataset. This has been applied with success in medical image analysis [131], which are atypical images in their uniformity of scale and aspect. It is less common in other image domains, likely due to the availability of other more straightforward methods such as applying transformational filters that are not applicable to medical images (a skewed or scaled chest x-ray, for example, loses information that is relevant to the prediction task).

It is possible to augment continuous time-series data in an analogous way, where noise can be added and filters applied in order to generate additional training samples that can

improve model generalisability [99] — see examples in Figure 5.1. Discrete data, however, are more challenging to modify in this manner, as noise and multiplication factors are meaningless. The problem of finding a generalised solution for discrete, ordered tokens (as found in text or EMR data) is a known challenge [132]. We propose instead, a domain-specific method of augmentation, which makes clinically relevant assumptions about the way data is entered into the source system.

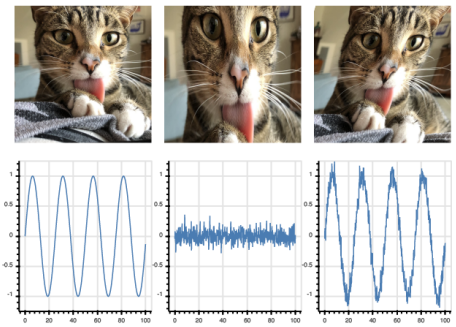


FIGURE 5.1. Examples of typical augmentation strategies for image and continuous time-series data. Top L-R: original data, scale, shear transformations; Bottom L-R: original data, generated noise, transformed signal.

5.4 Methods

5.4.1 Data and code

The source data for this work is an excerpt of the MIMIC-III Clinical Database [124, 133]. This dataset was accessed using the Amazon Web Services Athena Cloud Formation scripts provided by MIT-LCP [134]. Code that builds on these scripts to produce the results in this paper can be found in the Github repository https://github.com/CBDRH/PaTMan. These models were built using the TensorFlow library [135], version 2.0.

5.4.2 Input data

This dataset contains 61,500 ICU admissions across 57,773 hospital admissions, belonging to 46,646 patients. Hospital admissions without an 'admit' record in the transfers table are excluded, as these represent either newborns or incomplete records.

As input, we generate predictions only for the first ICU admission in each hospitalisation. Hospitalisations where the patient was discharged to general wards within 6 hours of their index ICU admission and where the patient died within the first 6 hours of their index ICU admission were excluded, leaving a final total of 52,770 included index ICU admissions.

5.4.3 Endpoint targets

In order to demonstrate this technique, we selected three prediction targets, each having a differing level of endpoint imbalance.

Endpoint	**Count True**	**Count False**	**Class Imbalance**
Death within this ICU admission	3346	49,424	0.94
Death within this hospital admission	5304	47,466	0.90
ICU admission duration $>$ 7 days	7915	44,855	0.85

TABLE 5.1. Endpoint distributions

5.4.4 Tokenisation

Discrete clinical events were gathered for patient demographics, historical admissions, historical diagnoses and historical ICU admissions. Pathology results were converted to discrete tokens according to their decile within all input data i.e. [test type]-[decile], or by [test type] alone for non-numeric results.

These tokens were concatenated in as a temporally ordered list, which describes the patient trajectory over time, e.g. [Admission, Female, 75, BUN-9, GFR-2, Ultrasound-Kidney, ...,

Discharge, N17.8, ..., Admission, Female, 77, ..., ICU-Admission, ...] describes a patient with one prior admission and a history of kidney failure. Each trajectory contains the most recent 500 events that occur prior to prediction time. Diagnoses from the current admission are not included, as coded diagnostic information is not available in real-time. This description of patient trajectories in tokenised form is equivalent to the pre-processing described in Chapter 4.

5.4.5 Model architecture

A simple model architecture was implemented, with a small set of hyperparameters tested for each prediction task. Two versions of the model network were implemented with either LSTM or GRU bidirectional recurrent layers of 5, 10 or 15 nodes, in order to observe the robustness of the technique across simple architecture changes. This set of piloted architectures was held the same across all prediction tasks, as the purpose of this work is to demonstrate the effect of the augmentation strategy, rather than to produce the most precisely accurate classifier for each endpoint.

5.4.6 Augmentation strategies

We make a number of assumptions about data within the electronic medical record that allow the creation of augmented samples that can be used to improve model accuracy.

Temporal ordering is of course significant when determining whether or not the patient trajectory is trending towards recovery or deterioration, however it is unlikely to matter at a resolution shorter than one hour in duration. The data entry workflow is not instantaneous, and can be modulated by systems that are outside of the scope of the target patient's condition, e.g. the precise time that a pathology result is returned or manual data entry is completed may be heavily affected by the overall workload of the hospital on a given day. We therefore bucket data into time windows and randomly shuffle events in each of these buckets before reassembling the trajectory in order to increase the number of available samples.

We also assume that the length of available patient history is only somewhat related to patient outcomes. A more complex history of interactions with the healthcare system can be expected to indicate a more severely ill patient, however this dataset was not generated within a closed system of care, and therefore the lack of available history data does not strictly indicate that it does not exist, as patients may have interacted with numerous other providers prior to this admission. Thus, after bucketing and shuffling of data, we randomly truncate patient trajectories by dropping up to one third of the oldest events in each sample.

Finally, clinical data entry is a noisy process, affected by many external forces, and therefore we assume that up to half of each patient trajectory could be randomly masked without changing the clinical interpretation.

By combining these strategies multiple times, we generate additional samples proportional to the input distributions to train each model.

5.4.7 Time to event weighting

The closer a patient is to time of death when a prediction is made, the more extreme their deterioration risk. Similarly, the longer the eventual ICU admission, the higher impact that early intervention may have on their overall trajectory.

We expect that amplifying the signal for subjects with the strongest evidence of deterioration risk will improve the overall calibration of our models.

For death endpoints, time to event (TTE) was set to the number of days until death at prediction time and the weighting was inversely proportional to this value (i.e. more repetition of data for subjects with lower time to death). For the long ICU admission endpoint, the TTE parameter was eventual ICU admission duration in weeks, and the weighting was directly proportional (higher repetition for the longest overall ICU admissions).

5.4.8 Data balancing

5 balancing strategies were tested:

(1) **None:** Input data was fed to the model according to the original distribution.
(2) **Oversampling (simple):** Minority class samples were randomly repeated at a rate required to approximately balance the input data.
(3) **Oversampling (TTE):** As per oversampling strategy, except the rate of repetition is instead calculated based on the time to event for the minority class. The total repetition rate is equivalent to the rate for simple oversampling.
(4) **Augmentation (simple):** Minority class samples were randomly augmented (first shuffling, then either truncating or masking). For data augmentation, we augment both majority and minority class samples, holding the ratio of these rates equivalent to the same rate as per simple oversampling.
(5) **Augmentation (TTE):** As per augmentation strategy, weighted based on the time to event for minority class.

5.4.9 Evaluation Framework

5.4.9.1 Standard metrics

Given the rare targets of these prediction models, we follow our previous work in Chapter 4 in reporting additional metrics to provide the necessary context that can be obscured by reporting the AUROC in isolation [102]. Specifically we focus on the effect of different training strategies on the workup to detection ratio (WDR) versus sensitivity, as this gives a concrete measure of the excess workload on clinicians (i.e. how many patients they must assess for each one correctly targeted intervention) as compared to the potential benefit to the patient (i.e. what proportion of truly at-risk patients are correctly highlighted by the model).

5.4.9.2 Model calibration

In order to combat the known issue of poor calibration of deep-learning models [101], we follow the same calibration process demonstrated in Chapter 4. This strategy uses the distribution of predictions generated for a held-out calibration set to establish reference cut-off thresholds that reflect the expected distribution of the target event.

These quantiles are set using a stick-breaking process, which generates 10 thresholds that are then transformed to produce a risk score of between 1 and 10 for each input trajectory. The stick breaking process is defined such that approximately the same proportion of inputs are classified 'high risk' (risk score of 5 or more) as the observed proportion in the calibration set. In practice for the most rare events this makes the high-risk bands very narrow and the low-risk bands quite wide, reflecting the expectation that many more patients will be at low risk of experiencing these rare target events than will be at high risk.

5.4.9.3 Risk stratification

Setting a score of 5 or more as 'high risk' and a score of between 2 and 4 as 'medium risk', we report the sensitivity, specificity and workup to detection ratio at each of these thresholds. This represents a likely end-user workflow, where patients at high risk of deterioration can be triaged and attended to preemptively. It also gives a more clinically relevant and interpretable indication of model performance than area under the receiver operating characteristic curve (AUROC) alone, which can be insufficient for fully understanding model performance for low prevalence events.

5.5 Results

5.5.1 Predictive performance

Figure 5.2 summarises performance statistics for each model architecture as applied to each of the target endpoints.

The AUROC metric (row 1) shows that the original data without any up-sampling applied rapidly fits to the majority class, struggling to capture much of the data signal at all, plateauing with an AUROC of close to 0.5 (where 0.5 is the AUROC for random classification, seen as a diagonal line). Reviewing the precision-recall curve (row 2) in combination with the workup to detection ratio (row 3) shows that for such imbalanced targets, all of the up-sampling techniques improve the performance somewhat, with the augmentation strategies generally outperforming the basic oversampling strategies across all metrics. The alerts per 100 patients versus sensitivity (bottom row) shows that in order to achieve 50% sensitivity, models trained using the original data distribution have to generate alerts for between 30 and 40% of patients, where the augmented data can achieve the same sensitivity while generating alerts for 10% of patients or fewer.

For the prediction of death in ICU, time to event augmentation (AUROC=0.83) and basic data augmentation (AUROC=0.82) outperform time to event oversampling (AUROC=0.80) and basic oversampling (AUROC=0.73). Likewise for prediction of in-patient death, time to event and basic augmentation (AUROC=0.82, 0.81 respectively) outperform time to event and basic oversampling (AUROC=0.79, 0.80 respectively). For the less severely imbalanced prediction task of long ICU admissions this also holds, with time to event (AUROC=0.80) and basic (AUROC=0.81) augmentation showing significant improvement over time to event (AUROC=0.74) and basic (AUROC=0.77) oversampling.

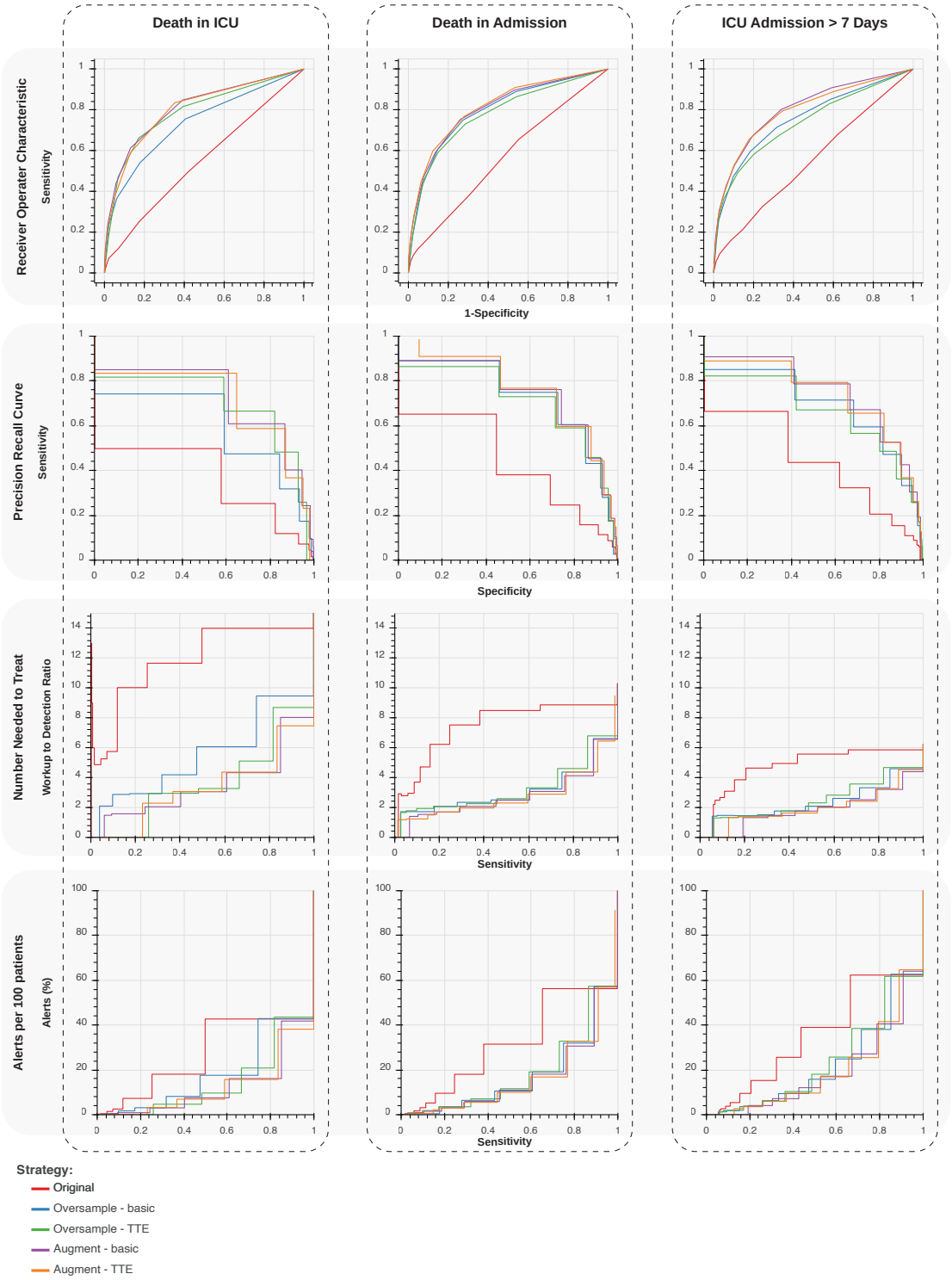


FIGURE 5.2. Comparing model statistics across endpoints and sampling strategies

5.5.2 Model calibration

In Figure 5.3, raw model output from the time-weighted augmentation strategy is compared with predictions that have been calibrated according to the expected target distribution and a more traditional isotonic recalibration technique [103]. In all cases, the distribution-based strategy is much closer to the line showing correctly calibrated risk, however the very low number of positive cases in the calibration set limits its utility for predicting death in ICU across the whole range of probabilities. It does, however, retain its qualities of improved calibration, despite being unable to reach higher levels of confidence.

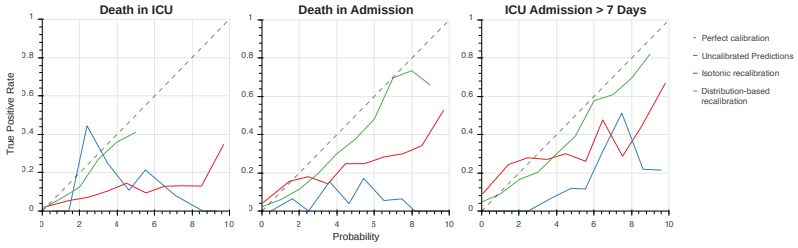


FIGURE 5.3. Effect of different model calibration strategies

Figure 5.5 compares the calibration curves for each of the piloted architectures. Ideal calibration is shown as a diagonal line. All of the original data distribution training strategies fall significantly below this ideal line, as they fit to the majority class and predict very few patients to be at high risk. The discrimination is poor, as there are a similar number of positive samples within those predicted to be at low risk as those predicted at high risk. Particularly of note are the almost horizontal portions of the graphs below 50% risk for both death in admission and long ICU stay.

The combination of time to event sampling and the data augmentation strategy has the most consistently acceptable calibration curves across all endpoints and architectures, meaning that there is less dependence on the model architecture itself, and the signal within the data is captured in a robust fashion. The LSTM architecture with width of 10 units had the most stable calibration across sampling strategies and end-points, so for the rest of the results section where architectures are not being compared, these are the results reported.

The relative stability of the calibration of models trained on augmented data versus over-sampled data provides evidence that the augmentation strategy described in this work does indeed achieve the stated goal of introducing temporal invariance through the modulation of bucketed event windows.

5.5.3 Time to event weighting strategies

When reviewing models produced under time-to-event weighting, this appears to have a different effect under the over-sampled and augmentation strategies. Applied to augmented data the improved stability of model calibration is quite clear, although the performance across other statistics is similar. This suggests that increasing attention to the most high-risk samples does indeed improve discrimination of patients at most imminent risk of deterioration from those at moderately elevated risk, and is likely to be a better decision with respect to clinical outcomes, rather than attending only to improvements in AUROC.

For over-sampled data under time-to-event weighting, although there is some improvement in discrimination for the high-risk categories, this improvement is less consistent and comes at the expense of a jump in the workup to detection ratio due to an increase in false-positive predictions.

This difference may be due to the fact that an augmentation strategy also acts as a sort of model ensembling, as all samples are augmented and therefore repeated multiple times, including those of the negative class. In the test set this means that all samples are repeated the same number of times with the prediction averaged, which can improve model performance in and of itself [136]. In addition, in the training set, those at extreme risk are augmented more frequently than those at elevated risk, but patients with elevated risk will still have significantly more samples than members of the negative class. If we aim to keep the overall distribution steady between oversampling under basic and time-to-event weighting in order to avoid overtraining to the minority class, an increase in the oversampling rate at the extremities will have the effect of decreasing the rate for positive class samples that are at less imminent risk, until they are only very slightly more prevalent than the negative class, and thus their signal is harder to capture.

Considering the events over the course of one week, where day 1 is the 24-hour period starting at prediction time, Figure 5.4 shows that the time to event sampling strategy does indeed behave as per this expectation, where the proportion of events occurring in the first day that are correctly predicted to be at high risk is much higher than for basic weighting, and that this is most clearly emphasised for the rarest endpoints.

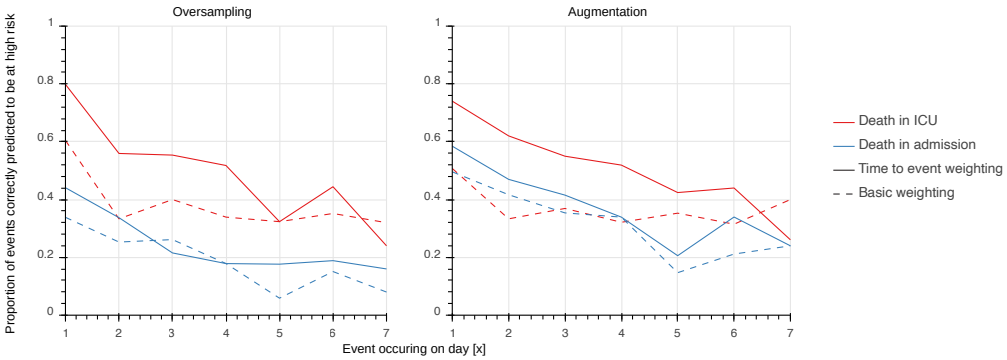


FIGURE 5.4. Proportion of cases correctly predicted as high risk for events occurring on days 1 to 7

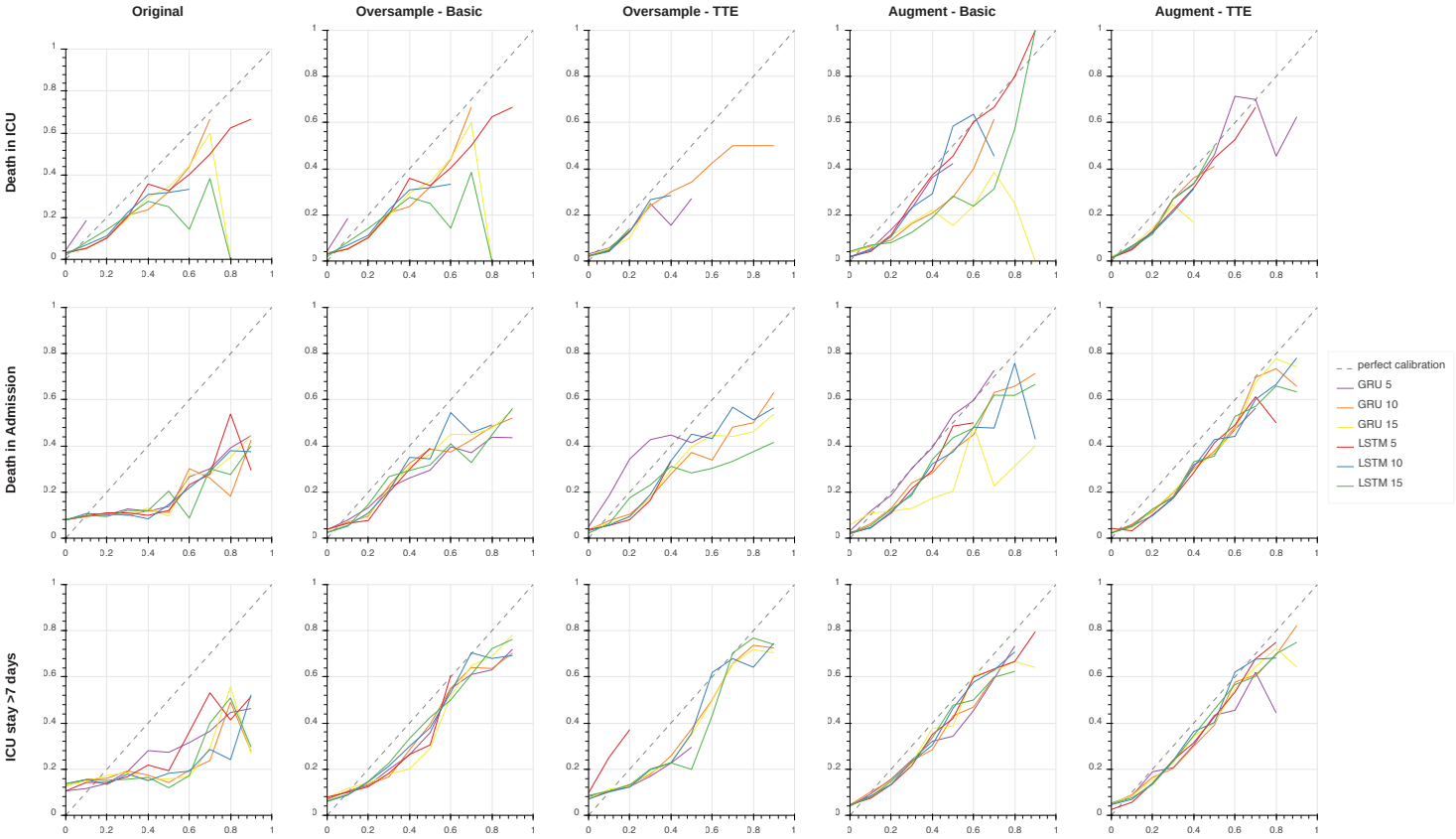


FIGURE 5.5. Calibration metrics across endpoints and architectures

5.5.4 Risk-stratification

In Table 5.2 we follow the reporting of model performance in [137] by stratification into high, medium and low risk categories, as this gives a concrete way for clinical end-users to anticipate the types of actions they would be likely to take relative to the number of alerts produced and the number of cases captured.

We define a score of 5 or above as high risk, which is specified according to the calibration set such that it should capture approximately the same proportion of the population as the expected event prevalence, and a score below 2 as low risk.

For death in ICU, the time to event augmentation strategy produced a well-calibrated result, but the discrimination at the high-end of the risk prediction was poor. This meant that for high scores, the risk buckets were excessively narrow, and thus failed to capture the targeted 6% of the population. Despite also only classifying 3.8% of the population as high risk, the basic augmentation strategy captured 135 events in this window, representing approximately one quarter of all true events. For every two events correctly classified in this risk stratum, approximately three additional cases were reviewed, for a workup to detection ratio of 2.39.

The clinical practicality of presenting results by risk category can be seen when comparing the time to event weighted oversampling and augmentation strategies for the death in ICU target. Although the sensitivity is higher for the oversampled data, capturing 141 events in the high risk stratum instead of 134, the workup to detection ratio drops more significantly. For these additional 7 cases correctly highlighted, there are 77 more alarms generated. It is unlikely to be considered prudent to generate that many additional alarms to capture a small number of events. It becomes easier to conceptualise such metrics with all of the practical clinical implications on both workload and patient outcomes when presented in this form.

Strategy	Original			Oversample - basic			Oversample - TTE			Augment - Basic			Augment - TTE		
Risk Stratum	**Low**	**Medium**	**High**	**Low**	**Medium**	**High**	**Low**	**Medium**	**High**	**Low**	**Medium**	**High**	**Low**	**Medium**	**High**
Death in ICU															
Number predicted	4881	2936	196	5028	2775	210	4673	2936	404	4557	3150	306	4669	3024	320
(%)	60.9	36.6	2.4	62.7	34.6	2.6	58.3	36.6	5	56.9	39.3	3.8	58.3	37.7	4
Number events in group	264	196	32	146	272	74	97	254	141	74	283	135	61	297	134
WDR	-	14.98	6.12	-	10.2	2.84	-	11.56	2.87	-	11.13	2.27	-	10.18	2.39
Sens.	-	0.4	0.07	-	0.55	0.15	-	0.52	0.29	-	0.58	0.27	-	0.6	0.27
NPV	0.95	-	-	0.97	-	-	0.98	-	-	0.98	-	-	0.99	-	-
Death in admission															
Number predicted	3731	3620	662	3598	3510	905	3816	3336	861	3634	3479	900	3527	3530	956
(%)	46.6	45.2	8.3	44.9	43.8	11.3	47.6	41.6	10.7	45.4	43.4	11.2	44	44.1	11.9
Number events in group	293	380	108	92	328	361	123	304	354	93	319	369	79	303	399
WDR	-	9.53	6.13	-	10.7	2.51	-	10.97	2.43	-	10.91	2.44	-	11.65	2.4
Sens.	-	0.49	0.14	-	0.42	0.46	-	0.39	0.45	-	0.41	0.47	-	0.39	0.51
NPV	0.92	-	-	0.97	-	-	0.97	-	-	0.97	-	-	0.98	-	-
ICU stay >7 days															
Number predicted	3202	3669	1142	3019	3717	1277	3063	3707	1243	3096	3613	1304	3063	3722	1228
(%)	40	45.8	14.3	37.7	46.4	15.9	38.2	46.3	15.5	38.6	45.1	16.3	38.2	46.4	15.3
Number events in group	438	560	251	181	471	597	215	470	564	128	475	646	142	475	632
WDR	-	6.55	4.55	-	7.89	2.14	-	7.89	2.2	-	7.61	2.02	-	7.84	1.94
Sens.	-	0.45	0.2	-	0.38	0.48	-	0.38	0.45	-	0.38	0.52	-	0.38	0.51
NPV	0.86	-	-	0.94	-	-	0.93	-	-	0.96	-	-	0.95	-	-

TABLE 5.2. Stratified prediction performance of data sampling strategies across endpoints. Note that for high and medium risk strata, predictions in this category are treated as positive. For these groups, the in-group event count, workup to detection ratio and sensitivity are calculated for that risk category alone (i.e. these values are not provided as 'greater than or equal to' for medium risk classification), and the negative predictive value cannot be calculated. Likewise low risk classification is treated as a negative prediction and therefore sensitivity and workup to detection ratio are not available.

5.6 Discussion

In this chapter, we do not implement a fully tuned architecture that is targeted to each specific endpoint of interest, as was demonstrated in Chapter 4, instead building a very simple, shallow network that can make explicit the effect of manipulating the data sampling strategy alone. In particular, this technique is specific to recurrent data, and thus we do not include the key component of the densely connected sub-model that ingests patient demographic factors. This fact notwithstanding, we still manage to produce a model that can predict half of inpatient deaths as high risk with a workup to detection ratio of 2.4 (for every 5 patients highlighted by the model, on average 2 will in fact die before discharge). Importantly, model calibration is greatly improved through the application of this sampling strategy, in a manner that is robust across different model architectures.

Traditional oversampling methods allow one to boost the signal of the minority class only, with a straightforward copy of each minority class sample. Using an augmentation strategy instead allows for more flexibility, where both the minority and majority class data may be strengthened by resampling each individual patient trajectory in a knowledge-driven fashion in order to create a much richer dataset for both classes.

This strategy is common in imaging and continuous time-series datasets, but the results presented here show that by making certain assumptions about the data collection methodology, it is possible to implement an equivalent strategy in discrete time-series data. This strategy has been designed around assumptions that are relevant to data entry in the electronic health record and proven against that data, however there are many equivalent input token-based datasets that may benefit from such treatment, for example consumer behaviour on websites that can be used to drive recommender systems.

Although generative models have been proposed for the purpose of creating augmented datasets for training models based in EMR data, they typically focus on generating aggregate data [138]. SMOTE is another alternative for adding synthetic data samples of the minority class [139], however this takes as its input tabular data, which limits its applicability to time-series data. Other methods of generating synthetic EMR data are knowledge-based and therefore restricted to specific disease domains [140, 141]. This is the only method to the authors' knowledge that is driven by known factors of the EMR data entry paradigm as opposed to the data itself, and therefore generalisable across all patient classes and robust to unseen combinations of patient characteristics.

A recent review of data augmentation for time-series data describes many strategies for augmenting continuous signals [142]. Although the masking, windowing and perturbation augmentation modalities are generally analogous to the strategies described here, their application to discrete tokens and specifically the EMR domain is novel. No comparable studies were found that addressed discrete time-series augmentation in other domains either.

In addition, this method is computationally and logically inexpensive in comparison to other generative methods. This factor not only reduces the cost of creating the input

data (both time and financial), but also increases the applicability of model introspection techniques such as LIME [110] or SHAP [111]. These algorithms for model explainability output the factors of highest importance with respect to a specific prediction, which may be obfuscated by the use of truly synthetic data.

By weighting model input according to the time-to-event parameter, we can ensure that risk immanency is captured and thereby robustly improve model calibration.

The strategy of basing an augmentation strategy around knowledge of the data-entry paradigm, rather than implementing a more heavily-engineered disease-specific method is also strongly related to one of the recurring themes seen in Chapter 4 of model scalability and productionalisation capacity. In a similar fashion the decision to retain all available data in the clinical record across all patients, rather than hand-curating a mapping to some required subset or population, is one that demonstrates a more realistic application of these technologies that are commonplace in domains other than the EMR.

5.7 Conclusion

The pattern of improvement seen from applying the data augmentation strategy described in this work is conclusive - improving prediction results across the board for three distinct end-points, each with a different level of data imbalance. Time to event sampling improves model calibration for all endpoints, although its effect on other metrics is less consistent.

CHAPTER 6

Watch-list User Interface

6.1 Objective

In Chapters 3–5 the case has been made that it is indeed technically feasible to implement a system that meets the stated requirements of the critical care outreach workflow, i.e. one that predicts short-term patient deterioration to an acceptable level of accuracy, using only data available in real-time at the target institution.

In order to address the next research question that is posed in this work, and measure the potential success of this system as perceived by likely stakeholders, it is necessary to envisage and describe a prototype user interface for this system.

Within the scope of this thesis, it is infeasible to include a complete and rigorous design phase for this user interface. There are, however, a number of key stakeholders who have shaped the direction of this work. Across a series of 1-1 interviews, a set of pilot use-cases were captured and used to define a realistic user interface that could be presented for an evaluation phase, which will be described in Chapter 7.

6.2 Use-Cases

The following four use-cases describe the most highly desired functionalities of a pilot implementation of the previously described models. There was significant alignment of priorities between the interview subjects, however it should be noted that the interviews

took place in an informal setting, at different stages throughout the project. Three out of four interviewees reviewed and endorsed the interface that was proposed as a result of these sessions, but one changed roles and was no longer available to the project team before this was possible.

6.2.1 UC-1: Prioritise Current Risk Across Hospital

This use-case was the primary driver and the genesis of this project, where users want to have a quick and accessible overview of current deterioration risk status for all patients, for the purposes of prioritising outreach interventions. This overview must be easily digested, and it was not felt to be important to provide detailed explanations for each risk assessment provided if it came at the expense of additional complexity.

6.2.2 UC-2: Individual Case Risk Trend

Although at a high-level, model introspection was secondary to usability, it was a non-negotiable requirement that the detailed model inputs would be available if a user drilled down to see further information for a particular case.

The models described do not have any inherent explainability or attention mechanisms. Instead we build upon the assumption that the system is not implemented as a single point-in-time assessment of risk, rather that for a given patient, their risk is evaluated at numerous times throughout their entire admission. If this time-line view of patient risk is taken, it is possible to present these results in such a way that one can see which clinical events align with an increase or decrease in predicted risk.

Figure 6.1 shows the mean risk prediction (line) and standard deviation (shaded) for the first 100 hours of admission for all admissions in the test set that are at least this long. They are grouped here by actual outcome after the 100^{th} hour. It is not possible to give an actual individual example of how this change in risk corresponds to specific events due to privacy and confidentiality restrictions on the source data, but this population-level

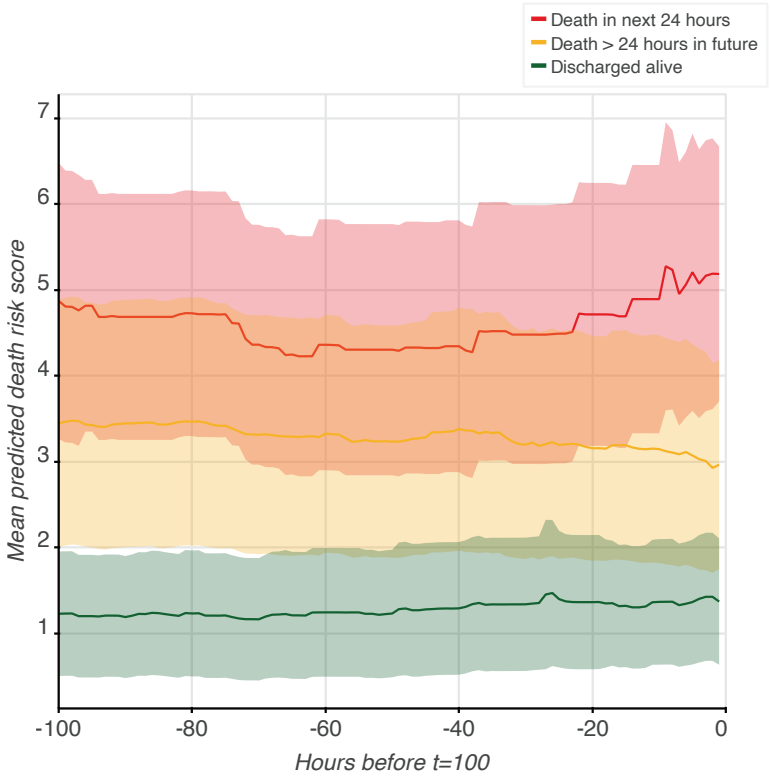


FIGURE 6.1. Risk changes up to prediction time

summary shows how the risk profile changes as more information is available. By aligning the trend line for a patient with the clinical data, a user may investigate the source of a change of interest to whatever level is useful to them.

6.2.3 UC-3: Manually Override

It was imperative to all interviewees that the workflow encompassing palliative and end-of-life (EOL) care be carefully considered in the design of this interface. This requires that an EOL patient that was surfaced as high deterioration risk by the system could be proactively marked as ‘risk reviewed and accepted’ or ‘comfort care only’, so that the high-risk summary would remain useful. A secondary effect of this would be to streamline the current EOL processes, allowing staff to easily see which patients do not yet have in place the necessary advance care directives.

6.2.4 UC-4: Sort, Filter and Summarise

There were a number of user groups proposed who do not fit the scope of the critical care outreach user, but who would also benefit from understanding the risk profile of inpatients in a summary fashion.

Medical officers performing shift handover: Filtering by assigned clinician and/or ward allows the generation of a priority list for handover discussions. Outside of formal handover processes, a clinician starting a shift may also filter for their own assigned patients to get a summary of acuity changes overnight.

Hospital admissions staff: A sortable ward-level summary of current patient deterioration risk is a better proxy measure for hospital case-mix and workload than any currently available in real-time. This has the potential benefit of anticipating overload and bed-block before it occurs, allowing staff to make informed decisions regarding scheduling, staffing and discretionary admissions/transfers/discharges.

Patient safety and quality departments: If one has a good idea of patient risk trajectories over time, it is possible to identify those deaths that were most and least expected. Of course there will always be deaths that are not anticipated by the system but which are entirely explainable by a clinician, as a limitation to the finite training set. This system must therefore not entirely replace any manual review of unacceptable mortality and morbidity, however it could rationally be used as one of numerous inputs into the patient safety governance procedures.

Risk summaries tracked over time also have potential uses as metrics to measure the impact of institution-level interventions, particularly in cases where other endpoints are rare, so it can be difficult to reach statistical significance. An overall decrease in acuity of target groups may give a more fine-grained view of improved outcomes.

FIGURE 6.2. The watch-list as envisaged

6.3 Interface Description

Figure 6.2 shows the resultant watch-list application as envisaged, based on the described use-cases.

The solution that is proposed takes visual and operational cues from common clinical dashboard applications in order to reduce as much as possible the burden of the application learning curve. A simple three-column design is used, which can support adaptive web-based implementations, expected to be suitable for mass-market tablet and mobile devices, reducing the risk of vendor dependence.

The three-column design is hierarchical from left to right, where the core use-cases (*UC-1, UC-2*) take the most prominent position, to support the goal of rapid-synthesis and overview. In a responsive design, this also means that in the face of reduced screen real-estate (such as on a mobile device), the core use-cases will still be prominent.

(A) Summary panel. Use the summary panel to see a high-level view of cases according to a simple traffic-light schema. This allows an immediate overview of the risk profile of the hospital and working priority list for outreach, meeting the goals of *UC-1*.

(B) Detail view. In the detail view, a user can see predictions for individual cases over time to understand trends, and highlight events co-occurring with risk changes over time by interacting with both the graphed patient trajectory and the time-stamped list of input events (*UC-2*). The precise user interface controls would require extensive piloting to navigate this complex relationship, but it is expected to be advantageous to link click events on the graphed risk with a change in focus or highlight in the event list and vice-versa.

The detail view also holds the controls required to enact *UC-3*, taking the form of a context menu to set patient-level characteristics. This manual override would then be reflected in the summary panel, where a user may switch between deterioration risk (default) and EOL care priority queues.

(C) Control panel. Controls available to sort and filter by risk category, location, assigned clinician or other relevant parameters, as required for all users described in *UC-4*.

Both figures 6.1 and 6.2, together with a summary of Chapter 4 results and a simplified description of the above use-cases were used as the input to the following chapter. The full text that was presented to survey participants can be found in the Appendix to Chapter 7: Proposed application and use-cases.

CHAPTER 7

Clinician Readiness to Adopt A.I. for Critical Care Prioritisation

---

7.1 Preamble

This chapter has been submitted for publication as [19] and is reproduced exactly as submitted, with the exception of this preamble.

In this chapter, the proposed watch-list application was presented to a series of likely stakeholders in order to determine their readiness to adopt it within their workflow. In the context of this thesis, this work is intended to take the core technical work described in Chapter 4 and provide the necessary data required to relate it back to the clinical requirements expressed in Chapter 2.

7.2 Abstract

There is a wide chasm between what has been shown to be feasible in the application of artificial intelligence to data from the electronic medical record, and what is currently available. The reasons for this are complex and understudied, and vary across technical, ethical and sociocultural domains. This work addresses the gap in the literature for studies that determine the readiness of clinical end-users to adopt such tools and the way in which they are perceived to affect clinical practice itself.

In this study, we present a novel, credible AI system for predicting in-patient deterioration to likely end users. We gauge their readiness to adopt this technology using a modified version of the technology adoption model.

Users are found to be moderately positive towards the potential introduction of this technology in their workflow, although they demonstrate particular concern for the appropriateness of the clinical setting into which it is deployed.

7.3 Background

Within the recent proliferation of reviews and summaries of the state of artificial intelligence (AI) in clinical settings [116, 143–147], there is evidence of successful translation of AI research into actual clinical practice for the analysis of images [148–152][1]. However, when it comes to the domain of AI decision support based on data from the electronic medical record (EMR), despite much interest and a number of viable models [81, 153, 154], there are very few signs of mature real-world implementations of such systems.

Asides from the technical and procedural challenges of data harmonisation and integration, model generalisability and ethical safeguards [144, 155, 156], and the high bar of achieving approval of software as a medical device [157], there is of course also a human and cultural barrier to entry that must be crossed in order to successfully implement these tools in practice. Despite this, there is no prior work to our knowledge that presents a credible EMR-based AI decision support system to likely clinician users to assess their opinion of its suitability within their daily work.

The work presented in Chapter 4 describes an AI system for predicting risk of in-patient deterioration, targeted to the specific use-case of prioritising the work of the critical care outreach team. This system uses only real-time available data and is designed with the goal of being integratable within the EMR.

[1]Note that these examples are limited to those demonstrating successful implementation and real-world use, as opposed to studies describing model development, which are plentiful.

7.3.1 Objective

The aim of this study is to assess the readiness of intensive care clinicians and leaders to adopt an AI-based decision support system for the prioritisation of patients at risk of deterioration on the wards.

The critical care outreach workflow is a reasonable one in which to pilot this emergent technology because it is a challenging and fast-paced role that requires rapidly updating awareness of events across the whole hospital. It is therefore a natural fit for any tool that can reliably synthesise a large amount of real-time data to augment clinical judgement. In addition, it already typically relies on track-and-trigger early warning systems (EWS) [80], and therefore the progression to what is effectively a risk model based on a broader selection of data is not as great a leap as it may be in other contexts.

We propose that this is a necessary next step towards completion of an appropriate impact study, as the complexity of such an implementation (even in pilot phase) cannot be understated. A full and theoretically grounded account of stakeholder readiness and understanding of potential pain points will be a powerful tool in navigating an intricately balanced set of clinical, cultural and technological priorities required for its successful execution.

7.3.2 Prior Work

The original technology acceptance model (TAM) [158] and its derivatives [15, 159] have been used to understand barriers to technology adoption in the context of many disruptive technologies such as wireless internet [160], e-commerce [161], and personal computing [162], and importantly have been demonstrated for healthcare applications such as telemedicine [163], electronic medical records [164] and mobile healthcare systems [165].

The TAM framework explains the behavioural intention of an individual to use a new technology as a factor of their attitude towards its use. Their attitude towards a technology

is affected directly by the measures of its *perceived usefulness* (PU) and *perceived ease of use* (PEOU). The premise of this model is that an individual's attitude to using a system is a good predictor of their behavioural intention to use it (BI), and in turn, their actual eventual use.

The review papers [166, 167] found that whilst the TAM has general capacity to predict technology acceptance in a clinical or healthcare setting, it is necessary to include additional context-specific explanatory variables.

We reviewed the literature for expanded versions of the TAM that had been validated in a healthcare setting [168, 169] in order to adopt a minimal set of relevant additional variables for this use-case. The proposed additional variables were kept to a minimal set of demographic questions in order to reduce the burden of response on the target subjects, in particular given the timing of the study during the outbreak of COVID-19 and the extraordinary pressure on intensive care teams during this time. It was deemed unlikely that within the highly targeted (and therefore small) population of probable users, sufficient responses could be gathered to validate any new constructs, so added variables were restricted to modulating factors only.

7.3.3 Research Model and Hypotheses

The TAM is comprised of the hypotheses summarised as H1–5 in Table 7.1. We further expanded the model with three additional context-specific hypotheses. We do not include the TAM2 second and third order factor antecedents of perceived usefulness [15] due to the challenge of meaningfully capturing these factors for a proposed (not implemented) system.

It is hypothesised that the perception of usefulness will be the most significant factor in the successful adoption of this technology (*H3b*), as the capacity for a model to improve the workflow itself is a critical factor for clinician buy-in. This is particularly true in this early stage, where it is necessary to confirm whether the model is even addressing a real need [16]. This hypothesis is supported by the observation made by [167] that in 100% of

reviewed healthcare TAM studies, the PU→BI relationship was significant, compared to just over half of reviewed studies finding a statistically significant PEOU→BI relationship.

This study expands the existing TAM hypotheses with three additional modulating factors:

- The perception of usefulness will be modulated by how well a potential user can envisage such a system supporting their own workflow, and therefore will be dependent upon their role (*H6*).
- A user who spends most of their working week assigned to relevant patient deterioration tasks will see greater value in automated data synthesis to support the workflow, increasing their perception of its usefulness (*H7*). Once the system is in regular use, it is likely that higher time devoted to the task would also increase a user's perception of ease of use, but in this hypothetical phase (prior to any learning curve), it is not anticipated that such an effect would be observed.
- Senior patient deterioration staff may be more separated from the day to day challenges imposed by interaction with clinical information management systems, and therefore it will affect their view of how practical an AI system would be in practice (*H8*).

	Hypothesis	**As Equation**
H1	Behavioural intent predicts actual use	BI→AU
H2	Attitude to use explains behavioural intent	ATT→BI
H3	Perceived usefulness explains attitude	PU→ATT
H4	Perceived ease of use explains attitude	PEOU→ATT
H5	Perceived ease of use explains perceived usefulness	PEOU→PU
H3b	Perceived usefulness will be the most important model factor in driving attitude	PU→ATT > PEOU→ATT
H6	Clinical role modulates perceived usefulness	ROLE∼PU
H7	Time devoted to task modulates perceived usefulness	TIME∼PU
H8	Experience modulates perceived usefulness	EXP∼PU

TABLE 7.1. Hypotheses proposed by the TAM (H1-5), extended with H3b and H6-8 which are specific to this context

7.4 Methods

7.4.1 Data Collection

The questionnaire consisted of two parts. The first section captured variables that described the subject's role, experience and demography. After responding to the demography questions, subjects were asked to review a prototype user interface, alongside descriptions of the tasks being performed (reproduced in Appendix to Chapter 7: Proposed application and use-cases). In addition, some high-level summary information was provided on the development of the predictive model itself (data sources, mode of operation and accuracy — as described in Chapter 4).

The main portion of the questionnaire was defined by adapting the measures in the original validated model [15] to the target context (refer to Appendix to Chapter 7: Measures for a full list). A single free-text response was also captured that allowed respondents to make any additional comments or suggestions that they felt were pertinent to the potential roll-out of such a tool. Finally, in order to identify disingenuous responses, a question to test malingering with a trivially correct response was included.

The survey was distributed by email and completed online using the Qualtrics survey platform. All TAM measures were captured as a 7 item Likert scale from 1 = Strongly Disagree to 7 = Strongly Agree.

7.4.2 Population

This questionnaire was distributed to members of the New South Wales Deteriorating Patient Advisory Group (NSW-DPAG) (n$\sim$170), which is made up of highly engaged critical care staff, including nurses, physicians and administrators across the public healthcare system. This represents the key decision-making group whose advocacy and leadership would be a necessary prerequisite for a successful pilot implementation of this system within their respective institutions.

7.4.3 Data Analysis

A structural equation modeling approach was used for this analysis, based on the lavaan library [170] in R.

7.5 Results

7.5.1 Descriptive Statistics

The questionnaire was sent out to the NSW-DPAG member email list in July of 2020 and was open for a period of one month. 59 responses were captured, giving a 34.7% response rate. 14 response sets were removed for not proceeding beyond review of the proposed application. None of the remaining sets provided an incorrect or missing response on the malingering item, and therefore 45 response sets were retained for analysis. 91% of the remaining respondents (41) completed all measures, giving an overall completion rate of 96% for the analysis data.

Overall the respondents were highly experienced, with the majority coming from a nursing background (60%). This is expected within the included group, and accurately represents the key decision makers who would be responsible for overseeing the implementation of such a system. Managerial levels are over-represented in this sample, however 67% had at least some regular assignment to relevant clinical tasks and therefore fit the profile of an expected end-user as well.

7.5.2 Model Validation

7.5.2.1 Construct Reliability

Re-validation of the model constructs was necessary, due to their modification to fit the research context.

TABLE 7.2. Respondent Demography

Measure	Response	n	%	Group
Gender	Female	33	73.3	
	Male	12	26.7	
Age	<30 years of age	7	15.6	
	30-39	11	24.4	
	40-49	6	13.3	
	50-59	18	40.0	
	60+	3	6.7	
Speciality	Nursing	28	62.2	Nursing
	Medical	9	20.0	Non-nursing
	Allied Health	4	8.9	
	Administrative	4	8.9	
Experience post-graduation	0-5 years	2	4.4	Less experienced
	6-10 years	9	20.0	
	11-15	6	13.3	Very experienced
	16+	28	62.2	
Level of assignment to clinical patient deterioration related tasks in a typical work-week	None	7	15.6	Low time
	Occasional or cover assignment only	8	17.8	
	Part-time but regular assignment	13	28.9	
	Majority of working week	9	20.0	High time
	Full-time or dedicated assignment	8	17.8	

After applying a mean imputation strategy, we cannot reject the null hypothesis of data non-normality (tested with [171] due to small sample size), and therefore are unable to use maximum likelihood estimation to fit our model. A robust unweighted least squares strategy [172] was thus used instead.

Despite a moderate response rate, the final number of responses was insufficient to assess the reliability of the original 4 factors comprising the TAM (model non-convergence). We therefore follow the lead of the updated TAM2 model [15] in merging the BI and ATT latent factors. In doing this, some of the nuance between a potential user believing the system overall is a good idea and it translating into their actual intention to use it is lost, but a simpler and more robust model is produced, which can withstand analysis under

Construct	Items	Factor Loading	AVE	Composite Reliability
PU	PU1	0.970	0.885	0.938
	PU2	0.910		
PEOU	PEOU1	0.915	0.607	0.820
	PEOU2	0.678		
	PEOU3	0.723		
BI/ATT	BI1	0.847	0.694	0.901
	BI2	0.862		
	ATT1	0.768		
	ATT2	0.853		

TABLE 7.3. Confirmatory Factor Analysis

a small data set. The high composite reliability for this merged factor (higher than the average CR for each individual construct in [173]) also supports this action.

The model was further simplified by removing one PU item (PU3) and one PEOU item (PEOU4) due to their high degree of collinearity with other items in their respective constructs.

In all instances, the resultant construct composite reliability (CR) meets the 0.8 standard for generally acceptable reliability [174, 175] (see Table 7.3). The average variance extracted (AVE) for each construct also exceeds the threshold of 0.5, required as per [176] to ensure that variance due to measurement error does not exceed the variance of the construct itself.

7.5.2.2 Model Fit and Assessment of Hypotheses

The Satorra-Bentler scaled chi-squared test statistic allows the generation of goodness of fit indices that do not make any assumptions about the normality of the underlying data [177]. Applying this method, we do not reject the null hypothesis of good model fit (p=0.281). Other model fit indices were also indicative of good fit: CFI = 0.989, TLI = 0.983, SRMR = 0.058.

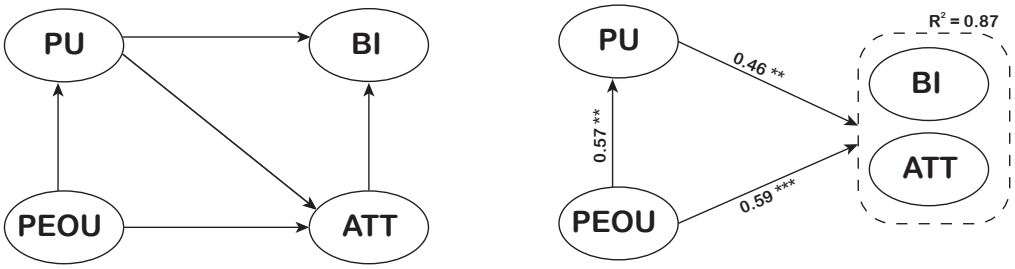


FIGURE 7.1. Original TAM (left) with adjustments for this context (right) (standardised factor weightings)

Most importantly, those relationships that have been retained in the simplified model are theoretically grounded. We therefore accept PU and PEOU as antecedents of a user's intention to use this system, and accept that the measures in this survey tool are sufficient to quantify PU, PEOU and BI constructs for these exploratory purposes.

Given the hypothetical nature of this target system, it is naturally impossible to test *H1* in advance of any meaningful pilot, although it is retained here for completeness, and as per section 7.5.2.1, we had insufficient statistical power to test *H2*.

Consistent with the existing literature, both perceived usefulness and perceived ease of use were significant determinants of behavioural intent to use (Table 7.4), and therefore *H3* and *H4* were supported. Perceived ease of use was also a significant antecedent of perceived usefulness, supporting *H5*.

The standardized estimates of the relationship between PU and BI/ATT and PEOU and BI/ATT are similar, and in fact PEOU is found to be somewhat higher in weighting. This is the opposite effect as proposed by *H3b*, so it is not supported.

7.5.2.3 Weighted Results

Table 7.5 reports the mean and standard deviation for each factor (weighted and unweighted). The ratio of weighted mean to the theoretical weighted maximum is above 0.5 for each construct, showing that overall the potential users rated the proposed application as relatively useful and relatively easy to use, leading to an overall positive attitude and intention to use.

Hypothesis		Estimate	SE	P	Support
H1	BI→AU	Untestable in this context			Untested
H2	ATT→BI	Untestable in this context			Untested
H3	PU→BI/ATT	0.457	0.151	**	Supported
H4	PEOU→BI/ATT	0.590	0.165	***	Supported
H5	PEOU→PU	0.570	0.152	***	Supported

Estimate: Standardized
SE: Standard Error
*** p < 0.001, ** p < 0.01
Note: SE and p vals estimated with 500 bootstrap samples

TABLE 7.4. Support for Hypotheses (TAM)

Construct (items)	Unweighted Mean	Unweighted SD	Weighted Mean	Weighted SD	Range - actual (theoretical)	WM/TM
PU (2)	10.42	2.32	9.79	2.18	2.85-13.16 (1.88-13.16)	0.74
PEOU (3)	15.23	3.30	11.68	2.59	5.94-16.21 (2.32-16.21)	0.72
BI/ATT (4)	20.62	4.32	17.14	3.61	6.66-22.45 (3.33-23.31)	0.76

Weightings: Standardized
WM/TM - ratio of weighted mean to theoretical maximum for this construct

TABLE 7.5. Weighted factors

Both PU and PEOU saw ceiling effects, where at least one respondent answered maximally positively (strongly agree) across the whole measure. No respondent answered maximally negatively (strongly disagree) across any measure. This is particularly true for PEOU, which had a minimum unweighted score of 8 out of a possible 21. The ranges of responses overall, however, were quite wide.

7.5.3 Multi-group Analysis

7.5.3.1 Model Invariance

To assess hypotheses that predict mediation of relationships between groups (H6, H7, H8), measurement invariance must first be confirmed. We first establish a baseline model to

Group	Model	χ^2	df	CFI	RMSEA
Nurses vs. non-nurses	Configural	10.94	34	0.999	0.043
	Metric	26.53	39	0.996	0.065
	Scalar	30.59	44	0.993	0.077
	Strict	57.39	52	**0.976**	**0.141**
High vs. Low task time	Configural	20.78	34	0.992	0.106
	Metric	33.55	39	0.993	0.097
	Scalar	36.17	44	0.992	0.095
	Strict	59.20	52	**0.981**	**0.133**

TABLE 7.6. Inter-group model comparisons

confirm the factor-loading pattern between groups (configural model), where the factor structure is the same, but all other elements of the model are allowed to vary freely between groups. This is compared to models where invariance is enforced for (1) factor loadings (metric model), (2) both factor loadings and model intercepts (scalar model) and (3) factor loadings, model intercepts and item variances (strict model) [178].

When comparing the *Nursing vs. Non-Nursing* and *High time vs. Low time* inter-group relationships, note that in each case, a single negative variance was produced, which is potentially indicative of model misspecification. In this case, however, it is likely to be due to the low group sizes, as in each instance the estimate plus the respective standard error was positive [179]. Conversely, when comparing the *Highly Experienced* group to the *Moderately or Less Experienced group*, more samples are required in order to define a fully-specified model, so it was not possible to test H8. As an alternative expression of H8, we note that experience and age are naturally highly dependent — $\chi^2(12, N = 45) = 52.1, p < 0.005$ and therefore compare older ($50+$) to younger ($\leqslant 49$) respondents as well, however additional samples are also required to validly test this version. This inability to define a validly weighted model for both forms of H8 implies that the relationship is unlikely to be a straightforward one.

Table 7.6 shows that there is good model fit for the configural, metric and scalar models, and therefore latent mean analysis is valid, as the latent factors can be assumed to be invariant in configuration, factor loadings and scale.

For the strict model, there is a change in the CFI that exceeds the threshold of -0.010 combined with an increase in RMSEA by more than 0.010, as suggested in [180]. The strict model change in fit implies that the variance in responses may differ across groups, despite the overall model structure remaining valid. To estimate the differences in latent-factor means across groups, we constrain the factor mean in a reference group to zero, and then estimate the mean in the comparison group to produce the difference in factor mean [181]. This was done individually for each latent factor in the model and is reported in Table 7.7.

7.5.3.2 Effect of Role

H6 is supported, as there is a statistically significant difference between the perceived usefulness between nurses and non-nurse respondents. A more meaningful inter-group comparison for this hypothesis would be nursing staff compared to medical staff, as there would be more uniformity in the scope of the roles being analysed, however once again we are limited by the small data set.

7.5.3.3 Effect of Workload

There is no statistically significant difference in the PU latent factor when comparing the group of respondents who spend all or the majority of their working week assigned to relevant clinical patient deterioration tasks, versus those who spend only part of their time or ad-hoc assignment in this role, so *H7* is not supported.

7.5.4 Free-text responses

23 subjects provided input to the optional free-text comment at the end of the study. These responses were typically very short (mean 175 characters, s.d. 144), but despite this,

Hypothesis		Reference	Comparison	Factor	Difference	Support
H6	ROLE∼PU	Nurses	Not Nurses	PU	0.61*	Supported
				PEOU	0.52	n/a
				BI	0.59	n/a
H7	TIME∼PU	High task time	Low task time	PU	0.52	Not supported
				PEOU	0.14	n/a
				BI	-0.24	n/a
H8	EXP∼PU	Untestable in this context				Untested

TABLE 7.7. Hypothesis testing for modulating factors

consistent themes were strongly evident. Only three comments were too general in their nature to fit into at least one of the identified themes. We report here the results of this abstraction for the purposes of driving the direction of future exploratory analyses [2].

The most common response type (10) referred to a specific setting (physical or logical) — either in support of the utility of this tool in a given context, or where there was some setting-specific limitation in its use.

Supported Settings	Settings with Limitations
A state-wide or universal EMR	Rural hospitals (rotating staff/training) (2)
Settings with large numbers of patients	Paediatrics (unsuitable endpoints)
Emergency departments	End-of-life care (treat EOL patients differently)
For junior clinicians	Community health (lack of applicability)
	Multipurpose services (lack of applicability)

Four responses referred to the impact of this tool on quality of care, two of which repeated the same concern that such automation must not be allowed to impact or supersede face to face care. The other two were more positive with respect to the potential effect that the tool may have on patient outcomes by introducing timely and specific alerts.

Workload was also mentioned in four responses, specifically: that any manual steps will increase the load of already overburdened critical care teams (2 responses); the concern that availability must be straightforward and flexible for it to be useful; and that a formalisation

[2]Note that it is outside of the bounds of the ethical approval of this study to report any quotes directly.

such as this may generate data to support implementing dedicated response teams where they do not already exist.

Three subjects commented on the value of synthesising large volumes of data from numerous sources, and the additional benefit that this may provide in terms of a high-level overview of current acuity levels.

Finally, three subjects provided general caveats or considerations that they considered key to the successful implementation of this tool. These were: for this to be useful, it must be possible to know the reason behind the deterioration; the necessity of specialist clinician informatician involvement, particularly with respect to privacy and security; and the importance of a carefully designed roll-out phase.

7.6 Discussion

7.6.1 Study Measure

This study demonstrated the applicability of the TAM measures to describe the attitudes and intentions of clinicians to adopt the proposed AI system for the purpose of decision support in a patient deterioration context. Although it was not possible to fully validate the ATT and BI measures, the retained relationships between latent factors were consistent with prior literature, and explained a large proportion of the variance in the overall opinions of the target users.

It is possible that the need to merge the BI and ATT factors was due to the composition of the study population, which is over-representative of very senior clinicians. BI items (when interpreted in their literal sense) ask a subject to reflect on the likelihood of an action on their part that may fall outside of the scope of a managerial role (e.g. *I would be a frequent user*), and thus breaks the directness of translation of a positive overall attitude into a behavioural intention to use. In future studies, it would be illustrative to identify the

role of decision-maker as distinct from likely user, and adjust the BI items to account for this difference in remit.

7.6.2 Clinician Readiness

As seen in Table 7.5, there is an overall positive view of all the latent factors in the research model. This means that a subject is more likely than not to perceive that the system under review is both somewhat useful and somewhat easy to use, although this is not universally true. The readiness to adopt this technology would be best described as 'moderate', with nurses somewhat more likely than other clinicians to have a positive view of its potential utility.

The favourable view of nurses as to the usefulness of this system is a good indicator of support for a pilot implementation, as nurses are generally more burdened with administrative tasks introduced by hospital information management systems. It would be reasonable therefore to expect them to be a more skeptical user group for novel hospital IT programs. Their positive assessment should be taken as evidence that this system has potential to fit in well with this workflow, and that as a user group they are open to the idea of automated information synthesis and risk assessments that could augment their patient care. There is also evidence that nurses do not always find the patient care escalation process to be without friction when based on intuition alone [182], so it would be informative to explore in what capacity nursing staff perceive this system to be useful - whether for its information-synthesis capacity, risk assessment, or as an additional measure that can be used to make the case for patient prioritisation.

The effects of the PU and PEOU factors upon the combined BI/ATT endogenous factor are fairly equally weighted in this model, which goes against *H3b*. This may be due to the novelty of this system, where it is easier for a user to assess ease of use against their mental models of existing clinical software than it is for them to fully imagine how it will assist their practice. Based on experience, however, we would posit that in order to take this system from the research to the clinical realm PU will in fact be a far higher barrier to cross. This was evident in the focus of the free-text responses on specific settings where

the system would be most useful or where it may have limitations to its usefulness. The hypothetical nature of this system may require the introduction of an additional factor to capture the friction between a user's general positive attitude towards the use of a system and the behavioural intention to not only use a system but also to overcoming technical and procedural changes necessary for its implementation.

Relationships between the latent factors were consistent between study groups, both in scale and factor loadings, although the variance showed some differences. More data would be required in order to further investigate these differences in variance and to infer anything about the patterns of which individual measures showed a statistically significant inter-group variation. This may also be an outcome of the seniority bias evident in the study sample.

7.6.3 Limitations

The most obvious limitation of this study is in its small sample size. We chose to prioritise the relevance and expertise of the subjects, at the expense of the available population. The results here are therefore challenging to generalise, although they give a solid basis upon which to build.

In addition, the novelty of the target system makes it difficult for subjects to meaningfully evaluate in this limited context. This is seen in the relatively high percentage of respondents who filled in the demography measures completely, but did not proceed with the questionnaire after reviewing the prototype application. Until a controlled experiment demonstrating this system in practice (or better: a working prototype) is available, any judgements of PU in particular will be insufficient to draw broad conclusions.

7.6.4 Future Work

The free-text responses were illustrative of the general concerns and objectives of this group of potential decision-makers and users, however they were insufficiently formal to

draw any significant conclusions. A semi-structured interview format would be best to further explore these themes, in order to identify the specific barriers between the moderate readiness identified in this work and an actual pilot implementation.

7.7 Conclusion

Clinicians were found to be moderately favourable towards the AI decision support system that was presented as a potential prototype for the support of managing critical care outreach workloads. Nurses were somewhat more likely than other clinicians to perceive the system as useful in their practice.

CHAPTER 8

Discussion

8.1 Review of background and objectives

The overarching objective of this thesis has been to deliver a body of work that seeks a technical solution to the clinical problem of predicting deterioration in an acute-care setting. By setting out both technical and clinical research questions to be addressed in parallel, the aim was to develop a watch-list application that allows critical care outreach staff to accurately identify patients at high risk of deterioration in a fashion that has capacity for successful translation into clinical practice.

This is a problem that has been addressed previously in the literature, typically by the use of an early warning score that is based on the detection of vital sign observations outside of pre-defined thresholds such as NEWS [80], although the evidence of impact of such systems is inconsistent [183, 184]. The approach described herein differs in two important ways. Pragmatically, the desire for automation at the target institution is hampered by the unavailability of physiological observations in the clinical record, which has led them to seek an alternative solution — the catalyst for this work. Further, the focus of this effort is on the workflow of the critical care outreach staff. By centring a specific clinical function and user group, it is possible not only to explore the mechanisms and performance of the predictive models themselves, but also to consider the ways in which their usage will affect clinical users and anticipate roadblocks and challenges in their eventual implementation.

This chapter will assess the proposed system in light of its ability to meet these stated clinical goals, in particular as measured against the perceptions of its end-users. The entirety

of the research output will also be assessed as a cohesive and complete methodology for delivering such systems.

Due to the desire to present a translatable solution in particular, this application will also be evaluated under the NASSS framework [185], which studies the root causes of technological non-adoption, abandonment, and issues of failure to scale, spread, and reach sustainability in clinical settings.

Finally, conclusions of this work will be summarised such as it provides answers to the research questions posed in Chapter 1.

8.2 Summary of main findings

8.2.1 Technical contributions

The work presented in Chapter 4 details the development of a set of models that compare favourably with existing methods for predicting in-patient death and comparably for unplanned ICU admission. For the target of death within 24 hours, an area under the receiver operating curve (AUROC) of 0.93 was achieved, and an AUROC of 0.78 for unplanned ICU admission in the same time-frame. As a point of comparison, the calling criteria set by NEWS [80] reports AUROC metrics of 0.89 and 0.86 respectively for these same end-points.

This was achieved despite the limiting factor of the absence of vital sign data in the clinical record in this data set. This limitation is of note, as vital sign observations form the crux of current best-practice for detection of inpatient deterioration at the target institution and around the world [1, 80]. In particular, the performance of the models show that the lack of vital sign observations is more limiting for the prediction of unplanned ICU admissions than it is for predicting in-patient death.

A thorough exposition of model performance that goes beyond the reporting of AUROC alone (which is often misleading in imbalanced prediction problems) reveals a well-calibrated prediction of risk for both targets across multiple points in time. These models therefore offer a satisfactory candidate input around which to form the basis of the watch-list application.

These models were based on elements of existing work applying deep neural networks to generate predictions from electronic medical record (EMR) data, such as [81, 82, 94], which treat the patient trajectory as a series of discrete tokens representing clinical events over time, to which it is possible to apply language-modelling techniques. This was extended with novel pre-processing methods that improve the performance of these models in a data set that is relatively small, both in terms of volume (moderate in number of patients, but more importantly low instance of positive-class samples) and breadth (absence of vital signs, primarily administrative data).

Chapter 5 confirms the generalisability of the core innovative techniques that were used to implement these models. In particular, under the highly imbalanced end-points that define many of the targets of interest for such data, a combination of data augmentation and modulation techniques greatly improved the predictive capacity of even very simple model architectures. The use of time-to-event sampling strategies further improved the robustness of the model calibration. The potential impact of the technical contribution of this work is wide-ranging, due to the fact that these techniques are not specific to any one clinical domain, rather they are predicated on the nature of EMR data entry processes themselves, and thus have applications in many similar prediction problems.

These results confirm that as a proof of concept, it is possible to reliably produce an adequately calibrated risk of imminent patient deterioration that can be expected to improve the capacity of critical care outreach staff in prioritising their workload.

8.2.2 Evaluation of overall research approach

8.2.2.1 Contributions

The decision to combine the technical research target with a parallel focus on ensuring its eventual translatability naturally increases the scope of inputs into this work. In doing so, the requirements become broad relative to what is achievable in a single doctoral thesis, and thus there are elements that must be compromised in depth in order to complete a cohesive unit of research, such as the lack of more detailed user interviews despite the clearly useful content gathered in even a preliminary open-response question in Chapter 7. This trade-off does, however, affirm the overarching research target, as it is demonstrative of the additional resources and interdisciplinary methods that are necessary to properly evaluate not only clinical predictive rule design but also their use in practice and the way that they impact end-users. The evidence in Chapter 2 shows the paucity in the literature of comprehensive research programs that tackle this task in a meaningful way. Although there has been a push for external validation of prediction rules in clinical contexts insofar as they affect health outcomes, few reach this level of evaluation, and the consideration of usability of the developed system and utility of the clinical target itself is often undertaken as a separate post-hoc effort, rather than shaping the body of work in its entirety [186].

As the capacity of the technology advances, and as the availability of real-time clinical data continues to increase in breadth and volume, it is safe to expect that there will be a consequent proliferation of efforts to implement prediction models driven by medical record data that are already available as a by-product of clinical care. Real-world implementation of models based on the full patient record absolutely must be delivered as an integrated solution in order to have any realistic chance of acceptance, making their technical roll-out a high-cost endeavour. As a consequence, the value not only of understanding the capacity of a given model to technically achieve its stated purpose but also being able to anticipate the needs and preferences of end-users is clear, as the potential waste due to unsuccessful implementations and/or non-adoption is vast. This is true even before considering the clinical and cultural barriers to implementation, which are also significant. This work takes a foundational step in considering such models from the perspective of the end user as a core tenet, rather than an afterthought.

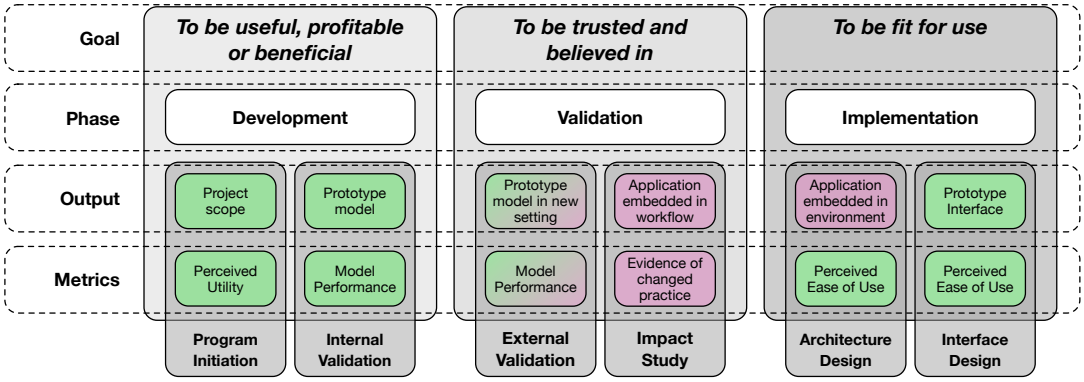


FIGURE 8.1. Mapping of research artifacts to identified themes: Green - present in research artifact; Pink - unavailable; Gradient - partially available

8.2.2.2 Limitations

Figure 8.1 expands on Figure 2.4 by aligning the output of this research project with the relevant themes and phases that emerged from the qualitative literature. This mapping exposes some of the limitations of this body of work in trying to address all the concerns of end-users.

Firstly, while the use of the MIMIC-III dataset to demonstrate the novelty of the domain-specific data augmentation strategy does show the general transferability of the techniques, this version of the model was heavily simplified. It therefore did not reach the same performance levels and thus cannot be interpreted as a true external validation study. This re-implementation of the model in publicly available data was primarily used to demonstrate the impact specific to the data augmentation strategy, which necessitated applicability across multiple end-points, and the additional cost to optimise the architecture to these multiple end-points and re-train to completeness made this a necessary compromise.

In addition, it is clear that this work cannot be considered complete with respect to clinician opinions and preferences, as there is no way to comment on its actual use in practice, and importantly it does not touch on the training, education and support materials that were seen to be highly impactful in the source literature [36, 39, 40, 52]. Where possible, implementation-relevant considerations have been included, such as integrated architecture and user interface, but with such a broad scope of work, even an implementation in an experimental setting was out of reach.

When compared with the development studies of other models however, such as [81, 153], it is evident that current best practice does not consider implementation details in a systemic fashion, nor does it require concern for the experience of the end-user. These cited examples come from groups with a mature practice of delivering consumer-facing applications and almost infinite capacity to resource a study of user experience, so the omission of this analysis is all the more glaring. As the trend for studies of technology adoption to become more pragmatic continues [187], the holistic approach demonstrated herein shows both the importance and feasibility of considering how a novel technology is situated within modern clinical praxis (with all its associated complexities) at the point of development, rather than as part of a post-hoc (pessimistically, post-mortem) analysis.

8.2.3 Addressing healthcare provider opinions and preferences

Since technical feasibility has been established per Chapters 4 and 5, the assessments that were refined from the literature analysis in Chapter 2 can be used to inform the delivery of predictive models that truly support the needs of clinical end-users. In this section, these goals (see the top row of Figure 8.1) are used to reflect on the healthcare provider opinions and preferences concerning clinical prediction rules, and to what end there is evidence of their consideration in the watch-list application itself, as supported by the responses to the technology adoption model questionnaire (TAM).

8.2.3.1 To be useful, profitable or beneficial

In some ways, this project was conferred an advantage from the point of its inception, due to the fact that it is addressing a concrete, specific, user-defined need, which is currently unmet by existing technologies.

This user-driven definition of the problem statement is in contrast to a research-driven or model-driven scope, where a theoretical capacity to predict an outcome has limited clinical utility. This was observed in models predicting post-operative nausea and vomiting, which were not found to be useful by clinicians, due to the relatively low burden on patients compared with the complexity of applying the model [51].

It is also in contrast to a 'top-down' approach, where there is a propensity for leadership to implement an innovation for theoretical benefit, which then struggles to be realised in practice, such as was the case in the roll-out of speech recognition technology for electronic medical records in New South Wales [188].

On the other hand, the scope of this project also presented a significant challenge with respect to the utility of the proposed model, in that the request was specifically for a generalised, numeric risk score, which is naturally limited in its capacity to have direct and actionable output. Deference is given here to the experience of the end-users who have specified that for this particular use, a risk score will provide sufficient benefit to be able to deprioritise this factor [51, 63, 66, 68].

It is possible that in this workflow, given the heterogeneity of the deterioration target (as distinct from a workflow that reviews the risk of a single patient at a time, such as [55]), that the action can be reframed as clinical review versus no clinical review, in which case actionability from a priority queue of numeric risk alone is restored. According to the users interviewed for the use-case development in Chapter 6, it is their belief that they can and would intend to take action based on such a numerical output in this workflow, as per the escalation procedure described in Chapter 1. This would of course need to be verified by testing of a real prototype in clinical practice.

Also heeding this desire for actionability, special attention was paid to the calibration of the models, so that it could be assumed to be reliable across not only the binary of risk/no-risk, but in identifying those at slightly elevated risk [40, 58, 74].

Chapter 7 shows that the group of prospective likely users on average perceived a moderate level of utility in this system. The strong focus on setting in the open-ended response from the TAM shows that while the potential for utility of the proposed application is generally accepted, this acceptance is restricted to appropriate clinical environments. It is not surprising to see that the supported settings (high volume hospitals with emergency departments, and teaching hospitals with a prevalence of junior clinicians) are congruent with the target institution, where the majority of queried settings (rural, community and multipurpose services) are not. Paediatric departments and end-of-life (EOL) care are also

both identified as domains requiring specialised end-points and actions. No paediatric data were included in the development of this model, reflecting the same assumption that those cases would indeed have different requirements, and EOL care is handled under a targeted use-case allowing those patients to be included or excluded as deemed appropriate by staff.

8.2.3.2 To be trusted and believed in

The fact that the external validation and impact study for this model is limited by necessity has already been addressed in Section 8.2.2. In addition to this, the credibility of this model must be judged on its ability to meet an appropriate level of face-validity in terms of both model scrutability and alignment with clinical best practice.

This model makes no attempt to use a technological solution for attention or explainability that can mathematically attribute the importance of input factors on the overall assessment. Instead, it takes advantage of the fact that it is core to the design of this solution that multiple risk assessments are produced, which change over time. By time-stamping each risk assessment, and allowing the user an intuitive mechanism for exploring the events that co-occur with any change in risk, the necessity for a more complex mechanism is avoided, which would bring with it more assumptions than this simple alignment of factors. A weakness of this approach is that it is not as directly applicable to patients just entering the system, but as seen in Chapter 4, the strength of this model increases as more data are collected, and therefore appropriately reflects the lower confidence that can be taken from any prediction made at those early stages. Although on average, the perceived utility (PU) of the system was moderately accepted, the PU measure had a wider range than the perceived ease of use (PEOU) measure, showing a level of scepticism remains that such a model could actually affect patient outcomes and clinical practice in a positive way. The only way to address this concern would be with an impact study, whether of a prototype in a controlled setting (such as described in [189]) or a comprehensive pre-post design to study the impact of an implemented system (such as [190]). Both of these approaches were infeasible within the scope of this work, but given the resource-intensive requirements for implementation into even a single institution, both will be required in order to pass over the barriers to local and wide-ranging implementations respectively.

8.2.3.3 To be fit for use

The most concrete way in which the proposed system addresses the identified usability requirements is in the strict adherence to the use of only real-time available data that is feasibly available in the clinical record, despite the technological impact of the integration requirement.

What it was not possible to deliver is an implementation plan that expands beyond the technological requirements, i.e. training and support materials, process change design, and the learning curve of these changes. The PEOU measure does give an indication of acceptable usability in the proposed system, where at least the interface design together with the description of the target use-cases is judged to be moderately easy to use. The free-text responses to the TAM also demonstrate in their consistency the ability of potential users to understand and identify with the core purpose of this intervention, despite only high-level description.

8.2.4 Evaluation under NASSS framework

Given the importance placed on qualities of translatability in this body of work, it is valuable to further assess the proposed application through the lens of implementation science. In doing this it becomes possible to anticipate the readiness of the target institution to successfully incorporate it within their clinical emergency response system.

The NASSS framework uses complexity theory to explain barriers between technological innovations and their eventual success or failure upon implementation. Importantly, it considers not only the obstacles to a successful initial system delivery, but also impediments to a sustainable, systemic, transforming change.

This framework asks users to assess the technological system under seven domains (see Table 8.1), through which it is possible to produce a narrative that reveals a comprehensive picture of the complexity of a technology within its setting. The core premise is that higher complexity systems are naturally less likely to see successful uptake. Each domain is

categorised as simple (few, predictable factors), complicated (many, predictably interacting factors), or complex (many factors, with fuzzy bounds and unpredictable or unknown interactions). Following from this is the supposition that by articulating areas of complexity, one may seek to simplify, or support organisations to handle complexities where this is not possible. In applying this framework, a note must be made that it was first released in 2018,

Domain	**Core question**
Condition	How well-defined is the condition that this technology addresses?
Technology	How challenging is the technology itself to both use and supply?
Value Proposition	Is the value proposition of the technology clear and plausible?
Adopters	Are the intended system users resistant to this innovation?
Organisation	Is the organisation one that can support innovation?
Wider System	How is this organisation's capacity to innovate affected by external forces?
Embedding & Adaptation over time	Does this technology have an intrinsic capacity to adapt to external, unexpected changes?

TABLE 8.1. NASSS Framework Domains

two years after searches were performed for the systematic review presented in Chapter 2, and thus does not inform any of the included studies. Despite this fact, it provides the sorely needed balance between theory and pragmatism that will bridge the gap between *the model in the journal* and *the model in practice*[1], and can be used post-hoc to explain some of the observed healthcare provider opinions.

This theoretical framework provides a useful lens through which to interpret the outputs of the *technology adoption model* (TAM) [158], which was applied in Chapter 7, as it addresses the translatability of a technological solution as it is situated within the highly complex system in which it is to be implemented. Although understanding the behavioural intentions and attitudes described by the TAM is of course a necessary precursor of successful implementation, and can inform a number of the NASSS domains (notably

[1]See: [187] for the parallel *patient in the guideline* vs. *the patient in the bed*

Technology and *Adopters*), they are not in and of themselves sufficient to predict whether a technological solution will deliver all the benefits that are expected or promised.

The experience of capturing concrete use-cases as described in Chapter 6 demonstrates clearly the need to consider the watch-list to be a complex adaptive system. This can be seen in particular in UC-3 and UC-4, where stakeholders anticipated second-order use-cases for end of life care and clinical administrative tasks respectively. Although these tasks did not form part of the initial project scope, through the iterative design process it rapidly became clear that the ways in which these user groups would also want to interact with the system must be accommodated. This complexity is thus evident before consideration is even expanded from the scope of the watch-list application itself to encompass the deep integration required with the clinical record. The integrated architecture brings with it not only the obvious technological interdependencies but also the impacts from any potential future process changes, which are difficult to predict, and can already be observed in the source data as described in Chapter 3.

Greenhalgh states that *"...complexity tends to be inherent in healthcare programmes, [and] the key challenge is often to develop ways of 'running with' complexity, rather than seeking to eliminate it"* [185]. To this end, in this section each domain of the NASSS framework is described as it applies to the proposed watch-list application, including ways in which the design of this system has either accommodated or mitigated those inherent or introduced complexities.

There are some logical limits to the application of this framework, given the as-yet hypothetical nature of the watch-list, and thus an inability to produce empirical evidence of the capacity of the organisation and wider system to adapt to this innovation. The example case studies provided by the framework authors and their colleagues [191, 192] are evaluative, and therefore have the advantage of hindsight; however the NASSS is also designed to help predict program success, before such complete information can be collected. One strength of this framing is its characterisation of systems as open, dynamic and adapting, so this form is followed to make assumptions around future uncertainty and unpredictable factors based on the available evidence. A concrete example of the impact of unanticipated future changes was seen in the sudden reallocation of funding and human

resources (including direct stakeholders in this project) within the critical care department due to the impact of COVID-19. Whether these personnel changes persist, are reverted, or some as yet undefined configuration is implemented remains to be seen. Nonetheless, it remains clear that any study of technological innovation of sufficiently significant scale that assumes a static system with known bounds will struggle to be effective.

8.2.4.1 The condition

The condition that is the target of the watch-list technology is definitionally complex. In-patient deterioration is often caused by the interaction of many factors. Although the root cause of a given individual's path to deterioration may be well defined, the broad scope of the critical care outreach role means that in any one shift these clinicians are likely to be dealing with many different care pathways and treatment options, including ones that are much less well understood. The conditions leading to deterioration are frequently also volatile and associated with multi-comorbidities as well as complex socio-cultural factors.

Here there is a tension between the stated requirements of this project's stakeholders, where the use of a watch-list is desirable specifically for its anticipated capacity to reduce existing complexity, and the NASSS framework which attributes failure of a technological innovation in part to its application to a complex condition. It may be more helpful therefore to describe the complexity of the intended clinical action of the technology (which is typically, but not universally, tied to the complexity of the condition), instead of relating this domain strictly to the condition itself.

In this instance, the clinical action from the output of the watch-list is specifically designed for its capacity to simplify. There is no attempt to diagnose or direct treatment, except in its ability to summarise large quantities of data over time and produce a priority queue. Although the inputs are many, varied, and unpredictable, the outputs are straightforward, and have the capacity to reduce the complexity of the status-quo, wherein individuals are required to keep track of this high volume of data in an ad-hoc fashion. Although this expected reduction in complexity is untested in actual clinical practice, it forms the core rationale behind the initiation of this work, and is further supported by numerous free-text responses captured in Chapter 7.

This can be compared to the analysis in [191], which saw large differences in uptake of the same video consultation technology for routine diabetes management versus management of gestational diabetes (20% and 3% respectively). This was explained by the different levels of complexity in the conditions, but perhaps a more nuanced explanation would be to compare the level of complexity in action instead — i.e. monitoring a stable, long-term patient, in contrast to educating and monitoring a metabolic instability condition for a patient who is not used to managing their disease, in particular, new to the physical act of insulin administration.

8.2.4.2 The technology

Application architecture. The proposed application is designed to be tightly integrated with the electronic medical record, which is a highly complex requirement even in a single institution. EMR systems are driven by idiosyncratic data models that are typically heavily customised for each implementation, requiring significant domain knowledge (both of the EMR application itself, but also specific to an institution's instance) to enact even simple outbound integrations.

Through initial model development, external model validation, and demonstrating the novel data augmentation techniques a version of the application architecture was developed for three different source data systems. Each time, it was possible to reconstruct the pipeline from the new source data model with relatively few adjustments, particularly considering the completely different source data models, coding systems, and scope of data collected. Of note, the scope of the development work was less than that required for a more traditional ML model with a comparable number of features, where precise item by item mapping is required. In order to generalise this innovation to multiple sites, it must be assumed that the integration layer will be bespoke for each deployment, noting however that in a realistic commercial implementation it would be feasible to abstract away many of these challenges with a robust ontological model covering each subset of the source data.

The complexity of this architecture is also somewhat ameliorated with the implementation of an Object Relational Model (ORM). This additional level of abstraction acts as a

buffer between the source data and the validated application and allows the necessary customisation to be achieved by re-configuration, as opposed to re-development.

In addition to the complex overall application architecture, the decision to implement a deep learning model at the core of the proposed solution must be considered. There is high complexity in the specialised knowledge required to design, implement and host the final model, and furthermore, there is very limited prior work to draw upon in anticipating second and third order consequences of its deployment.

Performance impacts. It is typical in modern hospitals to implement some form of data mirror, whereby the tasks of reporting and analysis are decoupled from the production system to reduce the risk of introducing performance and data integrity issues. Any potential for similar impact from the deployment of a deep learning model must be considered, in particular one that requires access to near-live clinical data.

A hypothetically sound implementation could be the implementation of a dedicated application server within the secure network which can host the trained models. New or updated clinical data can be pushed from the EMR to this server either as it becomes available or batched and pushed according to a schedule that is pre-determined to have an acceptable balance between responsiveness of the downstream application and burden on the production system.

This would likely be sufficient for a pilot implementation, but for a truly productionalised roll-out, this complexity must be further extended to include a strategy for model re-training, and the implementation of modern messaging paradigms (FHIR, HL7) that are supported in such settings.

User interface. The user interface itself is by comparison a far simpler consideration that can be built upon well-understood and flexible technologies that are straightforward to develop and deploy, and have very few resultant dependencies. Its implementation is envisaged as a web-app, which reflects current practice in the target institution for all clinical user interfaces, and is further typical of many EMR-based applications.

Here the PEOU construct from the TAM must be examined. Despite the complete novelty of the watch-list system, with no directly comparable precursor, users on average agreed with statements that described the system as easy to use, and only one user moderately disagreed with these statements across the board. The interface was designed to reflect affordances and workflows that are common in clinical applications, drawing on general familiarity and comfort with simple, modern application elements. The use of colour in the traffic-light style risk stratification and visualisation of risk changes over time was also highly valued in a number of studies [39, 43, 70]. The eventual burden of the application learning curve remains to be borne out in practice, however it is anticipated that it will be relatively simple for end-users to include in their workflow.

This technology stack can be imagined as akin to the proverbial duck, where the vast majority of its complexity is 'under the water', at least one step removed from the end-user. If a careful and considered roll-out that pays sufficient attention to the back-end complexity can be achieved, from the point of view of the clinical users the behaviour of the technology can be expected to fall within the bounds of acceptable complexity. From the point of view of the organisation itself, however, the level of complexity is extreme.

8.2.4.3 The value proposition

Although the core value proposition of this application is somewhat speculative at this nascent stage, it aligns with previous works [193, 194] that demonstrate a reduction in cardiac arrest events as a result of early review of the deteriorating patient. Whether this can be extrapolated to even earlier interventions leading to more conclusively positive outcomes remains to be seen (and measured).

Internal reports into the critical care outreach role itself [195] also point to the generally positive impact that this role is having, although low incident counts for key end-points preclude the drawing of statistically significant conclusions. An additional value proposition that was suggested by the clinical stakeholders in this project is based on the supposition that the watch-list is able to function as a self-monitoring system. This means that patient risk and outcomes over time are all stored within the application state itself and therefore any value realised from this intervention will be concrete and directly attributable,

which is a significant improvement over the ad-hoc monitoring processes that are the status-quo. This suggestion is also directly and spontaneously supported by one of the free-text responses to the TAM questionnaire.

The accommodation of the secondary end of life and administrative use-cases of these models also strengthens the argument for the value of this application. Despite the speculative nature of the value proposition, it has multiple paths by which it can be expected to streamline processes and be of material benefit to the healthcare system by reducing costs and/or improving outcomes. In particular, end of life care and advance care planning is an area that reportedly suffers from inconsistent and largely untracked application in the target institution, despite known benefits to the patient, surviving family members, and health system [196]. A straightforward and centralised way of identifying patients who should be under the care of a palliative-care specialist and/or have an advanced care directive in place has the potential to not only improve consistency of care, but also creates an opportunity for significant developments in the field of end of life research.

8.2.4.4 The intended adopters

The intended adopters are the staff who are expected to use the watch-list application for one or more of the identified use-cases. The way in which the implementation of this technology affects these users has formed the core driver behind much of this thesis, and as such, has already been addressed in the most detail of any of the NASSS domains. For a full exposition of this domain, and how it relates to the gathered evidence-base, please refer to Section 8.2.3.

8.2.4.5 The organisation

Without insider-level access to the management structures and culture at the target organisation, it is not possible to provide a full evaluation of its capacity to support innovation. Although some of the stakeholders in this project did indeed have such access, an analysis of organisational structure and climate were not within the agreed research scope, and therefore it was not appropriate to capture these details directly.

There are some externally known or observable factors, however, from which it is possible to infer some of the structural determinants that affect the organisation's readiness to pilot the watch-list.

St Vincent's hospital has a history of technological innovation, and were early adopters of electronic medical records, relative to other large Australian hospitals. The internally-developed deLacy application was first deployed in 1993 [197], before being modernised as Web deLacy in 2006. Of note is the way that this development effort was rooted in nursing process theory, and how nursing staff in particular transitioned into leadership informatics roles. This echoes the genesis of the watch-list project, which is very much nursing-led.

This practice of early adoption brings with it some challenges, however, and there have been periods where being out of step with the bulk of comparable hospitals (even in the positive direction) puts an institution at risk of incompatibilities and unsupported requirements. Despite this, deLacy has continued to be used to break new ground, being the first system to provide integration with the Australian National Personally Controlled Electronic Health Record (PCEHR) [198].

This appetite for innovation must be balanced against the lack of organisational slack and inflexibility of resourcing at this hospital [199]. This is evident in the roll-out of the critical care outreach role, which has remained only partially funded for the duration of this project, and has been granted insufficient resources to effectively evaluate the program.

8.2.4.6 The wider system

St Vincent's Health operates a mix of public and private hospitals, as well as numerous aged-care communities and two dedicated palliative-care facilities across the east coast of Australia. Although the scope of this project applies currently only to the public hospital in Darlinghurst, Sydney, this heterogeneity of geography and remit must be considered when planning a sustainable roll-out, in particular the effect of complex governance structures that span multiple state-level health authorities.

Based on the organisational appetite for innovation described in the previous section, however, it is possible that despite the increased complexity that comes from operating in multiple jurisdictions, this may actually break down the silos that are characteristic of the relationships between state healthcare systems in Australia and thus act as a facilitator of innovation as opposed to a barrier. There is an appetite in Australian research funding bodies to give weight to applications that cross state lines, so this facilitation may be as simple as the capacity of this institution to access funding, although such conclusions cannot be more than speculative.

To date, there is no attempt to specifically regulate the implementation of predictive models in healthcare in Australia beyond the existent privacy, security and discrimination obligations (under which the hospital systems already operate) although there is a road map towards AI standards [200] and a national AI ethical framework [201], both of which must inform such systems. This lags behind the European Union, which specifies requirements for model explainability under the General Data Protection Regulation [202] and the United States, which is starting to consider the role of clinical regulatory bodies as applied to these systems [157]. It is reasonable to expect that Australia will fall into step with at least some elements of these regulatory models however, as the AI standards discussion paper already highlights the need to be operating under internationally harmonised standards [200], so it would be wise to assume that similar controls will be put in place in the near future.

8.2.4.7 Evolution and adaptation over time

System adaptation. The tightly integrated architecture introduces a brittleness where changes in the data source can introduce downstream software bugs. This is a necessary trade-off between the desire of clinical end-users to avoid all instances of double data entry [48, 71] and the inherent complexity of any integrated software system.

The implementation of an ORM abstraction as described previously goes some of the way to reduce this impact, and further than just the decoupling of source and target system, the particular selection of SQLAlchemy offers a mature ecosystem of data migration tools. The scale of such migration projects in an enterprise-grade system (in particular one with auditable elements) must nonetheless not be underestimated.

Rajkomar et al. [81] use the FHIR messaging standard as the basis of an alternative architecture, as this data standard can be expected to be a persistent feature of the source data system, regardless of any future updates. Given the serialised processing that forms much of the core logic of the final models, this has the additional benefit of reducing the computational burden on the downstream system, which is not insignificant. At such time as this messaging standard becomes pervasive within the Australian system, this strategy should be considered for its multiple benefits in reducing the complexity of the technology and adaptability domains. The pre-processing techniques described in Chapter 3 can equally be applied to such messages, so this modification is technically feasible, and does not affect the core proposal.

Note that the equivalent HL7 messages are not persisted in the source system, so were not available for this proof-of-concept work.

Procedural adaptation. Finally, in Chapter 3, it is already possible to see evidence of ways in which a solution must be robust not only to technological changes, but also able to accommodate and reflect process change. This is especially true in a setting such as this, where there is an expectation that use of the system will itself be a driver of changing practice over time. Davis et al [203] suggest a procedure for updating clinical models to combat model performance drift that takes into account changes in population, end-point rates and changes in the relationships between predictors and outcomes. Any or all of these changes are highly likely to occur within the timescales of such an implementation.

Assuming that these process changes are implemented at a smooth pace, and with a sufficiently robust feedback loop in place, it is theoretically possible to update model training iteratively, thereby reflecting current practice, although a mature and updating model that can be trusted to adapt in this way is still many steps away from implementation, and represents a highly complex proposition.

The capacity of a version of the prototype model (and a highly simplified version at that) to demonstrate transferability of the core concepts into a data set that is fundamentally quite different (changed setting in particular) between Chapters 4 and 5 gives support to its technical feasibility, although not its procedural implementation.

8.2.5 Future research directions

In Sections 8.2.3 and 8.2.4, a number of important research directions are discussed, namely the expansion of external validation, impact study and the development of a principled roll-out strategy. The careful study of the fairness and ethical impact of any implemented system must be added to these. It is clear from the error analysis in Chapter 4 that there is a difference in the types of cases that cause false-positive and false-negative assessments from the model. If this is true, the capacity for systemic bias must be assumed to follow. The consistent highlighting of certain patient groups to be either prioritised or deprioritised relative to other groups will have wide-ranging impacts on patient outcomes that can easily be obscured in population-level analyses.

8.3 Conclusion

8.3.1 Account of research questions

As the conclusion, the four original research questions that were stated in Chapter 1 are evaluated.

T.1 Determine an appropriate modelling architecture that can, in principle, identify patients at risk of deterioration in the short term from the clinical record, in real time and without access to vital signs data.

Chapter 4 demonstrates that the lack of vital signs data, although significant, does not prevent the delivery of a model that can predict death and unplanned ICU admission within 24 hours with accuracy that is comparable to baseline deterioration models, despite their advantage of being able to include important physiological variables.

T.2 Measure how well such an architecture can generalise within the target institution.

Although in this discussion chapter the proposed solution was critiqued for not having reached the level of external validation, in fact the original proposal was to deliver a system that generalised locally, which is achieved through the use of hold-out test data as the performance benchmark. The validation of generalisability of the high-level techniques, and delivery of published code as applicable to a publicly available data set in Chapter 5, goes much further than this stated aim.

C.1 Understand the qualities of predictive models that are most valued by clinical end-users.

The review of qualitative literature in Chapter 2 provides a proxy for understanding how clinicians value and want to interact with predictive models in practice. There remains a significant gap between what is possible and what is available in this realm, so there is no perfect measure of this yet. From the consistency of results between what was expected from the literature and what was observed through the requirements-gathering phase in Chapter 6, and in the application of the technology adoption model in Chapter 7, it is reasonable to be confident that these observations provide a sufficiently strong foundation until such further examination is possible.

C.2 Apply these qualities to the delivered model as a prototype, and measure the success of this application as perceived by likely stakeholders.

Using the framework provided by the NASSS model, the general capacity for success of the proposed model is assessed. Although there are significant complexities observed, in most cases they can be justified as unavoidable in the provision of the core objectives of this work. Notably, in the technology and adaptability domains, there exists capacity to reduce complexity in the future through the adoption of widely accepted messaging standards.

In presenting the prototype interface to likely stakeholders in this pre-adoption state, its reception is warm across the domains of perceived usefulness, perceived ease of use, and in behavioural intention to use. Although this success is somewhat equivocal, until a ‘hands-on’ experience can be delivered, it is natural to expect some uncertainty.

8.3.2 Key contributions

Throughout this work it has been demonstrated that current best practice for clinical prediction system development (whether simple rules-based or bleeding-edge AI) does not take into account the way that it affects its intended end users. The current 'best-case' scenario is an implementation study showing an effect on clinical outcomes. This is a necessary but insufficient analysis if the end-goal is robust, systemic modernisation of the use of predictive technologies applied to real-time observational data in clinical practice.

This work demonstrates the feasibility of a more holistic approach, considering the perspectives of clinical end users in parallel with the technological development phase. By considering the clinical factors prior to implementation, it is possible to adapt to user needs much earlier in the process, reducing the risk of failure upon delivery.

The technical feasibility of a prototype system using the limited real-time available data at the target institution by applying novel context-specific processing techniques is also shown, and importantly, the key technical innovations of this proposed system were found to be transferable into other settings.

Bibliography

1. Hillman, K. *et al.* Antecedents to hospital deaths. *Internal medicine journal* **31,** 343–348 (2001).
2. Schein, R. M., Hazday, N., Pena, M., Ruben, B. H. & Sprung, C. L. Clinical antecedents to in-hospital cardiopulmonary arrest. *Chest* **98,** 1388–1392 (1990).
3. Kause, J. *et al.* A comparison of antecedents to cardiac arrests, deaths and emergency intensive care admissions in Australia and New Zealand, and the United Kingdom—the ACADEMIA study. *Resuscitation* **62,** 275–282 (2004).
4. Buist, M., Bernard, S., Nguyen, T. V., Moore, G. & Anderson, J. Association between clinically abnormal observations and subsequent in-hospital mortality: A prospective study. *Resuscitation* **62,** 137–141 (2004).
5. McQuillan, P. *et al.* Confidential inquiry into quality of care before admission to intensive care. *BMJ* **316,** 1853–1858 (1998).
6. Dubois, R. W. & Brook, R. H. Preventable deaths: who, how often, and why? *Annals of Internal Medicine* **109,** 582–589 (1988).
7. Clinical Excellence Commission. *Recognition and Management of Patients who are Clinically Deteriorating PD2013:049* https://www1.health.nsw.gov.au/pds/ActivePDSDocuments/PD2013_049.pdf. Accessed: 2019-03-10. 2013.
8. Hourihan, F., Bishop, G., Hillman, K. & Daffurn, K. The medical emergency team: A new strategy to identify and intervene in high-risk patients. *Clinical Intensive Care* **6,** 269–272 (1995).
9. DeVita, M. A. *et al.* "Identifying the hospitalised patient in crisis"—a consensus conference on the afferent limb of rapid response systems. *Resuscitation* **81,** 375–382 (2010).
10. Lee, A., Bishop, G., Hillman, K. & Daffurn, K. The medical emergency team. *Anaesthesia and intensive care* **23,** 183–186 (1995).

11. St Vincents Hospital Network — Safe, Harm-Free Care Committee. *Clinical Emergency Response System Protocol (SVH Internal)* tech. rep. (2018).
12. Department of Health (United Kingdom). Comprehensive critical care – A review of adult critical care services (May 2000).
13. McDonnell, A. *et al.* The provision of critical care outreach services in England: findings from a national survey. *Journal of critical care* **22,** 212–218 (2007).
14. Jones, D., Mitchell, I., Hillman, K. & Story, D. Defining clinical deterioration. *Resuscitation* **84,** 1029–1034 (2013).
15. Venkatesh, V. & Davis, F. D. A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management science* **46,** 186–204 (2000).
16. Kennedy, G. & Gallego, B. Clinical prediction rules: A systematic review of healthcare provider opinions and preferences. *International journal of medical informatics* **123,** 1–10 (2018).
17. Kennedy, G., Rihari-Thomas, J., Dras, M. & Gallego, B. Developing a deep learning system to drive the work of the critical care outreach team. *medRxiv* (2020).
18. Kennedy, G., Dras, M. & Gallego, B. Augmentation of electronic medical record data for deep learning. *medRxiv* (2021).
19. Kennedy, G. & Gallego, B. Clinical readiness to adopt A.I. for critical care prioritisation. *medRxiv* (2021).
20. Groves, P., Kayyali, B., Knott, D. & Kuiken, S. V. The 'big data' revolution in healthcare: Accelerating value and innovation (2016).
21. Raghupathi, W. & Raghupathi, V. Big data analytics in healthcare: Promise and potential. *Health information science and systems* **2,** 3 (2014).
22. Murdoch, T. B. & Detsky, A. S. The inevitable application of big data to health care. *JAMA* **309,** 1351–1352 (2013).
23. Kellermann, A. L. & Jones, S. S. What it will take to achieve the as-yet-unfulfilled promises of health information technology. *Health Affairs* **32,** 63–68 (2013).
24. Boonstra, A. & Broekhuis, M. Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions. *BMC health services research* **10,** 1 (2010).
25. Najaftorkaman, M., Ghapanchi, A. H., Talaei-Khoei, A. & Ray, P. A taxonomy of antecedents to user adoption of health information systems: A synthesis of thirty

years of research. *Journal of the Association for Information Science and Technology* **66,** 576–598 (2015).

26. Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports* **6,** 26094 (2016).
27. Sackett, D. L. Evidence-based medicine. *Wiley Online Library* (2000).
28. Longhurst, C. A., Harrington, R. A. & Shah, N. H. A 'green button' for using aggregate patient data at the point of care. *Health Affairs* **33,** 1229–1235 (2014).
29. Reilly, B. M. & Evans, A. T. Translating clinical research into clinical practice: Impact of using prediction rules to make decisions. *Annals of internal medicine* **144,** 201–209 (2006).
30. Ingui, B. J. & Rogers, M. A. Searching for clinical prediction rules in MEDLINE. *Journal of the American Medical Informatics Association* **8,** 391 (2001).
31. Geersing, G.-J. *et al.* Search filters for finding prognostic and diagnostic prediction studies in MEDLINE to enhance systematic reviews. *PLoS One* **7,** e32844 (2012).
32. Perry, J. J. *et al.* National survey of Canadian neurologists' current practice for transient ischemic attack and the need for a clinical decision rule. *Stroke* **41,** 987–991 (2010).
33. Oostema, J. A., Brown, M. D. & Reeves, M. Emergency department management of transient ischemic attack: A survey of emergency physicians. *Journal of Stroke and Cerebrovascular Diseases* **25,** 1517–1523 (2016).
34. Lai, F., Macmillan, J., Daudelin, D. H. & Kent, D. M. The potential of training to increase acceptance and use of computerized decision support systems for medical diagnosis. *Human Factors: The Journal of the Human Factors and Ergonomics Society* **48,** 95–108 (2006).
35. Engelhardt, E. G. *et al.* Breast cancer specialists' views on and use of risk prediction models in clinical practice: A mixed methods approach. *Acta Oncologica* **54,** 361–367 (2015).
36. Weber, S. A qualitative analysis of how advanced practice nurses use clinical decision support systems. *Journal of the American Academy of Nurse Practitioners* **19,** 652–667 (2007).

37. Mertz, E. *et al.* Provider attitudes toward the implementation of clinical decision support tools in dental practice. *Journal of Evidence Based Dental Practice* **15,** 152–163 (2015).

38. Zakhari, R. & Sterrett, S. E. Attitudes toward evidence-based clinical decision support tools to reduce exposure to ionizing radiation: The Canadian CT Head Rule. *Journal of the American Association of Nurse Practitioners* **28,** 659–667 (2016).

39. Dikomitis, L., Green, T. & Macleod, U. Dealing with uncertainty: A qualitative evaluation of the usability and acceptability of an electronic risk assessment tool to aid cancer diagnosis in general practice. *Macmillan Electronic Risk Assessment Tool Summary,* 1–19 (2012).

40. Green, T. *et al.* Exploring GPs' experiences of using diagnostic tools for cancer: A qualitative study in primary care. *Family practice* **32,** 101–105 (2015).

41. Norton, W. E. *et al.* Acceptability of the decision support for safer surgery tool. *The American Journal of Surgery* **209,** 977–984 (2015).

42. Carroll, C. *et al.* Involving users in the design and usability evaluation of a clinical decision support system. *Computer methods and programs in biomedicine* **69,** 123–135 (2002).

43. Peiris, D. *et al.* An electronic clinical decision support tool to assist primary care providers in cardiovascular disease risk management: Development and mixed methods evaluation. *Journal of medical internet research* **11,** e51 (2009).

44. Peiris, D., Usherwood, T., Weeramanthri, T., Cass, A. & Patel, A. New tools for an old trade: A socio-technical appraisal of how electronic decision support is used by primary care practitioners. *Sociology of health & illness* **33,** 1002–1018 (2011).

45. Elustondo, S. G., Aguado, P. N., de La Rasilla Cooper, C. G., Manzanet, J. P. & Sendin, D. S. Cardiovascular risk tables: Opinion and degree of use of Primary Care doctors from Madrid, Spain. *Journal of evaluation in clinical practice* **19,** 148–152 (2013).

46. Liew, S. M., Blacklock, C., Hislop, J., Glasziou, P. & Mant, D. Cardiovascular risk scores: Qualitative study of how primary care practitioners understand and use them. *Br J Gen Pract* **63,** e401–e407 (2013).

47. Braithwaite, D., Sutton, S., Smithson, W. H. & Emery, J. Internet-based risk assessment and decision support for the management of familial cancer in primary care: A survey of GPs' attitudes and intentions. *Family Practice* **19,** 587–590 (2002).

48. Hoonakker, P., Khunlertkit, A., Tattersall, M., Keevil, J. & Smith, P. D. Computer decision support tools in primary care. *Work* **41,** 4474–4478 (2012).

49. Pearson, S. D., Goldman, L., Garcia, T. B., Cook, E. F. & Lee, T. H. Physician response to a prediction rule for the triage of emergency department patients with chest pain. *Journal of general internal medicine* **9,** 241–247 (1994).

50. Van Oort, L. *et al.* Evaluation of the usefulness of 2 prediction models of clinical prediction models in physical therapy: A qualitative process evaluation. *Journal of manipulative and physiological therapeutics* **37,** 334–341 (2014).

51. Kappen, T. H. *et al.* Barriers and facilitators perceived by physicians when using prediction models in practice. *Journal of Clinical Epidemiology* **70,** 136–145 (2016).

52. Short, D., Frischer, M. & Bashford, J. Barriers to the adoption of computerised decision support systems in general practice consultations: A qualitative study of GPs' perspectives. *International journal of medical informatics* **73,** 357–362 (2004).

53. Boutis, K. *et al.* Pediatric emergency physician opinions on ankle radiograph clinical decision rules. *Academic Emergency Medicine* **17,** 709–717 (2010).

54. Zwar, N., Comino, E., Harris, M., Torley, D. *et al.* GPs' views of absolute cardiovascular risk and its role in primary prevention. *Australian family physician* **34,** 503 (2005).

55. Eichler, K., Zoller, M., Tschudi, P. & Steurer, J. Barriers to apply cardiovascular prediction rules in primary care: A postal survey. *BMC family practice* **8,** 1 (2007).

56. Müller-Riemenschneider, F. *et al.* Barriers to routine risk-score use for healthy primary care patients: survey and qualitative study. *Archives of Internal Medicine* **170,** 719–724 (2010).

57. Sarazin, M., Chiappe, S. G., Kasprzyk, M., Mismetti, P. & Lasserre, A. A survey of French general practitioners and a qualitative study on their use and assessment of predictive clinical scores. *International journal of general medicine* **6,** 419 (2013).

58. Bonner, C. *et al.* General practitioners' use of different cardiovascular risk assessment strategies: A qualitative study. *Medical Journal of Australia* **199,** 485–489 (2013).

59. Haskins, R., Osmotherly, P. G., Southgate, E. & Rivett, D. A. Physiotherapists' knowledge, attitudes and practices regarding clinical prediction rules for low back pain. *Manual therapy* **19,** 142–151 (2014).

60. Plüddemann, A. *et al.* Clinical prediction rules in practice: Review of clinical guidelines and survey of GPs. *Br J Gen Pract* **64,** e233–e242 (2014).

61. Knox, G. M., Snodgrass, S. J. & Rivett, D. A. Physiotherapy clinical educators' perceptions and experiences of clinical prediction rules. *Physiotherapy* **101,** 364–372 (2015).

62. Richardson, S. *et al.* Healthcare provider perceptions of clinical prediction rules. *BMJ open* **5,** e008461 (2015).

63. Brown, B. *et al.* Understanding clinical prediction models as 'innovations': A mixed methods study in UK family practice. *BMC medical informatics and decision making* **16,** 106 (2016).

64. Feder, S. L. *et al.* Risk stratification in older patients with acute myocardial infarction: Physicians' perspectives. *Journal of aging and health* **28,** 387–402 (2016).

65. Graham, I. D., Stiell, I. G., Laupacis, A., O'Connor, A. M. & Wells, G. A. Emergency physicians' attitudes toward and use of clinical decision rules for radiography. *Academic Emergency Medicine* **5,** 134–140 (1998).

66. Graham, I. D. *et al.* Awareness and use of the Ottawa ankle and knee rules in 5 countries: Can publication alone be enough to change practice? *Annals of emergency medicine* **37,** 259–266 (2001).

67. Brehaut, J. C., Stiell, I. G., Visentin, L. & Graham, I. D. Clinical decision rules "in the real world": How a widely disseminated rule is used in everyday practice. *Academic emergency medicine* **12,** 948–956 (2005).

68. Ballard, D. W. *et al.* Emergency physicians' knowledge and attitudes of clinical decision support in the electronic health record: A survey-based study. *Academic Emergency Medicine* **20,** 352–360 (2013).

69. Porter, A. *et al.* It could be a 'Golden Goose': A qualitative study of views in primary care on an emergency admission risk prediction tool prior to implementation. *BMC family practice* **17,** 1 (2016).

70. Chiang, P. P., Glance, D., Walker, J., Walter, F. & Emery, J. Implementing a QCancer risk tool into general practice consultations: An exploratory study using simulated consultations with Australian general practitioners. *British journal of cancer* **112,** S77–S83 (2015).

71. Crawford, F., Bekker, H., Young, M. & Sheikh, A. General practitioners' and nurses' experiences of using computerised decision support in screening for diabetic

foot disease: Implementing Scottish Clinical Information-Diabetes Care in routine clinical practice. *Journal of Innovation in Health Informatics* **18,** 259–268 (2010).

72. Lautenbach, E., Localio, R. & Nachamkin, I. Clinicians required very high sensitivity of a bacteremia prediction rule. *Journal of clinical epidemiology* **57,** 1104–1106 (2004).
73. Perry, J. J. *et al.* Emergency physicians' management of transient ischemic attack and desired sensitivity of a clinical decision rule for stroke in three countries. *CJEM* **13,** 19–27 (2011).
74. Collins, I. M. *et al.* Assessing and managing breast cancer risk: Clinicians' current practice and future needs. *The Breast* **23,** 644–650 (2014).
75. Haskins, R., Osmotherly, P. G., Southgate, E. & Rivett, D. A. Australian physiotherapists' priorities for the development of clinical prediction rules for low back pain: A qualitative study. *Physiotherapy* **101,** 44–49 (2015).
76. Li, A. C. *et al.* Integrating usability testing and think-aloud protocol analysis with 'near-live' clinical simulations in evaluating clinical decision support. *International journal of medical informatics* **81,** 761–772 (2012).
77. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC medicine* **13,** 1 (2015).
78. Shickel, B., Tighe, P. J., Bihorac, A. & Rashidi, P. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics* (2017).
79. Richardson, S. *et al.* 'Think aloud' and 'Near live' usability testing of two complex clinical decision support tools. *International journal of medical informatics* **106,** 1–8 (2017).
80. Smith, G. B., Prytherch, D. R., Meredith, P., Schmidt, P. E. & Featherstone, P. I. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation* **84,** 465–470 (2013).
81. Rajkomar, A. *et al.* Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine* **1,** 18 (2018).

82. Baytas, I. M. *et al. Patient subtyping via time-aware LSTM networks* in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (2017), 65–74.

83. Wang, L., Wang, H., Song, Y. & Wang, Q. MCPL-Based FT-LSTM: Medical Representation Learning-Based Clinical Prediction Model for Time Series Events. *IEEE Accesss* **7,** 70253–70264 (2019).

84. Guthrie, D., Allison, B., Liu, W., Guthrie, L. & Wilks, Y. *A Closer Look at Skip-gram Modelling* in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)* (European Language Resources Association (ELRA), Genoa, Italy, May 2006).

85. Beam, A. L. *et al.* Clinical concept embeddings learned from massive sources of multimodal medical data. *arXiv preprint arXiv:1804.01486* (2018).

86. Laksana, E. *et al.* The impact of extraneous variables on the performance of recurrent neural network models in clinical tasks. *arXiv preprint arXiv:1904.01125* (2019).

87. Gupta, P., Malhotra, P., Vig, L. & Shroff, G. Transfer learning for clinical time series analysis using recurrent neural networks. *arXiv preprint arXiv:1807.01705* (2018).

88. Alvarez, C. A. *et al.* Predicting out of intensive care unit cardiopulmonary arrest or death using electronic medical record data. *BMC medical informatics and decision making* **13,** 28 (2013).

89. Churpek, M. M., Yuen, T. C., Park, S. Y., Gibbons, R. & Edelson, D. P. Using electronic health record data to develop and validate a prediction model for adverse outcomes on the wards. *Critical care medicine* **42,** 841 (2014).

90. Churpek, M. M. *et al.* Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Critical care medicine* **44,** 368 (2016).

91. Kipnis, P. *et al.* Development and validation of an electronic medical record-based alert score for detection of inpatient deterioration outside the ICU. *Journal of biomedical informatics* **64,** 10–19 (2016).

92. Green, M. *et al.* Comparison of the Between the Flags calling criteria to the MEWS, NEWS and the electronic Cardiac Arrest Risk Triage (eCART) score for the identification of deteriorating ward patients. *Resuscitation* **123,** 86–91 (2018).

93. Skyttberg, N., Chen, R., Blomqvist, H. & Koch, S. Exploring vital sign data quality in electronic health records with focus on emergency care warning scores. *Applied clinical informatics* **8,** 880 (2017).

94. Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F. & Sun, J. *Doctor AI: Predicting clinical events via recurrent neural networks* in *Machine Learning for Healthcare Conference* (2016), 301–318.

95. Hettige, B., Li, Y.-F., Wang, W., Le, S. & Buntine, W. MedGraph: Structural and temporal representation learning of electronic medical records. *arXiv preprint ArXiv:1912.03703* (2019).

96. Hong, S., Wu, M., Li, H. & Wu, Z. *Event2vec: Learning representations of events on temporal sequences* in *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data* (2017), 33–47.

97. Farhan, W. *et al.* A predictive model for medical events based on contextual embedding of temporal sequences. *JMIR medical informatics* **4,** e39 (2016).

98. Branco, P., Torgo, L. & Ribeiro, R. P. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)* **49,** 31 (2016).

99. Le Guennec, A., Malinowski, S. & Tavenard, R. *Data augmentation for time series classification using convolutional neural networks* in (2016).

100. Zadrozny, B. & Elkan, C. *Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers* in *Icml* **1** (2001), 609–616.

101. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. *On calibration of modern neural networks* in *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (2017), 1321–1330.

102. Raeder, T., Forman, G. & Chawla, N. V. in *Data mining: Foundations and intelligent paradigms* (eds Holmes, D. E. & Jain, L. C.) 315–331 (Springer, Berlin, Heidelberg, 2012).

103. Zadrozny, B. & Elkan, C. *Transforming classifier scores into accurate multiclass probability estimates* in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (2002), 694–699.

104. Platt, J. *et al.* Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* **10,** 61–74 (1999).

105. Kumar, A., Sarawagi, S. & Jain, U. *Trainable calibration measures for neural networks from kernel mean embeddings* in *International Conference on Machine Learning* (2018), 2805–2814.

106. Johnson, A. E., Pollard, T. J., Horng, S., Celi, L. A. & Mark, R. MIMIC-IV, version 1.0. *PhysioNet* (2021).

107. Franklin, C. & Mathew, J. Developing strategies to prevent inhospital cardiac arrest: Analyzing responses of physicians and nurses in the hours before the event. *Critical care medicine* **22,** 244–247 (1994).

108. McGloin, H., Adam, S. K. & Singer, M. Unexpected deaths and referrals to intensive care of patients on general wards. Are some cases potentially avoidable? *Journal of the Royal College of Physicians of London* **33,** 255–259 (1999).

109. Hughes, C., Pain, C., Braithwaite, J. & Hillman, K. 'Between the flags': Implementing a rapid response system at scale. *BMJ Qual Saf* **23,** 714–717 (2014).

110. Ribeiro, M. T., Singh, S. & Guestrin, C. "Why should I trust you?": Explaining the predictions of any classifier. *CoRR* **abs/1602.04938.** arXiv: 1602.04938 (2016).

111. Lundberg, S. M. & Lee, S.-I. *A unified approach to interpreting model predictions* in *Advances in neural information processing systems* (2017), 4765–4774.

112. Osborne, S., Douglas, C., Reid, C., Jones, L., Gardner, G. *et al.* The primacy of vital signs–acute care nurses' and midwives' use of physical assessment skills: A cross sectional study. *International Journal of Nursing Studies* **52,** 951–962 (2015).

113. Douglas, C. *et al.* Nursing and medical perceptions of a hospital rapid response system. *Journal of nursing care quality* **31,** E1–E10 (2016).

114. Douw, G. *et al.* Nurses' worry or concern and early recognition of deteriorating patients on general wards in acute care hospitals: A systematic review. *Critical care* **19,** 230 (2015).

115. Linnen, D. T. *et al.* Statistical modeling and aggregate-weighted scoring systems in prediction of mortality and ICU transfer: A systematic review. *Journal of hospital medicine* **14,** 161 (2019).

116. Xiao, C., Choi, E. & Sun, J. Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review. *Journal of the American Medical Informatics Association* **25,** 1419–1428 (2018).

117. Osmani, V. *et al.* Processing of electronic health records using deep learning: A review. *arXiv preprint arXiv:1804.01758* (2018).

118. Mayampurath, A., Sanchez-Pinto, L. N., Carey, K. A., Venable, L.-R. & Churpek, M. Combining patient visual timelines with deep learning to predict mortality. *PloS one* **14** (2019).

119. Avati, A. *et al.* Improving palliative care with deep learning. *BMC medical informatics and decision making* **18,** 122 (2018).

120. Cheng, Y., Wang, F., Zhang, P. & Hu, J. *Risk prediction with electronic health records: A deep learning approach* in *Proceedings of the 2016 SIAM International Conference on Data Mining* (2016), 432–440.

121. Kam, H. J. & Kim, H. Y. Learning representations for the early detection of sepsis with deep neural networks. *Computers in biology and medicine* **89,** 248–255 (2017).

122. Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G. & Galstyan, A. Multitask learning and benchmarking with clinical time series data. *Scientific data* **6,** 96 (2019).

123. Purushotham, S., Meng, C., Che, Z. & Liu, Y. Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics* **83,** 112–134 (2018).

124. Johnson, A. E. *et al.* MIMIC-III, a freely accessible critical care database. *Scientific data* **3,** 160035 (2016).

125. Cai, X. *et al.* Real-time prediction of mortality, readmission, and length of stay using electronic health record data. *Journal of the American Medical Informatics Association* **23,** 553–561 (2015).

126. Hosmer, D. W. & Lemesbow, S. Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods* **9,** 1043–1069 (1980).

127. Kramer, A. A. & Zimmerman, J. E. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Critical care medicine* **35,** 2052–2056 (2007).

128. Austin, P. C. & Steyerberg, E. W. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Statistics in medicine* **38,** 4051–4065 (2019).

129. Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of Big Data* **6,** 60 (2019).

130. Elisa, C. *et al.* Generative adversarial networks applied to observational health data. *arXiv preprint arXiv:2005.13510* (2020).

131. Yi, X., Walia, E. & Babyn, P. Generative adversarial network in medical imaging: A review. *Medical image analysis* **58,** 101552 (2019).

132. Wei, J. & Zou, K. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196* (2019).

133. Pollard, T. J. & Johnson, A. E. *The MIMIC-III clinical database* 2016.

134. MIT Laboratory for Computational Physiology. *MIMIC Source Code* https://github.com/MIT-LCP/mimic-code/tree/master/buildmimic/aws-athena. Accessed: 2020-02-01.

135. Abadi, M. *et al.* Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).

136. Ju, C., Bibaut, A. & van der Laan, M. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *Journal of Applied Statistics* **45,** 2800–2818 (2018).

137. Lo-Ciganic, W.-H. *et al.* Evaluation of machine-learning algorithms for predicting opioid overdose risk among medicare beneficiaries with opioid prescriptions. *JAMA network open* **2,** e190968–e190968 (2019).

138. Choi, E. *et al.* Generating multi-label discrete patient records using generative adversarial networks. *arXiv preprint arXiv:1703.06490* (2017).

139. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* **16,** 321–357 (2002).

140. Walonoski, J. *et al.* Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association* **25,** 230–238 (2018).

141. McLachlan, S., Dube, K. & Gallagher, T. *Using the caremap with health incidents statistics for generating the realistic synthetic electronic healthcare record* in *2016 IEEE International Conference on Healthcare Informatics (ICHI)* (2016), 439–448.

142. Wen, Q. *et al.* Time series data augmentation for deep learning: A survey. *arXiv preprint arXiv:2002.12478* (2020).

143. Esteva, A. *et al.* A guide to deep learning in healthcare. *Nature medicine* **25,** 24–29 (2019).

144. He, J. *et al.* The practical implementation of artificial intelligence technologies in medicine. *Nature medicine* **25,** 30–36 (2019).

145. Liang, Z., Zhang, G., Huang, J. X. & Hu, Q. V. *Deep learning for healthcare decision making with EMRs* in *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (2014), 556–559.

146. Miotto, R., Wang, F., Wang, S., Jiang, X. & Dudley, J. T. Deep learning for healthcare: Review, opportunities and challenges. *Briefings in bioinformatics* **19,** 1236–1246 (2018).

147. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Medical image analysis* **42,** 60–88 (2017).

148. Lehman, C. D. *et al.* Mammographic breast density assessment using deep learning: Clinical implementation. *Radiology* **290,** 52–58 (2019).

149. Rajalakshmi, R., Subashini, R., Anjana, R. M. & Mohan, V. Automated diabetic retinopathy detection in smartphone-based fundus photography using artificial intelligence. *Eye* **32,** 1138–1144 (2018).

150. Van Der Heijden, A. A. *et al.* Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System. *Acta ophthalmologica* **96,** 63–68 (2018).

151. Bluemke, D. A. Radiology in 2018: Are you working with AI or being replaced by AI? *Radiology* **287,** 365–366 (2018).

152. *Statement from FDA Commissioner Scott Gottlieb, M.D. on steps toward a new, tailored review framework for artificial intelligence-based medical devices* https://www.fda.gov/news-events/press-announcements/statement-fda-commissioner-scott-gottlieb-md-steps-toward-new-tailored-review-framework-artificial. Accessed: 2020-09-15.

153. Tomašev, N. *et al.* A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **572,** 116–119 (2019).

154. Lipton, Z. C., Kale, D. C., Elkan, C. & Wetzel, R. Learning to diagnose with LSTM recurrent neural networks. *arXiv preprint arXiv:1511.03677* (2015).

155. Panch, T., Mattie, H. & Celi, L. A. The "inconvenient truth" about AI in healthcare. *NPJ Digital Medicine* **2,** 1–3 (2019).

156. Schönberger, D. Artificial intelligence in healthcare: A critical analysis of the legal and ethical implications. *International Journal of Law and Information Technology* **27,** 171–203 (2019).

157. *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* tech. rep. (US Food and Drug Administration, Apr. 2019).

158. Davis, F. D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly,* 319–340 (1989).

159. Venkatesh, V., Morris, M. G., Davis, G. B. & Davis, F. D. User acceptance of information technology: Toward a unified view. *MIS quarterly,* 425–478 (2003).

160. Lu, J., Yu, C.-S., Liu, C. & Yao, J. E. Technology acceptance model for wireless Internet. *Internet research* (2003).

161. Pavlou, P. A. Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model. *International journal of electronic commerce* **7,** 101–134 (2003).

162. Igbaria, M., Zinatelli, N., Cragg, P. & Cavaye, A. L. Personal computing acceptance factors in small firms: A structural equation model. *MIS quarterly,* 279–305 (1997).

163. Chau, P. Y. & Hu, P. J.-H. Investigating healthcare professionals' decisions to accept telemedicine technology: An empirical test of competing theories. *Information & management* **39,** 297–311 (2002).

164. Liu, L. & Ma, Q. Perceived system performance: A test of an extended technology acceptance model. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems* **37,** 51–59 (2006).

165. Wu, J.-H., Wang, S.-C. & Lin, L.-M. Mobile computing acceptance factors in the healthcare industry: A structural equation model. *International journal of medical informatics* **76,** 66–77 (2007).

166. Yarbrough, A. K. & Smith, T. B. Technology acceptance among physicians: A new take on TAM. *Medical Care Research and Review* **64,** 650–672 (2007).

167. Holden, R. J. & Karsh, B.-T. The technology acceptance model: Its past and its future in health care. *Journal of biomedical informatics* **43,** 159–172 (2010).

168. Rho, M. J., young Choi, I. & Lee, J. Predictive factors of telemedicine service acceptance and behavioral intention of physicians. *International journal of medical informatics* **83,** 559–571 (2014).

169. Melas, C. D., Zampetakis, L. A., Dimopoulou, A. & Moustakis, V. Modeling the acceptance of clinical information systems among hospital medical staff: An extended TAM model. *Journal of biomedical informatics* **44,** 553–564 (2011).

170. Rosseel, Y. Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of statistical software* **48,** 1–36 (2012).

171. Mecklin, C. J. & Mundfrom, D. J. A Monte Carlo comparison of the Type I and Type II error rates of tests of multivariate normality. *Journal of Statistical Computation and Simulation* **75,** 93–107 (2005).

172. Savalei, V. Understanding robust corrections in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal* **21,** 149–160 (2014).

173. King, W. R. & He, J. A meta-analysis of the technology acceptance model. *Information & management* **43,** 740–755 (2006).

174. Nunnally, J. C. *Psychometric theory, 3rd ed.* (McGraw-Hill, New York, 1994).

175. Lance, C. E., Butts, M. M. & Michels, L. C. The sources of four commonly reported cutoff criteria: What did they really say? *Organizational research methods* **9,** 202–220 (2006).

176. Fornell, C. & Larcker, D. F. Evaluating structural equation models with unobservable variables and measurement error. *Journal of marketing research* **18,** 39–50 (1981).

177. Satorra, A. & Bentler, P. M. A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika* **66,** 507–514 (2001).

178. Meredith, W. Measurement invariance, factor analysis and factorial invariance. *Psychometrika* **58,** 525–543 (1993).

179. Dillon, W. R., Kumar, A. & Mulani, N. Offending estimates in covariance structure analysis: Comments on the causes of and solutions to Heywood cases. *Psychological bulletin* **101,** 126 (1987).

180. Chen, F. F. Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural equation modeling: A multidisciplinary journal* **14,** 464–504 (2007).

181. Hong, S., Malik, M. L. & Lee, M.-K. Testing configural, metric, scalar, and latent mean invariance across genders in sociotropy and autonomy using a non-Western sample. *Educational and psychological measurement* **63,** 636–654 (2003).

182. Chua, W. L. *et al.* A call for better doctor–nurse collaboration: A qualitative study of the experiences of junior doctors and nurses in escalating care for deteriorating ward patients. *Australian Critical Care* **33,** 54–61 (2020).

183. Bedoya, A. D. *et al.* Minimal impact of implemented early warning score and best practice alert for patient deterioration. *Critical care medicine* **47,** 49 (2019).

184. Alam, N. *et al.* The impact of the use of the Early Warning Score (EWS) on patient outcomes: A systematic review. *Resuscitation* **85,** 587–594 (2014).

185. Greenhalgh, T. & Abimbola, S. The NASSS framework—a synthesis of multiple theories of technology implementation. *Stud. Health Technol. Inform* **263,** 193–204 (2019).

186. Khalifa, M., Magrabi, F. & Gallego, B. Developing a framework for evidence-based grading and assessment of predictive tools for clinical decision support. *BMC medical informatics and decision making* **19,** 1–17 (2019).

187. Greenhalgh, T. & Papoutsi, C. *Studying complexity in health services research: desperately seeking an overdue paradigm shift* 2018.

188. Hodgson, T., Magrabi, F. & Coiera, E. Evaluating the usability of speech recognition to create clinical documentation using a commercial electronic health record. *International Journal of Medical Informatics* **113,** 38–42 (2018).

189. Reilly, B. M., Evans, A. T., Schaider, J. J. & Wang, Y. Triage of patients with chest pain in the emergency department: A comparative study of physicians' decisions. *The American journal of medicine* **112,** 95–103 (2002).

190. Hodgson, L. E. *et al.* The ICE-AKI study: Impact analysis of a Clinical prediction rule and Electronic AKI alert in general medical patients. *PloS one* **13,** e0200584 (2018).

191. Greenhalgh, T. *et al.* Analysing the role of complexity in explaining the fortunes of technology programmes: Empirical application of the NASSS framework. *BMC medicine* **16,** 66 (2018).

192. Greenhalgh, T. *et al.* Real-world implementation of video outpatient consultations at macro, meso, and micro levels: Mixed-method study. *Journal of medical Internet research* **20,** e150 (2018).

193. Chen, J. *et al.* The relationship between early emergency team calls and serious adverse events. *Critical care medicine* **37,** 148–153 (2009).

194. Chan, P. S., Jain, R., Nallmothu, B. K., Berg, R. A. & Sasson, C. Rapid response teams: A systematic review and meta-analysis. *Archives of internal medicine* **170,** 18–26 (2010).

195. Rihari-Thomas, J. & Kennedy, R. *Evaluation report: Critical Care Outreach (SVH Internal)* tech. rep. (2018).

196. Brinkman-Stoppelenburg, A., Rietjens, J. A. & Van der Heide, A. The effects of advance care planning on end-of-life care: A systematic review. *Palliative medicine* **28,** 1000–1025 (2014).

197. Nursing and Clinical Services Directorate. *St Vincent's Private Hospital Nursing Report (SVH Internal)* tech. rep. (2016).

198. *Emerging Systems Success Stories* Accessed: 2020-10-15. https://www.emerging.com.au/success.asp.

199. Greenhalgh, T., Robert, G., Macfarlane, F., Bate, P. & Kyriakidou, O. Diffusion of innovations in service organizations: systematic review and recommendations. *The Milbank Quarterly* **82,** 581–629 (2004).

200. Standards Australia. *Developing Standards for Artificial Intelligence: Hearing Australia's Voice* tech. rep. (2019).

201. Dawson, D and Schleiger, E and Horton, J and McLaughlin, J and Robinson, C and Quezada, G and Scowcroft, J and Hajkowicz S. *Artificial Intelligence: Australia's Ethics Framework* tech. rep. (2019).

202. Forcier, M. B., Gallois, H., Mullan, S. & Joly, Y. Integrating artificial intelligence into health care through data access: can the GDPR act as a beacon for policymakers? *Journal of Law and the Biosciences* **6,** 317 (2019).

203. Davis, S. E. *et al.* A nonparametric updating method to correct clinical prediction model drift. *Journal of the American Medical Informatics Association* **26,** 1448–1457 (2019).

Appendices

Acknowledgements of open domain image sources

Figure 1.1: This diagram has been designed using resources from Freepik.com

Figure 7.1: This interface was designed using resources from Vecteezy.com

Appendix to Chapter 2: Literature Review

Search Strings. MEDLINE: 1. Practice Patterns, Physicians'/; 2. Practice Patterns, Nurses'/; 3. Attitude of Health Personnel/; 4. 1 or 2 or 3; 5. Decision Support Techniques/; 6. Decision Support Systems, Clinical/; 7. prediction model*.mp.; 8. risk predict*.mp.; 9. clinical predict*.mp.; 10. decision rule*.mp.; 11. prediction rule*.mp.; 12. prediction tool*.mp.; 13. 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12; 14. 4 and 13

Scopus: (KEY(practice patterns, physicians) or KEY(practice patterns, nurses) or KEY(attitude of health personnel)) and (KEY(decision support techniques) or KEY(decision support systems, clinical) or TITLE-ABS(prediction model) or TITLE-ABS(risk predict) or TITLE-ABS(clinical predict) or TITLE-ABS(decision rule) or TITLE-ABS(prediction rule) or TITLE-ABS(prediction tool))

CINAHL: (practice pattern* or attitudes of health*) AND (decision support* or prediction model* or risk predict* or clinical predict* or decision rule* or prediction rule* or prediction tool*)

EMBASE: 1. clinical practice/; 2. health personnel attitude/; 3. nursing practice/; 4. 1 or 2 or 3; 5. decision support system/; 6. prediction model*.mp.; 7. risk predict*.mp.; 8. clinical predict*.mp.; 9. decision rule*.mp.; 10. prediction rule*.mp.; 11. prediction tool*.mp.; 12. 5 or 6 or 7 or 8 or 9 or 10 or 11; 13. 4 and 12

DARE: 1. MeSH Descriptor: [Practice Patterns, Physicians']; 2. MeSH Descriptor: [Practice Patterns, Nurses']; 3. MeSH Descriptor: [Attitude of Health Personnel']; 4. 1 or 2 or 3; 5. MeSH Descriptor: [Decision Support Techniques']; 6. MeSH Descriptor: [Decision Support Systems, Clinical']; 7. prediction model*; 8. risk predict*; 9. clinical predict*; 10. decision rule*; 11. prediction rule*; 12. prediction tool*; 13. 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12; 14. 4 and 13

Please note that during the piloting phase, lists of known decision rules were included in the initial searches, however produced no additional eligible results. It seems that qualitative studies are sufficiently likely to use the more generic 'rule' terminology and thus be captured by the final search strings described here.

Funding. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Summary.

- Clinician opinions and uptake of clinical prediction rules is inconsistent
- Technical capacity for prediction in healthcare is higher than what is implemented in practice
- Clinicians require actionable output from prediction rules
- Face validity is important for translation of risk scores
- Increasingly, clinicians desire fully integrated risk-score calculators
- Utility, credibility and usability are necessary for acceptance of CPR

Appendix to Chapter 7: Proposed application and use-cases

Text presented to survey respondents

Background

This application has been developed in order to assist the work of the Critical Care Outreach team by presenting assessments of patient deterioration risk over time.

You will be presented with a system mock-up and text describing the intended use-cases.

The purpose of the questions that follow is to determine the potential usefulness, ease of use and attitudes towards such an application. In responding to these questions, please assume that the risk assessments presented have been validated to the following level of accuracy for in-patient deaths and unplanned ICU admissions.

Application Overview

Data sources. The watch-list is envisaged as an application that can be run automatically, from data that is available in the clinical record in real time, i.e. pathology orders and results, medication administration records, surgical theatre bookings, ward movements and administrative data.

This list will update regularly as new data is available, producing a risk assessment for each patient.

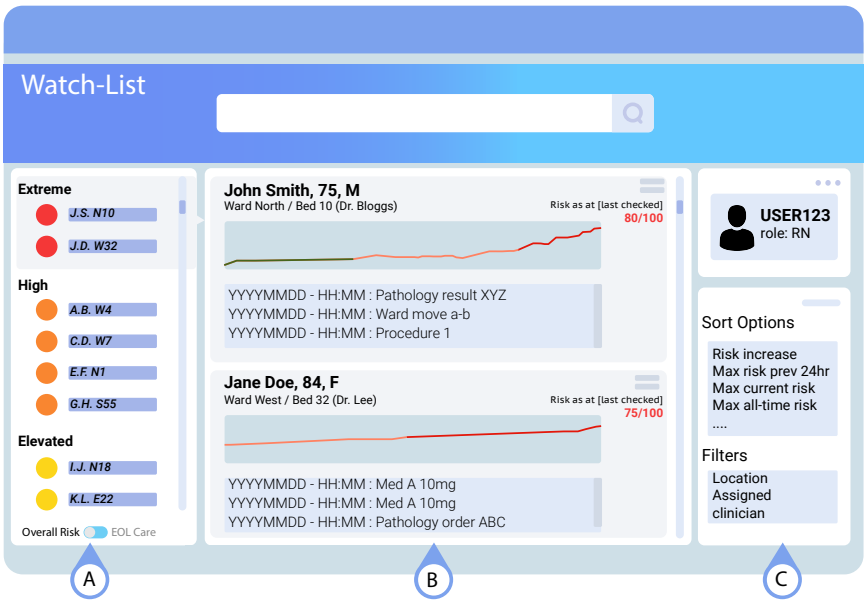
Prediction of in-patient death. The prototype system is able to predict death within 24 hours with an accuracy that compares favourably to the commonly used early warning score NEWS (AUROC 0.93 vs 0.89).

Prediction of unplanned ICU admission. This system is able to predict unplanned ICU admission within 24 hours (defined here as admissions to ICU from any source other than direct from surgical theatres) with an AUROC of 0.77.

This accuracy and clinical applicability will be tested and reported upon separately. Your responses should focus on the usability of the application itself, assuming the accuracy described above.

Note that the prototype system does not make use of vital signs data or clinical notes, and accuracy is expected to increase significantly (particularly in regards to unplanned ICU admission) when this data is made available.

Application Usage



(A) Summary panel - see high-level view of cases allowing an immediate overview of the risk profile of the hospital

(B) Detail view - see predictions for individual cases over time to understand trends, and highlight events co-occurring with risk changes over time

(C) Control panel - sort and summarise by risk category, location, assigned clinician or other relevant parameters

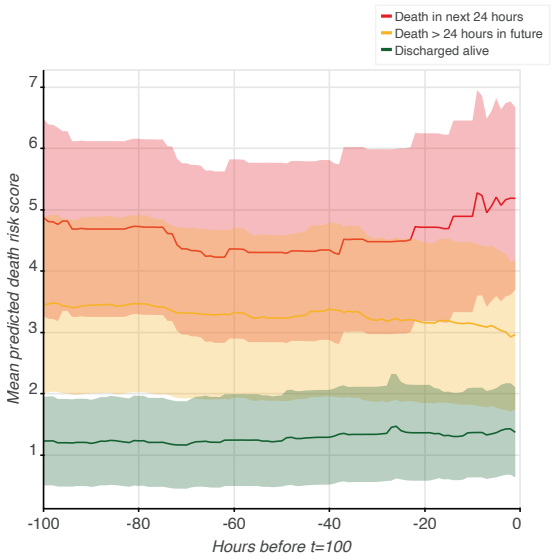


FIGURE 8.2. Risk changes up to prediction time

It is not possible to give a real example for the detail panel due to privacy restrictions on the source data, so here we present the average risk trend for the first 100 hours of admission for admissions > 100 hours in length in the test set (grouped by actual outcome).

Application Use Cases

A non-exhaustive list of expected use-cases includes:

1. Using the control panel to sort by location and risk, the critical care outreach nurse begins their shift and is able to rapidly assess where to visit first.

2. A medical officer is handing over at the end of their shift. They filter by assigned clinician and/or location and use the output as a priority list to work through for the handover conversation.

3. A rapid-response is called on a patient. The responding medical officer is not familiar with this patient, however they are able to use the view of risk over time, together with the listed events summarised below, to help them get up to speed on this case more rapidly.

4. Clinicians can manually override a high-risk prediction in the instance that the patient in question is receiving end-of-life care.

5. Post-hoc analysis of outlier predictions can be used to classify and identify unexpected deterioration in order to drive institution-level policies.

Appendix to Chapter 7: Measures

Question	Responses
Age	Under 30 years of age
	30-39 years
	40-49 years
	50-59 years
	60+ years
	Prefer not to respond
Gender	Male
	Female
	Prefer not to respond
	Other
Level of assignment to clinical patient deterioration related tasks (including outreach and emergency response) in a typical work-week	None
	Occasional or cover assignment only
	Part-time but regular assignment
	Majority of working week
	Full-time or dedicated assignment
Role	Nursing
	Medical
	Administrative
	Other
Years of experience post-graduation	0-5
	6-10
	11-15
	16+

TABLE 8.2. Demography Measures

Measure	**Questions**	
ATT	ATT1	Providing risk assessments in such a format to prioritise Critical Care Outreach work is a good idea
	ATT2	I am positive toward the idea of assessing patient deterioration risk in such a fashion
BI	BI1	Were this system offered in my practice, I believe I would be a frequent user
	BI2	Were this system available, but not offered in my practice, I would advocate for its use
PEOU	PEOU1	It would be easy for me to use this application in my clinical practice
	PEOU2	It would be easy to become skillful at using this application
	PEOU3	It is easy for me to understand the patient deterioration risk assessments that are presented by this application
	PEOU4	I feel confident that I could find the information I wanted in this application
PU	PU1	Death and unplanned ICU admission risk assessments presented in this manner would enhance the effectiveness of the Critical Care Outreach team
	PU2	This application would make it easier to prioritise Critical Care Outreach resources
	PU3	This application would be useful to the Critical Care Outreach team
Free-text		Do you have any additional comments that you would like to make regarding the clinical utility of the proposed tool?

TABLE 8.3. TAM Measures