



The
University
Of
Sheffield.

Predicting an avoidable conveyance to the Emergency
Department for ambulance patients on scene

Jamie Miles

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

The University of Sheffield
Faculty of Medicine, Dentistry and Health
School of Health and Related Research

Submitted on the 25th July 2022

Abstract

One of the main problems currently facing the delivery of safe and effective emergency care is excess demand, which causes congestion at different time points in a patient's journey. The modern case-mix of prehospital patients is broad and complex, diverging from the traditional 'time critical accident and emergency' patients. It now includes many low-acuity patients and those with social care and mental health needs. In the ambulance service, transport decisions are the hardest to make and paramedics decide to take more patients to the ED than would have a clinical benefit. As such, this thesis asks the following research questions:

In adult patients attending the ED by ambulance, can prehospital information predict an avoidable attendance?

Can the model derived from the primary outcome be spatially transported?

A linked dataset of 101,522 ambulance service and ED data from the whole of Yorkshire between July 2019 and February 2020 was used as the sample for this study. A machine learning method known as XGBoost was applied to the data in a novel way called Internal-External Cross Validation (IECV) to build the model. The results showed great discrimination with a C-statistic of 0.81 (95%CI 0.79-0.83) and excellent calibration with an O:E ratio was 0.995 (95% CI 0.97 – 1.03), with the most important variables being a patient's mobility, their physiological observations and clinical impression with psychiatric problems, allergic reactions, cardiac chest pain, head injury, non-traumatic back pain, and minor cuts and bruising being the most important.

This thesis has successfully developed a decision-support model that can be transformed into a tool that could help paramedics make better transport decisions on scene, known as the SINEPOST model. It is accurate, and spatially validated across multiple geographies including rural, urban, and coastal. It is a fair algorithm that does not discriminate new patients based on their age, gender, ethnicity, or decile of deprivation. It can be embedded into an electronic Patient Care Record system and automatically calculate the probability that a patient will have an avoidable attendance at the ED, if they were transported.

Acknowledgements

This study was funded by Health Education England and the National Institute of Health Research (HEE/NIHR ICA Programme Clinical Doctoral Research Fellowship, Mr. Jamie Miles, ICA-CDRF-2018-04- ST2-044).

I would like to start by thanking Professor Suzanne Mason for being such a supportive and amazing primary supervisor. Becoming a clinical academic has its challenges, but I felt these were all achievable to overcome thanks to Sue's support both at the University and in the Emergency Department. I would also like to thank Dr. Richard Jacques, Janette Turner, and Professor Julia Williams for their roles in supervising me on this project. Richard was brilliant at providing expertise on the design and the interpretation of the project, particularly the more technical aspects. Janette was able to support me in ensuring the project was placed within the wider healthcare context and was great at checking I had explained concepts in understandable ways. Julia's kind support and mentorship helped me with the evolution into a clinical academic and made sure the project was always centred around the patient benefit.

This project could not have gone ahead without the support and sponsorship of Yorkshire Ambulance Service NHS Trust. Within the trust, I wish to extend my special thanks to Dr. Fiona Bell for her support and for helping me with the contractual, financial, and organisational management of the project.

For parts of this project, I relied on the knowledge and skills of data experts to extract and link the analysis datasets. I wish to show my appreciation to Tony Stone, Richard Campbell, and Richard Pilbery for undertaking these actions, and for providing me with help in the data applications to NHS Digital.

Undertaking a PhD is a stressful experience at times, and the one person who was always there for me, supporting me in the lows, and celebrating with me in the highs was my fiancée Kathryn who I cannot thank enough. I would also like to thank our three children Mia, Ben, and Rory for being very patient with me. Whilst my head was down, they gave me a reason to look up.

Table of contents

| | |
|--|------------|
| Table of figures | 4 |
| Table of tables | 5 |
| Declarations, Contributions, and permissions | 6 |
| Glossary of important acronyms and concepts | 9 |
| 1. Introduction | 14 |
| 1.1 Introduction to the thesis | 15 |
| 1.2 Thesis structure | 19 |
| 1.3 Conclusion | 20 |
| 2. Background | 22 |
| 2.1 Introduction | 23 |
| 2.2 Global picture of emergency care | 23 |
| 2.3 Demand for prehospital care | 26 |
| 2.4 Causes and consequences of prehospital demand | 28 |
| 2.5 Ambulance offload delay and ramping | 31 |
| 2.6 Low-acuity navigation into emergency care | 37 |
| 2.7 The quantity of avoidable emergency care contacts | 43 |
| 2.8 Strategies to optimise care | 46 |
| 2.9 Decision support for paramedics navigating the prehospital case-mix | 56 |
| 2.10 Conclusion | 64 |
| 3. Systematic review of machine learning risk prediction models to triage emergency care patients | 66 |
| 3.1 Introduction | 67 |
| 3.2 Purpose of the review | 67 |
| 3.3 Published manuscript | 68 |
| 3.4 Supporting information | 80 |
| 3.5 Conclusion | 83 |
| 4. Aims and objectives of this thesis | 84 |
| 4.1 Introduction | 85 |
| 4.2 Primary aims and objectives | 86 |
| 4.3 Secondary aims and objectives | 88 |
| 4.4 Conclusion | 89 |
| 5. Theoretical considerations and algorithm selection | 90 |
| 5.1 Introduction | 91 |
| 5.2 Theory of knowledge | 91 |
| 5.3 Theoretical considerations | 96 |
| 5.4 Conclusion | 110 |
| 6. Protocol | 111 |
| 6.1 Introduction | 112 |
| 7. Protocol expansion | 122 |
| 7.1 Introduction | 123 |
| 7.2 Study design and setting | 123 |
| 7.4 Dataset creation | 127 |

| | |
|--|------------|
| 7.5 Outcome variable | 130 |
| 7.6 Candidate predictors | 140 |
| 7.7 Sample size | 154 |
| 7.8 Internal-external cross-validation (IECV) | 155 |
| 7.9 Evaluation of model performance | 157 |
| 7.10 Further data preparation | 161 |
| 7.11 Protocol deviations | 167 |
| 7.12 Ethical considerations | 168 |
| 7.13 Patient and public involvement | 170 |
| 8. Results (presented as a manuscript) | 172 |
| 8.1 Introduction | 173 |
| 8.2 Methods | 175 |
| 8.3 Results | 179 |
| 8.4 Discussion | 188 |
| 8.5 Conclusion | 192 |
| 9. Further results | 193 |
| 9.1 Introduction | 194 |
| 9.2 Data modelling | 194 |
| 9.3 Cluster results | 196 |
| 9.4 Fair Machine Learning analysis | 205 |
| 9.5 Evaluation on two different cohorts | 209 |
| 9.6 Conclusion | 210 |
| 10. Discussion | 211 |
| 10.1 Introduction | 212 |
| 10.2 Principal findings of the study | 212 |
| 10.3 Limitations | 226 |
| 10.4 Future research | 230 |
| 10.5 Conclusion | 231 |
| 11. Personal reflection | 232 |
| 11.1 Personal reflection | 233 |
| 12. Conclusion | 235 |
| 12.1 Introduction | 236 |
| 12.2 Summary of findings | 237 |
| 12.3 Contributions to knowledge | 238 |
| 12.4 Recommendations for policy | 238 |
| 12.5 Recommendations for future research | 239 |
| 12.6 Conclusion | 239 |
| 13. References | 240 |
| 14. Appendices | 260 |
| Appendix A: Licenses and permissions | 261 |
| Appendix B: International triage scales | 273 |
| Appendix C: Supplementary information for the systematic review | 275 |
| Appendix D: Technical elaboration on algorithm methods | 281 |
| Appendix E: List of Emergency Departments included in this study | 295 |
| Appendix F: NEWS Score | 297 |

| | |
|---|-----|
| Appendix G: HRA, REC and CAG Approval | 298 |
| Appendix H: Included variables in the model | 317 |
| Appendix I: Hyperparameter values per cluster | 325 |
| Appendix J: ROC and calibration curves for the IECVmodels | 326 |

Table of figures

| | |
|--|-----|
| Figure 1: The on-scene stage of the ECSF | 24 |
| Figure 2: The transport stage of the ECSF | 25 |
| Figure 3: Facility stage of the ECSF | 25 |
| Figure 4: Ambulance demand in England 2018-2021..... | 27 |
| Figure 5: Proportion of patients not conveyed to hospital after a face-to-face ambulance assessment between 2017 and 2021..... | 49 |
| Figure 6: Demonstration of overfitting (reproduced from Badillo et al. with permission) ¹²⁶ | 99 |
| Figure 7: NEWS score and its relationship with risk of hospital admission from Cameron et al. ¹³³ | 102 |
| Figure 8: Neural network example | 103 |
| Figure 9: Example of a tree-based model using Steyerberg AAA data (n=238) ¹⁹⁶ | 105 |
| Figure 10: Data flow diagram for dataset creation | 129 |
| Figure 11: Confusion Matrix..... | 157 |
| Figure 12: Frequency distribution by confusion matrix category | 168 |
| Figure 13: Full model calibration plot | 183 |
| Figure 14: ROC curve of the full model..... | 184 |
| Figure 15: Meta-analysis of cluster discrimination | 185 |
| Figure 16: Top 20 variables with the highest relative contribution to the full model (gain) | 187 |
| Figure 17: Top 20 variables used in the full model by frequency..... | 187 |
| Figure 18: Top 20 variables with the greatest number of instances when splitting (cover) | 188 |
| Figure 19: The fourth tree in the full XGBoost model | 196 |
| Figure 20: ROC curves grouped by geographical area..... | 204 |
| Figure 21: Fair machine learning: Gender | 205 |
| Figure 22: Fair machine learning: Age..... | 206 |
| Figure 23: Fair machine learning: Indices of Deprivation | 207 |
| Figure 24: Fair machine learning: Ethnicity..... | 208 |
| Figure 25: Probability densities of conveyed vs non conveyed | 209 |

Table of tables

| | |
|---|-----|
| Table 1: Summary of repeated access to healthcare following EMS discharge in the general population from Ebben et al. ⁶⁴ | 50 |
| Table 2: Barriers and facilitators to implementing EARP models in primary care, reproduced from Snooks et al. ¹⁰¹ | 62 |
| Table 3: NHS Definition of low acuity attendance | 132 |
| Table 4: Investigation codes | 134 |
| Table 5: Treatment codes..... | 135 |
| Table 6: Discharge status codes | 136 |
| Table 7: Full definition for this study | 137 |
| Table 8: Demographic candidate variables | 146 |
| Table 9: Social candidate variables..... | 147 |
| Table 10: Clinical candidate variables | 148 |
| Table 11: Interventional candidate variables | 153 |
| Table 12: Characteristics of participants | 179 |
| Table 13: Model performance measures | 182 |
| Table 14: Hyperparameter grid | 194 |

Declarations, Contributions, and permissions

I, the author, confirm that the Thesis is my own work. I am aware of the University's Guidance on the Use of Unfair Means (www.sheffield.ac.uk/ssid/unfair-means). This work has not been previously presented for an award at this, or any other, university.

This study was funded by Health Education England and the National Institute of Health Research. The funders have not inputted into the collection, analysis, or interpretation of data in writing this manuscript. This report is independent research supported by Health Education England and the National Institute of Health Research (HEE/NIHR ICA Programme Clinical Doctoral Research Fellowship, Mr. Jamie Miles, ICA-CDRF-2018-04- ST2-044). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care.

Chapter 1: *Figures 1-3*

The World Health Organisation (WHO) have granted permission to alter their infographic for this thesis. The original image and license can be found in appendix A. This license applies to figures 1 – 3.¹

Chapter 2: “Using machine-learning risk prediction models to triage the acuity of undifferentiated patients entering the emergency care system: a systematic review”

The development of the search strategy, searching, screening, analysis, figure production and manuscript writing was all completed by me (Jamie Miles). The co-authors are PhD supervisors for this thesis and contributed to the concept and

steer of the systematic review. In addition, Janette Turner performed secondary screening, which is required for scientific integrity.

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution, and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third-party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0>. In accordance with the above terms, no changes were made to the publication in reproduction for this thesis.

Chapter 5: *Figure 6*

This image was taken directly from Badillo et al. (Badillo S, Banfai B, Birzele F, Davydov II, Hutchinson L, Kam-Thong T, et al. An Introduction to Machine Learning. Clin Pharmacol Ther. 2020;107(4):871–85.). The image was not changed in any way, and was used under Section 32 of the Copyright, Designs and Patents Act 1988, which allows for the copying of a work for the purpose of illustrating a teaching point.² This is not limited to teaching within an educational establishment. The image was used in the context of explaining the bias-variance trade off and model overfitting. This is the same context the authors used in their original paper. The image was used with permission, which can be found in appendix A2.

Chapter 6: “The Safety INdEx of Prehospital On Scene Triage (SINEPOST) study: the development and validation of a risk prediction model to support ambulance clinical transport decisions on-scene—a protocol”

JM is the study lead and drafted the manuscript. SM is the lead supervisor for this study, contributed to the development of the research question and its overall design. SM has contributed to the drafting of this manuscript. RJ is a co-supervisor and informed the statistical analysis plan in this manuscript. JT is a co-supervisor and informed the clinical importance of the study in this manuscript.

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution, and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third-party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0>. In accordance with the above terms, no changes were made to the publication in reproduction for this thesis.

Pagination notice

This thesis contains journal publications that have been inserted in their publication format. As a result, they have their own page numbers and referencing. To avoid confusion, the thesis page numbers appear central at the bottom of each page and the reference list has been included as per the publication.

Glossary of important acronyms and concepts

| Term | Abbreviation | Definition |
|--|--------------|--|
| 999 | - | The phone number in the UK that is dialled for emergencies. |
| Ambulance ramping | - | When patients are conveyed to the ED and held in a queue until the department's clinical staff can receive the patient. |
| Area Under the Receiver Operating Characteristic Curve | AUC | A graphical representation of discrimination which plots sensitivity and 1-specificity. A perfect AUC would be 1, and an AUC of 0.5 means the risk prediction model is no better than chance. See chapter 7, section 7.9.2 |
| Association of Ambulance Chief Executives | AACE | AACE provide ambulance services with a central organisation that supports, coordinates, and implements nationally agreed policy. It also provides the public and other stakeholders with a central resource of information about NHS ambulance services. |
| Australian Triage Scale | ATS | A five-level scale to assess a patient's clinical acuity, from ATS 1 (immediate need for treatment) to ATS 5 (Treatment within 120 minutes). See appendix B. |
| Avoidable ED attendance | - | A first attendance to the ED with some recorded treatments or investigations all of which may have reasonably been provided in a non-emergency care setting, followed by discharge home or to GP care. |
| Avoidable conveyance | | As above but transported by the Ambulance Service. |
| C-statistic | - | See AUC. |
| Calibration | - | A performance measure of a risk prediction model to ensure that the model provides a good description of system behaviour. It is commonly measured using the O:E ratio and Spiegelhalter's Z-test. See chapter 7, section 7.9.1. |

| | | |
|--|------|--|
| Candidate variables | - | Also known as input variables and are used as predictors in a model. |
| Commissioning Data Set | CDS | The former routine dataset that was replaced by the ECDS (see ECDS). |
| Computerised Clinical Decision Support | CCDS | A form of decision support that can be integrated into digital technology or is derived by computer technology. |
| Confidence interval | CI | A range of values so defined that there is a specified probability (usually 95%) that the true value of a parameter lies within it. |
| COVID-19 | - | A disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus. Caused a global pandemic in 2019-2022. |
| Discrimination | - | A performance measure of a risk prediction model that evaluates whether it can differentiate between a random instance with the event, and one without. Commonly measured using the C-statistic. See chapter 7, section 7.9.2. |
| Double Crewed Ambulance | DCA | The most common type of ambulance that often consists of 2 staff members – a paramedic and an emergency care assistant. |
| Early Warning Score | EWS | A generic term for a clinical scoring tool to decide how poorly a patient is, and to detect any deterioration in a patient's condition. |
| Electronic Health Records | EHR | A digital health record used by clinicians to record all aspects of care about their patients. |
| electronic Patient Care Record | ePCR | Synonymous with EHR. See above. |
| Emergency Care Data Set | ECDS | A specific subset of SNOMED-CT codes that describe all activity found within the ED. Is coded in the EHR. |
| Emergency Care System Framework | ECSF | Captures essential emergency care functions at the scene of injury or illness, during transport, and through to the emergency unit. |
| Emergency Department | ED | The care setting found within a hospital that treats medical and traumatic emergencies. |

| | | |
|------------------------------------|------|---|
| Emergency Medical Service | EMS | Synonymous with ambulance service. |
| Emergency Severity Index | ESI | A five-level scale to assess a patient's clinical acuity, from ESI 1 (immediate need for treatment) to ESI 5 (Non-urgent need for treatment). See appendix B. |
| Fair machine learning | - | Consciously attempting to reduce or eliminate bias in machine learning models. |
| False Negative | FN | see chapter 7, section 7.8 |
| False Positive | FP | see chapter 7, section 7.8 |
| Hazardous Area Response Team | HART | A specialist ambulance response that can rescue patients from difficult to access and/or dangerous areas. |
| Hyperparameter | - | A prespecified rule placed on an algorithm prior to model development. It dictates how the algorithm learns. |
| Incident Rate Ratios | IRR | Similar to an odds ratio, an incidence rate ratio compares the incident rate between two different groups. |
| Internal-External Cross Validation | IECV | See chapter 7, section 7.8. For an illustration on IECV see chapter 6. |
| Low acuity | - | Patients with urgent and complex care needs as opposed to life threatening. |
| Manchester Triage System | MTS | An acuity triage system developed in the UK. |
| National Early Warning Score | NEWS | The current EWS used in the UK and the latest version can be found in appendix F. |
| Negative Predictive Value | NPV | See chapter 7, section 7.8. |
| NHS111 | 111 | A free-to-call single non-emergency number medical helpline operating in England, Scotland, and parts of Wales. |
| non-conveyance | | When an ambulance clinician assesses a patient and decides they do not need to be transported anywhere. |
| Observed vs Expected ratio | O:E | see chapter 7, section 7.9.1. |
| Odds ratio | OR | The ratio of the odds of A in the presence of B and the odds of A in the absence of B. |
| Offload delay | | A consequence of ambulance ramping which results in long delays in patient handover between the ambulance and the ED. |
| one-hot encoding | | Transforming a categorical variable with n categories into n binary variables. |

| | | |
|--|-----------|---|
| overfitting | | A model is fit well to the data, but not representative of the target system. |
| Positive Predictive Value | PPV | see chapter 7, section 7.8. |
| Prehospital | - | The section of clinical care that takes place within an ambulance service. Urgent and emergency care before the ED. |
| Patient and Public Involvement | PPI | See chapter 7, section 7.13. |
| Rapid response Vehicle | RRV | Contrary to a DCA, this is a fast car with a solo responder. |
| Realist synthesis | - | Aims to not just appraise the evidence, but also account for the context as well as the outcome. |
| Recursive Feature Elimination | RFE | Removing variables from a dataset that have a weak/ no association with the outcome. |
| Restricted grid search | - | See chapter 7, section 7.10.4. |
| Royal College of Emergency Medicine | RCEM | Represents clinicians who practice emergency medicine in the UK. |
| Sensitivity | - | see chapter 7, section 7.8. |
| Simulated transportability | - | Uses the same dataset as the training data to test whether the model could be developed under different geographical circumstances. |
| Specificity | - | see chapter 7, section 7.8. |
| Spiegelhalter's Z-test | - | see chapter 7, section 7.9.1. |
| Systemised Nomenclature Of Medicine Clinical Terms | SNOMED CT | Systematically organized computer-processable collection of medical terms. Used to code routine datasets such as ECDS. |
| True Negative | TN | see chapter 7, section 7.8 |
| True Positive | TP | see chapter 7, section 7.8 |
| Type 1 Emergency Department | | A consultant led 24-hour ED with full resuscitation facilities and designated accommodation for the reception of accident and emergency patients. |
| Urgent Treatment Centre | UTC | A care setting like an ED but can provide specialist urgent care to mid- and low-acuity patients. |
| XGBoost | - | An extreme gradient boosted decision tree algorithm. |
| Yorkshire Ambulance Service | YAS | The ambulance service that provides care for patients throughout the English county of Yorkshire. |

Chapter 1

The Introduction

1.1 Introduction to the thesis

Emergency care plays a crucial role in reducing the global burden of disease. Not only does it directly affect the mortality and morbidity of the populations it serves, but it is also a human right.³ There are some basic principles which underpin delivering an emergency care system and these form a skeletal model of emergency care from call, to intervention.⁴ However, it is important to acknowledge that not all countries provide this basic model. Higher-income countries have more complex models than those of lower or lower-middle income countries. There are different components in the emergency care system, such as the ambulance service, Emergency Department (ED) and primary care. These are all affected by demand, but this thesis focuses on the ambulance service perspective.

One of the main problems with excess demand is ambulances having to wait to handover a patient once they have transported them to the Emergency Department (ED). This is the concept of ‘offload delay’ and results in many patients being held queuing in the ambulance until the crews can offload the patient into the ED (also known as “ambulance ramping”). More details of offload delay can be found in chapter 2, section 2.5. Offload delay can cause problems downstream for the ED, as they may divert resources away from waiting room patients to prioritise queuing ambulances if they perceive patients arriving by ambulance are higher acuity. It also causes upstream problems, as the ambulance cannot respond to another potentially ill patient waiting in the community.

Studies have shown that not all patients transported to hospital by ambulance require such a high level of care, and these studies will be discussed in chapter 2, section 2.6. The modern case-mix of prehospital patients is broad and complex, which diverges from a traditional model of ‘time critical accident and emergency’ patients and now includes many low-acuity patients and those with social care and mental health needs.

The uniqueness of the prehospital environment is that it is a remote and portable healthcare setting. This places pressure on paramedics to have judicious decision-making skills when dealing with the clinical diversity of patients that call for an ambulance. However, this decision making is not always accurate. Chapter 2, section 2.9 details how transport decisions are the hardest to make and paramedics decide to take more patients to the ED than are required.^{5,6} Decision support systems (such as the paramedic pathfinder) have proven no better than the humans using them.⁷⁻⁹ Computer-based decision support systems have shown promise though.

A systematic review in chapter 3 has found that it is possible to build risk prediction models of patient acuity based on statistical and in-silico modelling of patient data. Their common limitations are that they have either focussed on predicting high acuity patients or are not based in the prehospital setting, which makes their applicability to reducing avoidable conveyances limited. The opportunity in this thesis is to examine the possibility of developing a model which can predict an avoidable conveyance whilst paramedics are still on scene with their patients. This is achieved by using prehospital information to predict an ED outcome, which is a near unique situation whereby the prediction is offering novel information to clinicians on scene. This study is restricted to just adult patients as there were ambulance policies around conveying children that could have confounded the modelling. Therefore, the research questions that are posed in this thesis and expanded on later in chapter 4, section 4.1 and 4.2 are:

In adult patients attending the ED by ambulance, can prehospital information predict an avoidable attendance?

Can the model derived from the primary outcome be spatially transported?

Transportability is being able to apply the model to different types of population (for example a different geography) and it still showing acceptable levels of

performance. Simulated transportability is using the existing population the model was derived on and simulating whether it could be transported. It is an important research question to ask when developing a new model using a large dataset. Poor transportability is a barrier to implementation that can be easily overcome if it is considered in the initial model development.

A linked dataset of 101,522 ambulance service and ED patient episodes from the whole of Yorkshire between July 2019 and February 2020 was used as the sample for this study. Each instance had all prehospital ambulance care record data, which was created by the clinician on scene. It also had the outcome of whether they had an avoidable attendance at ED, created using a modified data driven definition adopted by NHS Digital and elaborated on in chapter 7, section 7.5. A machine learning method known as XGBoost was applied to the data to build the model. This was evaluated for its performance by calculating statistics for calibration (O:E ratio and Spiegelhalter's Z-test) and discrimination (C-statistic). More information on the algorithm and evaluation can be found in chapter 7, sections 7.8-7.10. To answer the second research question, the data was split several ways. Each iteration had a test set which was all the data from a single ED. The rest of the data was used to train the model using the exact same procedures as the full model. There were 17 EDs in the study, therefore there were 17 models built. Each of these models were then meta-analysed and the performance measures were used to update the full model. The reason for this was the original model would have been optimistic as it was evaluated on the same data it was trained on. The meta-analysed results are a more realistic truth on model performance.

The *raison d'être* of this study was to match patient clinical presentations with the most appropriate care setting for their need. This was to ensure that patients were accessing the right care setting for them as expeditiously as possible. The novel contribution of this thesis to the scientific field includes using a machine learning algorithm to develop a new clinical prediction model that has been

appropriately validated and assessed as evidenced in chapter 8. It also brings an example of how machine learning can be fair and reduce the risk of bias when used for clinical decision making. Fairness is a concept that will be explained further in chapter 7, section 7.6.1, but in essence is ensuring an algorithm does not discriminate against protected characteristics.

This thesis has successfully developed a decision-support model that can potentially help paramedics make better transport decisions on scene, known as the SINEPOST model. It has good calibration and discrimination and could be described as an accurate model. These concepts are defined and elaborated further in chapter 7, section 7.9.1 and 7.9.2 respectively. The model is also spatially validated across multiple geographies including rural, urban, and coastal. It is a fair algorithm that does not discriminate new patients based on their age, gender, ethnicity, or decile of deprivation. This can be seen in chapter 9, section 9.3. The pragmatic research design of using ambulance ePCR data as candidate variables means it could easily be embedded into an electronic Patient Care Record system and automatically calculate the probability that a patient will have an avoidable attendance at the ED, if they were transported.

The results in this study could lead to important and original advancements across the urgent and emergency care system. In prehospital care, the SINEPOST model could support paramedic decision making as to whether their patient requires transportation. More discernible decisions by paramedics could then potentially reduce ambulances queueing at the Emergency Department, waiting to hand their patients over. This has a subsequent effect on the demands placed on those in emergency medicine. In General Practice, the model will improve patient's access to primary care, by identifying primary care needs within ambulance service contacts.

1.2 Thesis structure

This is the first chapter in this thesis, the remaining chapters are arranged in the following way:

The Background (chapters 2, 3 and 4)

The second chapter introduces the emergency care system under investigation in this study. It discusses the causes and consequences of ambulance offload delay and appraises strategies that have been tested to reducing it. It extends the discussion to paramedic decision making and how this has been helped with the use of decision support tools. In chapter 3, the systematic review aims to examine all available evidence to ascertain whether prehospital prediction models of acuity have already been developed. This would have led to an external validation of an existing model, as opposed to the creation of a new one. The methodologies used to develop models included in the review are examined for their feasibility and success for this study. The systematic review was published in BMC Diagnostic and Prognostic Research. Chapter 4 outlines the research questions, aims and objectives of the study.

The Methods (chapters 5, 6, 7)

In chapter 5 there is a descriptive outline of theoretical considerations that act as the foundations for predictive modelling. This is then elaborated into a section on algorithm selection, where the new information gained from the systematic review is synthesised into selecting the best algorithm to solve the problem identified in chapter 4. Chapter 6 presents the protocol, which was published in BMC Diagnostic and Prognostic Research in 2020. The manuscript details in a succinct fashion the methods that are being used to answer the aims and objectives. Due to the concise nature of the manuscript, chapter 7 offers an expansion of the methods, where decisions are justified, and more detail is given.

The chapter includes deviations from the protocol since its publication, as well as ethical considerations and public and patient involvement.

The Results (chapters 8 and 9)

The results begin with chapter 8, which is a presentation of the main findings in publication format. This has been written for a clinical audience. Like chapter 6, due to the conciseness of the manuscript, chapter 9 offers an expansion of the results with a more statistics/ computer science lens. This includes the individual performance of each cluster analysis and the fair machine learning analysis (which is conceptually defined in chapter 7, section 7.6.1).

The Discussion (chapters 10, 11, and 12)

The discussion chapter starts with a summary of whether the results answered the research questions identified in the background chapters. It then expands to critically place the new knowledge within the context of existing knowledge and clinical practice. Limitations are acknowledged and next steps are proposed. There is a personal reflection in chapter 11, before the thesis draws to a close with a conclusion in chapter 12.

1.3 Conclusion

This chapter has provided a brief overview of what is contained in this thesis, as well as highlighting the key contributions to knowledge. It has introduced the format and structure of the thesis, to help the reader navigate different sections. The next chapter provides the background information which will frame the research question but also illustrate the landscape for which the new evidence will be situated.

Chapter 2

The Background

2.1 Introduction

The aim of this chapter is to illustrate how the emergency care system operates both globally and in the UK. One of the most important challenges that is faced in this system is the demand of patients that are accessing the high acuity level of care provided by the ambulance services and ED. This comes at a direct contrast to the case mix of prehospital and emergency patients and section 2.4 of this chapter will examine the causes and consequences of trying to meet excess demand. The chapter frames the problem, but also scrutinises solutions at the macro, meso, and micro level. Policies that have aimed to both meet and reduce demand will be appraised in section 2.8, but the chapter will focus further on a specific cause of demand. The transport of high volumes of patients, some of which are low acuity contribute to ambulances queuing at the ED. Section 2.9 takes a closer look at paramedic decision making around transportation as well as exploring what decision support systems are successful in helping paramedics make this decision.

2.2 Global picture of emergency care

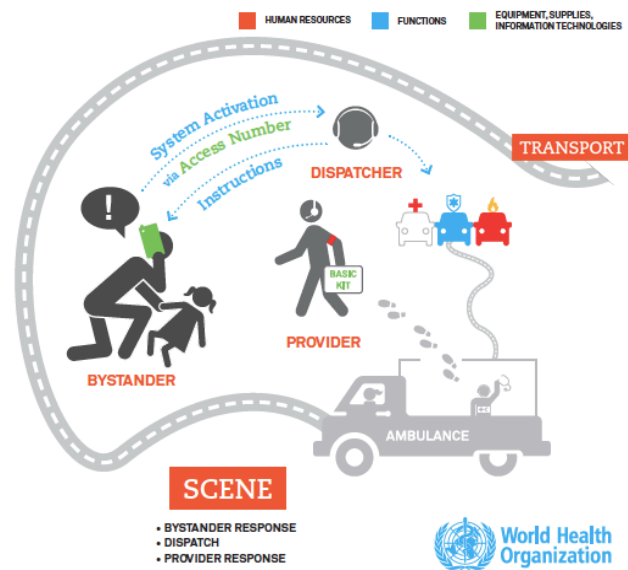
Emergency care focuses on reducing preventable mortality (death), morbidity (suffering) and disability from time-sensitive disease processes.¹⁰ The World Health Assembly used this definition in the 2007 resolution 60.22, which sets out a framework for all countries to develop an effective emergency care system. Emergency care spans across medical disciplines and has a crucial opportunity in reducing the global burden of disease.⁴ However, the current state of emergency care varies dramatically across the globe. The World Health Assemblies resolution 60.22 provided detail on the most basic system that every country should aim to invest in known as the Emergency Care System Framework (ECSF). Sections 2.2.1-2.2.3 below summarise what the WHO expects should happen in a basic system, and one that Lower-Middle Income Countries (LMICs) should strive to achieve. Within the system, there are three stages.

2.2.1 The 'On Scene' Stage

The first stage in the ECSF is the 'on scene' stage. This occurs at the site of the emergency, such as at someone's home or in the street (illustrated in figure 1).⁴ When a person has an accident or a medical emergency, they will call a dedicated emergency phone number (for example in the UK, a person would call 999). The person will talk to the dispatcher who can then send an ambulance with a healthcare provider to the scene.

Figure 1: The on-scene stage of the ECSF

The dispatcher also gives instructions to the person. The healthcare provider will arrive on scene in their ambulance with a driver. They may only have basic equipment, but they fulfil the function of transporting the patient to the nearest healthcare facility.



2.2.2 The 'Transportation' stage

The second stage is transportation, which is when the person is in the ambulance and the healthcare provider is with them. As the person is being cared for by a healthcare provider, they will be referred to as a patient from now on. The healthcare provider will monitor and care for the patient whilst the driver of the ambulance transports the patient to the nearest facility. There is often a communication channel with the original dispatch resource and the receiving facility. The second stage ends at the 'handover gate'. This is where an

⁴ Figures 1-3 have been modified from the World Health Organisation with permission (license found in appendix A, section A1). World Health Organization. WHO Emergency Care Systems Framework. 2015 [cited 2022 Apr 26]; Available from: <https://www.who.int/publications/i/item/who-emergency-care-system-framework>

ambulance has arrived at the receiving facility and moves the patient from the ambulance onto a hospital bed. The ambulance healthcare provider will also give a clinical handover to the receiving provider, detailing what has happened.

2.2.3 The 'Healthcare Facility' stage

The third stage is the healthcare facility stage as illustrated by figure 3. There are more expertise

and resources at the facility stage, and this is often the definitive care for the patient. When patients enter this system from the ambulance service, they will be further triaged for their acuity, and this will decide what level of care the

patient needs. They may be sent to a different area of the department to be closely monitored. Those with less time-critical emergencies may enter a queue and will have to wait to be seen and treated.

There comes a point during the facility stage when a decision must be made about where the patient needs to be directed to

next. If the patient receives all the care they need in the ED they will be discharged, but if they require further care they will be admitted to a hospital bed for appropriate specialist care.

Figure 2: The transport stage of the ECSF

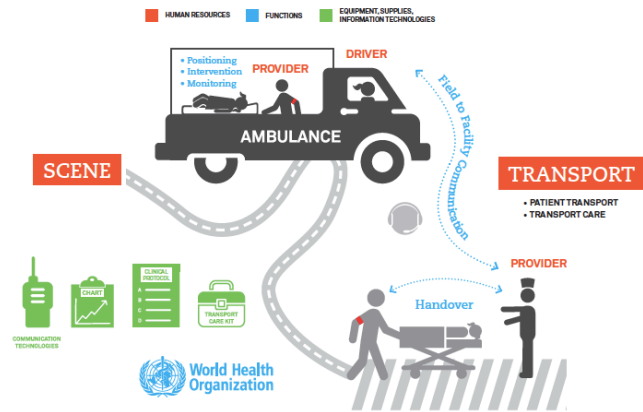
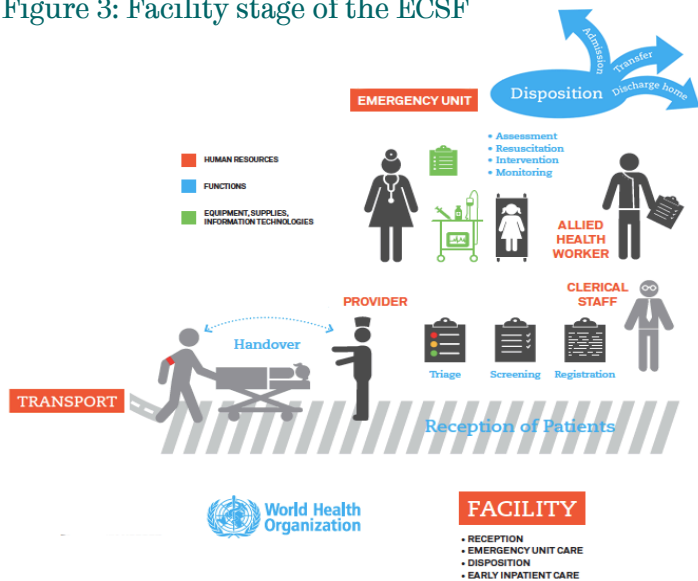


Figure 3: Facility stage of the ECSF



complicated to address a different set of challenges. These systems will be the focus of the thesis from now on, with a particular emphasis on the UK system.

The on-scene stage contains numerous types of healthcare resources that could be eligible for dispatch to an emergency. This includes a Double-Crewed Ambulance (DCA), which could contain two paramedics, or a mix of a paramedic and a non-registered healthcare professional trained in emergency care. Other resources include a Rapid Response Vehicle (RRV) containing a paramedic or a doctor; a helicopter (air ambulance) with a pilot, doctor and paramedic mix or a specialist resource such as the Hazardous Area Response Team (HART), which is a team of specialist paramedics with extra equipment. All these resources would arrive on scene with advanced equipment including drugs and a defibrillator, forming a mobile healthcare clinic. They are then able to triage the patient and make decisions around whether they need any further care and if so, where. The underlying principles of triage are to stratify patients according to acuity to ensure the sickest are treated first. The transportation stage and the facility stage appear to remain largely the same globally, from the perspective of the ambulance service. The exception is hospital diversion where paramedics will transport a patient to a specialist facility for management of major trauma, heart attack or stroke, which may be further away than the nearest ED.^{11,12}

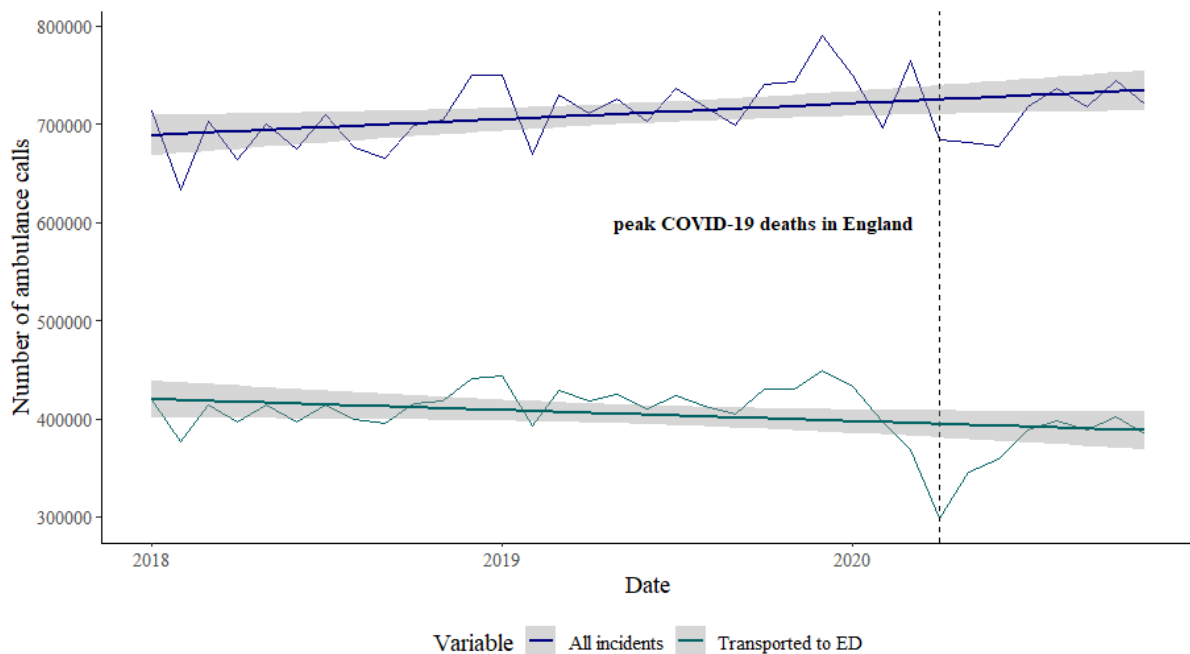
The purpose of a diverse and skilled workforce with advanced equipment is to deal with the increasing demand and complicated case-mix that presents in prehospital care. The next section will discuss this further.

2.3 Demand for prehospital care

The utilisation of prehospital care is rising every year. Studies have shown that there is an annual growth of both ambulance service use and ED attendances. An Australian study used retrospective routine data over an 8-year period (2008-2015) and found in a sample of 2,443,952 records that there was an increase in demand of 1.4% per annum.¹³ A key strength to this study is that they adjusted for

seasonality and population. In the UK, growth appears much higher, with a National Audit Office (NAO) report showing an increase in ambulance demand of 5.2% per annum over a seven year period between 2010-2017.¹⁴ A limitation with the report was a lack of information on the sources of data. It also did not publish raw data or describe how the data was handled in the analysis. A key difference in outcomes between the two studies was that the Australian study only included face-to-face assessments by emergency ambulances, whereas the NAO report measured all calls to the ambulance service. This report included calls passed from the non-emergency phone line known as ‘NHS111’.^B A report published in 2017 by Turner et al. examined a longer time frame between 1994 and 2016.¹⁵ It was found that there was an increase in demand of 56% from the start of the study (1994) to the end in 2016. An unfortunate limitation of this study was the methodology was not specifically designed to model growth, but to re-evaluate response-time targets. This meant only the aggregated total growth over a period was reported for illustrative purposes. However, NHS England

Figure 4: Ambulance demand in England 2018-2021



^B NHS 111 is an alternative public telephone number to support urgent care needs. The service provides triage, and can help patients access healthcare services, clinicians and advice.²⁴⁸

publish annual data on English ambulance services and from 2018 to 2020, there was an annual increase in the number of ambulance calls of 5.09%. This can be seen in figure 4, with the grey shadows representing 95% confidence intervals.¹⁶ During this time-period, the global pandemic of the SARS-COV-2 virus began. The first case of this in England was on January 28th, 2020, with the first death being on March 2nd, 2020. The peak daily deaths of the first ‘wave’ was on April 8th, 2020.¹⁷ This peak is shown in figure 4.^c There have been subsequent waves, not included in figure 4. During the first peak, there was a decrease in ambulance incidents and ambulance conveyances to the ED. This could be due to policy changes and public health guidance; however, at the time of writing there is not any robust evidence to confirm this. The global pandemic appears to cause anomalous data points much lower than the previous trend. Nevertheless, all studies that have used longitudinal retrospective data have reported high demand in the system, and an annual growth of around 5% in this demand.

2.4 Causes and consequences of prehospital demand

The Australian study referenced above was by Andrew et al. and their focus was on the drivers of increased ambulance demand in 2020. Throughout their analysis, they presented Incident Rate Ratios (IRRs)^d for annual growth over the eight-year study period. An IRR of 1.0 translates to no growth. For overall demand, there was an annual growth of 1.4%, which is an IRR of 1.014 (95% CI 1.011 – 1.017). Patient factors that were associated with demand included the Charlson Comorbidity Index (CCI). This is a weighted index of comorbidities which are then calculated into a risk stratification score.¹⁸ More than half of the patients in their sample had no pre-existing health conditions according to the CCI, and this

^c Figure 4 has been created by Jamie Miles (the author) for the purposes of illustration using R v.4.1.2. The data used has been appropriately referenced within the text.

^d The IRR can be calculated by dividing two proportions of two different groups. In an exposed group (E) the proportion with the outcome (EO) is divided by the proportion of the outcome (NO) in the non-exposed group (N). Thus the IRR is EO/NO .²⁴⁹

proportion increased over the study period. This means that patients were increasingly engaging in prehospital care who had fewer underlying health conditions. When patients in the sample were stratified by specific conditions, some grew at a faster rate (per annum) than others. These included mental health issues (IRR 1.058 (95% CI 1.054-1.062)), alcohol or drug abuse (IRR 1.061 (1.056-1.066)) and a Charlson Comorbidity Index score greater ≥ 4 (IRR 1.045 (95%CI 1.039-1.051)).¹³ There were also differences in the rate of growth amongst age categories, particularly those aged 40-59 years (IRR 1.025 (1.020-1.030) and those aged 20-39 years (IRR 1.022 (1.019- 1.025)). Socio-economic factors were also explored. The Socio-Economic Indexes for Australia (SEIFA) from the 2011 census was used. For education and employment, patients were stratified according to deciles. Higher deciles equate to higher levels of education and employment in that geographical area. The authors found that patients from lower- or middle-deciles had higher growth rates in demand, with deciles 1-6 experiencing growth of between 2.0-3.1% per annum. For socio-economic advantage, it appeared that the medium deciles 3-6 grew at the fastest IRR at 1.02 per annum. This can be interpreted as those with medium deciles of education, occupation or socio-economic status were contributing to the growth in the demand more than lower or higher deciles. There were also clinical factors that were associated with growing demand. Patients with the final working impression of alcohol or drug related condition had an IRR of 1.072 (1.056-1.086) and patients with pain had an IRR of 1.044 (1.039-1.050). Paramedics were medically intervening less. In 2008, 60% of patients had a paramedic-led medical intervention, compared with only 45.6% in 2015. Ambulance conveyances to the ED had also increased with an IRR of 1.012 (1.009-1.016), which is significantly different compared to the quantity of patients not requiring medical intervention (IRR 1.067 (1.063-1.072)). This means that patients were being conveyed to the ED but were not requiring medical intervention. An exposition into this cohort that perhaps did not receive a clinical benefit to ambulance transportation will be discussed later in section 2.6 of this chapter.

A systematic review in 2017 summarised all available international evidence about why people were driven to access the urgent and emergency care services.¹⁹ There were six strongly evidenced themes that emerged from the review that are not by definition distinct from each other and have possible overlaps between them. Twenty-six studies provided evidence that confidence in primary care services drove patients to seeking urgent and emergency care. Barriers to accessing primary care were not necessarily created by services themselves, but by patient's perceptions of the service. For example, patients felt they would not get a primary care appointment in a timely manner that suited them. Patients also felt health anxiety that led them to access a higher acuity service than was perhaps necessary. They needed reassurance sooner than primary care could provide. This contrasts with other evidence identified in the review, which claims patients are unable to assess the acuity of their conditions. One study found that 24% of patients presenting to the ED were classed as 'non urgent' but felt they needed to be admitted to hospital. This strongly links with another theme which narrates those patients had a perceived need for ambulance service or hospital treatments or investigations. A large theme of the review was urgent and emergency care being recommended to patients by healthcare professionals, family, or friends. One study found that 52% of all ED patients had been recommended to attend by one of these groups. Fifteen studies identified that convenience (location, no appointments, and 24-hour service) significantly impacted a patient's decision to access urgent and emergency care service.¹⁹

There are also system factors that contribute to demand such as provider-induced demand. This is where the demand rises due to the health system, insurance companies, and healthcare providers (clinicians). Qualitative evidence has explored factors that contribute to supplier-induced demand and have found that legal consequences and agency are key factors that induce demand such as increased investigations and treatments than is perhaps necessary.²⁰ Another study set in UK Emergency Departments found that 85% of their sample of 478 emergency physicians felt that too many diagnostic tests were ordered for

patients. They also found that 97% of the sample felt that at least some of the investigations they personally order were likely to be unnecessary.²¹

Demand is a significant factor for the ambulance service. When patients are conveyed to the ED, they can be held in a queue until the department's clinical staff can receive the patient. This queuing is also known as 'ambulance ramping'. This extends the handover time per ambulance crew and results in a process known as 'offload delay'. Both ambulance ramping and offload delay are consequences of increased demand in the system that has not been matched with increased resources. However, when examining the two concepts in more detail, offload delay appears to cause ramping. When ambulances are queueing to hand their patient over at the ED, they are unable to respond to prospective emergencies. Nehme et al. found that hospital turnaround time was associated with lengthened response times, but the extent of the association remains unclear.²² The next section will discuss ambulance ramping and offload delay in more detail.

2.5 Ambulance offload delay and ramping

A systematic review published in 2018 identified 137 articles relating to ambulance offload delay. 28 focussed on the causes, 14 on its effects and 89 on proposed solutions.²³ As this is the only systematic review on the subject, it will be used as a framework for this section. The limitation with the review is that it did not undertake a quality of bias assessment on the included articles, and most of the included studies are older than ten years. However, pertinent studies have been extracted and assessed for limitations.

2.5.1 Magnitude of offload delay

International evidence on the magnitude of offload delay varies. A 2005 Canadian study by Segal et al. combined time-motion data from paramedics presenting to ED and a prehospital call database of time data.²⁴ It was only a small sample of 152

calls, however 45% of the job cycle time for paramedics was in the Emergency Department. The job cycle time refers to the period starting from when a patient is assigned to a paramedic, to when the patient is handed over or discharged. Three other studies have examined the delay in handover. Silvestri et al. undertook a small prospective observational study in the USA and recorded the offload times of all patients attending a single level 1 ED between the hours of 11AM and 11PM for one week. In the 167 patients in the study, it was found that most (n=122) were triage category 'green', which is described as 'least severe'. The triage tool used is unknown for the study, but the implication is that these patients were the least acute patients in the study cohort. Despite 52% of patients being handed over in a timely manner (<15 minutes), there were 15% taking over an hour to hand over.²⁵ An Australian study in 2012 used a larger sample, but retrospective, of 141,381 ambulance transports. It was found that 12.5% of patients experienced a handover delay of 30-60 minutes, and 5% had a delay of >60 minutes. It was also found that larger hospitals in urban areas, especially in the winter were contributing factors. In the UK, NHS England publish a daily situational report for urgent and emergency care. This only occurs during the winter months but collects national data around demand. Data shows that between 2017 and 2021, 10% of all ambulance transports had a delay of between 30 and 60 minutes. NHS England also published that an average of 3.1% of all ambulance transports result in a delay of over an hour to handover.²⁶ A limitation of the report is that it does not include annual figures, which means the averages are likely to be skewed to an extreme as the included months occur in the winter period, where incidence of disease significantly rises, as does system demand.

2.5.2 Causes of offload delay

Li et al. in their review found 28 studies related to the causes of offload delay.²³ Fundamentally, the greatest cause of offload delay is the ED having too many patients that they struggle to manage, known as 'crowding'. Six studies directly linked crowding to offload delay, with a further three linking it with significant

negative effects for ambulance providers. The review identified that there were certain factors of ED crowding that contributed more to offload delay. These include resource shortages, high demand of high-acuity patients, diagnostic delays, language barriers and increased bureaucracy.

Not all studies agree with ED crowding being the largest contributing factor in offload delay. In 2015, Lee et al. from Korea published a study on ED crowding and ambulance turnaround time.²⁷ The study was large and used a prospective cohort of 163,659 patients transported to 28 EDs within the study area. Using multi-level regression modelling with random effects for EDs, a negative association between ED occupancy and ambulance turnaround time was found. This equated to a 1% increase in occupancy showing a 0.02-minute decrease in turnaround time (95% CI 0.01-0.03). However, this study had numerous limitations. The definition of turnaround time was not consistent with previous literature as they defined it as the time from arrival at ED to the return of the ambulance to the base station, and then adjusted for distance from the ED. They also did not examine confounders and adjust for them in the analyses. The clinical significance of their finding could be negligible in practice.²⁷

2.5.3 Consequences of offload delay

In the review by Li et al. they segregate the consequences of offload delay into its impact on patients, the EMS system, financial cost and legal implications.²³

Asplin et al. in 2003 developed a conceptual model of ED crowding to aid future research in understanding the cause, consequences and solutions of crowding in the ED.²⁸ They separate the issue into three interdependent components. These are input, throughput, and output. The input component examines all the factors that contribute to ED demand. The throughput component focuses on the patient experience within the ED and includes factors that disrupt timely flow whilst the patient is in the department. The output component contains elements that delay discharge from the ED. However, this thesis will break down the consequences of offload delay using the same framework modified for context, as

follows: the upstream is consequences for the ambulance service in terms of system delivery, midstream constitutes ambulance ramping and downstream explores the consequences of offload delay for the patients experience in the ED.

2.5.3.1 Upstream consequences of offload delay

There is a paucity of evidence for the upstream consequences of offload delay. Eckstein and Chan undertook a prospective, longitudinal study in 2004 examining the effects of ED crowding on paramedic ambulance availability. The study included all incidents that had a handover of over fifteen minutes at the ED over a one-year period. One in eight of their ambulance transports resulted in a delay of over 15 minutes and 8.4% took over an hour. A conclusion of the study was that these delays have an impact on ambulance service delivery, however there was no justification for this in the study.²⁹ A 2017 report by the Royal College of Emergency Medicine (RCEM), NHS Improvement and the Association of Ambulance Chief Executives (AACE) claimed that in England in 2016, over 41,000 12-hour ambulance shifts were lost due to offload delay.³⁰ Other studies have looked at the impact of ambulance diversion. This is an interventional strategy to ED crowding and offload delay, which sees ambulances diverted away from overcrowded EDs to less busy ones further away. Pham et al. undertook a systematic review into the effects of ambulance diversion. They included 107 studies, mainly of low quality. They found that studies which looked at the effect of diversion on ambulance flow had minimal impact on ED crowding. This ranged from 0.3-0.15 patients per hour being brought in by ambulance. It has also been demonstrated that diversion increases ambulance transport time and has economic implications. However, these were limited to American studies and the economic impact was in lost revenue to the hospital.³¹ A 2018 time-series analysis in the UK by Knowles et al. examined the effect of closing five district EDs. They found that ambulance call volume increased in geographical areas of ED closure, and also journey times to the ED.³²

Another upstream consequence identified in the literature is the idea of mutual aid. Ambulance services operate within geographical boundaries. However, when resources are scarce, because of offload delay, a neighbouring service may respond to the emergency if they are nearest. The limitation to this, as identified by both Majedi in 2008, and Cooney in 2011, is that this practice could leave rural areas uncovered and it then leaves the neighbouring service a resource down for a significant period of time.^{33,34} Majedi also used mathematical modelling to conclude that adding more ambulances into a fleet to compensate further exacerbates offload delay, describing the handover-gate as a ‘bottleneck’ system.³⁴

2.5.3.2 Midstream consequences of offload delay

An interpretive phenomenological study in 2015 was undertaken by Kingswell et al.³⁵ It aimed to elicit how patients felt about ambulance ramping and undertook semi-structured interviews with seven patients who had an offload delay of greater than thirty minutes. Patients who were waiting to be handed over for a prolonged length of time felt as though they were left ‘in the dark’, waiting in a queue without privacy and not knowing when they would be seen. Hammond et al. undertook an exploratory descriptive study to develop a definition of ambulance ramping. Using in-depth interviews, focus groups and chart audits within Queensland Ambulance Service (Australia) and ten EDs, they determined that ambulance ramping was:

“A practice primarily of the triage nurses in which patients brought to the ED by ambulance are not admitted to the ED because of overcrowding or insufficient staffing levels.”³⁶

However, publications about ambulance ramping do not comply with this definition. Kingswell et al. in 2017 published a scoping review into ambulance ramping and included thirteen studies in their review. When examining the

included studies, ten defined ambulance ramping as synonymous with offload delay and three did not define it at all.³⁷ This is disappointing as they are conceptually distinct phenomenon. The former, is a consequence of the latter, in the same way that ED boarding is not the same as a hospital with no inpatient beds. The ED is boarding because there are no beds, the ambulances are queueing because there is an offload delay. A hospital can have no beds, but the ED might be empty and therefore ED boarding would not exist. The department could be full, and handovers might be delayed by an hour (for example), but if there is only one ambulance there for two hours, there is no ramping. Conversely, it could be argued that a delayed handover with only one patient in the queue is still a ramped patient. This appears to be supported by the literature identified by Kingswell et al.³⁷ However, there is a paucity in the literature for upstream consequences to ambulance services. The harm of a single delayed handover versus a long queue of ambulance patients is unknown. Future studies into ambulance ramping should consider the unit of measurement as queue length and not just handover delay.

2.5.3.3 Downstream consequences of offload delay

Patients who do not have prolonged handover times have been shown to have a better experience than those who do. An Australian study by Crilly et al. in 2015 used twelve months of linked data (ambulance service and ED) to compare the characteristics of delayed handover patients and non-delayed. The study had a large sample of 40,783 ambulance patients. Patients who were not delayed had a shorter time to triage, ambulance turnaround time, time to see healthcare professional and ED length of stay. However, these findings could be linked to the degree of crowding in ED and the study did not adjust for this. It has been demonstrated that the main driver for offload delay is ED crowding.³⁸

As a possible intervention point to offload delay, one study has suggested that paramedics can stream patients with the same degree of accuracy as the triage nurse. This was an Australian prospective study of 500 ambulance cases. There

was a concordance of 86.4% (95%CI 83.1 – 89.1) with the triage nurse. Streaming dispositions included resuscitation room, a cubicle or fast track.³⁹

In section 2.3, evidence was presented that there is significant rising demand for prehospital and emergency care. This demand is not a blanket rise in all patients across all conditions, but peaks in specific areas. As Andrew et al. alluded to, there appears to be a growing cohort of patients who do not need medical intervention but find themselves seeking emergency care. The majority of ambulance service patients require fewer critical interventions and more community based care.^{40,41}

This chapter has so far examined the drivers of demand for all patients entering the system; however, there are patients that have found themselves accessing a higher level of care than perhaps is needed. Their acuity of illness could be treated in the community, as opposed to an ambulance or the ED.

2.6 Low-acuity navigation into emergency care

From a policy perspective, NHS England clearly define two populations according to their care need: emergency and urgent. These definitions appear to be distinct with clear boundaries; however, in practice there is much overlap between the two, especially when the access to urgent and emergency care is often patient led.

“Emergency: Life threatening illnesses or accidents which require immediate intensive treatment. Services that should be accessed in an emergency include ambulance (via 999) and emergency departments.

Urgent: An illness or injury that requires urgent attention but is not a life-threatening situation. Urgent care services include a phone consultation through the NHS111 Clinical Assessment Service,

pharmacy advice, [in hours] and out-of-hours GP appointments,
and/or referral to an urgent treatment centre (UTC).”⁴²

Pope et al. undertook a qualitative study with members of the public to elicit their perceptions of what defines ‘urgent’ and ‘emergency’.⁴³ Using citizens’ panels, and a large sample of semi-structured interviews, there was no clear separation between the two definitions. Instead, a theoretical model was described whereby patient choices were socially constructed, contingent, and informed by beliefs and experience. If a patient had rarely contacted any urgent and emergency care service, but they had previously attended the ED, they were more likely to associate this as an appropriate care setting for their urgent need. As a tangential point, it is reasonable to extend the work of Pope et al. to possible belief structures of paramedics. They have traditionally transported patients to the ED and the inception of alternative care pathways, referrals to primary care and discharging on scene are all modern advancements. The ED is very familiar to the paramedic and this experiential knowledge could feature in their decision-making. Pope et al. also found that one source of confusion for which care setting to access, was that patients saw the options as equivalent rather than hierarchical. The quote below illustrates the points made by Pope et al.

“We had a conversation here, didn’t we, about the confusion, and how do you know what to do. And actually, you know, if you’ve used services a lot you know what to do. But if you’ve had an urgent care incident, and you’ve only had one in the last 20 years, how do you know what to do?”

(Public panel)

P14: "Urgent care, I would think of, probably, well, an ambulance, A&E, you know, if it was urgent, yes. Otherwise it would be just a trip up the doctors to see what the problem is, you know.

Interviewer: "... emergency care, what do you think of?"

P14: "Emergency care is, well, the same thing, really. Yes. I mean, if I could see there was a major problem with anything ... if it really looked bad, you’d have to ring 999, I think."

(Younger interviewee)⁴³

In NHS England’s definitions, they dichotomise patients to optimise care. The emergency patients call an ambulance and/or attend ED, whilst those with urgent care needs are referred to community options such as the GP. This is not always successful in practice, and evidence has often focused on why urgent care patients have found their way into an emergency care setting, inducing demand. MacKichan et al. examined why primary care patients attend the ED. They found that the convoluted way to access primary care engendered a mistrust in the system. Furthermore, the increase in telephone triage in primary care to mitigate demand had led to patients becoming ambivalent to its benefit. They explained that speaking to a clinician over the phone was unsatisfactory and was inequitable as some groups were not able to ring on the ‘first-come first-served’ appointment system.⁴⁴ Attempts to help patients navigate the system using telephone triage have not demonstrated a clear benefit for similar reasons. In Pope et al., patients were reluctant to accept the advice of the NHS111 service, especially if there was a language barrier on the phone. This is supported by a recent quantitative study by Egan et al. who used a sample of 16,563,946 calls to the NHS111 service and found that for every 20 calls where callers were told not to

go to the ED, 1 resulted in an avoidable ED attendance within 24 hours.⁴⁴ There is a distinction between the telephone triage in general practice (which forms part of the consultation), and the NHS111 service which is primarily a signposting service. If patients struggle to define their own urgency and must grapple with a complicated system to find the right care setting, then patients can potentially end up presenting in the wrong place. If a patient attends the ED (a care setting for emergency patients) with a care need that could be defined as urgent, they are an avoidable attendance.

But the definition of urgent is abstruse in this context as it is defining what a patient is not, as opposed to what they are. The label is applied not because they have demonstrated a clear urgent care need, but because they have lacked a clear emergency care need. The term could be considered interchangeable with others such as 'avoidable', 'preventable', 'non-urgent', 'unnecessary', 'inappropriate' and 'primary care problems'.⁴⁵ This is on the assumption these terms are used in the context of describing emergency care patients in the wrong care setting for their need. The general principle is a patient does not have a qualifying clinical benefit for emergency care, but through complex processes finds themselves in the ED. Qualitative work by Parkinson et al. have attempted to break this concept down further.⁴⁵ Within the avoidable attendance concept, the authors propose three distinct groups.⁴⁵ The first are the clinically divertible attendances. This group of patients need access to healthcare, but they do not need the specialisation of the Emergency Department. The second are the clinically preventable attendances. These do require the specialisation of the Emergency Department, but this could have been prevented if their condition was managed better or if there was an earlier intervention. Thirdly, there are the clinically unnecessary attendances. These do not require any clinical care. The clinically preventable patients have an 'emergency care' problem. The divertible patients are synonymous with the urgent care patients that have been previously defined. The preventable were likely to have been urgent sooner in their disease manifestation but now find

themselves as emergencies. The unnecessary, are neither urgent, nor are they emergencies.

In 2020, O’Cathain et al. used mixed methods to explore the drivers of ‘clinically unnecessary’ use of emergency and urgent care. The term ‘clinically unnecessary’ is important to define. The authors define it as:

“The term ‘clinically unnecessary’ defines use that doctors, nurses and paramedics assess as not requiring the level or urgency of clinical care provided by their service.”²²

Through examination of a large mixed methods study by O’Cathain et al., the term is synonymous with ‘low acuity conditions’, ‘medically unnecessary’, ‘unnecessary use’, ‘non-urgent’, ‘low acuity’ and ‘potentially preventable use’.²² A limitation in creating this group of synonyms however, is that some terms are tied to different contexts. For example, the terms ‘low acuity’ and ‘non-urgent’ reference a label attached to how unwell a patient is, and how fast they need medical help. Conversely, terms like ‘unnecessary use’ and ‘potentially preventable use’ are related to the care setting the patient has found themselves in. The former labels would not change if the system was redesigned, whereas the latter would. The study by O’Cathain et al. spanned the three main clinical areas of urgent and emergency care, namely, primary care (GP), prehospital (ambulance service) and the emergency department (ED). This is appropriate when exploring why patients may not be successful in navigating the right place for their care, resulting in a ‘clinically unnecessary’ attendance in one setting, is likely to be entirely appropriate in one of the other two settings. There were five elements to

the study design. A realist synthesis^E, qualitative interviews, focus groups, a population survey and integration.

In the realist synthesis, the drivers of clinically unnecessary attendances from the patients' perspective were explored, and the findings identified six underlying mechanisms were explored. Patients found they were *minimising risk* by attending a higher healthcare setting than required.²² The uncertainty they felt around their symptoms created anxiety. Patients would use heuristic experience to inform future choices. If patients had experienced a delay in accessing care in the past and there were consequences, they were more likely to seek immediate healthcare in the future. This was further heightened in the context of caring for others and a fear of consequences if something went wrong. Another mechanism for urgency was the '*need for speed*'.²² If people were unable to continue their daily responsibilities such as paid work or childcare, they would often seek help at the ED to resolve the problem quicker. This was particularly the case with paediatric patients with working parents. Patients also felt they needed immediate care when it was for pain relief. If patients were uncomfortable but could not access a GP, they would attend ED or call an ambulance. There would be a delay in seeking healthcare up until a 'tipping point' where patients could no longer cope and would seek a higher level of care. The remaining four mechanisms focus on the benefit of an instant ED visit compared with attempting to access primary care. There is *low effort required for help seeking* in emergency care.²² People who have stressors on their internal and external resources such as time, or money can lack the extra capacity needed to cope with a new illness. This occurs subgroups such as those with low socioeconomic status, parents, people who are isolated, people with demanding jobs, and people with mental health problems. It leads to a feeling of helplessness, which drives them to seek immediate help. Patients were also *compliant* with advice from their trusted social network such as family or healthcare professionals. This means patients

^E A realist synthesis aims to not just appraise evidence, but also account for the context as well as outcomes.²²

would accept the recommendation of attending ED or calling an ambulance if it came from a trusted source, regardless of their own belief of where the most appropriate place to access care was. There is also the *availability and quality of care* provided by the ED.²² Patients felt that the ED would provide a better service than what is available in primary care and trusted this healthcare setting to meet their need. This leads to the final mechanism, which is a *frustration with access to the GP*. When people are not able to book a GP appointment in an acceptable period, it can give the perception of an inaccessible and uncaring service. This leads to patients feeling they have no choice but to access emergency care.²²

There are also wider considerations to why patients access healthcare, such as those described by Hannay in 1980.⁴⁶ They reference a difference between the patient reported 'iceberg' of illness, and the GP's perceived triviality of patient presentations. Patients felt that they only sought medical attention for around a third of illness, leaving most of the ill health hidden from clinical practice. This is known as the iceberg. However, when GPs were asked about the presentations that patients attended practice with, they perceived there were a significant amount that fell below the threshold of needing an ongoing referral. In chapter 7, section 7.6.1 there is an expansion on how social and demographic factors can be associated with a low-acuity patient.

It is understood from the literature discussed in section 2.7 why there is this cohort of patients, and the next section seeks to quantify how large this cohort is in the context of the emergency care case-mix.

2.7 The quantity of avoidable emergency care contacts

The previous section defined avoidable ED attendances and examined what drives this cohort to access healthcare at a higher level than what is required. The same context is existent in ambulance service patients as well. The two are not mutually exclusive cohorts and it is possible to be both. Indeed, it is the patients

that fit both criteria who would benefit most from improved navigation of services. When quantifying the amount of avoidable emergency care contacts, both the ambulance service and the ED are considered emergency care settings and so quantifying the magnitude for both are explored in this section. In 1998, Snooks et al. undertook a systematic review of studies aiming to quantify the amount of 'inappropriate users' of ambulance services. Combining the results of the ten included studies, there was a mean of 38.4% (standard deviation 11.6) considered inappropriate users in comparison with all ambulance users.⁴⁷ The main limitation of the study was each had their own definition of inappropriate user and this meant drawing conclusions was difficult. Furthermore, at the time of publication, non-conveyance was not always an option and only 17% of patients were not taken to hospital.⁴⁸ Patton and Thakore took a small sample of 910 ED attendances over a 7-day period at a single hospital in Scotland. The objective of their study was to quantify how many ambulance conveyances to the ED were inappropriate conveyances. The results show that 295 presented by ambulance, and 84 (32%) of these were considered inappropriate following a review of the ambulance patient report forms and ED notes. The definition of inappropriate was a judgement made by the ED consultant at the time of presentation. The majority of the inappropriate attendances had primary care needs and could have been managed in the community.⁴⁹ The sample size used in this study is low and the findings were not presented with confidence intervals. In addition, using subjective judgement from multiple experts has significant limitations if the concordance between them is not assessed. This is because different experts will have different thresholds of 'appropriateness'. However, a key conclusion drawn from the study was that a reduction in inappropriate ambulance conveyances to the ED would result in an 11% decrease in ED workload. Andrew et al. also found that over the eight-year study period, patients not requiring a medical intervention from paramedics had grown by 6.7% annually (IRR 1.067 (95%CI 1.063-1.072)).¹³ A prospective, multicentre study found that 19.4% (95%CI 18.0 – 20.8%) of ED attendances were avoidable. This was using a sample of 3044 case reviews.⁵⁰ However, it is recognised that the choice of definition has an impact

on the quantity of avoidable conveyances. McHale et al. used a large national dataset in the UK to examine all ED attendances between 2011 and 2012. Their definition of 'inappropriate attendances' were patients who were self-referred as a first presentation, received no investigation, had no treatments (or just guidance/advice) and were discharged with either no follow-up or just a GP follow-up. The study included 15,056,095 patients and found that 11.3% of patients were avoidable.⁵¹ In 2007, a Swedish study used a qualitative outcome measure of asking ambulance staff whether each patient required ambulance transport.⁵² This was a prospective study of 1977 patients. According to the ambulance staff, a third did not need ambulance transport. From the labels given by the ambulance staff, the study authors undertook a descriptive analysis of this cohort compared to those who were judged to require transport. A broad range of clinical presentations that did not need hospital transport or ambulance intervention were identified. Studies that are more recent have reframed the definition from avoidable to 'non-urgent'; however, the definition remains contextually the same. O'Keeffe et al. found that, in a linked data analysis of 3,667,601 patients, there were 554,564 (15.1%) who were avoidable attendances at the ED. Like the Swedish study, this one also characterised the avoidable attendances and found that these patients were more likely to present out of hours (OR 1.19 (95%CI 1.18 – 1.20)), and more likely to be younger (aged 16-44). Interestingly, it was found that there were 8.5% avoidable attendances that arrived by ambulance.⁵ A 2018 study by Miles et al. found that up to 16.9% of ambulance conveyances to the ED were potentially avoidable.⁵³ This study used the same definition as O'Keeffe et al. Even though this estimate of 16.9% is noticeably higher than other studies, it is important to recognise that there were limitations in the data that were not addressed, such as handling of missing data from specific sites.⁵⁴ Andrew et al. in their study of 2,443,952 ambulance journeys found that there was a 6.7% per year growth in patients who did not need paramedic intervention.¹³ This was a longitudinal study over seven years between 2008 and 2015.

Avoidable attendances and conveyances appear to contribute significantly to the input of ED demand. Evidence has demonstrated paramedics transport patients to the ED who potentially could be treated in the community. The acuity of prehospital care patients has changed from being mainly high acuity, life threatening emergencies, to mainly patients with urgent and complex care needs. It is important that their care be optimised to take place at the appropriate care setting. There are many different terms used to describe this population; however, for this thesis the term 'low acuity' will be adopted. This is because it feels patient centred and detaches the context of them being 'in the wrong place'. It is also the terminology which is adopted by policy in the UK. In this thesis, low acuity has been operationalised and this can be found in chapter 7, section 7.5.

2.8 Strategies to optimise care

In 2002, Snooks et al. summarised different strategies that could optimise the care of low acuity patients contacting the ambulance service.⁵⁵ The most relevant findings were that calibrating priority dispatch systems could triage patient acuity at the time of call and resources could be allocated accordingly with more accuracy. A priority dispatch system is an algorithm that makes decisions on how sick a patient is based on the answers they give to questions asked by an ambulance dispatcher. Sicker patients may have more resources allocated to them, and sooner. Also, Snooks et al. found that telephone triage systems led by nurses were effective in an out of hours setting. These interventions focus on the control room when patients call for an ambulance. A 2015 rapid review by Turner et al. collated all available evidence on the different strategies of delivering effective care. Ten systematic reviews and 44 original studies were identified as evaluating telephone triage. The results show that it provides safe decision making and patient satisfaction is high. However, most studies lacked a whole system perspective, which limits the evaluation of this aspect of urgent and emergency care.⁵⁶ More effective strategies have been identified on scene with the patient.

In the study by Snooks et al., the most prevalent strategies were non-conveyance and field triage and diagnosis made by paramedics. This was reiterated in 2015 by Turner et al. who undertook a review on management of urgent care needs in the community by ambulance staff as part of their rapid review. Seven systematic reviews and 12 primary studies supported the view that extending the knowledge and skills of ambulance staff to meet the case mix is an effective strategy.⁵⁶

The strategies of non-conveyance and increasing paramedic's skills to diagnose and discharge in the field have been a mainstay in UK policy for several years.^{12,57,58} It has been encouraging paramedics to make autonomous and dynamic decisions about their patient care before transporting them to the ED.^{12,41,59,60} This theme was particularly emphasised in the Urgent and Emergency Care Review Team (UECRT) report in 2013.⁵⁹ One of the main outcomes they aimed to achieve in the report was for those with an urgent need to access healthcare be provided with a highly responsive, effective, and personalised service outside of hospital. This has led to the growth of non-conveyance. In a 2013 follow up to the Snooks et al study, the focus was whether the strategies identified originally had been translated into practice. It was found that over the twelve year gap there had been a reduction in conveyance rates from 90% to 58% (equivalent to 2.7 billion fewer journeys).⁶¹ However, there had been a plateau in the amount of non-conveyance with recent policy stating that 90% of calls were not life-threatening but low-acuity, and nearly 60% resulted in a patient being conveyed to ED.⁵⁸

2.8.1 Non-conveyance

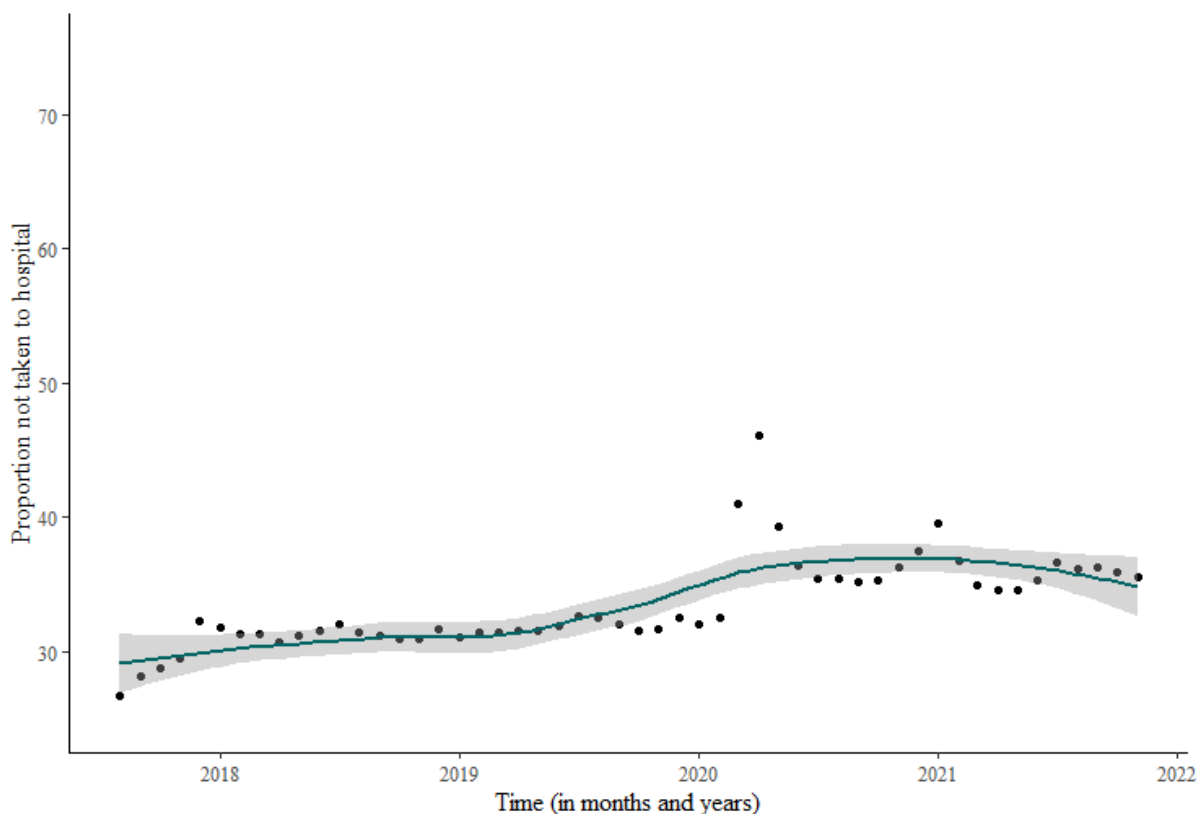
Non-conveyance can be defined as “an ambulance deployment as appropriate, where the patient after examination and/or treatment on-scene does not require conveyance with medical personnel and equipment to the healthcare facility”.⁶² Mikolaizak et al. in 2013 published a systematic review analysing the rates and outcomes of non-transported older patients who had fallen. Twelve studies were included, and their risk of bias were scored by the authors as ranked moderate-

low quality. In this sub-group of older patients who had fallen, the rate of non-conveyance was between 11 and 56%.⁶³ The review also explored the reasons behind non-conveyance. There were multiple factors; however, the most common was refusal to travel by the patient, which was mentioned in seven studies. Other reasons included the treatment on scene was sufficient or the patient was referred to the GP.⁶³ A later review in 2017 by Ebben et al. produced a more comprehensive systematic review aiming to describe ambulance non-conveyance rates, the characteristics of these patients, the follow-up care after non conveyance and influencing factors in the decision making process.⁶⁴ This systematic review appears to be more comprehensive in comparison to other reviews undertaken in the same subject area.^{63,65-67} Sixty- seven studies were included in the review with the majority being quantitative observational studies of moderate-low quality (in the context of risk of bias, as appraised by the authors). The rate of non- conveyance varied between 3.7–93.7% in the unselected population (not disease-specific). The wide range was due to the differences in prehospital models between countries, however the median rate was 15.7% and the mean was 24.4%. Heterogeneity between studies meant a meta-analysis was not possible. On reviewing the supplementary material of the study, it is apparent that there are differences in sample size, population of interest (even in unselected cohorts) and in system-set up. One of the contributors to the extremes in the range appears to be the sample size. If studies with < 2000 patients were excluded the range would be 3.7-31.7%, with a mean and median of 14.5%.⁶⁴

According to NHS England, the then (November 2021) non-conveyance rate experienced nationally in England was 31.3%.⁶⁸ The limitation of this source of data is that it is used as a performance measure for the ambulance service, which is linked to their funding. This could introduce a bias into the data collection stage and affect the accuracy of the report, as ambulance services might well be motivated to perform for funding reviews.⁶⁹ Figure 5 shows longitudinal data from NHS England on non-conveyance proportions of all face-to-face ambulance assessment with the grey shadows plotting the 95% confidence interval. The

figure was created by the thesis author (Jamie Miles) using publicly available NHS data.¹⁶ The graph shows two spikes in non-conveyance, and these relate to peaks in cases during the COVID-19 pandemic. Beyond these spikes it is demonstrated that the non-conveyance in England has barely changed over time. Note the Y-axis on the figure has been truncated to between 25% and 75% to facilitate interpretation.

Figure 5: Proportion of patients not conveyed to hospital after a face-to-face ambulance assessment between 2017 and 2021



Ebben et al. continued to describe the characteristics of non-conveyed patients. In their included studies, they found that age, gender, ethnicity, and geographic area had been mentioned as demographic characteristics. Age range was quite wide and ranged from 14-90 years old. Gender was predominantly male, and most non-conveyed patients were in an urban area. For ethnicity, one study claimed that 90.6% of non-conveyed patients were white, whereas another study reported 48.3% were African American. The limitation in this reporting was an absence of

describing the underlying population ethnicity. Vital signs were mentioned by three studies in the review. One study found that 14.9% of non-conveyed patients had abnormal vital signs. These could have been blood pressure, O₂ saturation, Glasgow Coma Scale and body temperature.⁷⁰ The two other included studies agree with this concept of non-conveyed patients not always having observations considered within normal parameters.^{71,72} A limitation with Ebben et al. is the age of the included studies. Eleven of the studies were published between 1990 and 1999. As studies become older, their validity in such a rapidly evolving field such as prehospital care becomes compromised.

2.8.1.2 The success of non-conveyance

Ebben et al. also included in their review a synthesis of evidence on the outcomes of patients who were not conveyed. They focussed on two specific outcomes of repeated access to care and patient outcomes. Repeated access is where a patient re-contacts an urgent or emergency care service following their initial patient episode. Services included in the study were the ED, EMS-system (ambulance service), GP or walk-in clinic. The periods included repeated access within <24 hours, <48 hours, <72 hours and <7 days. Seventeen studies used repeated access as their outcome measure and included a general population (as opposed to a specific disease population). Patient outcomes included mortality, hospitalisation and recurrence of symptoms.⁶⁴ Table 1 summarises their findings.

Table 1: Summary of repeated access to healthcare following EMS discharge in the general population from Ebben et al.⁶⁴

| | <24 hrs | < 48 hrs | < 72 hrs | < 7 days |
|---------------------------------|------------|------------|---------------|-------------|
| Repeat access to the ED | 4.6 - 7% | 19% | 6.4 - 25.8% | 8.1 - 80.1% |
| Repeat access to the EMS | 6.1% | 2.3 - 2.5% | - | 7.4 – 13.5% |
| Repeat access to the GP | 13% | - | 36.8 – 50% | 46.2% |
| Mortality | 0.2 – 3.5% | 0.3% | 0.3-6.1% | 0.3 – 0.7% |
| Hospitalisation | 3.3% | 1% | 4.5% - 12.1 % | 5 – 8.1% |

The results indicate that following discharge on-scene a significant amount of patient's re-enter into the healthcare system at a later point. However, the incidences are higher with urgent care services such as the GP. There is also a small amount of mortality and hospitalisation. This could be interpreted as these patients are all low-acuity and the decision to convey was correct, despite re-entering healthcare. Coster et al. undertook a data linkage study of 42,108 ambulance journeys and found that, within three days from discharge, there was a 9% re-contact rate with the ambulance service, 12.6% re-contact rate with the ED, and 6.3% admission to hospital. When the time frame was extended to seven days, all of these re-contacts increased.⁷³ However, there needs to be more information about the appropriateness of the re-contact as it is a crude marker without any context. Fraess-Phillips published a narrative review in 2016 on whether paramedics could safely leave patients at home. The review included eleven studies but did not appraise the quality. It was found there was insufficient evidence on safe non-conveyance as triage decisions were different between prehospital and ED staff.⁶⁵ The limitation of this review was that it did not follow a systematic methodology to identify studies. This could explain why they only found eleven studies and Ebben et al. found 67 studies a year later. Patients who are not conveyed due to the paramedic deciding they may not gain a benefit in the ED is significant, and it has been demonstrated above that not all patients who are conveyed, necessarily need the ED.

In 2018, O'Cathain et al. used data from ten ambulance trusts in the UK to explore why there could be variation in non-transport rates. The variation was associated by the skill level of paramedic (those with extended skills were more likely to not convey), and the risk attitude of senior managers. These were considered potentially modifiable factors.⁷⁴

2.8.2 Increasing the paramedic skill level

One of the strategies that has been shown to optimise care is to upskill paramedics for them to manage urgent care in the community effectively. In 2007, Mason et al. undertook a cluster randomised controlled trial (56 clusters) in a large urban area of England. They assessed the benefits of paramedic practitioners (paramedics with extra training) on treating older people in the community with minor illness or injury. The sample included 3018 patients aged over 60 and had three main outcome measures. These were ED attendance or hospital admission within 28 days, the job cycle time (call to discharge) and patient satisfaction. Patients were less likely to attend the ED (RR 0.72, 95%CI 0.68-0.75), less likely to be admitted (RR 0.87, 95%CI 0.81 – 0.94), had shorter job cycle times (235 minutes vs 278 minutes (95%CI for difference 60 minutes to -25 minutes) and were highly satisfied with care (RR 1.16, 95%CI 1.09-1.23). These findings were in comparison to the control group of standard paramedic practice. This study demonstrates that it is possible for the ambulance service to reduce transportations to the ED, and this benefits patients (increased satisfaction), the ambulance service (through decreased job cycle time) and the ED (reduced input). Tohira et al. in 2013 published a systematic review and meta-analysis into the impact of paramedic practitioners on ambulance transports to the ED. Many of the studies were set in the UK and were of medium quality (according to a risk of bias assessment by the study authors). Paramedics with advanced skills were more likely to discharge patients on scene (OR 10.5 (95%CI 5.8-19)) than standard paramedic practice. The authors could not pool study results for re-contact (either with the ambulance service, ED, admission) following discharge, however there was a signal in the evidence that there was reduced re-contact in shorter time-frames (24 hours) but possibly more within longer time frames (28 days).⁶⁶ A later study by Tohira et al. used linked data (ED and ambulance service) to quantify the risk of re-contact. The study had a large sample of 47,330 ambulance patients, of whom 19,732 were discharged on scene and 27,598 were discharged from the ED. They found that the patients discharged on-scene were more likely to request a subsequent ambulance, attend ED and be admitted to hospital

compared to the ED cohort. One consideration for these results is that they cannot be directly compared to the previously mentioned studies. This is because they are examining the non-conveyance decisions of regular paramedics whereas the above evidence is comparing advanced paramedics.⁷⁰ Upskilling ambulance staff is a patient and system benefit, and has been noted in the first ever quality standard specific to ambulance services by the National Institute for Health and Care Excellence (NICE). The quality standard (QS174) states that ambulance services should have specialist and advanced paramedic practitioners.⁷⁵ However, the limitation to this strategy is that it relies on either accurate triaging of low-acuity patients prior to ambulance dispatch, or availability of advanced paramedics. It would be too costly to upskill all paramedics to mitigate this. Paramedic decision making around non-conveyance and low-acuity patients remains difficult, and the next section examines the evidence as to why this is the case.

2.8.3 Paramedic navigation of the complex case-mix (decision making)

In 2014 O'Hara et al. identified that the most complex type of decision for paramedics is that of non-conveyance and discharging the patient on-scene.⁷⁶ They undertook a large multimethod qualitative study exploring paramedic decision-making and its system influence. A typology of paramedic decision making that included nine types of conveyance decision was synthesised. Seven of these related to system and contextual decisions to convey a patient to the ED, one referenced patient refusal and the remaining type was that of non-conveyance. It was considered that non-conveyance was the hardest and most complex decision to make. Paramedics referred to how isolated they felt and did not have senior clinical support (compared to ED colleagues). One interviewee noted that if just one of the crewmembers thought they should convey, then they would. Multiple observations revealed that, culturally, it was not considered good practice to argue for non-conveyance. Paramedics would also follow patient preferences and if the patient were expecting transport to ED, they would fulfil this request as opposed to challenging it when needed. O'Hara and colleagues

produced seven overarching system influences on paramedic decision making. The first system influence was meeting the increasing demand. This influence meant less exposure to emergencies and shifting the decision making to more primary care and psychosocial decisions. The second was performance regime and priorities. This recognises that assigning too many patients to an 8-minute response (ambulance time target for emergencies) has an impact on resources such as staff and vehicles. This idea of time targets for response time and time on scene conflict with paramedics taking longer to look at the wider patient context and make complex decisions. The third was access to appropriate care options and this refers to the disparate referral opportunities between geographies. This increases the complexity of decisions and can cause the paramedic to default to convey. The fourth was a disproportionate risk aversion. This is influenced by the staff perception of their own competence, confidence and negative experiences. There is a perceived blame culture and paramedics are unwilling to take any professional risk. The fifth was staff education, training and development. Due to operational demands, this is often forfeited and creates differences in the decision making of paramedics who have specialist skills, those who have undertaken training in their own time and those who only completed service training when available. The sixth was feedback to crews. This does not routinely happen and so crews cannot recalibrate their decision making as they rarely get an opportunity to find out the clinical outcome of patients. It only occurs when a mistake has been made, which precipitates the fourth system influence of disproportionate risk aversion. There is also a lack of clinical support for transport decisions or advice in general about the patient care. The final system influence was ambulance service resources. Like the first two system influences, the skills, resources and equipment differ between geographies and especially between services. When demand is high, it puts further strain on all these resources. The study by O'Hara et al. agrees with many other sources in that paramedic decision making is complex.⁷⁷⁻⁸⁰ A cause of this complexity is the perception of job role as illustrated by Hoikka et al.⁸¹ They elude to paramedic education being focussed on high acuity situations, which foster a culture that

could struggle in recognising and discharging low acuity patients.⁸¹ Simpson et al. also demonstrated that role perception was crucial in the decision-making process of paramedic's on-scene. They undertook a qualitative study with thirty-three paramedics examining decision-making in elderly people who had fallen. It was acknowledged that role perception was an important factor on how a paramedic approaches a decision.⁷⁹ This idea is reaffirmed by Brydges et al. who state:

“Many of the participants perceived their role as a paramedic to be defined by responding to emergency calls for help (i.e., consistent with their initial education and certification expectations), and referral programs represented a formal departure from that enduring view.” (p. 633)⁸²

When examining decision making in the context of non-conveyance, the decision to transport a patient to the ED is not always accurate. Snooks et al. identified nine studies exploring this in a literature review in 2004 and concluded that paramedics were not accurate at this and needed extra training.⁸³ For example, a small study with 313 patients found that paramedic decision making had a sensitivity of 81% and specificity of 34% in predicting requirement for ED care. The reference standard in this study was a data driven definition of whether the patient was 1) admitted, 2) needed a clinical review by a specialist, or 3) had advanced radiology such as an x-ray or computerised tomography (CT) scan.⁸⁴ The study used patient experience criterion as the reference standard. This was defined as an admittance from ED, if they required further assessment by a specialist doctor, or if they required advanced radiologic procedures. This is quite a liberal definition of a necessary ED attendance. Hauswald et al. in 2002 found there was poor agreement between paramedic decisions in their study when deciding whether the patient required ED care ($\kappa = 0.32$, 95% CI = 0.17-

0.46). Their study used a reference standard of whether the patient required an intervention not provided by an urgent care centre.⁸⁵ More recent studies have not shown a significant improvement on paramedic decision making. A 2019 study used vignettes of real patients and presented them to paramedics. They were examining whether paramedics were able to identify a low-acuity patient. In their sample of 143 paramedics, there was clear agreement between paramedics ($k=0.63$) with an overall accuracy in decision-making relating to transporting a patient to ED of 0.69 (95%CI 0.66-0.73). However, when this was broken down into sensitivity (0.89) and specificity (0.51) it was not a vast improvement from Silvestri et al. The sensitivity and specificity in this study used the reference standard of an avoidable attendance at ED, as defined by O'Keeffe et al.⁵ It also means that there were 49% false positive decisions in their sample.⁸⁶ This was a mixed method study that also explored the rationale behind the decisions. When paramedics made false positive decisions (i.e., they transported a patient that lacked a clinical benefit in the ED), they were treating the patient within the confines of the episode, as opposed to looking at the whole picture of patient care. There was also significant agreement with the findings of O'Hara et al. in relation to paramedic decisions being risk averse due to a fear of litigation and role confusion.^{76,86} For this thesis, the term 'avoidable conveyance' will be used, which is when a clinician has assessed the patient on scene and decided they need to be conveyed, but they would be classed as low-acuity and thus do not require the ED. An avoidable conveyance is a decision, and in the context of Miles et al., it would be classed as a 'false positive'. The idea of supporting paramedic decision making of patient acuity and transportation is not new and there have been studies that have tried to support this decision.

2.9 Decision support for paramedics navigating the prehospital case-mix

From section 2.8 of this chapter, it was deduced that paramedic decision making is complex, and this can lead to transporting patients to a care setting that is not

optimised for their care need. In this section, decision support tools are evaluated to discern whether their use provides an improvement compared to paramedic decision making alone.

2.9.1 Triage scales in ED

It has been commonplace globally to triage the acuity of patients entering the ED since the 1990s. This is often using a five-level triage system and derived using expert opinion. In Australia, it is the Australasian Triage Scale (ATS). This groups presenting clinical indicators of risk into five categories, and incorporates physiological parameters (such as blood pressure, pulse etc.) with presenting symptoms.⁸⁷ In the USA, it is the Emergency Severity Index (ESI) was developed.⁸⁸⁻⁹¹ This is a simpler triage tool that relies on more subjective measures of the user. It still incorporates physiological values; however, it is more concise than the ATS. One of the limitations of the ESI is that it struggles to identify low-acuity patients and over triages patients into 'level 2' (higher acuity). This contrasts with the ATS which can distribute patients more appropriately.⁹² A copy of both the ATS and the ESI can be found in appendix B. In the UK, it is the Manchester Triage System (MTS) tool that is commonly used to triage patients.⁹³ The MTS is more complex and requires attendance at paid training courses. The tool itself is subject to cyclical updates and mandate revalidation of training. The MTS was originally published in 1996 as a series of flow charts that could be used by EDs at the point of triage.⁹³ It was derived using a multidisciplinary consensus group; however, its subsequent evaluation of benefit has been poor. A systematic review of the validity and reliability of the MTS by Parenti et al. in 2014 identified twelve studies for inclusion. They found that only two studies reported enough to be judged a high grade of quality and reporting. The MTS was shown to have wide inter-rater reliability and over triaged lower acuity levels. This could be symptomatic of all triage systems as they focus on ensuring the people who are critically ill are not missed during triage, which leads to over triage of less acute patients.

2.9.2 Prehospital triage scales

The MTS provided the basis of a prehospital triage tool and led to the development of the paramedic pathfinder tool in North West Ambulance Service (NWS) in the UK. However, the derived methodology did not extend beyond working group consensus. The tool uses physiological parameters and 'red flag' discriminators to classify trauma or medical patients into emergency, urgent, community or self-care.^{8,94} It was user tested in 2014 when a sample of 481 patients had the tool applied to them with results showing sensitivity of 94.83% and specificity of 57.9% compared to a gold standard of expert clinical judgement.⁹⁵ The tool was highly risk averse with almost half of the patients being transported to the Emergency Department (ED), when the care could have been provided safely elsewhere. The study also lacked scientific rigour in the derivation of the model and reporting the outcomes. An evaluation by NWS compared the MTS to paramedic pathfinder.⁹ The MTS was first transformed into a portable document file with a flow diagram for each presenting complaint. There are three MTS products: MTS Emergency Triage 3e (for use in the ED), MTS TTE 1e (for use in telephone triage), and MTS NaRTle (for use in nursing and residential settings). The evaluation by NWS used the MTS TTE product for use in the field. This was then given to a sample of 177 paramedics to use. The evaluation found that ten of the MTS cards accounted for 73% of the 5858 applications of the MTS. There was a difference between the tool suggesting primary care and the clinicians following this advice. It recommended that 37% of patients were suitable for primary care and 48.6% required ED. However, the actual decisions were that 24.8% went to primary care and 54.5% went to ED. This could indicate either the tool is inaccurate at triaging lower acuity, or that the paramedic decision making is still risk averse even with the advent of a tool. It could also mean paramedics are taking other things into consideration, not accounted for by the evaluation. The comparison with the MTS versus the pathfinder revealed that the pathfinder tool significantly over triaged in rural areas, but not in urban areas. Most MTS patients were classed as 'level 3 or 4' (urgent or standard), whereas the pathfinder grouped most patients into 'level 1

or 2' (immediate or very urgent). The authors did highlight that the absence of exclusions within MTS appeared to improve paramedic decision making. An exclusion is a condition on the paramedic pathfinder that overrules the triage level.

Some studies have examined whether physiological parameters on their own can predict acuity. A systematic review by Patel et al. in 2018 explored whether an Early Warning Score (EWS), which is a composite score based on physiological variables, can identify deteriorating patients in the prehospital setting. At the time of the study, there were seventeen included studies (twelve of which were considered high risk of bias). If a patient had a high EWS score (generally > 7) they were more likely to deteriorate than patients with lower scores and could be labelled as high acuity. However, the EWS was not as successful at predicting mid- and low-acuity patients.⁹⁶ Challen and Walter in 2009 explored whether the use of prehospital EWS can predict an avoidable attendance at the ED. In a small sample of 215 patient care records, no patient with a EWS < 2 was admitted. If this threshold was used as the cut-off for deciding an appropriate ED attendance, it would have 100% sensitivity, 15% specificity, PPV of 68% and an NPV of 100% for ED care. However, this study has a very small sample and was retrospective in design, which is not pragmatic in the context of evaluating decision making. The scores for the study were also calculated retrospectively, and only accepted conditions coded as 'shortness of breath' or 'difficulty in breathing'. This is a promising signal though, that prehospital measurement of physiological variables can contribute to a useful prehospital acuity score. The limitation with the above tools is that they were derived using consensus opinion and often required labour on the part of the user. They are also aimed at triaging the high acuity patient successfully, which would skew the distribution of patients in a multi-level model towards higher acuity outcomes. One study, published in 2014, did explore the benefit of a triage model aimed at the lower acuity patient.⁹⁷ The 'Support and Assessment for Fall Emergency Referrals: SAFER1' was a clustered randomised controlled trial that used computerised clinical decision support

(CCDS) for use by paramedics.⁹⁷ The derivation of the CCDS intervention is not robustly defined. In the earlier published protocol, the definition extends to the following criteria:

“The CCDS prompts the assessment and examination of injuries associated with the fall, co-morbidity that may have contributed to the fall (e.g. Breathlessness or chest pain), psychosocial needs (e.g. cognitive state and ability to undertake activities of daily living) and assessment of environmental risk.”⁹⁸

In the randomised controlled trial, the intervention was used in 436 participants versus the control group of 343. When using the intervention, the odds ratio of a falls referral was 2.04 (95% CI 1.12-3.72) compared to standard practice. There was no difference in safety between the two groups. The limitation with the study was the small sample size, which led to non-significant differences in secondary objectives such as non-conveyance. However, the strength is that it shows computerised decision support for low acuity patients is effective and safe. The study relied on electronic data capture, which meant it limits implementation across ambulance services, a message that has been identified in recent policy.⁵⁸

2.9.3 Emerging opportunities in Computerised Clinical Decision Support (CCDS)

There is currently a digital revolution for UK ambulance services that sees both policy and research symbiotically accelerating the quality of care in the prehospital arena. One innovative example is using a machine-learning algorithm to identify a cardiac arrest at the point of telephone call to the ambulance service. This is not just a blue-sky idea, but a feasible solution to gain crucial seconds and hopefully save more lives. It was tested in a sample of 5242 arrests and, compared to the humans, the algorithm was more accurate and was

ten seconds quicker at identifying an arrest.⁹⁹ The potential benefit of such an algorithm has not gone unnoticed: a policy document known as 'the Carter review' used this example as a paragon for what could be achieved in the UK, if science and strategy work together towards a common objective.⁵⁸

This is promising: however, the same research team later undertook a randomised controlled trial, published in 2021, to detect whether the machine-learning algorithm in practice had an effect on decision support.¹⁰⁰ An intervention group used the algorithm alongside their decision-making, compared with a control group who did not. The results found there was no significant difference between the two groups. However, a further examination of the algorithm used on its own without a dispatcher, was able to identify more cardiac arrests than either dispatcher group. This could highlight that a decision support model may have barriers to implementation and use and concurs with the findings of others.

Snooks et al. undertook multiple work packages to evaluate a risk prediction model in the community known as an 'Early Admissions Risk Prediction' tool.¹⁰¹ It was designed to flag patients to their GP who would be at risk of an admission within 12 months. As part of this project, they identified barriers and facilitators to the implementation of a risk prediction model. This has been reproduced in table 2 below. It appears that successful adoption of a risk prediction model can be challenging as it can conflict with clinical judgement. However, this can possibly be overcome with training and development of staff. Their study overall found that risk prediction increased hospital admissions and ED attendances, which was the opposite of the tool's intention. This failing of the model function could be because the perception of the decision support by staff was that it was to identify high-risk patients and act sooner. The study could have been improved by examining clinical outcomes beyond patient destination.

Table 2: Barriers and facilitators to implementing EARP models in primary care, reproduced from Snooks et al.¹⁰¹

| Barriers | Facilitators |
|--|---|
| Not an organisational priority | Supportive organisational processes for risk prediction models |
| Difficult to fit use of model into reactive way of working | Individual and organisational support for a population management approach to primary care delivery |
| Low interest in using risk prediction models and new ways of working | Training for staff |
| Priority placed on personal and clinical knowledge over risk model information | Interest in new approaches to primary care delivery |
| Questions over accuracy and timeliness of risk model data | Confidence and skills in IT |
| Inadequate access to IT equipment | |

The barriers and facilitators identified by Snooks et al. are similar to those found by a qualitative evaluation as part of the SAFER1 study, comprising a focus into the experience of paramedics using a Computerised Clinical Decision Support (CCDS) for low acuity patients.¹⁰² They used strong structuration theory, which allows the analysis to account for the wider context under investigation, and tested this with a sample of twenty paramedics who had previously participated in SAFER1. The findings indicated a decay in use due to the labour involved in using the CCDS. However, they also found that paramedics felt the CCDS was a reassuring safety net to their own decision, which could overcome the risk aversion identified by O'Hara et al.⁷⁶ Both the quantitative trial and the qualitative evaluation demonstrate that there are clinical benefits to prehospital computerised decision support, but that there are barriers to implementation that may need to be overcome. The 'user barriers' have been discussed above, but there are also infrastructure barriers that need to be considered. In 2020, Porter et al. published a large multiple method study into electronic health records

(EHRs). It was an inquiry into how ambulance services and staff engage with EHRs and found that there was inconsistency in its use. Often the data was indirectly input into the system by using a different medium (such as writing on the glove) first, and then transferred over to the EHR. Their most important finding was that the potential of the EHR (the transfer of clinical information, supporting decision-making and changing patient care) had not been realised and the use was limited to just storing patient information. This is disappointing as there is an emerging evidence base to suggest that computerised clinical decision support can be useful in triaging low acuity patients. A limitation with SAFER1 was that it was undertaken in a sub-group of the prehospital case-mix (patients who had fallen). However, there is evidence that computer-based decision support can be used to triage undifferentiated patients in the ED.

Levin et al. used machine learning to develop an electronic triage system and compared it to the ESI in a cohort of 172,726 ED visits.¹⁰³ The electronic triage predicted three outcomes: critical care, emergent procedure and hospitalisation. The ESI level three (mid acuity) was the majority classification, but the e-triage only agreed in 77% of cases and classified the remaining patients equally to higher and lower triage levels. This means that the computer-based model can potentially overcome the limitations of the ESI by filtering the mid-acuity. Raita et al. also used multiple machine learning methods to predict clinical outcomes in the ED at the point of triage.¹⁰⁴ Using a reference model of logistic regression, they then compared it against four machine-learning models (Lasso regression, random forest, gradient boosted decision tree, and deep neural network). For more information on these methodologies, please see section 5.3.4 and appendix D. The machine learning models outperformed the reference model, which indicates that complex algorithms used to derive decision support could be more accurate than logistic regression alone. A limitation in the use of machine learning models is the 'black box' axiom that ties to clinical accountability. If a clinician cannot decipher how a support tool has reached a conclusion, nor were they involved in the evolution and training of the algorithm, it raises the

question as to where the liability falls. An article published in the *Annals of Emergency Medicine* in 2020 provided some reassurance for the clinical appetite of artificial intelligence in emergency medicine.¹⁰⁵ The limitations can be overcome by including checks and balances into algorithm development, and into clinical practice. Further discussion on this can be found in chapter 10, section 10.2.4. Modern algorithms can now provide an indication of the relative weights attached to features used to develop the model.

2.10 Conclusion

Demand in prehospital care is rising by at least 5% every year. The cause of this demand rests mainly on a changing case mix away from high acuity patients and towards the urgent care of low acuity patients. Policy interventions have focused around two main areas: non-conveyance and upskilling paramedics.^{57,58,60} The former is a strategic aim, whilst the latter is a mechanism to achieve it.

A limitation with this theory is that upskilling is costly, and, for a major impact, the whole paramedic workforce needs to be involved. Studies have shown that their decision making for low acuity presentations is risk averse, and the safety net of transporting to ED will always be attractive.^{6,76} In their training, they focus on ruling in 'red flag' emergencies as opposed to treating the patient in the wider context of their healthcare need. This would skew their interpretation of patient acuity towards triaging patients to a higher level than is necessary.

Decision support tools are available in urgent and emergency care globally and often triage patients into five triage levels of acuity. Nevertheless, their derivation is often by expert opinion and designed to identify high acuity patients. There are tools specific to prehospital such as paramedic pathfinder and a new modification to MTS TTA for use by paramedics on scene. These have been evaluated in service to see if they can support these complex decisions and improve patient care. They have seen limited success with the pathfinder

triaging patients into higher acuity levels and the MTS (although performed better) not demonstrating that the paramedic would trust the triage category.

Computerised decision support tools have also been trialled to identify low acuity patients who have had a fall (CCDS). The tool showed a clear benefit without compromising safety and it was economically viable. The derivation of the model was not clearly described as it was evaluating its use in practice, which is often the final step in model development. The qualitative evaluation of its use though hinted at barriers to using computer support. For example, extra labour is involved on behalf of the clinician in order to categorise the acuity, which can lead to a decay in use. There have been promising signals from ED triage models that have used machine-learning methods to mitigate some of these issues. This includes automatic calculation based on electronic health fields and deriving the model using computer science and statistics, which are largely blinded to clinical judgement. Computerised clinical decision support is an underexplored mechanism for managing the increased demand and ensuring low acuity patients exit the urgent and emergency system at the most appropriate point.

There is an opportunity to develop a tool as more ambulance services are adopting electronic health records for which decision support could be embedded. The next chapter will systematically review the evidence to examine what machine learning derived CCDS tools have already been developed to triage patients entering urgent and emergency care, and which methods were the most successful.

Chapter 3

The Systematic Review of Machine Learning
Risk Prediction Models to Triage Emergency
Care Patients

3.1 Introduction

As the previous chapter alluded to in section 2.9.3, there is an opportunity with the emerging digital healthcare infrastructure to explore whether complex statistical and in silico modelling can triage emergency care patients. However, there are broad ranges of techniques that can be used to build predictive models. This chapter includes a systematic review that was conducted and published in BMC Diagnostic and Prognostic Research in 2020 as part of this thesis.¹⁰⁶ This creates an informed approach to algorithm selection when designing a model. The manuscript is included as published, with supporting information appearing afterwards. The full text can be found here:

<https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-020-00084-1>

3.2 Purpose of the review

This review was conducted to ascertain whether a successful model has already been developed in prehospital or emergency care. If this were found to be the case, this thesis would stand to externally validate such a model. The review was also designed to identify which methods had the function of triaging patient acuity when entering the emergency care system.

3.3 Published manuscript

Miles et al. *Diagnostic and Prognostic Research* (2020) 4:16
<https://doi.org/10.1186/s41512-020-00084-1>

Diagnostic and
Prognostic Research

RESEARCH

Open Access

Using machine-learning risk prediction models to triage the acuity of undifferentiated patients entering the emergency care system: a systematic review



Jamie Miles^{1*}, Janette Turner², Richard Jacques², Julia Williams³ and Suzanne Mason²

Abstract

Background: The primary objective of this review is to assess the accuracy of machine learning methods in their application of triaging the acuity of patients presenting in the Emergency Care System (ECS). The population are patients that have contacted the ambulance service or turned up at the Emergency Department. The index test is a machine-learning algorithm that aims to stratify the acuity of incoming patients at initial triage. This is in comparison to either an existing decision support tool, clinical opinion or in the absence of these, no comparator. The outcome of this review is the calibration, discrimination and classification statistics.

Methods: Only derivation studies (with or without internal validation) were included. MEDLINE, CINAHL, PubMed and the grey literature were searched on the 14th December 2019. Risk of bias was assessed using the PROBAST tool and data was extracted using the CHARMS checklist. Discrimination (C-statistic) was a commonly reported model performance measure and therefore these statistics were represented as a range within each machine learning method. The majority of studies had poorly reported outcomes and thus a narrative synthesis of results was performed.

Results: There was a total of 92 models (from 25 studies) included in the review. There were two main triage outcomes: hospitalisation (56 models), and critical care need (25 models). For hospitalisation, neural networks and tree-based methods both had a median C-statistic of 0.81 (IQR 0.80-0.84, 0.79-0.82). Logistic regression had a median C-statistic of 0.80 (0.74-0.83). For critical care need, neural networks had a median C-statistic of 0.89 (0.86-0.91), tree based 0.85 (0.84-0.88), and logistic regression 0.83 (0.79-0.84).

Conclusions: Machine-learning methods appear accurate in triaging undifferentiated patients entering the Emergency Care System. There was no clear benefit of using one technique over another; however, models derived by logistic regression were more transparent in reporting model performance. Future studies should adhere to reporting guidelines and use these at the protocol design stage.

(Continued on next page)

* Correspondence: Jamie.miles@nhs.net; j.miles@sheffield.ac.uk

¹Yorkshire Ambulance Service, Brindley Way, Wakefield WF2 0XQ, UK

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

(Continued from previous page)

Registration and funding: This systematic review is registered on the International prospective register of systematic reviews (PROSPERO) and can be accessed online at the following URL: https://www.crd.york.ac.uk/PROSPERO/display_record.php?ID=CRD42020168696

This study was funded by the NIHR as part of a Clinical Doctoral Research Fellowship.

Keywords: Ambulance service, Emergency department, Machine learning, Triage, Patients

Introduction

Rationale

Machine learning (ML) can be defined as ‘a set of methods that can automatically detect patterns in data, and then use uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty’ [1]. To date, ML has already proven effective at predicting outcomes for disease specific patients such as predicting bronchiolitis in infants and predicting whether trauma patients require a computerised tomography scan (CT) or have a cranio-cervical junction injury [2–4]. Other models have outperformed existing tools such as the Global Registry of Acute coronary Events (GRACE) and Thrombolysis In Myocardial Infarction (TIMI) risk tools at predicting cardiovascular risk [5, 6].

Initial triage at any stage of the Emergency Care System (ECS) has become challenging due to the increase in patients with varying levels of acuity [7]. Patients in a modern ECS have complex needs, which can often span mental health and social care [8].

Recently, there has been increased interest in combining ‘artificial intelligence’ with the Emergency Department for the purpose of initial triage [9–12]. However, this has been largely through the use of supervised learning algorithms, a sub-category of ML techniques [9]. The benefit of using these ML methods is they can identify non-linear relationships between candidate predictors and the outcome [11]. Furthermore, they can be embedded into electronic Patient Care Records (ePCR), removing the labour involved in triage and allowing for more complex models to be integrated [12].

The application of non-ML triage algorithms has previously led to the majority of patients being identified as mid-acuity. The Emergency Severity Index (ESI) is one such example [10, 11]. These triage systems can often have a clinical time-cost in their application [7]. In order for the benefits of triage algorithms to be actualised, the patient benefit at every acuity level has to be shortened. This means those with high acuity needs are treated quicker, those who are likely to be admitted are identified sooner and those with low acuity needs are discharged faster [10].

Clinical role for the index test

The index test under investigation in this systematic review is any triage model that is applied by a clinician at

the point of entry in the ECS. There are three possible entry points for patients. The first is when a patient calls the emergency medical service and is triaged by the Emergency Operations Centre (EOC). The second entry point is a face-to-face assessment by a paramedic on-scene. The third is on arrival to the Emergency Department (ED) [13]. A patient may enter at any of these points and also move through them all, being triaged multiple times. However the objective at each stage is the same: to stratify the acuity of an individual patient and allow the result to modify an ongoing care plan.

Objectives

The primary objective of this review is to assess the accuracy of machine learning methods in their application of triaging the acuity of patients presenting in the Emergency Care System (ECS).

Methods

This review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement. It is registered with PROSPERO (CRD42020168696).

Eligibility criteria

Population

All patients presenting to the ECS who require a process of triage to discern the immediacy of care. The population cannot be differentiated by clinical severity or condition prior to the application of the triage tool. This is due to the index test under investigation being able to be applied to all incoming patients. The population can be differentiated by demographic variables such as age, as it is recognised, there is a difference in service need between younger and older populations [14–19].

Intervention (index test)

Machine learning algorithms that have been used to derive and internally validate a decision support tool. This includes commonly used methods such as logistic regression. However, for this review, the application of logistic regression must extend to making predictions in future data and not just uncovering patterns. The restriction to only derivation and internal validation studies is to ensure the method under investigation is clearly

defined as opposed to an existing tool being externally validated in a subsequent population.

Comparison (reference test)

The reference test in this review is hierarchical. Preferably, there would be a decision support tool already used in the clinical setting identified in the paper as a comparator. In the absence of such, the study would include a clinician judgement. However, studies that have no comparator would also be accepted because derivation studies can often lack performance comparison with existing practice.

Outcome

For clarification, outcome has been divided into two parts. Prediction outcome and accuracy outcome.

Prediction outcome

To be included in this review, the outcome has to be a triage acuity outcome for emergency care. Each included study is aiming to make a prediction about how ill a patient is, or how urgent their care need is. Because the methods of how these predictions are developed is under investigation in this systematic review, the prediction outcome was allowed to be broadened in order to capture all relevant studies. This may be strictly a triage level (such as the Emergency Severity Index 5–level) or a surrogate outcome such as predicting the need for critical care or hospitalisation.

Model performance

For all the prediction models that have been included, their performance is described in terms of accuracy metrics reported in the final model performance. This includes discrimination (C-statistic), calibration (calibration plot, calibration slope, Hosmer-Lemeshow) and classification statistics (sensitivity, specificity, accuracy, positive predictive value (PPV), negative predictive value (NPV), likelihood ratio +/-). Some studies have used synonyms such as 'precision' instead of PPV, or 'recall' instead of sensitivity. For clarity in this review, all terms have been aligned to classification statistics identified in Steyerberg et al. [20].

Information sources

On the 14th December 2019, the Medical Literature Analysis and Retrieval System Online (MEDLINE), the Cumulative Index to Nursing and Allied Health Literature (CINAHL), PubMed and the grey literature were searched. This included Google scholar and the IEEE arXiv.

Search

A search strategy was developed through iteration and piloting. It was adapted from key words identified in the research questions and can be found in the [supplementary](#)

[material](#). The search strategy was used for MEDLINE, CINAHL and PubMed. This can be found in the [supplementary material](#).

The search strategy was for the last 10 years only. This is due to clinical contexts and computer capabilities being rapidly changing industries and thus older studies have a higher risk of being void or outdated. The search also encompassed only those studies presented in the English language. This is due to limited access to interpretation services. Any non-English language studies were excluded at the selection stage.

Study selection

Title screening was performed directly on source sites by JM and then exported to Endnote (version X9 for Windows) for abstract screening. This was subsequently fully second screened by JT. Then full text screening was performed by JM, with JT independently reviewing a random sample of 30% of the chosen included texts. Results were then compared with any disagreements being resolved by a third reviewer (SM). The data was selected from the studies retrieved during the searches using a visual schema transposed from the inclusion and exclusion criteria. This can be found in the [supplementary material](#). There were four stages of selection based on the screening results of the studies. The first involved a population assessment, ensuring the study was set in the emergency care system and the patients are not differentiated clinically. The second stage involved intervention screening and ensuring the candidate variables were measured at triage (entry point). The third stage involved method screening, which in turn was subdivided into two sections, the first ensuring that machine learning was used to derive the model, and the second, ensuring that the methodological outcome was accuracy in prediction. The final stage involved outcome screening, ensuring that each selected study was setting out to risk-stratify patients. There was co-author validation of the included articles.

Data collection process

Data was extracted using the Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS) checklist [21]. This was completed in Microsoft Excel (2016) by JM. The total spreadsheet was reviewed by RJ. Any disagreements were mediated by a third reviewer (JT). Data extracted for each included study is provided in the [supplementary material](#), as well as details regarding study quality assessment.

Risk of bias and applicability

Risk of bias and applicability was undertaken using the Prediction model Risk of Bias Assessment Tool (PROBAST) [22]. A template was accessed at http://www.probast.org/wp-content/uploads/2020/02/PROBAST_20190515.pdf

It was completed for each model by JM and then checked by RJ. Any disagreements were mediated by a third reviewer (JT).

Diagnostic accuracy measures

The principle diagnostic accuracy measures will be broadly covering three key areas. These are calibration, discrimination and classification of the final model within each study.

Synthesis of results

The included studies were too heterogeneous to undertake a robust meta-analysis; therefore, a narrative synthesis was performed. This centred on discrimination as the most reported summary statistic of model performance. Where derivation and internal validation results have been presented separately in a model, only the internally validated performance is included in this review and not the apparent performance.

The included models were sub-grouped by outcome, and further by method. Median and IQR was used to illustrate the spread of the C-statistics within each method. The analysis plan was informed by the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy [23].

Results

Study selection

All databases were searched on the 14th December 2019. There was a total of 712 studies identified from the database searching. This included 257 from MEDLINE, 298 from CINAHL and 150 from PubMed. Seven other sources from the grey literature were found. After title and abstract screening, 55 studies were taken through to eligibility screening. Thirty articles were excluded for the following reasons: 3 were external validation only, 6 were not machine learning, 2 were protocol only, 3 were studying the wrong population, 1 was a prognostic factor study, 13 had patients that were already triaged and 2 studies were not related to triage. This left a total of 25 studies included in this review. A PRISMA schematic diagram can be found below, and the PRISMA checklist can be found in the [supplementary material](#) [24]. Many studies investigated more than one machine learning technique, which meant that contained within the included studies was a total of 92 models to examine in this review (Fig. 1).

Study characteristics

The three most common methods were logistic regression (36 models), tree-based methods (23) and neural networks (20). Other models included support vector machines (6), Bayesian models (5), a K-nearest neighbour model and a unique artificial neuro-fuzzy inference system. Of the 92 models, there were only 13 that were set in the pre-hospital setting. The rest were set in the ED at the point

of triage. The two main outcomes that were being predicted by the studies were admission to hospital (53 models) or critical care outcome (28 models). Less common outcomes that appeared in these studies were predicting existing triage structures (9 models), and the prediction of whether a patient would be discharged from ED (3 models). Table 1 below summarises the key features of the included studies.

There were 44 models derived in the USA, 18 in Korea, 14 in Australia, 5 in Spain, 4 in India, 2 in Malaysia, 2 in Israel and 1 in Taiwan, Scotland and the Netherlands. Eighty-four models were purely retrospective using existing registry or cohort data. Only 4 models included data collection that was prospective and there were 4 models that did not include whether the data source was retrospective or prospective. Sixty-three models were derived using data from multiple sites, whilst 25 models were developed using a single centre. Four models failed to publish this information.

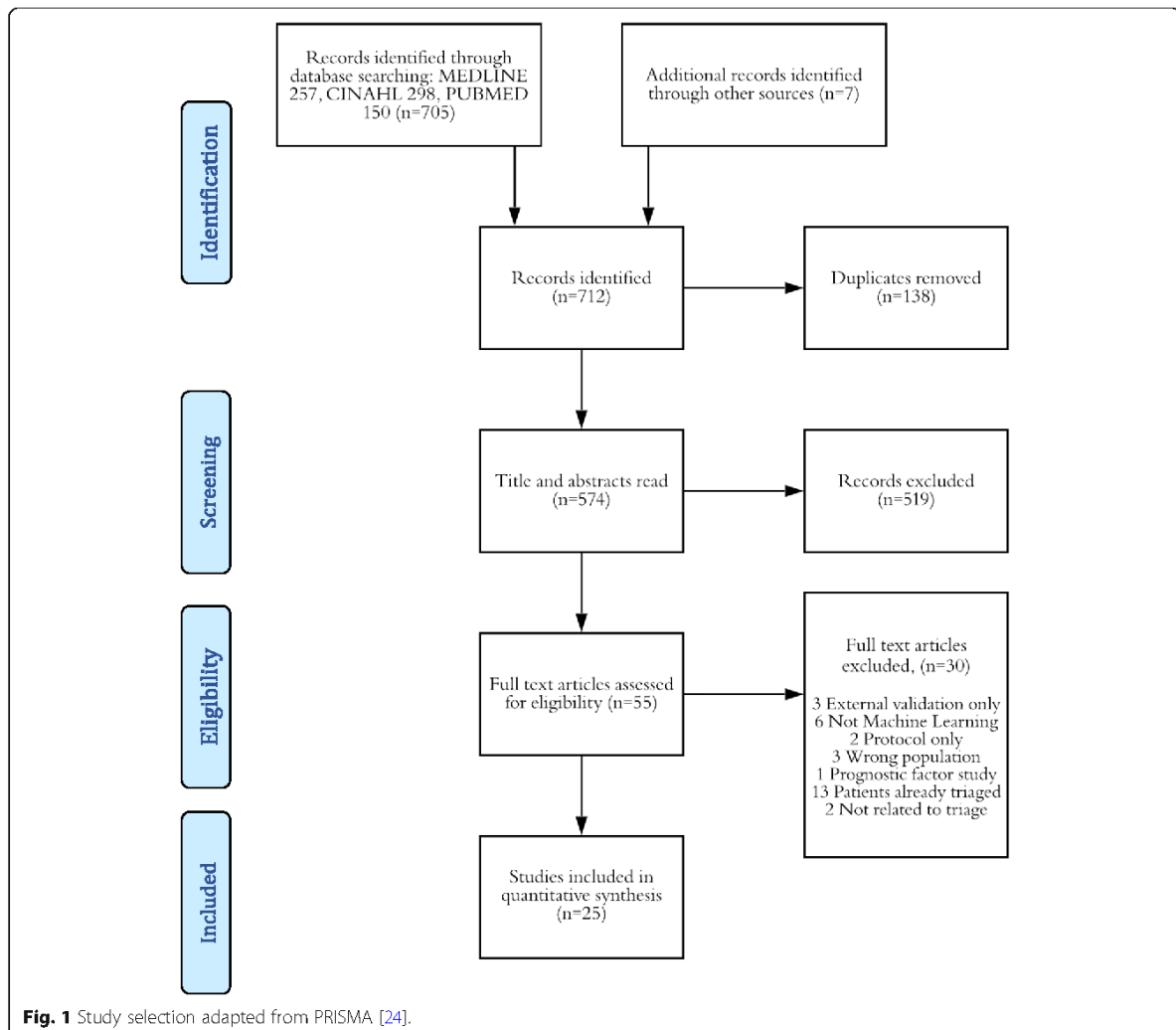
Risk of bias and applicability

There was a significant amount of incomplete reporting within the results. Only four models reported any calibration, mainly using the Hosmer-Lemeshow statistic [27, 44, 48]. One reported the *p* value of this, but not the statistic itself [27]. In terms of discrimination, there were 81 models that reported a concordance statistic (C-statistic), but of these, only 74 generated confidence intervals around this statistic. Only 47 models described classification statistics; however, these were incongruous between studies and only 1 study included the amount of true positive, true negative, false positive and false negative results. This makes it unfeasible to meta-analyse models which share the same population and outcome. A summary of the PROBAST assessment can be found in Fig. 2 and was adapted from Debray et al. [50]. When applying the PROBAST tool, there were only three studies which could be considered a low risk of bias [31, 33, 42]. This limits the benefit of grouping high vs low risk of bias studies. Most studies had low applicability concern, except for six studies [26, 30, 38, 41, 46, 49].

Synthesis of results

Hospitalisation outcome

There was a total of 56 models which were predicting whether the patient was likely to be hospitalised as the outcome. Of these, 27 used logistic regression (two used the LASSO penalty term). Twelve studies used a neural network, 10 used a tree-based design, 3 Bayesian methods, 3 support vector machine models and one K-nearest neighbour. Only three models reported calibration in this outcome group [28, 48]. The most reported result was model discrimination using the C-statistic (also known as the area under the ROC curve, or AUC). Whilst the heterogeneity



between models is too severe to undertake a meta-analysis, it was possible to cluster results by outcome and method. Figure 3 illustrates which machine learning methods were most able to differentiate between those with a positive outcome and those with a negative. The size of the data points is a normalised transformation of the sample size used to derive each model. Neural networks and tree-based methods both had a median *C*-statistic of 0.81 with their interquartile ranges (IQR) being 0.80-0.84 and 0.79-0.82 respectively. This compares to logistic regression which had a median *C*-statistic of 0.80 (IQR 0.74-0.83). The larger sample sizes generated smaller *C*-statistics. The three support vector machine models did not report the *C*-statistic. Classification was poorly reported with only 19 models publishing sensitivity and specificity, and only 10 of these also reporting confidence intervals. Twenty-one models reported accuracy, but only four of these had

confidence intervals. Please refer to the CHARMS supplement for more details.

Critical illness

There were 28 models that used critical illness as an outcome measure. Eleven were logistic regression (one with LASSO penalty), 11 were tree-based and 6 were neural networks. There was an incongruity with the precise definition of critical illness, Table 2 highlights the differences within the definitions. Only one model in this group reported any calibration. They found that deep neural networks were the most discriminate with a *C*-statistic of 0.89 (95% CI 0.88-0.89). This compared to logistic regression and random forest modelling which both had the same result of 0.87 (95% CI 0.86-0.87).

The most common statistic was the *C*-statistic for discrimination. Figure 4 illustrates which methods were

Table 1 Study characteristics

| Author | Year | Country | Population | Outcome | Methods used | Predictors | Sample size | EPV | Method of testing |
|----------------------------|------|-------------|--------------|-----------------------------------|--------------------------|------------|-------------|-----------|---|
| Azeez et al. [25] | 2014 | Malaysia | ED | Triage level | NN, ANFIS | 20 | 2223 | | Random split sample (70:30) |
| Caicedo-Torres et al. [26] | 2016 | Spain | ED | Discharge | LR, SVM, NN | 147 | 1205 | | Random split sample (80:20), 10-fCV |
| Cameron et al. [27] | 2015 | Scotland | ED | Hospitalisation | LR | 9 | 215231 | | Random split sample (66:33), bootstrapping (10,000) |
| Dinh et al. [28] | 2016 | Australia | ED | Hospitalisation | LR | 10 | 860832 | 9470 | Random split sample (50:50) |
| Dugas et al. [29] | 2016 | USA | ED | Critical illness | LR | 9 | 97000000 | 525 | Random split sample (90:10), 10f-CV |
| Golmohammadi [30] | 2016 | USA | ED | Hospitalisation | LR, NN | 8 | 7266 | 460.25 | Split sample (70:30) |
| Goto et al. [31] | 2019 | USA | ED | Critical illness, hospitalisation | LR, LASSO, RF, GBDT, DNN | 5 | 52037 | 32.60 | Random split sample (70:30) |
| Hong et al. [32] | 2018 | USA | ED | Hospitalisation | LR, GBDT, DNN | 972 | 560486 | 171.44 | Random split sample (90:10) |
| Kim, D et al. [33] | 2018 | Korea | Prehospital | Critical illness | LR, RF, DNN | 5 | 460865 | 3583.60 | 10f-CV |
| Kim, S et al. [34] | 2014 | Australia | ED | Hospitalisation | LR | 8 | 100123 | 1074.86 | Apparent performance |
| Kwon et al. (1) [35] | 2018 | Korea | ED | Critical illness, hospitalisation | DNN, RF | 7 | 10967518 | 133667.89 | Split sample (50:50), + external validation dataset |
| Kwon et al. (2) [36] | 2019 | Korea | ED | Critical illness, hospitalisation | DNN, RF, LR | 8 | 2937078 | 14047.57 | Split sample (50:50) |
| Levin et al. [37] | 2018 | USA | ED | Critical illness, hospitalisation | RF | 6 | 172726 | 56.74 | Random split sample (66:33), bootstrapping |
| Li et al. [38] | 2009 | USA | Pre-hospital | Hospitalisation | LR, NB, DT, SVM | 6 | 2784 | | 10f-CV |
| Meisel et al. [39] | 2008 | USA | Pre-hospital | Hospitalisation | LR | 9 | 401 | | Bootstrap resampling (1000) |
| Newgard et al. [40] | 2013 | USA | Prehospital | Critical illness | CART | 40 | 89261 | | Cross-validation |
| Olivia et al. [41] | 2018 | India | ED | Triage level | DT, SVM, NN, NB | 8 | | | 10f-CV |
| Raita et al. [42] | 2019 | USA | ED | Critical illness, hospitalisation | LR, LASSO, RF, GBDT, DNN | 6 | 135470 | 107 | Random split sample (70:30) |
| Rendell et al. [43] | 2019 | Australia | ED | Hospitalisation | B, DT, LR, NN, NB, KNN | 11 | 1721294 | 5521 | 10f-CV |
| Seymour et al. [44] | 2010 | USA | Prehospital | Critical illness | LR | 12 | 144913 | 156 | Random split sample (60:40) |
| van Rein et al. [45] | 2019 | Netherlands | Prehospital | Critical illness | LR | 48 | 6859 | 3.4375 | Separate external validation |
| Wang et al. [46] | 2013 | Taiwan | ED | Triage level | SVM | 6 | 3000 | | 10f-CV |
| Zhang et al. [47] | 2017 | USA | ED | Hospitalisation | LR, NN | 25 | 47200 | 91.8 | 10f-CV |
| Zlotnik et al. [48] | 2016 | Spain | ED | Hospitalisation | NN | 9 | 153970 | 614.5 | 10f-CV |
| Zmiri et al. [49] | 2012 | Israel | ED | Triage level | NB, C4.5 | 4 | 402 | | 10f-CV |

ANFIS Adaptive Neuro-Fuzzy Inference System, B Bayesian Network, CART Classification and Regression Tree, DT Decision Tree, DNN Deep Neural Network, EPV Events Per Variable, GBDT Gradient Boosted Decision Tree, KNN K-Nearest Neighbours, LR logistic regression, LASSO Least Absolute Shrinkage and Selection Operator, NB Naive Bayes, NN Neural Network, RF Random Forest, SVM Support Vector Machine

most discriminative at predicting a critical care outcome. As above, the sample size is represented by the size of the data point. Neural networks had a median of 0.89 (IQR 0.87-0.90) tree based had a median of 0.85 (IQR

0.84-0.88) and logistic regression had a median of 0.83 (IQR 0.80-0.85).

There were only 10 models from two studies that included classification metrics such as sensitivity and

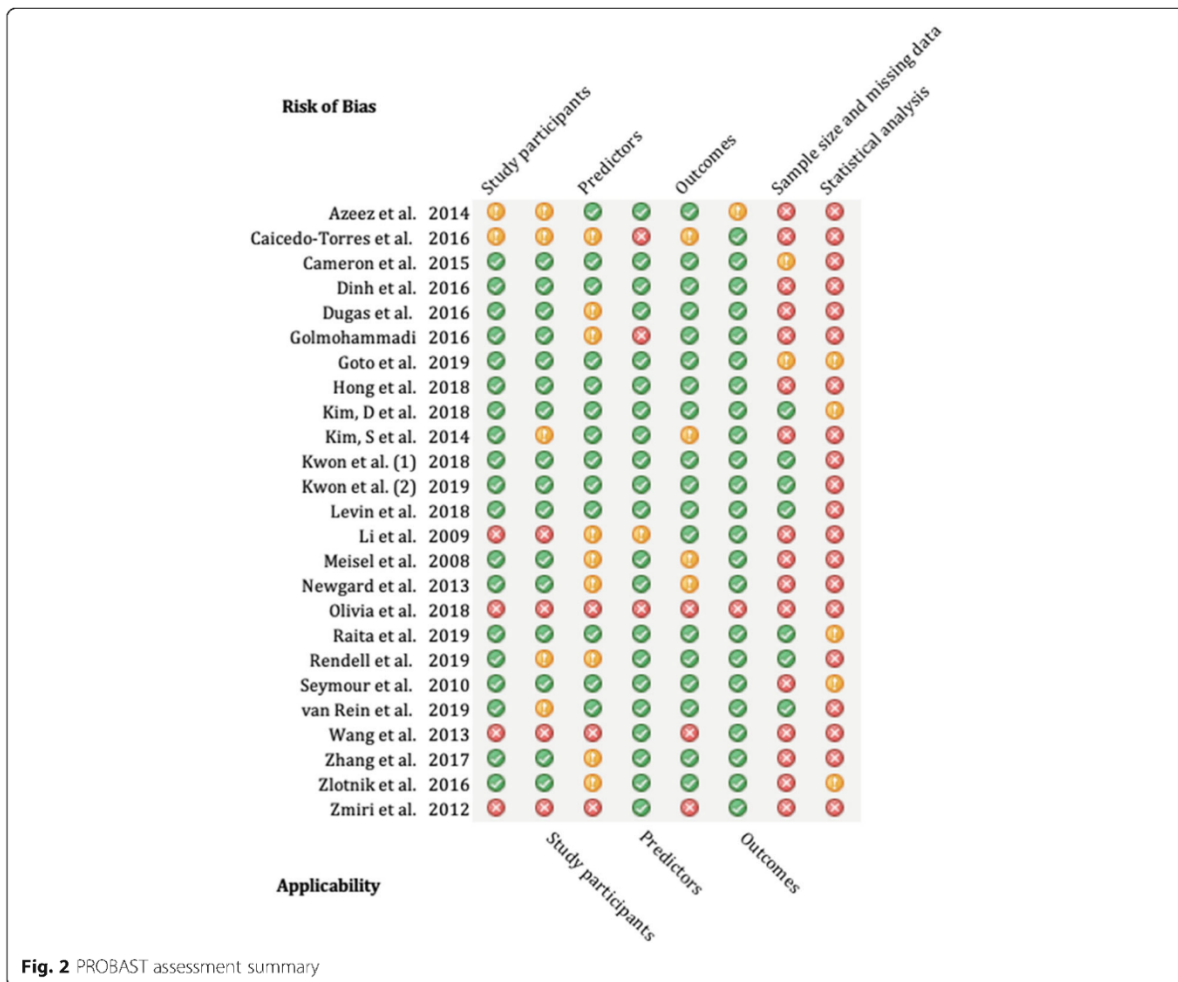


Fig. 2 PROBAST assessment summary

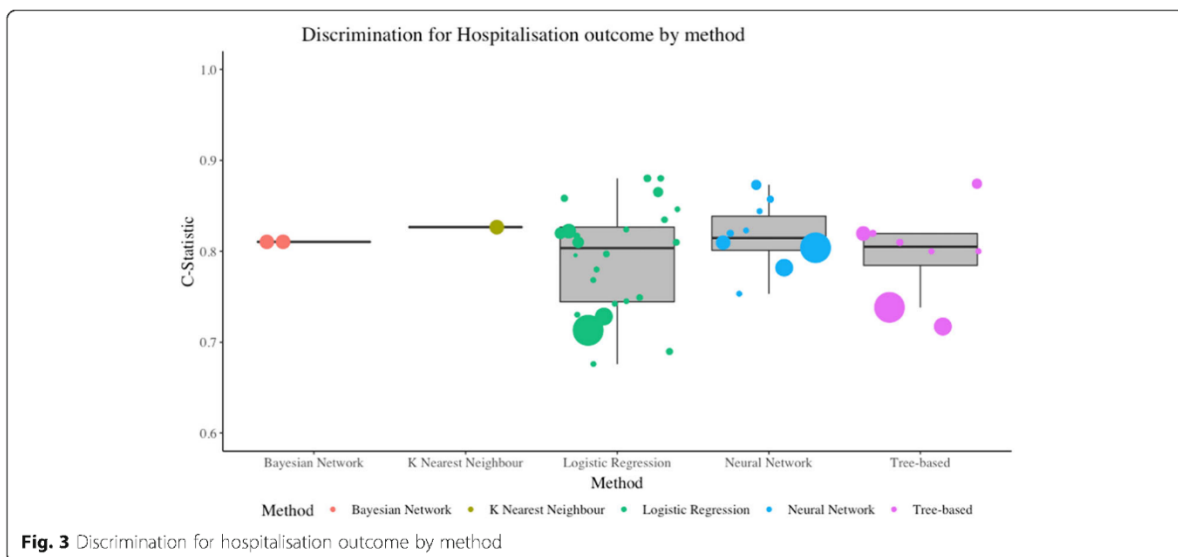


Fig. 3 Discrimination for hospitalisation outcome by method

Table 2 Critical care outcome definitions between studies

| Study | Direct ICU | Death | Direct theatre | Direct pPCI | Severe sepsis | Mechanical intervention | ISS > 15 | ISS > 16 |
|-----------------|------------|-------|----------------|-------------|---------------|-------------------------|----------|----------|
| Dugas et al. | ✓ | ✓ | ✓ | ✓ | | | | |
| Goto et al. | ✓ | ✓ | | | | | | |
| Kim D et al. | | ✓ | | | | | | |
| Kwon et al. | ✓ | ✓ | | | | | | |
| Kwon et al. (2) | ✓ | | | | | | | |
| Levin et al. | ✓ | ✓ | | | | | | |
| Newgard et al. | | | | | | | | ✓ |
| Raita et al. | ✓ | ✓ | | | | | | |
| Seymour et al. | | ✓ | | | ✓ | ✓ | | |
| van Rein et al. | | | | | | | ✓ | |

ICU Intensive Care Unit, pPCI primary Percutaneous Coronary Intervention, ISS injury severity score

specificity with their associated confidence intervals [31, 42]. This makes comparison limited.

Discharge outcome

Three models from a single study used discharge related outcome measures [26]. The study focussed on predicting patients that would be discharged from the ED, and diverting them to a fast track service. They used logistic regression, support vector machines and neural networks for comparison. They did not report discrimination and only reported limited classification statistics [26]. They found that the neural network had the most precise estimates with a PPV (0.85) compared to the support vector machine and logistic regression (0.83 and 0.82). However, when examining the reported F1 score (PPV* sensitivity/PPV + sensitivity), logistic regression reported the most accurate estimate with an F1 score of 0.85, compared to the support vector machine (0.82) and the neural network (0.82).

Triage level outcome

Three studies that used machine learning to stratify patients into existing triage tools, all of which had a high risk of bias [25, 46, 49]. One focused on the Objective Primary Triage Scale (OPTS) in Malaysia [25]. This is a three tiered triage scale of emergent, urgent and non-urgent. They used neural networks and an artificial neuro-fuzzy inference system (ANFIS) to make predictions. There was no model calibration performed and the C-statistic did not have any confidence intervals. They did report accuracy and PPV for both methods and found the neural network had an accuracy of 0.84 (PPV 0.87) which was better performing than the ANFIS method (accuracy 0.6, PPV 0.61) [25]. Two studies used a local four level triage scale [46, 49]. One used Support Vector Machines with a Principle Component Analysis and a back propagated neural network, reporting an accuracy of 1.0 and 0.97 respectively [46]. The results in this study are likely to be biased. The

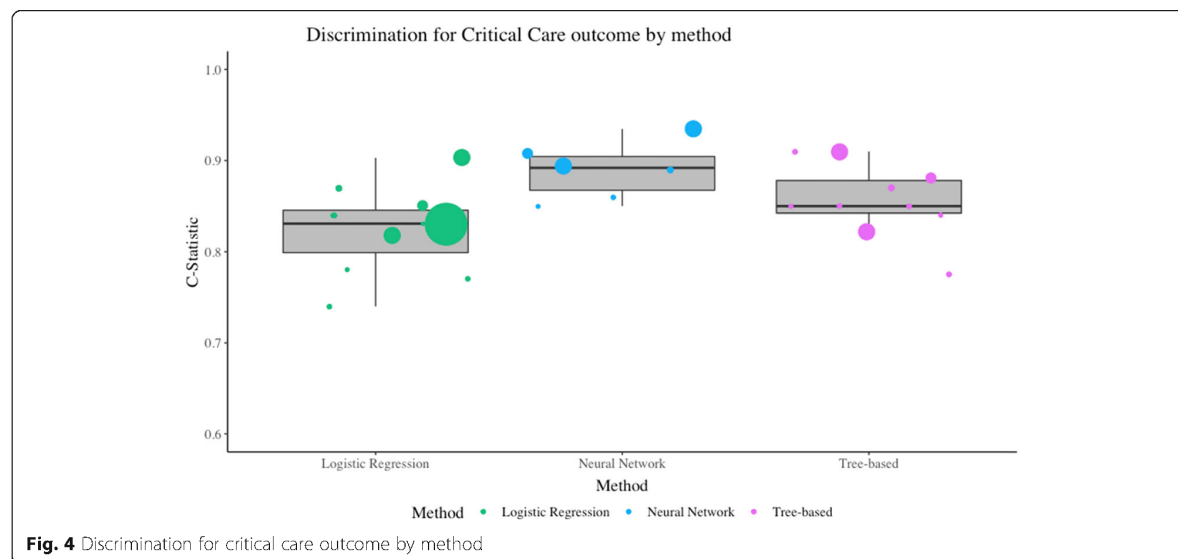


Fig. 4 Discrimination for critical care outcome by method

other study which examined a four-tiered triage scale used a naïve Bayes and a C4.5 tree-based classifier [49]. They only reported accuracy; however, they found that when they simplified the scale to be two grades, both models had higher prediction (average accuracy 71.37) than when it was four grades (52.94).

Discussion

Summary of included studies

In the last 10 years, there has been an increase in the number of prediction models that have utilised already existing methods in statistics and computer science. This may be due to the widespread availability of data worldwide. This systematic review identified 25 studies which aimed to derive a risk prediction model for triaging the acuity of undifferentiated patients in the emergency care system. The most common method was logistic regression with 36 models, but this was followed closely by both tree-based methods and neural networks. Most studies used hospital admission as an outcome for prediction. The objective of this review was to assess the accuracy of different machine learning methods. This was challenging due to differences in reporting how models were developed and evaluated. Furthermore, the reporting of the majority of models did not give enough information on model development, validation and performance which makes a critical appraisal difficult and a meta-analysis of accuracy stratified by the method almost impossible.

There have been common pitfalls amongst the included studies which will be discussed including the reference standard, the handling of candidate variables, and the analysis of performance.

The reference standard

In evaluating the performance of a diagnostic model, it is important to compare the index test (the new model) with a 'gold standard', known as the reference standard. In practice, this could be subjective such as a clinician making a decision or deciding a triage level. Alternatively, it could be an objective standard such as an ICD-10 classification of disease, mortality or a clearly defined event [51].

Most studies that are determining the cross-sectional acuity of any given patient in emergency care have subjective reference standards. To illustrate, the Emergency Severity Index (ESI) 5-level triage is almost exclusively subjective and depends on the clinician undertaking the triage. A limitation of using this as a reference standard is inter-rater reliability can widely vary. A meta-analysis has shown that the inter-rater reliability of the ESI had an unweighted kappa of 0.786 (95% CI 0.745-0.821) [52]. Using subjective reference standards could lead to inherent problems maintaining the accuracy when transporting the model.

In contrast, Liu et al. undertook a study predicting cardiac arrest within 72 h of ED attendance [53]. A cardiac arrest is an empirical outcome measure and can be defined as "the abrupt loss of heart function in a person who may or may not have been diagnosed with heart disease" [54]. Liu prospectively collected data on 1386 participants and recorded whether or not they had a cardiac arrest within 72 h. In this example, the reference standard is a clearly defined outcome, which is not open for interpretation or subjectivity, and thus would provide a reliable benchmark to compare a derived model.

Handling of candidate variables

Prior to developing a diagnostic model, it is important to consider which variables in the data are candidates for the final model. These candidate variables can be identified not only through subject knowledge or literature searching but also through statistical methods of examining the distribution or weighting [20]. A common problem with the included studies was how they reported the identification of candidate variables. Fifteen of the included studies provided a clear rationale, with data available at triage being the most common reason. Two studies used all the variables in the dataset, and eight studies provided no rationale at all.

It is also important to rationalise why there is a need to transform continuous variables given that it can be statistically inappropriate when developing prognostic models and leads to a significant loss of information [55]. Only 6 studies kept variables in their original format, whilst the remaining studies either categorised the variables (such as age) or did not describe the variables in a level of detail that an assessment could be made. Furthermore, no study elaborated on the linearity of the continuous variables and reported how they would model non-linear relationships (such as using fractional polynomials or restricted cubic splines) [56].

One of the benefits of using machine learning is the ability of performing feature selection during analysis [1]. The methods of undertaking feature selection can vary according to method, but the principle is beneficial to creating a simple model that can be embedded into practice. Methods such as deep neural networks can allow for fitting complex non-linear relationships through their architecture. The more hidden layers, the more complex the relationships. Univariable screening is not recommended as it does not account for any important collinearities between other candidate variables [57]. Despite this, it was used in 5 of the included studies.

Reporting

The concordance statistic (C-statistic) was the most commonly reported and appeared in 81 out of 97 models. The C-statistic evaluates how discriminative a model is. For example, if a pair of subjects were selected at random (one with the outcome and one without), how often would the

model classify both subjects correctly [58]. There were no significant differences in discrimination between methods, and all reported a range of C-statistics that performed well (above 0.7). However, reporting how discriminative a model is does not provide a full picture and the performance of the model should account for calibration. This is an assessment of accuracy or more specifically, how well the predictions matched the observed outcomes in the data [56]. If studies only report discrimination, then it does not help troubleshoot poor performance in a transported model. This is when the model is adopted in a new setting, such as a new hospital, or new country. Only five models reported any calibration, and two of these used the Hosmer-Lemeshow statistic [44, 48]. This is prone to poor interpretability and can be sensitive to sample size and grouping [56]. With machine learning methods, miscalibration can be adjusted when transporting a model to a different setting. Further ways to present accuracy are classification statistics. These include accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and likelihood ratios.

No studies reported classification statistics in full. If they had published the true positive, false positive, true negative and false negative results for their model performance, a meta-analysis could have been performed [23].

Nearly all the studies had the potential of a high risk of bias due to the results being incomplete. More information is needed in order to make a robust judgement. The PROBAST statement recommends transparency in reporting and the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) gives clear guidance on how to achieve this. Even though machine learning can be perceived as 'black box', this axiom is not entirely true. For example, DNN can obtain a matrix of parameter values and this can then be subsequently transformed into the ranking of variable importance. The reporting of model performance can still be generated [59, 60].

Limitations in this review

This review identified and appraised all available literature; however, it did not directly contact authors for original data or further statistics. As such, the level of missing data in reporting which prevented the generation of a summary statistic remained throughout. This also had an impact on the risk of bias assessment. The Excerpta Medica database (EMBASE) was not used in this review as it was deemed too similar to MEDLINE.

Conclusion

This systematic review has found that machine learning methods such as neural networks, tree-based, and logistic regression designs appear equal at triaging undifferentiated patients. However, the inconsistency and absence of

information has significant implications on the risk of bias in all studies. Therefore no definitive answer can be drawn about the most accurate method. Future studies need to conform to reporting guidelines to ensure transparency and integrity of the models.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s41512-020-00084-1>.

Additional file 1:

Abbreviations

ANFIS: Adaptive Neuro-Fuzzy Inference System; AUC: Area under curve; B: Bayesian network; CART: Classification and regression trees; CHAR MS: Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies; CINAHL: The Cumulative Index to Nursing and Allied Health Literature; CT: Computerised tomography; DNN: Deep neural network; DT: Decision tree; ECS: Emergency Care System; ED: Emergency Department; EOC: Emergency Operations Centre; EPCR: Electronic patient care record; EPV: Events per variable; ES: Emergency severity index; GBDT: Gradient boosted decision tree; GRACE: Global Registry of Acute Coronary Event; ICD-10: International Classification of Disease-10; IQR: Interquartile range; KNN: K-nearest neighbours; LR: Logistic regression; MEDL INE: Medical Literature Analysis and Retrieval System; ML: Machine learning; NIHR: National Institute for Health Research; NN: Neural network; NPV: Negative predictive value; OPTS: Objective Primary Triage Scale; PPV: Positive predictive value; PROBAST: Prediction model Risk Of Bias Assessment Tool; PROSPERO: The International Prospective Register of Systematic Reviews; RF: Random Forest; ROC: Receiver operating characteristic; SVM: Support vector machine; TIMI: Thrombolysis in Myocardial Infarction; TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

Acknowledgements

Not applicable

Authors' contributions

JM developed the search strategy, ran the search, did initial screening, analysed the data and was the primary author of the manuscript. JT second screened the data and reviewed eligible studies, also steered the clinical perspective of the project. RJ co-developed the statistical analysis plan and reviewed the analysis. JW shaped the clinical importance of the review and checked the clinical context of the included studies. SM provided oversight for the review, steered the search strategy to ensure clinical context was maintained. The authors read and approved the final manuscript.

Authors' information

JM is a paramedic and currently leading a project called the Safety InDEx of Prehospital On Scene Triage (SINEPOST) which is a prediction study aiming to derive and internally validate a model which can help paramedics make transport decisions whilst still with the patient. For more information on this project, and the lead author visit: <https://www.sheffield.ac.uk/scharr/people/pgr-students/jamie-miles>

Funding

This study was funded by Health Education England and the National Institute of Health Research. The funders have not inputted into the collection, analysis or interpretation of data in writing this manuscript. This report is independent research supported by Health Education England and the National Institute of Health Research (HEE/NIHR ICA Programme Clinical Doctoral Research Fellowship, Mr. Jamie Miles, ICA-CDRF-2018-04-ST2-044). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests

Author details

¹Yorkshire Ambulance Service, Brindley Way, Wakefield WF2 0XQ, UK. ²School of Health and Related Research, 3rd Floor, Regent Court (ScHARR), 30 Regent Street, Sheffield S1 4DA, UK. ³University of Herfordshire, Hatfield, Herfordshire, UK

Received: 30 July 2020 Accepted: 11 September 2020

Published online: 02 October 2020

References

- Murphy KP. Machine learning: a probabilistic perspective. London: the MIT press; 2012.
- Bektas F, Eken C, Soyuncu S, Kilicaslan I, Cete Y. Artificial neural network in predicting craniocervical junction injury: an alternative approach to trauma patients. *Eur J Emerg Med*. 2008 Dec;15(6):318–23.
- Walsh P, Cunningham P, Rothenberg SJ, O'Doherty S, Hoey H, Healy R. An artificial neural network ensemble to predict disposition and length of stay in children presenting with bronchiolitis. *Eur J Emerg Med*. 2004;11(5):259–64.
- Molaei S, Korley FK, Soroushmehr SMR, Falk H, Sair H, Ward K, et al. A machine learning based approach for identifying traumatic brain injury patients for whom a head CT scan can be avoided. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS. Institute of Electrical and Electronics Engineers Inc; 2016. p. 2258–61.
- Vanhouten JP, Stammer JM, Lorenzi NM, Maron DJ, Lasko TA. Machine learning for risk prediction of acute coronary syndrome.
- Harrison RF, Kennedy RL. Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation. *Ann Emerg Med*. 2005 Nov 1;46(5):431–9.
- Weber EJ. Triage: Making the simple complex? *Emerg Med J*. 2018;36(2):64–5.
- O'Keeffe C, Mason S, Jacques R, Nicholl J. Characterising non-urgent users of the emergency department (ED): a retrospective analysis of routine ED data. *PLoS One*. 2018;13(2):1–14.
- Stewart J, Sprivilis P, Dwivedi G. Artificial intelligence and machine learning in emergency medicine. *Emerg Med Australas* [Internet]. 2018 Dec;30(6): 870–4. Available from: <http://doi.wiley.com/10.1111/1742-6723.13145>.
- Berlyand Y, Raja AS, Dorner SC, Prabhakar AM, Sonis JD, Gottumukkala R V, et al. How artificial intelligence could transform emergency department operations. *Am J Emerg Med* [Internet]. 2018 Aug;36(8):1515–7. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0735675718300184>.
- Grant K, McParland A. Applications of artificial intelligence in emergency medicine. *Univ Toronto Med J*. 2019;96(1):37–9.
- Liu N, Zhang Z, Wah Ho AF, Ong MEH. Artificial intelligence in emergency medicine. *J Emerg Crit Care Med*. 2018;2(4):82–82.
- Aacharya RP, Gastmans C, Denier Y. Emergency department triage: an ethical analysis. *BMC Emerg Med*. 2011 Oct 7;11:16.
- Brousseau DC, Hoffmann RG, Nattinger AB, Flores G, Zhang Y, Gorelick M. Quality of primary care and subsequent pediatric emergency department utilization. *Pediatrics*. 2007 Jun 1;119(6):1131–8.
- Simpson R, Croft S, O'Keeffe C, Jacques R, Stone T, Ahmed N, et al. Exploring the characteristics, acuity and management of adult ED patients at night-time. *Emerg Med J*. 2019 Sep 1;36(9):554–7.
- McCusker J, Karp I, Cardin S, Durand P, Morin J. Determinants of emergency department visits by older adults: a systematic review. *Acad Emerg Med* [Internet]. 2003 Dec 1 [cited 2020 Mar 6];10(12):1362–70. Available from: [http://doi.wiley.com/10.1197/S1069-6563\(03\)00539-6](http://doi.wiley.com/10.1197/S1069-6563(03)00539-6).
- Latham LP, Ackroyd-Stolarz S. Emergency department utilization by older adults: a descriptive study. *Can Geriatr J*. 2014 Dec 1;17(4):118–25.
- Lehmann CU, Barr J, Kelly PJ. Emergency department utilization by adolescents. *J Adolesc Heal*. 1994;15(6):485–90.
- Ziv A, Boulet JR, Slap GB. Emergency department utilization by adolescents in the United States. *Pediatrics*. 1998 Jun 1;101(6):987–94.
- Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating [Internet]. New York, NY: Springer New York; 2009. (Statistics for Biology and Health). Available from: <http://link.springer.com/10.1007/978-0-387-77244-8>.
- Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* [Internet]. 2014 Oct 14 [cited 2020 Mar 1];11(10):e1001744. Available from: <https://dx.plos.org/10.1371/journal.pmed.1001744>.
- Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019 Jan 1;170(1):51–8.
- Macaskill P, Gatsonis C, Deeks J, Harbord R, Takwoingi Y. Analysing and presenting results. In: The Cochrane collaboration, editor. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* [Internet]. Version 1. 2010. Available from: <http://srdta.cochrane.org/>.
- Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* [Internet]. 2009 Jul 21 [cited 2020 Mar 6];6(7):e1000097. Available from: <https://dx.plos.org/10.1371/journal.pmed.1000097>.
- Azeez D, Ali MAM, Gan KB, Saiboon I. Comparison of adaptive neuro-fuzzy inference system and artificial neural networks model to categorize patients in the emergency department. *Springerplus*. 2013;2(1):1–10.
- Caicedo-Torres W, Hernando Pinzon G. A machine learning model for triage in lean paediatric emergency departments. Montes y Gómez M, Escalante HJ, Segura A, Murillo J de D, editors. 2016;10022(November 2016):259–70. Available from: <http://link.springer.com/10.1007/978-3-319-47955-2>.
- Cameron A, Rodgers K, Ireland A, Jamdar R, McKay GA. A simple tool to predict admission at the time of triage. *Emerg Med J* [Internet]. 2015;32(3): 174–9. Available from: <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med11&NEWS=N&AN=24421344>.
- Dinh MM, Russell SB, Bein KJ, Rogers K, Muscatello D, Paoloni R, et al. The Sydney Triage to Admission Risk Tool (START) to predict emergency department disposition: a derivation and internal validation study using retrospective state-wide data from New South Wales, Australia. *BMC Emerg Med* [Internet]. 2016;16(1):1–7 Available from: <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med12&NEWS=N&AN=27912757>.
- Dugas AF, Kirsch TD, Toerper M, Korley F, Yenokyan G, France D, et al. An electronic emergency triage system to improve patient distribution by critical outcomes. *J Emerg Med*. 2016;50(6):910–8.
- Golmohammadi D. Predicting hospital admissions to reduce emergency department boarding. *Int J Prod Econ*. 2016;182(September):535–44.
- Goto T, Carnago CAJ, Faridi MK, Freishtat RJ, Hasegawa K. Machine learning-based prediction of clinical outcomes for children during emergency department triage. *JAMA Netw open* [Internet]. 2019;2(1): e186937. Available from: <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=pem&NEWS=N&AN=30646206>.
- Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency department triage using machine learning. *PLoS One* [Internet]. 2018;13(7):1–13 Available from: <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med1&NEWS=N&AN=30028888>.
- Kim D, You S, So S, Lee J, Yook S, Jang DP, et al. A data-driven artificial intelligence model for remote triage in the prehospital environment. *PLoS One*. 2018;13(10):1–14.
- Kim SW, Li JY, Hakendorf P, Teubner DJJO, Ben-Tovim DI, Thompson CH. Predicting admission of patients by their presentation to the emergency department. *EMA - Emerg Med Australas*. 2014 Aug;26(4):361–7.
- Kwon Jmyoung, Lee YY, Lee YY, Lee S, Park H, Park J. Validation of deep-learning-based triage and acuity score using a large national dataset. *PLoS One* [Internet]. 2018;13(10):1–10. Available from: <https://doi.org/10.1371/journal.pone.0205836>.
- Kwon J, Jeon K-H, Lee M, Kim K-H, Park J, Oh B-H. Deep learning algorithm to predict need for critical care in pediatric emergency departments. *Pediatr Emerg Care* [Internet]. 2019 Jul;1. Available from: <http://insights.ovid.com/crossref?an=00006565-900000000-98117>.

37. Levin S, Toerper M, Hamrock E, Hinson JS, Barnes S, Gardner H, et al. Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index. *Ann Emerg Med*. 2018;71(5):565-574.e2.
38. Li J, Guo L, Handly N. Hospital admission prediction using pre-hospital variables. 2009 IEEE Int Conf Bioinforma Biomed BIBM 2009. 2009;283-286.
39. Meisel ZF, Pollack CV, Mechem CC, Pines JM. Derivation and internal validation of a rule to predict hospital admission in prehospital patients. *Prehospital Emerg Care*. 2008;12(3):314-9.
40. Newgard CD, Hsia RY, Mann NC, Schmidt T, Sahni R, Bulger EM, et al. The trade-offs in field trauma triage. *J Trauma Acute Care Surg* [Internet]. 2013; 74(5):1298-306 Available from: <http://insights.ovid.com/crossref?an=01586154-201305000-00017>.
41. Olivia D, Nayak A, Balachandra M. Machine learning based electronic triage for emergency department. In 2018. p. 215-21. Available from: http://link.springer.com/10.1007/978-981-13-2907-4_19.
42. Raita Y, Goto T, Faridi MK, Brown DFM, Camargo CA, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit Care* [Internet]. 2019 22;23(1):64. Available from: <https://ccforum.biomedcentral.com/articles/10.1186/s13054-019-2351-7>.
43. Rendell K, Koprinska I, Kyrme A, Ebker-White AA, Dinh MM. The Sydney Triage to Admission Risk Tool (START2) using machine learning techniques to support disposition decision-making. *EMA - Emerg Med Australas*. 2019; 31(3):429-35.
44. Seymour CW, Kahn JM, Cooke CR, Watkins TR, Rea TD. During out-of-hospital emergency care. 2010;304(7):747-54.
45. van Rein EAJ, van der Sluijs R, Voskens FJ, Lansink KWW, Houwert RM, Lichtveld RA, et al. Development and Validation of a Prediction Model for Prehospital Triage of Trauma Patients. *JAMA Surg* [Internet]. 2019;154(5): 421-9 Available from: <http://search.ebscohost.com/login.aspx?direct=true&db=cin20&AN=136501962&site=ehost-live>.
46. Wang S-T. Construct an optimal triage prediction model: a case study of the emergency department of a teaching hospital in Taiwan. *J Med Syst* [Internet]. 2013; 29(5):9968. Available from: <http://link.springer.com/10.1007/s10916-013-9968-x>.
47. Zhang X, Kim J, Patzer RE, Pitts SR, Patzer A, Schragger JD. Prediction of emergency department hospital admission based on natural language processing and neural networks. *Methods Inf Med* [Internet]. 2017;56(05): 377-89 Available from: <http://www.thieme-connect.de/DOI/DOI?10.3414/ME17-01-0024>.
48. Zlotnik A, Alfaro MC, Pérez MCP, Gallardo-Antolín A, Martínez JMM. Building a decision support system for inpatient admission prediction with the Manchester triage system and administrative check-in variables. *CIN Comput Informatics, Nurs* [Internet]. 2016 May;34(5):224-30. Available from: <http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00024665-201605000-00006>.
49. Zmiri D, Shahar Y, Taieb-Maimon M. Classification of patients by severity grades during triage in the emergency department using data mining methods. *J Eval Clin Pract*. 2012;18(2):378-88 Available from: <http://doi.wiley.com/10.1111/j.1365-2753.2010.01592.x>.
50. Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ*. 2017;356.
51. Takwoingi Y, Quinn TJ. Review of diagnostic test accuracy (DTA) studies in older people. *Age Ageing*. 2018;47(3):349-55.
52. Mirhaghi A, Heydari A, Mazlom R, Hasanzadeh F. Reliability of the emergency severity index: meta-analysis. *Sultan Qaboos Univ Med J*. 2015; 15(1):e71-7.
53. Liu N, Lin Z, Cao J, Koh Z, Zhang T, Bin HG, et al. An intelligent scoring system and its application to cardiac arrest prediction. *IEEE Trans Inf Technol Biomed*. 2012;16(6):1324-31.
54. American Heart Association. Cardiac arrest [Internet]. 2020 [cited 2020 Jun 2]. Available from: <https://www.heart.org/en/health-topics/cardiac-arrest#:~:text=About Cardiac Arrest,the wake of other symptoms>.
55. Collins GS, Ogundimu EO, Cook JA, Le Manach Y, Altman DG. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. 2016.
56. Riley RD, van der Windt D, Croft P, Moons KGM, editors. Prognosis research in health care [Internet]. Oxford University Press; 2019. Available from: <http://www.oxfordmedicine.com/view/10.1093/med/9780198796619.001.0001/med-9780198796619>.
57. Sun G-W, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. Vol. 49, *J Clin Epidemiol*. 1996.
58. Caetano SJ, Sonpavde G, Pond GR. C-statistic: a brief explanation of its construction, interpretation and limitations. *Eur J Cancer*. 2018;90:130-2.
59. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015 Jan 6;162(1):W1-73.
60. Moons KGM, Wolff RF, Riley RD, Penny ; Whiting F, Westwood M, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: Explanation and Elaboration *Annals of Internal Medicine RESEARCH AND REPORTING METHODS*. *Ann Intern Med* [Internet]. 2019 [cited 2020 Mar 8];170:1-33. Available from: www.probast.org.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more blomedcentral.com/submissions



3.4 Supporting information

The supplementary information for the manuscript including the search strategy can be found in appendix C.

The previous chapter concluded that the current prehospital triage models had limitations in their development and implementation. Using an expert consensus group for model derivations is useful for identifying sensible candidate predictors, but these then need testing for their actual association with the desired outcome. The systematic review only included derivation studies that used mathematics to develop the risk prediction model for this reason. The most common method was logistic regression, which was used for thirty-six models.

Prehospital modelling in this area is scarce, with only six models identified in the systematic review, all aiming to predict the risk of a high acuity outcome. The quality of these models could be considered questionable when assessing their risk of bias.

A small study predicting hospital admission only used 401 patients over a two-month period. This is a surprisingly low sample size and the results reflected this. The final model predictors had wide confidence intervals that sometimes appeared statistically non-significant. For example, a history of cancer gave an odds ratio of 3.9 (95%CI 0.5 – 30.4).¹⁰⁷ Larger samples predicting the same outcome have also been methodologically flawed. Li et al. had a sample size of 2784 patients over a single month time frame.¹⁰⁸ This is a temporal limitation as it is almost apodictic that there are seasonal differences in healthcare demand and acuity. This study was also limited in its evaluation of the model performance. Only accuracy, sensitivity and specificity were reported, which means that the model's ability to discriminate was absent, as was calibration. A possible cause for this was the study appeared to be methodology-driven as

opposed to context-driven. Numerous machine-learning algorithms were developed on the same data, predicting the same outcome, and then evaluated for the most accurate model.

Whilst it does not portray the full picture in an analysis, it is an important theoretical concept that sensitivity and specificity are trade-offs. Unstable models will struggle to find the optimal threshold for decision making. An illustrative example is Seymour et al. who aimed to predict critical illness in the prehospital setting. Once the authors had built the model, it was transposed into a point scoring tool.¹⁰⁹ If a score of four or more was used to predict critical illness, the model had a sensitivity of 0.22 and specificity of 0.98. This meant that there were hardly any false positives and therefore it could rule in critical illness by way of a positive result. However, there was a high number of false negatives, which means they would miss a significant quantity of critical illness with a negative result. When the model threshold was changed to one, the model had a sensitivity of 0.98 and a specificity of 0.17. This has the opposite effect to the 'four or more' threshold and meant that a negative result could effectively rule critical illness out (as there were few false negatives), but a positive result could easily be false. The latter threshold is perceived to be a safer option as it does not miss critical illness. However, this does mean that demand will be created in the care of critical patients until the false positives are re-triaged, limiting the value of the tool. The choices of optimising sensitivity, specificity, or balancing the two (for example taking the Youden index as the threshold) is less of a mathematical debate, and more allied to the clinical context.

In the clinical context of triaging rare outcomes (i.e., the tails of acuity distribution), model instability is common. Out of the six prehospital models that have been identified, three were not impervious to the challenge of optimising the cut-off points for classification. One trauma triage study found that the optimum cut-point gave a specificity of 50% and a sensitivity of 89%.¹¹⁰ It is worth noting that trauma triage is an interesting conundrum as patients often require

body imaging to reveal the full extent of the trauma, which adds complexity in prehospital prediction. The result is a large number of false positives in order to avoid missing any life-threatening (major) trauma. Another trauma triage study concluded that the only way to counter the challenges of accurate triage in practice is to accept lower sensitivity thresholds. This means allowing a greater number of false negatives to occur, i.e., accepting greater risk.¹¹¹

Kim et al. used a limited amount of candidate predictors that could be collected using a wearable device.¹¹² The aim of the model was to predict the acuity of medical patients based on their blood pressure, respiratory rate, pulse rate and the simplified consciousness score (see the candidate predictors section in chapter 5 of this thesis for more information on these). Whilst it appeared to be successful, it was a hypothetical situation as there was no implementation strategy or mention of which wearable device could be used.¹¹² Furthermore, using outcomes that specifically focus on either medical or trauma patients can have limitations in transportability as both are present in the prehospital environment.

Of all the prehospital risk prediction models that have been mathematically developed so far, none of them would be suitable for external validation and therefore a new model needs to be developed and appropriately evaluated. The existing models identified above are designed around high acuity patients and thus have a high sensitivity and low specificity. High acuity patients are more likely to have physiological changes such as abnormal blood pressure and pulse, which is why these features make good candidate variables. However, they are often within normal limits in the mid- and low-acuity patients.

Studies set in the ED have used machine learning to predict outcomes based on the five-level ESI. Raita et al. used a panel of algorithms including logistic regression, LASSO regression, random forest, gradient boosted decision tree, and deep neural network.¹⁰⁴ The outcomes used were similar to the prehospital

studies (critical care and hospitalisation). The latter three algorithms were all able to outperform logistic regression in triaging patients at all five acuity levels. Furthermore, these machine-learning methods were able to increase both sensitivity and specificity. The best performing algorithm for critical care was the deep neural network, which had a C-statistic of 0.86, sensitivity of 0.80 and specificity of 0.76. For hospitalisation it was the gradient boosted decision tree with a C-statistic of 0.82, sensitivity and specificity both 0.75. A similar study that included only paediatric patients had the same findings, with decision tree models and neural networks significantly outperforming logistic regression at triaging patients.¹¹³

3.5 Conclusion

Machine learning algorithms have shown success in their ability to triage undifferentiated patients entering the emergency care system. However, with the limitations outlined in this section, there are opportunities to develop a new algorithm for predicting low acuity. The next chapter outlines the primary aim of the thesis and the objectives needed to achieve the aim. A secondary aim is also reported, and the rationale for the second aim is discussed in more detail.

Chapter 4

The Aims and Objectives of this thesis

4.1 Introduction

The evidence in chapter one has highlighted that there is a clear need for decision support in prehospital care. In particular, decisions that involve transporting the patient to the ED (see section 2.9). There is a growing cohort of patients that do not need the level of care that the ED provides but are still transported there by ambulance (see section 2.6). Currently, there is a lack of decision support in this area for paramedics still on scene with their patients (see section 2.9). The problem reduces to a matter of patient acuity. The new evidence generated in the systematic review in chapter three highlights that it is possible to use statistical and in silico modelling to accurately triage patients according to their acuity. In addition, it is possible that supervised machine learning methods such as logistic regression, tree-based models and neural networks can produce accurate models. But most of the models in the systematic review were created in the ED, and there is a gap in evidence for prehospital care.

This study is triaging acuity, but five-level models (such as the Manchester Triage System (MTS), Emergency Severity Index (ESI)) are often subjective, convoluted and triage patients towards higher acuity outcomes. This study aims to optimise the care of mid- and lower-acuity patients who may not need the level of care the ED provides.

Developing a model that outputs at five-levels does not translate easily to paramedic decision-support around ambulance transportation. For example, let us assume a hypothetical five-level model is created, with one being absolute transport and five being absolute 'leave on scene'. A patient triaged at level 'three' provides little meaningful information on whether to transport the patient. The purpose and rationale of this study is to support paramedics with a binary decision '*to transport the patient, or discharge on scene*'. Therefore, it would be of most benefit to present the paramedic with a probability of an outcome if they did transport the patient.

The training framework of paramedics is to identify potential red flag conditions in patients and transport them to the ED. Qualitative studies supported this idea and have shown that paramedics currently decide to convey some patients who do not need the expertise of the ED and could be treated in primary care. Tools that have been developed so far have aimed to identify high-acuity patients, but it appears that paramedics can do this themselves effectively. The difficulty in decision-making occurs when there is a mid-acuity or low-acuity patient, such as levels three to five on the MTS or ESI. If these could be identified whilst the paramedic was on-scene with them, it could yield the most benefit to the patient.

Accounting for the arguments above, this thesis has the overarching aim of developing a model that is simple, pragmatic, and intuitive, whilst able to identify mid- and low-acuity patients that may not need the expertise of the emergency department. These factors lead to the following research questions:

Primary research question

In adult patients attending the ED by ambulance, can prehospital information predict an avoidable attendance?

4.2 Primary aims and objectives

The aim of this study is to ascertain whether prehospital variables found in the ambulance electronic patient care record (ePCR) can identify a patient who *would* have an avoidable attendance at the ED, *if* they were conveyed, i.e., they were conveyed but it was not necessary. This can be broken down into the following objectives:

1. Extract prehospital variables from ambulance service electronic patient care records.

In the first objective, a data collection period will be defined, and inclusion and exclusion criteria formalised. These set the parameters for data extraction. The data will then be retrieved and collated into a single dataset. Prehospital variables are all those that are collected in the ambulance service electronic Patient Care Record (ePCR). These will contain demographic, clinical and interventional information on each patient.

2. Link the data with ED electronic patient care records.

Each patient episode will then be linked to their corresponding electronic healthcare record from the ED, if they were transported. This will expand the dataset to show complete patient journeys from phone call to ED disposition. Not all patient episodes will have a complete journey that spans this length of care. Part of the data extraction in the first objective will contain patients who were not conveyed to hospital, these will be removed from the dataset but retained in a separate dataset as they are useful for objective 5.

3. Identify low acuity patients in the dataset using the ED information.

For those instances in the dataset where the prehospital instance is linked to ED, a predetermined data-driven definition of an avoidable attendance will be engineered. This will be a new variable, and once created the rest of the ED data can be removed from the dataset, leaving a final dataset where each patient episode contains all prehospital variables and a single outcome measure.

4. Build a predictive model using prehospital variables.

Risk prediction methodology will be used to derive a model which takes prehospital variables as the inputs, to predict an avoidable attendance at the ED.

5. Measure the success of the model in predicting an avoidable attendance using prehospital variables.

Keeping with the risk prediction methodology, the model derived from objective 4 will be assessed for its performance. It can then be applied to the non-conveyed sample removed in objective 1 for sense-checking.

One of the limitations in model development is that accuracy can decrease when the model is applied to a new patient and a process known as external validation is required to ensure that the developed model is truly accurate. Traditionally, this should be done using a different dataset, in a different area, by different researchers.¹¹⁴ However, this is challenging to achieve for this thesis and therefore the simulated transportability will be determined instead. This leads to the 2nd research question below

Secondary research questions

Can the model derived from the primary outcome be spatially transported?

4.3 Secondary aims and objectives

The aim is to determine whether the model will work in practice, but mindful of the fact no further data is available and an interventional study would not be feasible. Barriers to implementation would be that the model works well in one geography but behaves differently in another. This equates to poor spatial validation. Another barrier would be that the model disadvantages patients according to their characteristics. The secondary objectives are therefore as follows:

6. Test spatial validation

The modelling technique used in objective 4 will be repeated using different subgroups of data, which separate instances according to geography. For example, rural, urban and coastal. In objective 2, each ED is a natural clustering unit as each instance only goes to one ED, and there are inherent differences between EDs. Therefore, the data will be trained on all the data excluding one ED and tested on the excluded ED data. This will then be repeated for each ED.

7. Test model discrimination of protected characteristics.

A post-production analysis of patients will be undertaken according to their characteristics to identify whether there are any groups of people that would be disadvantaged from the model if it was to be implemented. The model will then be corrected for any discrepancies identified. If objectives 1 to 7 are successful, then both research questions will be answered, and this will complete the study.

4.4 Conclusion

This chapter has outlined that the primary research question in this study centres on patients attending the ED by ambulance and trying to predict those who may not need emergency level care. The secondary research question is an extension of the first, aiming to simulate whether a prediction model would be successful if it was derived on different types of geographies. In the next chapter, the theoretical considerations of designing the study are outlined, prior to any specific methodology being decided. This then moves on to a critical argument for the type of algorithm that will be used to answer these research questions.

Chapter 5

The Theoretical Considerations and Algorithm Selection

5.1 Introduction

In the systematic review in chapter 3, many algorithms were identified and used in predicting the acuity of undifferentiated patients. However, there was no single dominant algorithm that outperformed the others. This chapter provides a brief overview of the function of predictive modelling, followed by some important concepts that have underpinned the algorithm decisions, and the methods that were used in this thesis. There is also a critical argument on the theoretical considerations for a desirable algorithm, followed by the algorithm selection itself. The first part of the chapter though, explores the epistemology, ontology and axiology that anchors where and how this new knowledge will be created.

5.2 Theory of knowledge

5.2.1 Epistemology

In this thesis, empiricism is the central philosophical notion for which new information is being created. Empiricism subscribes to the idea that knowledge is acquired through sensed experience. This occurs in the absence of a priori knowledge structures. Empiricism can extend beyond the realms of perceptual knowledge, and can use experiential acquisition of knowledge to induce conceptual beliefs to be true.¹¹⁵ This is in contrast to the idea of rationalism, which argues some prior knowledge at the origin of knowledge generation.¹¹⁵ A limitation in adopting the empiricist approach is the wider contexts are not accounted for in the generation of new knowledge, which they would if rationalism was adopted. Modern ideas in empiricism extend beyond the concept of 'seeing is believing' and accommodate the augmentation, extrapolation, and conversion arguments that technology (especially computers) have created.¹¹⁶ Extrapolation allows us to extend our existing modality, for example vision when employing a microscope or telescope to see beyond the human eye. Conversion translates one modality into another, which can be observed empirically, such as sonar devices representing a visual picture. Augmentation allows us to

experience phenomenon that would otherwise be unknown to us, such as detecting alpha particles.¹¹⁶ Rationalism can accommodate these in its ideology, however the interpretation of the results is framed differently. An empiricist would see the result as a measurement of the measured, accepting that the result being a 'good result', or a 'bad result' is open to interpretation between observers. Conversely, the rationalist would place the result within a wider context, and this requires a greater degree of subjectivity, which will vary between observers. In this thesis, the research questions in chapter 4 have been asked due to existing knowledge summarised in chapter 2 and refined in chapter 3. However, this does not qualify as rationalism. There are a few assumptions being made. The study is designed to encompass all that is needed to answer the very specific questions, but it remains an inductive approach. The question is not generating evidence to prove an existing theory, such as the deductive approach. Such a question might appear as: 'Can the data fields in the ePCR match or exceed the known prediction abilities of paramedics?'. Contrastingly, this thesis is asking whether prehospital data is predictive of an avoidable attendance at all. One of the assumptions being made is that this research question accounts for all the intrinsic and extrinsic factors that influence ambulance conveyance and paramedic decision making. The reality is this is not the case, and there are subjective factors that cannot readily be collected in the data. For example, the mood of the paramedic during a conveyance decision, and whether this affects their decision. This becomes a limitation of the study, and a more rationalist approach would incorporate research questions that may go some way to accounting for the subjective. However, if this study demonstrates that it is possible to use data to predict an avoidable conveyance, then subsequent studies into the implementation science of such a model could be undertaken. More information on this is provided in chapter 10, section 10.3.4.

5.2.2 Ontology

Ontology concerns itself with the theory of being and concentrates on the taxonomies of how the world is constructed. In the context of this thesis, the

main ontological consideration is ordering the concept of urgent and emergency care, followed by the placement of any given patient within this concept. The construction of urgent and emergency care is weighted towards a flat ontology within this thesis, much akin to the theory proposed by Quine.^{117,118} This suggests that the existence of things is categorically equal and not dependent on more fundamental (more important) categories which is defined as a hierarchical ontology.¹¹⁷ In this context, it would suggest that the urgent and emergency care settings (primary care, ambulance service and emergency department) contain equally weighted definitions, and it is possible to explicitly define each so that any given patient entering such a system could be 'sorted' by logic into the right care setting. However, there are limitations in applying this theory in the context of urgent and emergency care and the ideas of Carnap and, more recently, Harare expose these.^{117,119} Carnap suggests that the way in which the world is constructed is limited by the language that exists to describe such a reality. However, through a process of explication (conceptual engineering), new linguistic frameworks can be developed that better explain reality. In the context here, the concepts of urgent and emergency care (and the care settings contained within) are constructed from policy and health strategy, and their existence is bound by linguistic frameworks known to these. If the concepts were created by the service users (patients), or the practitioners (clinicians), then the system may be defined and organised differently. This relates to the suggestion by Harare which broadly aligns with the metaphysical distinction between abstract and concrete entities. Harare posits that there are the physical and perceptual entities such as people, buildings, and objects (the concrete). More importantly, there are the imagined realities or abstract entities. It is a form of realist social ontology in that society is arranged around concepts that do not really exist except in the human mind and can only exist in the mass subscription of many human minds. For example, religion, money, nations, corporations are all human inventions. Urgent and emergency care is an imagined reality and the care settings such as primary care, community care, ambulance service and emergency department care are all constructed from the abstract as opposed to the concrete. As such, their

definition and purpose can change over time, or in different realities.

Furthermore, patients entering such a system can have abstract concepts of disease and disease process attached to their perceptual self. This influences the interaction with the imagined reality of the urgent and emergency care system.

This thesis assumes the existence of low-acuity patients not needing the ED, and because they exist there is merit in being able to identify them before they arrive at the ED. However, it is accepted as a limitation (and elaborated upon in chapter 7, section 7.5.1) that they only exist because of how the urgent and emergency care system is currently constructed.

The data used in this thesis is coded using the Systemised Nomenclature of Medicine – Clinical Terms (SNOMED-CT), and more information about this can be found in chapter 7, section 7.5. SNOMED-CT is a polyhierarchical ontology, which allows for a child term to have multiple parent terms, but without creating ambiguity. To illustrate, ‘viral meningitis’ is a label attached to the SNOMED-CT code ‘58170007’. This can be a member of the parent concept ‘infective meningitis - 312216007’ but also be a member of ‘viral infections of the central nervous system - 302810003’.¹²⁰ Having an exhaustive polyhierarchical ontology in data coding has significant benefits, especially when it comes to healthcare funding and being able to identify cohorts of patients with certain medical conditions, or who may have had a clinical procedure. The limitation occurs during the pursuit of exhaustion. Eventually, there becomes a widening detachment between the patient in front of the clinician, and the ‘coded’ version of the patient in front of the computer. For example, imagine a patient has a condition called ‘Chronic Obstructive Pulmonary Disorder’. In brief, this is a long-term respiratory condition, which changes how their lungs work and can render the patient with difficulty breathing and getting oxygen around their blood. This disease has a single SNOMED-CT code, with child concepts describing the severity (such as mild or severe). The data would not describe the condition well enough to inform

clinical judgement or decision making. It fails to account for how the patient lives with the disease and how the disease affects the patient at different times. The data assumes that everyone with the same diagnosis has the same experience with the disease. However, this is not the case, and is compounded by evidence that human coders (such as clinicians) have poor inter-coder reliability when coding the same patient.¹²¹

5.2.3 Axiology

Axiology is the theory of value and is often dichotomised into intrinsic and extrinsic value.^{122,123} An entity has intrinsic value if it is good, or if it is good for its own sake. This contrasts with extrinsic value which is where an entity is only valuable if it ascribes value to something else.^{122,123} In this thesis, the ideas of pluralism and emotivism are adopted. Pluralism is where value can be expressed in multiple forms including pleasure or knowledge.¹²⁴ Emotivism does not consider value statements to be objectively ascribing value to an object, instead it is translating the emotion (or sentiment) of the speaker about the object.¹²³ In urgent and emergency care, all care settings can be described as having intrinsic value as the same patient would gain value in attending any given setting. However, the quantity and form of the value may be different in each setting. For example, imagine if a patient fell from a tree and suffered a broken arm. If they attended the GP practice, they would gain the value of knowledge that they need to go to hospital. They may also be able to get interim pain relief (analgesia) and a clinical assessment. Compare this to if the patient attended the ED first. They would get analgesia, an x-ray, a treatment of the wound, and an ongoing fracture appointment. The value is greater by attending the ED. To incorporate emotivism into the context of this thesis, the patient may attribute certain values to each care setting, but this may differ according to the value objectively defined in this thesis. A low acuity patient identified in the data (as defined in chapter 7, section 7.5.3) may disagree with being classified as low acuity as they believe there was both intrinsic and extrinsic value to their attendance. Therefore, the

axiological limitation is that the low acuity definition is steered by the emotivism of the author and therefore the definition may not be considered ubiquitously valuable to everyone.

5.3 Theoretical considerations

5.3.1 The objective function

Fundamentally in traditional clinical prediction modelling, irrespective of context or data, it reduces to finding a function (f), given a set of inputs (x) to identify the value of an output (y). In other words: $y = f(x)$. In the case of this thesis, x are the ambulance service variables, and y is the ED experience. The ability of finding the function is most often done using a labelled training data set. The training data set (\mathcal{D}), is composed of a set of candidate variables (x) for each instance (x_i) of N patients, with the outcome variable (y) associated for each instance (y_i).¹²⁵ The training set can be represented by the following equation:

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$$

This training set is used to estimate an objective function (\hat{f}) that will make accurate predictions of the outcome (\hat{y}) in new data. The circumflex indicates it is an estimate. So, it is an estimated function that will lead to estimated predictions in new data, or new patients. The theoretical idea that finding the objective function and making accurate predictions being probabilistic is central to predictive modelling. Even in the context of binary classification (as in this study), it is more interpretable to indicate the probability of class membership. This philosophy therefore requires an expansion to the simplistic $\hat{y} = \hat{f}(x)$ to the following:

$$\hat{y} = \hat{f}(x) = \underset{c = 1}{\overset{C}{\operatorname{argmax}}} p(y = c|x, \mathcal{D})$$

Here, the predicted probability that the outcome belongs to a given class (\hat{y}) is calculated by creating a vector of C , which is the length of classification groups. In binary classification, C is equal to 2 as it can only accept 2 possible states ($c =$

1, $c = 0$), but a single number can be returned as the probabilities of both classes would sum to 1. Using the positive class ($c = 1$), the right-hand side of the function can be therefore described as calculating the probability that $y = 1$, given the set of input variables (x) that are the same as those found in the training dataset (\mathcal{D}). This translates to the outcome being a best guess at the true label.¹²⁵

In this study, the candidate variables (x) were all found within the ambulance service ePCR data and are detailed further in chapter 7, section 6. They all have an associated (y) label which was derived from the ED data and is the defined outcome measure of an avoidable ambulance conveyance to the ED (also found in chapter 7). The quantity of patients (N) is large, which means that the only efficient analytical way of deriving the objective function was to use automated methods. This forms a definition of machine learning. From the literature, it has been defined as the following:

“A set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty”¹²⁵

In supervised machine learning for risk prediction modelling, the model is initially derived on a training dataset with labelled outcomes. A good prediction model will make accurate predictions in new data. As a starting point, the model can be developed on the whole dataset and then the performance evaluated by testing on the same dataset. This gives what is known as the apparent validity. It is a stable measure of model performance as it is being tested on the whole dataset. It has drawbacks though, and the results of the evaluation will be optimistic. Optimism is the idea that the model will perform too well when evaluated in the same data it was derived on. It does not reflect the true model performance. It occurs because the derived model has overfit the training data,

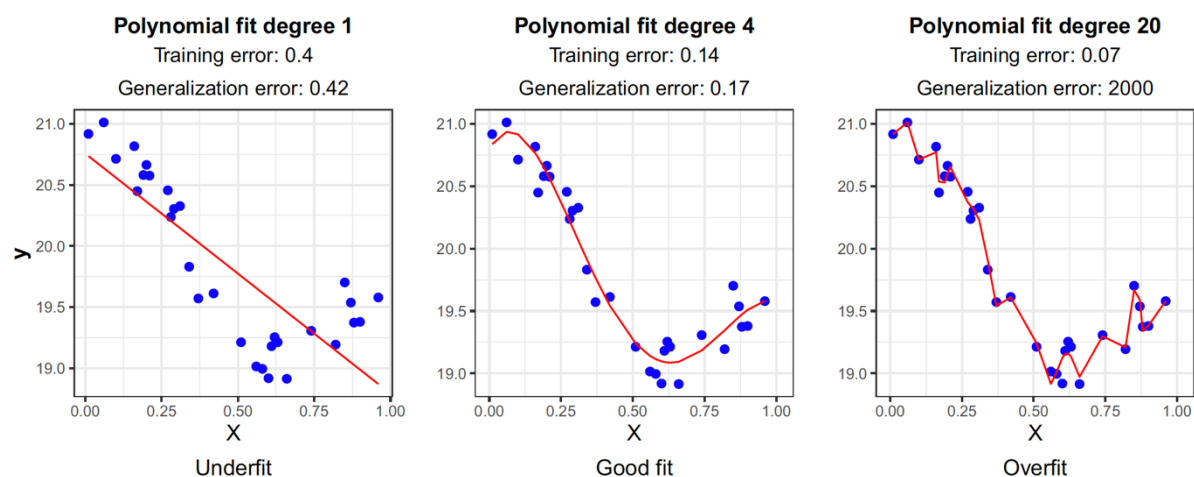
which means it cannot generalise to new instances. Overfitting is an important concept in risk prediction modelling and can be expanded to the bias variance trade-off.

5.3.2 The bias variance trade-off

In prediction modelling, a good model will have low distances between predicted values and observed values. However, when developing a model in the training data it is important that the model complexity is just right. If the complexity is too low, or too high then the model will fail to capture the true relationship between the candidate predictors and the outcome, and the model will fail.

Figure 6, reproduced with permission from Badillo et al. shows three examples of fitting a model.¹²⁶ The permission has been included in appendix A, section A2. The left panel is showing an underfit model, where the regression line is unable to represent the data and therefore leaves large residuals, especially when the X variable is at 0.5. The right panel is the other extreme and represents overfitting. The regression line moves through all the data points in the training data, but when tested, results in huge errors. The ideal model has a regression line that can represent the shape of the training data but can generalise to the test set (middle panel). The great mathematician John von Neumann famously said “With four parameters I can fit an elephant and with five I can make him wiggle his trunk.”¹²⁷

Figure 6: Demonstration of overfitting (reproduced from Badillo et al. with permission)¹²⁶



An overfit model has high error in the test set. The error of a model can be decomposed into three parts: the bias, the variance, and the irreducible error. Bias occurs when the model has failed to capture the true relationship between the predictors and the outcome, rendering all predictions systematically distant from their true values. Bias is high when a model is not complex enough to capture the relationship. The irreducible error cannot be changed and is an extraneous bias that is captured in the model. Bias and variance are inversely proportionate to each other, which is why there is a trade-off between the two. In a model with high variance (over fit) the model is too complex, which is why it has failed to capture the true relationship of the data and fails to make good predictions in new instances. However, the bias will be low as the complexity matches the relationship in the training data. The solution to reduce variance in a model is to rebuild the model in the training dataset, increasing the bias by reducing model complexity. If the variance between the training and test set is low, but the error is high, this is due to a bias in the prediction. The model is not complex enough. The ideal model has low variance and bias, to reduce the optimism found in the apparent validity, bias needs to be applied. This is achieved through model validation.¹²⁸

5.3.3 The law of parsimony

The overwhelming taxonomy of machine learning algorithms can make selection difficult. Many studies have used a panel of algorithms on the same dataset to compare and select a ‘winner’. This competitive technique can be a normal construction for how computer scientists will approach solving a problem. This study has benefitted from having a preliminary dataset that could be used to support algorithm choice. The dataset, known as the DS2 dataset, contained 100,000 YAS ePCR’s. These were non-conveyed patients that were attended by a YAS clinician between the 1st of July 2019 and the 29th February 2020.

In healthcare data, there can be many independent variables that are being examined with their association with the dependent. This causes the problem of high dimensional datasets. The issue arises as the quantity of candidate variables counter intuitively has a negative impact on prediction with the addition of further variables. This is known as the ‘curse of dimensionality’. Due to additional variables causing increased dimensionality, the resultant necessary sample size increases by a disproportionate amount. This leads to an increase in computational expense for the analysis. Therefore, an essential objective is to reduce the dimensionality down, without compromising predictive performance.^{129,130}

This approach has been called the principle of *novacula occami* (Occam’s razor), or the law of parsimony. Fundamentally, it states that entities should not be multiplied beyond necessity. The eponymous maxim is attributed to William of Ockam in the 13th century, although ancient philosophers had already described this approach. Ptolemy wrote that “*We may assume the superiority ceteris paribus of the demonstration which derives from fewer postulates or hypotheses*”.¹³¹ Even when complexity is required, the idea is that the minimum should be exercised, without the explanation losing meaning. Reducing dimensionality is subscribing to this philosophy whilst gaining from the statistical benefits of a parsimonious model; but there are also practical benefits.

In the model implementation stage where real-time prediction is required, it would be of great importance to have a model that did not require a user to input 60 variables (for example) before a prediction could be made. This argument of reducing dimensionality is central for identifying a desirable algorithm.

Other important algorithm features that are desirable are the ability to handle missing data and collinearity. Missing data is often found in large datasets and can come in many forms from completely random, to specific patterns.

Collinearity is when two or more variables are strongly associated with each other. This is a problem when the association is between candidate (input) variables.

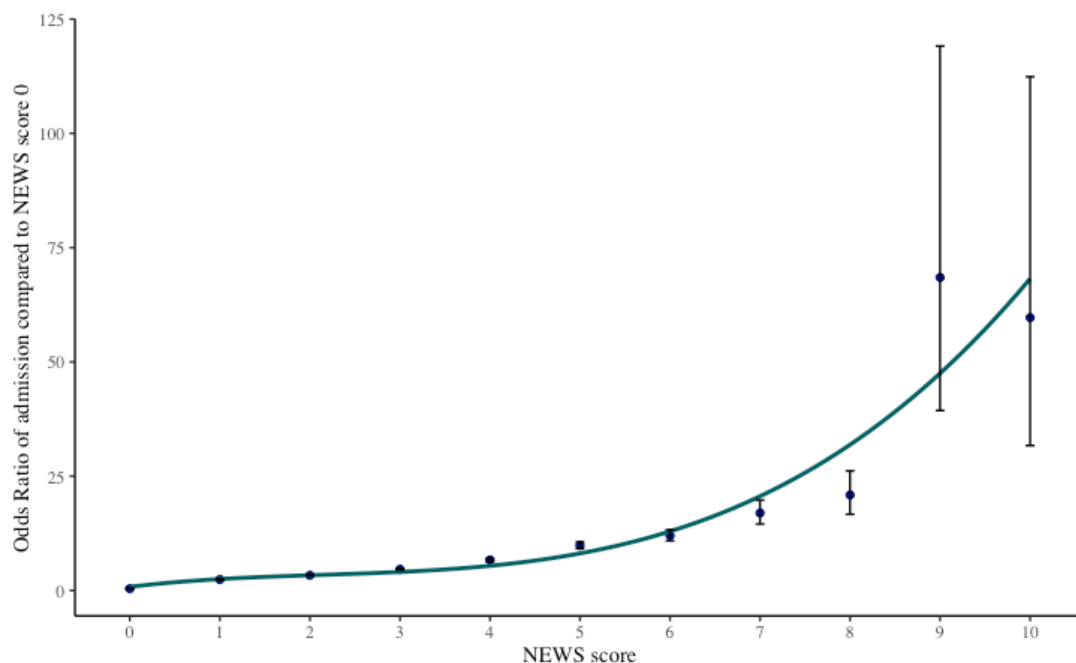
5.3.4 Algorithm of choice

Prespecifying the algorithm of choice has the disadvantage of inhibiting a competitive space where different algorithms can be tested on the data and the best algorithm selected. However, it does have the advantage of customising a modelling framework around the most appropriate algorithm. The lens used in choosing the algorithm was to decide which would most suit the data being used. The algorithms were assessed on their ability to handle missing data, non-linearity and how they incorporate feature selection methods. They were also assessed on their computational expense and how they performed solving a similar problem identified in the systematic review. The technical details on how these algorithms function have been omitted from the main thesis, but can be found in appendix D.

The review in chapter 3 identified that the most common method (and therefore the industry standard algorithm) for risk prediction modelling of patient acuity is logistic regression. This has the advantage of transparency as it is explanatory as well as predictive, offering more information as an end 'product'. However, logistic regression has a limitation in modelling non-linearity as it has a linear decision surface.¹³² In brief, non-linearity is when two associated variables (x and

y) do not uniformly have a relationship that is not uniform, when plotted on a graph. A change in one variable at one point on the x-axis will demonstrate a change in the other variable on the y-axis, but at another point on the x-axis the same change will be associated with a different change in the other variable (on the y-axis). For example, taking original data from one of the studies in the systematic review, figure 7 has been created for this thesis and shows the non-linear relationship between a patient's National Early Warning Score (NEWS) and their odds ratio of admission.

Figure 7: NEWS score and its relationship with risk of hospital admission from Cameron et al.¹³³

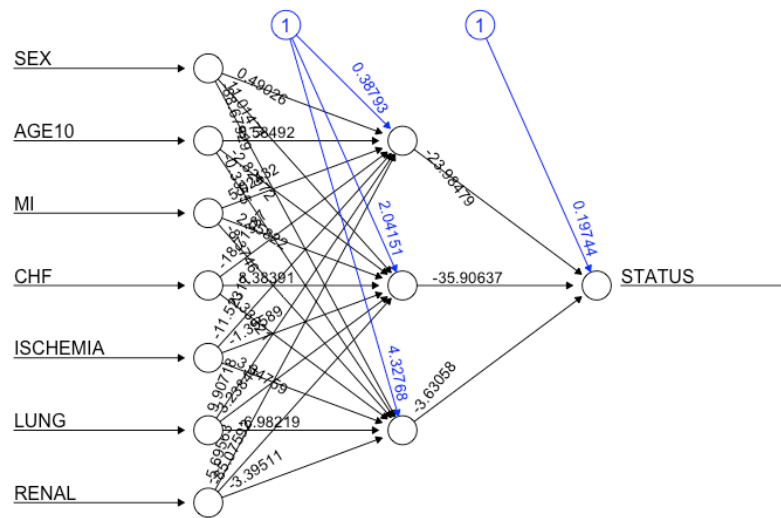


Non-linearity is common in healthcare and so an algorithm that can naturally handle complex relationships between variables would be a better fit to answer the research questions. Although logistic regression is not computationally expensive, there is more pre-processing labour than other algorithms. The method does not handle missing values and so statistical methods such as multiple imputation must be employed to complete the dataset.

One of the best performing algorithms in the systematic review was the neural network. In simplified terms, a neural network is a collection of interlinking nodes that are representations of biological neurons. Each neuron takes the values of incoming candidate variables and assigns a weight according to the value. The weights are summed across all the candidate variables and a bias term is also applied, leaving a single numerical figure. A predetermined 'activation threshold' is decided and if the figure is greater than the threshold, then the neuron 'activates'. Activation could mean the next neuron in the sequence is used, or a classification is made. A more detailed description of neural networks can be found in appendix D.

Figure 7 has been created for illustrative use in this thesis using publicly available data designed to predict the risk of perioperative mortality following elective abdominal aortic aneurysm surgery.^{134,135} Input variables were sex, age (categorised into ten-year groups), a previous myocardial infarction (MI), a medical history of congestive heart failure (CHF), ischaemia, respiratory conditions, or renal conditions. In figure 9, the weighted values of each input variable are labelled in black, with the bias terms shown in blue. Neural networks can be comprised of many hidden layers where the output of the neuron acts as the input for the next layer.

Figure 8: Neural network example



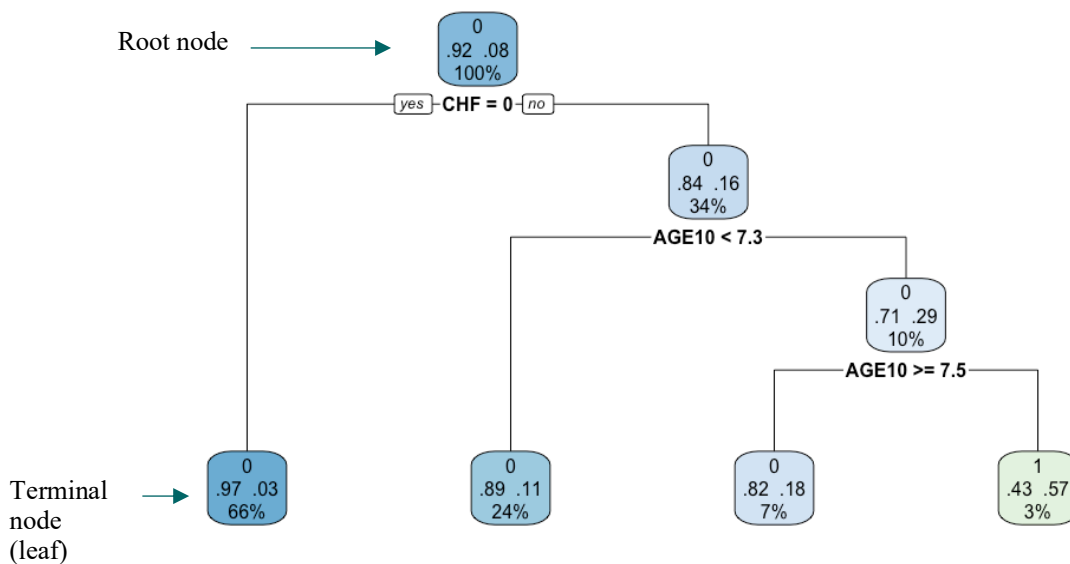
Error: 5.851236 Steps: 903

The neural network can handle non-linearity effectively through the activation function; however, the algorithm design struggles to overcome collinearity without having first undergone procedures such as Principal Component Analysis (PCA) in the data preparation stage. The process of PCA transforms the data into fewer variables that are a representation of the relationships contained within the original variables but are not the original variables. This means that the management of collineated variables comes at the expense of model interpretability.¹³⁶ More information on principal component analysis can be found in section The algorithm, by the nature of it being an extension of the generalised linear model, means it can inherit the same issues with missing data. Solutions that overcome this for the neural network include modelling the uncertainty of attributes with probability density functions.¹³⁷ However, a neural network can be computationally greedy, and it can take many hidden layers in order to create an accurate model. There are also limitations in how neural networks handle structured data. In classification problems that use unstructured data, such as images or sounds, canonical architectures translate these forms into meaningful inputs for neural networks. However, this is a difficult task with structured, tabular data.¹³⁸ The systematic review also identified decision trees as a popular method. These are diverse in their methods

but can overcome some of the limitations of logistic regression and neural networks.

In the systematic review, decision trees featured in 23 models. The basic components of a decision tree are nodes. These are simple filters that take an input and split it into two or more outputs based on the split criteria. Trees include a root node, which is the first or starting node. Figure 9 has been created for this thesis as an example of a simple decision tree and is based on the same AAA data by Steyerberg et al.^{134,139}

Figure 9: Example of a tree-based model using Steyerberg AAA data (n=238)¹⁹⁶



The root node is the top node on the illustration and the population going into the node is being split down two branches depending on whether they have a medical history of congestive heart failure (CHF). The left branch (those with CHF) leads to a terminal node or leaf. The right leads to another node that is making a split on whether the AGE10 is less than 7.3. This is equivalent to patients being under the age of 73. This node is the child node of the root node, but also the parent node of the right branch coming off. In the illustration, each node displays certain information. It has the prediction made at the top (0 or 1 as the outcome is binary). It also shows the probability of the predicted outcome

(and the inverse) and proportion of the population going into the node as a percentage. For example, it can be interpreted that 66% of the sample had CHF and their probability of the outcome (0) was 0.97. Decision trees are simple to interpret and can be explanatory for simple models. They operate differently to the methods above in that they use recursive partitioning of the data in order to classify. Recursive partitioning is where the sample is split based on a value within a feature. For example, if age was the variable used as a root node, recursive partitioning finds the optimum cut point and moves all those participants to one side of the tree or the other, depending on the optimum age threshold. If the split results in a child node, the partition made at the node is not using all the dataset, only a subsample of participants that qualified through the parent node. This can be seen in figure 9 as the percentage decreases down the branches. The way a decision tree decides on which variables to use as the parent node, or any internal node (internal nodes are any node situated between the root and the leaf) is not random but mathematically calculated as a measure of information gain. More detail on these calculations can be found in appendix D.

In decision tree modelling, because of the recursive partitioning, a categorical variable can only be used once, however a continuous variable can be used multiple times providing its subsequent use operates at a different threshold to one that has already been used. Decision trees have the advantage of being able to overcome the weaknesses of logistic regression and neural networks such as non-linearity and collinearity. However, they have their own limitations. The way a decision tree handles continuous variables is to categorise them. This leads to a loss of information and is discouraged in prediction modelling.¹²⁸ Simple decision trees also can generalise beyond the data. For example, if a binary variable is being split into two leaf nodes but all the sample are in the positive class, the tree will automatically assign an outcome for the negative class. This is decided by automatically choosing whichever outcome was the majority at the parent node. Decision trees can also be prone to overfitting to the training data.

If there are no safeguards in place, the model will keep splitting until there is either no more variables left to select, or there is no more information gained from splitting. This makes a greedy algorithm, and as the tree becomes deeper with more splits, it will fit the training data with less error but will increase the error when applied to a different sample (the test set). There are methods to reduce overfitting that are effective and easy to implement. One method is tree pruning.

Pruning is the removal of sub-branches and replacement with leaf nodes.¹⁴⁰ Even with the addition of pruning, a single tree classifier is considered a weak learner on its own and can easily be overfit to the dataset used to train it. Modern approaches use ensemble methods which constitute a whole forest of decision trees. In the systematic review, only two models were simple decision trees, the remaining 21 models were ensemble learners.

Ensemble decision tree models were described in detail by Leo Breiman in the 1990's and are an extension of the simple recursive partitioning tree detailed above.¹⁴¹ Ensembles create many trees that are individually diverse in their decision making. These diverse models are aggregated in a voting system to make a final prediction. The two well-known techniques of creating a forest of trees are known as bagging and boosting.

Bagging is an abbreviation of bootstrap aggregation. The purpose of bootstrap aggregation is to prevent a forest of trees from group decision making. This would occur if all the trees were created using the same data. In bagging, each tree is technically built on a different dataset. A limitation with even a simple ensemble model using tree bagging is collinearity between strong predictors. Bagging will de-correlate to an extent by deriving the model on different training sets. However, strongly correlated predictors will consistently yield the highest entropy regardless of the sample space within the ensemble. This is because bootstrapping creates new datasets, but the distributions are largely the same over the variables themselves. If there is a noisy variable, it will be noisy in all the

trees and will encourage the group decision making in the forest. To counter this, a common method is to build the bagging ensemble model but use a random subset of predictor variables in each classifier. This is known as random forest modelling. If a dataset had 100 variables, the architect of a random forest model could specify how many variables out of the 100 should be randomly selected each time for inclusion in the tree. This does not necessarily mean all randomly selected variables will be included in the tree, but it does mean that strongly correlated variables with the outcome will not be in every tree model.^{125,140,141}

Novel approaches to ensembles have managed to combine features of bagging and boosting into computationally efficient algorithms. Extreme Gradient Boosting (XGBoost) is an algorithm developed by Chen and Guestrin in 2016.¹⁴² The extreme nature of the algorithm is primarily placed in its computational efficiency. It operates ten times faster than other gradient boosted algorithms. This occurs because the algorithm splits the data into subsamples that are then sent to different computer cores. These are scanned in parallel for the best splits and thresholds in the continuous variables. Instead of using the exact greedy approach mentioned above, it uses an approximate greedy algorithm, running parallel in every subgroup. The data is scanned, and quantiles are created. These are then combined in a histogram to identify the optimum threshold. The quantiles are the thresholds, and these are weighted so that the sum of the weights between each quantile are the same. This is a method known as the weighted quantile sketch. The major advantage of the XGBoost algorithm is its ability to handle missing data. One of the limitations of many tree algorithms is that they work best with dense matrices. This is where the dataset has most of the variables complete, or very few missing values. A sparse matrix has mostly zeros in the columns with the alternative being 1. In XGBoost, a dense dataset is transformed into a sparse, by one-hot encoding all variables. As an example, if the variable 'Location found' had three values (home, public place, and prison), one hot encoding would widen the variable into three binary variables. It can then calculate how to handle the zeros in the data, in a process known as sparsity

aware split finding. In this process, when the algorithm encounters a sparse value (0), it models placing all the non-missing values down the left branch of a split and then calculates the gain. It then repeats the process, moving the instance down the right branch. The direction with the most gain is then set as the default direction and all sparse values move in the default direction. The XGBoost algorithm also contains the elements of bagging in the hyperparameter 'colsample_bytree' which takes a random sample of columns for inclusion in subsequent tree development. It also can incorporate regularisation penalties on the loss function, such as those that can be applied in logistic regression.¹⁴² Recursive Feature Elimination (RFE) can be used with these decision trees to select the optimum combination of features that would generate the most accurate model. In RFE, a cross-validated model is built using all the features. Each feature has its relative importance calculated. The least important feature is eliminated, and the procedure repeats itself, minus the eliminated feature. The evaluation score is calculated for each iteration, with the highest performing model (and its associated features) selected. One of the drawbacks with this method is that it does not safeguard against eliminating weak features that may combine with others to create significantly improved results.¹⁴³⁻¹⁴⁵ The steps are similar to backwards elimination in regression. Tree-based models can be computationally expensive, especially in split finding and identifying optimal thresholds for dichotomising continuous variables. However, the XGBoost algorithm has managed to find a solution to this problem. XGBoost can handle missing data through sparsity aware split-finding, it naturally handles nonlinearity, and it is computationally efficient.¹⁴² It has mechanisms to adjust bias and variance through hyperparameters, including regularisation. For these reasons, XGBoost was selected as the algorithm in this thesis. No study identified in the systematic review has used XGBoost for triaging the acuity of emergency care patients and so this thesis is the first to use it.

5.4 Conclusion

In this chapter, it has been detailed that the new knowledge is anchored from an empiricist epistemological perspective, on a flat ontology with pluralism and emotivism providing the perspective of value. The objective function has been outlined from a mathematical point of view and considerations such as the bias-variance trade-off and the law of parsimony have been explored. There were numerous algorithms identified in the systematic review in chapter 3, which were methodological candidates. After a critical argument, it was decided that the best algorithm would be the XGBoost. The next chapter is a published protocol, which outlines the procedures for undertaking this research.

Chapter 6

The Protocol

6.1 Introduction

In the previous chapter, there was a critical argument for algorithm selection that concluded with the use of a more neoteric algorithm known as XGBoost. Building on the work of the systematic review (Chapter 3) and algorithm identification (Chapter 5), a protocol was published in 2021. This sets out the methods that were used in the thesis to develop and validate the model. The protocol follows the reporting guidelines of the Transparent Reporting of a multivariable prediction model for Individual prognosis or Diagnosis (TRIPOD).¹⁴⁶ The purpose of publishing the protocol in advance of the study enhances the transparency of the research and also allows for early peer review to ensure the methods are suitable to answer the research questions. The study was given the following long and short titles:

Short title: The SINEPOST Study

Long title: The Safety INdEx of Prehospital On Scene Triage (SINEPOST): The development and validation of a risk prediction model to support ambulance clinical transport decisions on-scene

The full text can be found here:

<https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-021-00108-4>


Following on from the protocol chapter, there is an expansion where more information is provided, and justifications are made for certain methodological choices.

PROTOCOL

Open Access

The Safety INdEx of Prehospital On Scene Triage (SINEPOST) study: the development and validation of a risk prediction model to support ambulance clinical transport decisions on-scene—a protocol



Jamie Miles^{1,2*} , Richard Jacques³, Janette Turner¹ and Suzanne Mason¹

Abstract

Background: Demand for both the ambulance service and the emergency department (ED) is rising every year and when this demand is excessive in both systems, ambulance crews queue at the ED waiting to hand patients over. Some transported ambulance patients are 'low-acuity' and do not require the treatment of the ED. However, paramedics can find it challenging to identify these patients accurately. Decision support tools have been developed using expert opinion to help identify these low acuity patients but have failed to show a benefit beyond regular decision-making. Predictive algorithms may be able to build accurate models, which can be used in the field to support the decision not to take a low-acuity patient to an ED.

Methods and analysis: All patients in Yorkshire who were transported to the ED by ambulance between July 2019 and February 2020 will be included. Ambulance electronic patient care record (ePCR) clinical data will be used as candidate predictors for the model. These will then be linked to the corresponding ED record, which holds the outcome of a 'non-urgent attendance'. The estimated sample size is 52,958, with 4767 events and an EPP of 7.48. An XGBoost algorithm will be used for model development. Initially, a model will be derived using all the data and the apparent performance will be assessed. Then internal-external validation will use non-random nested cross-validation (CV) with test sets held out for each ED (spatial validation). After all models are created, a random-effects meta-analysis will be undertaken. This will pool performance measures such as goodness of fit, discrimination and calibration. It will also generate a prediction interval and measure heterogeneity between clusters. The performance of the full model will be updated with the pooled results.

Discussion: Creating a risk prediction model in this area will lead to further development of a clinical decision support tool that ensures every ambulance patient can get to the right place of care, first time. If this study is successful, it could help paramedics evaluate the benefit of transporting a patient to the ED before they leave the scene. It could also reduce congestion in the urgent and emergency care system.

Trial Registration: This study was retrospectively registered with the [ISRCTN: 12121281](https://www.isrctn.com/12121281)

Keywords: Ambulance, EMS, Emergency, Triage, Acuity, Machine learning, Logistic regression, XGBoost

* Correspondence: j.miles@sheffield.ac.uk; jamie.miles@nhs.net

¹CURE Group, School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent Street, Sheffield S1 4DA, UK

²Yorkshire Ambulance Service, Brindley Way, Wakefield WF2 0XQ, UK

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Background

Demand in the emergency care system is increasing. In prehospital care, this translates to an increase of around 5% per annum and in the emergency department (ED) is around 3–6% [1, 2]. When the ED is busy, ambulance crews can be held in a queue at the ED and this is known as offload delay. In the winter of 2019/2020 in England, there were 137,009 offload delays of between 30 and 60 min and 39,304 delays of over an hour [3]. With crews held at the ED, it reduces the prehospital fleet capacity to respond to emergencies and subsequently puts patients in the community at risk.

One of the contributors to demand in the system is the case-mix of patients that access emergency care. The majority of ambulance service patients require fewer critical interventions and more community-based care [4, 5].

This appears to be at a juxtaposition to the training of paramedics. Numerous studies have found that there is role confusion when paramedics are presented with a low-acuity patient, as their foundational knowledge and education was rooted in emergency care [6–11]. This meant that decisions to leave a patient at home (non-conveyance) are the most complex to make and this was further compounded by a perceived lack of managerial support [6].

As a result, transport decisions are not always accurate and there could be between 9 and 32% avoidable conveyances to the ED [4, 12–14]. Miles et al. used vignettes of real patient journeys and asked paramedics to make transport decisions. They found that there was clear agreement between the sample paramedics ($k=0.63$), and the overall accuracy in decision-making was 0.69 (95% confidence interval (CI) 0.66–0.73). Reassuringly, the sensitivity for transport decisions was high (0.89, 95% CI 0.86–0.92) meaning that there were few decisions not to convey a true emergency. However, the specificity was 0.51 (95% CI 0.46–0.56) meaning that almost half of the sample decided to transport a low-acuity patient [15].

There is a paucity of evidence for transport decision-support tools for paramedics. One example, which has been adopted by numerous ambulance services, is the paramedic pathfinder tool [16, 17]. This was developed using a Delphi approach with a multidisciplinary team of experts. The tool was user tested in 2014 on a sample of 481 patients (361 medical patients and 114 trauma). Results for medical patients showed a sensitivity of 0.94 (95% CI 0.91–0.97) and specificity of 0.58 (95% CI 0.49–0.66). For trauma sensitivity was 0.96 (95% CI 0.88–0.99) and specificity 0.6 (95% CI 0.48–0.72). These results are not a significant improvement on paramedics making their own decisions, which limits the usefulness of the pathfinder tool.

A recent systematic review by Miles et al. looked at whether computer algorithms could triage the acuity of all patients entering emergency care and support decision making [18]. They found 92 models from 25 studies. The review demonstrated that it is possible to triage patients accurately using machine-learning algorithms but only six studies had a prehospital focus. Two studies demonstrated that prehospital variables could predict hospital admission. Meisel et al. used logistic regression to create an admission prediction score with a C-statistic of 0.80 [19]. Li and Handy used a panel of algorithms, with the most successful being a modified support vector machine, which had an accuracy of 0.81 [20].

Seymour et al. used logistic regression to derive a risk score to predict critical illness in prehospital patients. Their model had a C-statistic of 0.77 (95% CI 0.76–0.78) [21]. van Rein developed a triage model for trauma patients and found that the model had a C-statistic of 0.82 (95% CI 0.81–0.83) [22].

These studies have demonstrated that it is possible to develop accurate models prehospital for triaging patients using clinical data. However, they have been developed to predict high-acuity patients as opposed to low-acuity.

Objectives

Primary research question

Can ambulance service clinical data predict an avoidable attendance at the ED in adults?

Primary objective

To build risk prediction models using prehospital clinical data as input candidate variables, and ED experience as the output variable.

Primary outcome measure

An avoidable attendance at ED as defined by O’Keeffe et al. (2018). This is described as ‘First attendance with some recorded treatments or investigations, all of which may have reasonably been provided in a non-emergency care setting, followed by discharge home or to GP care’ [13].

Secondary research questions

What is the simulated transportability of the model derived from the primary outcome?

Secondary objectives

Evaluate model test performance under different spatial test sets.

Compare the different models for accuracy and feasibility to embed in practice.

Secondary outcome measure

There are no secondary outcome measures

Methods and analysis plan

This protocol has used the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) guidelines in its structure [23]. The final study publication will also adhere to these guidelines.

Source of data

This study is an observational cohort study using retrospective data. All patients attended by Yorkshire Ambulance Service (YAS) have an electronic patient care record (ePCR) completed by the paramedic treating them. This contains all demographic and clinical data relating to that episode. If the patient is transported to an emergency department (ED), a similar record is created for their ED episode containing all demographic and clinical information. These two records will be linked together to create a single patient journey for each patient from the moment the paramedic arrived on scene, to their outcome at ED. This cohort is the primary analysis cohort and will be used for model development, and internal-external validation.

The data collection period started on 1st July 2019, as this was the earliest date that Yorkshire Ambulance service had a region-wide rollout of the electronic patient care record (ePCR). The end date was the 29th February 2020. The end date was chosen for a maximum sample size, without the data being compromised by the COVID-19 pandemic. The data was not extracted until after the end date.

Participants

The study is set in pre-hospital care but uses ED experience as the outcome. There is one ambulance service involved (YAS) and sixteen EDs throughout Yorkshire.

Patients were eligible for inclusion if they were over 18 years old at the time of attendance and had a completed record in the ambulance service data, and the ED data (if they were transported). The patients can be described as largely 'unselected'. This means all patients are eligible, irrespective of any demographic or disease process. The only restriction in selection is age being

over eighteen. This is due to ambulance service policies often mandating the transport of children to hospital.

Outcome

The outcome of the model is a non-urgent attendance at the ED. The reference standard is described by O'Keeffe et al. who state: "first attendance with some recorded treatments or investigations all of which may have reasonably been provided in a non-emergency care setting, followed by discharge home or to GP care." [13]. This definition has been transformed into a data-driven coded definition and is found in the routinely collected Emergency Care Data Set (ECDS), and the former Commissioning Data Set (CDS) in the UK [24]. The full coded definition can be found in the [supplementary material](#). The definition is calculated by examining each patient's ED experience across six variables. These are department type, attendance category, arrival mode, investigations, treatments and discharge status. For a patient to be coded as non-urgent, they need to only have experienced the values recorded in the definition. As an illustrative example, please see Table 1.

The justification for this reference standard is that it has been adopted by National Health Service (NHS) Digital as the accepted definition of non-urgent attendance at the ED. There are two modifications to this standard for this study in that arrival mode was defined as non-ambulance arrival, but this has been changed to ambulance arrival only. The included investigations and treatments have been expanded to reflect the practice of the ambulance service and the provision of primary care. The modifications were decided by an expert group.

Candidate predictors

In order to inform the protocol and the sample size calculation, a combination of previously published literature and an exploration of prehospital data was used (not used in model development). Previous prediction modelling studies of emergency triage have published variables that were significant in their models. Physiological variables for example pulse rate and blood pressure appear to be the most significant predictors of

Table 1 Illustrative example of how the definition is applied to patients

| Variable | Patient 1 | Patient 2 | Patient 3 |
|-----------------------|-----------------|-------------------------------------|-------------------------|
| Department type | Type 1 | Type 1 | Type 1 |
| Arrival mode | Ambulance | Ambulance | Ambulance |
| Attendance category | First | First | First |
| Investigations | None | Urinalysis, pregnancy test | Urinalysis, chest X-ray |
| Treatments | Guidance/advice | Recording vital signs, prescription | None |
| Discharge status | Discharged | Discharged | Discharged |
| Non-urgent attendance | Yes | Yes | No |

acuity. This is followed by patient comorbidities and whether the case originated from a non-residential setting [25, 26]. A sample of ambulance ePCRs (114,715) was used to identify clinically useful candidate variables in the ambulance data. The model is designed to be pragmatic so if a candidate predictor had more than 30% missing data it was removed. If a variable was likely to contain missing data as it did not occur (judged by evidence of a positive class within the variable) then 'none' was imputed. For example, the variable 'drug_name' only gets completed if a drug is given. In the sample data, there were 106,052 (93%) missing values in this field, and in rest a specific drug was named (e.g. Adrenaline 1: 1000). Therefore, 'no drug' can be imputed into the missing values as it is assumed nothing was administered. This is the same process that NHS Digital use in their definition of the outcome. In the sample, there were 503 variables in the data. Four-hundred and forty-three variables had more than 30% missing data and were excluded from the analysis. This left 60 variables available for analysis, physiological variables, interventions, treatments and source of call (residential home, care home etc.) were all included.

Statistical analysis methods

An XGBoost algorithm will be used to develop the models. This has been chosen as it can accept missing data in the candidate variables during model development, which may have an advantage when transforming it into an electronic decision support tool. Another strength of a gradient boosted algorithm is that it can increase the cost of errors on a minority class being predicted, which is a benefit in a dataset with a class imbalance. It also has a strength over neural networks when handling tabular data, which is how the data will be structured in the analysis, and finally, it is fast at processing data compared with other machine learning algorithms. This is important when it comes to the number of models required in a grid search (discussed later).

Sample size

Minimum sample size was derived using 'pmsampsize v1.1.0' package for R v3.6.1 for Windows [27]. This package is based on the work of Riley et al. for calculating sample sizes for prediction models [28, 29]. A systematic review of similar outcomes including discharge from ED, critical care requirement and hospitalisation informed the sample size [18]. From these studies, the average C-statistic was 0.80. Candidate variables were examined in the non-conveyed data to estimate parameters. A limitation with XGBoost is the handling of categorical variables. This requires each category within a variable to become its own binary variable which has a

single degree of freedom. There was a total of 637 parameters identified in the data. The total parameters per variable can be found in the [supplementary material](#). A study examining avoidable conveyances reported a conservative estimate of 9% avoidable conveyances in the same population as this study [13]. The C-statistic was transformed into a Cox-Snell R^2 via the *pmsampsize* package [30]. The arguments used in *pmsampsize* were therefore type = binary, C-statistic = 0.80, parameters = 637 and prevalence = 0.09. This gave an estimated minimum sample size of 52,958, with an anticipated 4767 events and an EPP of 7.48. A frequency analysis of the actual ePCR dataset shows there were 328,763 patients eligible for inclusion. However, the outcome measure requires data linkage, with unsuccessful linkage causing cases to be excluded [31]. This will likely result in fewer incidents to be included in the study.

Missing data

Missing values within the candidate variables will be handled as described above. If a variable contains missing values, it will be assessed as to whether they are the negative class within the variable as opposed to missing. This will be done by analysing the variable in the context of the ePCR to check if the field is only completed if the event happened. If this is the case, the missing values will be imputed with 'none'. Once this has been completed, any variable with more than 30% missing data will be excluded from analysis, as this provides evidence that the variable is not routinely collected and could cause model failure in practice, if included. Once the candidate predictors have been assessed for missing values, missing fields in each case will be examined. If any case does not have the outcome variable, but an ED record present, they will be excluded from the analysis. During model development, missing data will be handled via sparsity-aware split finding. This happens as part of the XGBoost algorithm. It uses non-missing data at each split to generate a default split. Then if there is missing information at the node, the algorithm defaults down the branch [32].

Variable handling

Nominal, ordinal and binary will be treated in the same way and will be one-hot encoded into binary variables. Continuous variables will remain in their natural format. Feature engineering of a previous attendance within 24 h of the current incident will also be engineered into a binary variable. The rationale to create this variable is so the model is aware of a second contact with the emergency service (ensuring it accounts for repeat presentations, which can indicate a missed problem the first time). All variables will be included in the model development initially. Then, the model will undergo Recursive

Feature Elimination (RFE). A feature importance score will be assigned to each feature and the least important stripped from the model. The model will be developed again with the same default hyperparameters but with one less feature. This repeats with the accuracy being recorded each time. The optimum set of features to take forward into model development will be identified by the model with the highest *C*-statistic with the default parameters. The data will be subset to only the features that yielded the optimum *C*-statistic and this subset will be used for all further modelling.

Hyperparameters

To prevent model overfitting, there will be tuning of hyperparameters before developing each model. This will be done with a fixed set of values for certain hyperparameters within a restricted search space. In order to optimise the search space for the grid search, individual hyperparameters will be tuned on the whole dataset sequentially and the 3 best performing values within each hyperparameter will be taken forward to create the restricted grid search space for all subsequent modelling. The following hyperparameters will be tuned:

To control model complexity the following hyperparameters will be tuned:

max_depth—The maximum depth of each tree. The initial search space will be between 2 and 10, with intervals of 1.

min_child_weight—This is a threshold for whether to continue partitioning a tree based on the sum of instance weight, with larger numbers creating a more conservative model. The initial search space will be 1 and 10, with intervals of 1.

gamma—Also known as *min_split_loss*. Like *min_child_weight*, it is a threshold for further partitions, but is based on the minimum loss reduction. Initial search space will be between 0 and 10 with intervals of 0.5.

To introduce randomness, making the training data more robust to noise, the following hyperparameters will also be tuned.

subsample—This is the percentage of the training data that is randomly sampled at each boosting iteration. Initial search space will be between 0.5 and 1, with intervals of 0.1.

colsample_bytree—Indicates what fraction of columns (features) are selected for tree development per tree. Initial search space will be between 0 and 1, with intervals of 0.1.

eta—step size shrinkage. The initial search space will be between 0 and 1, with intervals of 0.1.

Once the restricted search space has been defined, each time the modelling process requires hyperparameter tuning, the grid search will run a total of 729 iterations to find the optimum set of hyperparameters.

All other parameters will be fixed at the default value.

Development of the model

Conventional modelling strategies involve developing an unadjusted model on the dataset and then evaluating the apparent validity by testing the performance on the same dataset it was developed on. Then, through a process of resampling multiple times, models for each 'resample' can be developed by following the exact same modelling steps as in the apparent model. Once this has been completed, the average performance can then act as a penalty on the original model, creating an optimism-adjusted model. This is known as internally validating a model as it has been developed using resampling samples, but from the same data [33]. External validation should occur in a different sample from the development data, and preferably in a different geography and/or time frame [33].

Apparent validation

In the strategy proposed here, the algorithm does not create an unpenalized model to begin with. This is because tuneable hyperparameters are used to determine *how* the algorithm is developed on the data, prior to model development. In this way, the resultant model is already penalised at the point of development. To obtain the apparent validity of the full model, the three-step process of tuning hyperparameters, building a model on the optimal hyperparameters, and then evaluating the performance will occur on the full dataset. This will be the final model, as it has used the most information of the underlying population in development. The performance however will still be optimistic, even with the tuned hyperparameters as it has been evaluated in the same data it was derived from.

Internal-external validation

This study benefits from using individual patient data (IPD) from regional datasets clustered by ED. This provides an opportunity for internal-external cross-validation (IECV).

Cross-validation (CV) is a method whereby the data is split into *K* number of partitions (folds) and one-fold is left out as a 'test set'. The remaining folds are used collectively to train a model. Once the trained model has been applied to the test set, performance measures are recorded, and the set is placed back in the data. The next fold is then held out and the process repeated. This repeats until all folds have been held out. The benefit of cross-validation is that it provides a spread of performance instead of a point estimate. This is useful for indicating model stability.

Nested cross-validation is a variant of CV and consists of an inner-loop and an outer-loop. Like the CV

procedure above, the data is split into ten random parts. Then, one tenth is removed as an outer loop test set and the remaining nine tenths are split into random folds again. Due to the quantity of models being developed using this method, this is likely to be 5 random folds. One fifth of the inner loop is removed (inner loop test set), and hyperparameters can be grid-scanned using the data from the remaining four fifths. Optimum hyperparameters are then applied to the inner test set for

performance checking. The inner test set is then replaced, and the next fifth removed. The process repeats until all five folds have been removed and tested. The best performing inner loop model has its hyperparameter values applied to the whole inner loop (in-effect, outer-loop training set) to develop a model. This is then applied to the outer-loop test set for model performance. This outer loop tenth is then replaced and the process repeated. Performing nested CV internally validates the

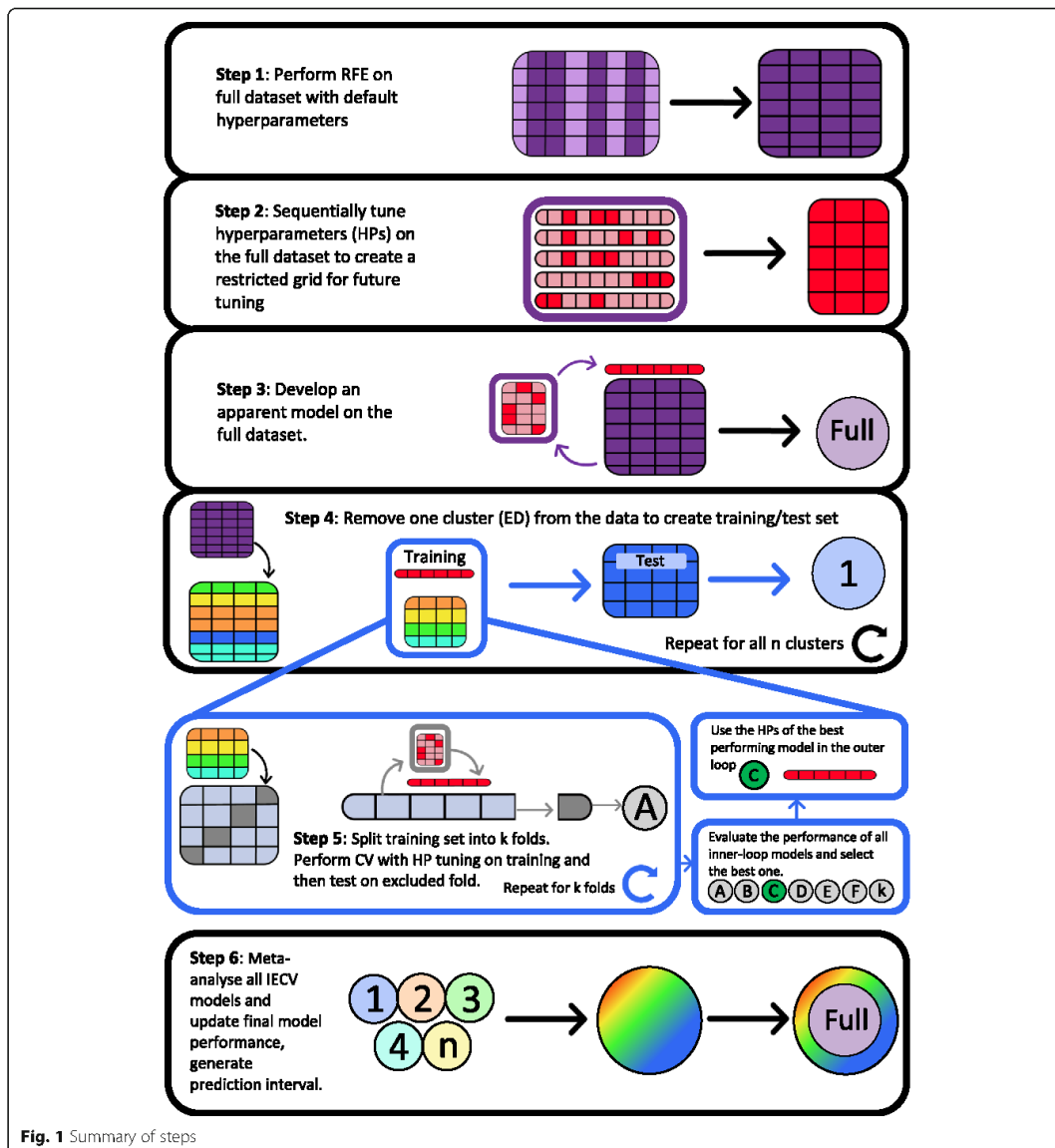


Fig. 1 Summary of steps

model as it is resampling; however, the random splits mean it is not being tested in a new geography.

As a way of simulating this, outer loop test sets are not random but in fact spatial clusters. In this way, the model is being internally-externally validated as it is resampling from the same data but testing it in a new population.

For spatial validation, a different ED will be used for each outer loop holdout. For example, 'the Sheffield ED model' will be trained on all EDs except for Sheffield, and then the performance tested on the Sheffield data. There are sixteen EDs in Yorkshire and therefore there will be sixteen spatial clusters.

A limitation of this modelling strategy is the computational expense. For every model, there needs to be 729 models built to identify the optimal hyperparameters. This is then repeated 5 times in each of the inner loops. With 29 clusters (including the full dataset), this means that there will be ~102,060 models required to be developed. If it becomes too expensive, then the number of inner loop folds can be reduced from 5 to 3, and any hyperparameter that has the default value as the optimum value in the preliminary search space will become fixed.

The different cluster results will then be pooled into a random-effects meta-analysis [34]. This is to estimate the average performance, the magnitude of heterogeneity between clusters and the range of performance across settings [35]. The predictor effects will not change from the internally validated model, but the performance measures will be updated according to the results of the meta-analysis. It would also be possible to derive a prediction interval for how the model would be expected to perform in a similar population.

Evaluating the model performance

For hyperparameter tuning, the *C*-statistic will be used to measure performance. For the apparent and IECV models, there will be three evaluations. The first is the goodness-of-fit as a general measure of model performance. This will be the Cox-Snell pseudo R^2 . For discrimination, the *C*-statistic will be used and receiver operating characteristic (ROC) curve plotted. For calibration, the plot, intercept and slope will be calculated. All the evaluation metrics will be entered into the meta-analysis to pool and update the performance of the final (full) model. Below is a figure graphically representing the modelling steps (Fig. 1).

Discussion

Benefit of a new tool

This study aims to develop a prediction model that can be used to create a tool supporting paramedics in making appropriate and effective decisions for patients who

may not require the level of care provided by a hospital. It is important as it is aiming to navigate care decisions that will safely provide patients with the right care, first time. If a paramedic can see the probability that their patient may have an avoidable attendance, it opens an opportunity to explore community options. It also empowers the patient to be an active partner in developing a self-care plan.

It could also have secondary benefits such as freeing ambulance fleet capacity to respond to a patient still waiting for help. With less patients being transported to the ED with low-acuity problems, it could also contribute to minimising delays in care for those who do need specialist ED interventions.

Presenting the model as a tool

It is anticipated that the prediction model can be presented as a probability of the positive class to the clinician. As an illustrative example, once all predictor variables are inputted into the ePCR by the clinician, it may display the following message—"The probability of this patient having an avoidable attendance at ED is 32%".

Abbreviations

ECDS: Emergency Care Data Set; ED: Emergency department; CAG: Confidentiality Advisory Group; CDS: Commissioning Data Set; CI: Confidence interval; CV: Cross-validation; ED: Emergency department; EMS: Emergency medical service; ePCR: electronic patient care record; GP: General Practitioner; IECV: Internal-external cross-validation; IPD: Individual patient data; NHS: National Health Service; REC: Research Ethics Committee; ROC: Receiver operating characteristic; TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis; YAS: Yorkshire Ambulance Service

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s41512-021-00108-4>.

Additional file 1.

Acknowledgements

Julia Williams has significantly contributed to the development of the project and its relevance to prehospital care and clinical practice. JW is also a co-supervisor on the project.

Authors' contributions

JM is the study lead and drafted the manuscript. SM is the lead supervisor for this study, contributed to the development of the research question and its overall design. SM has contributed to the drafting of this manuscript. RJ is a co-supervisor and informed the statistical analysis plan in this manuscript. JT is a co-supervisor and informed the clinical importance of the study in this manuscript.

Funding

This report is independent research supported by Health Education England and the National Institute for Health Research (HEE/NIHR ICA Programme Clinical Doctoral Research Fellowship, Mr Jamie Miles, ICA-CDRF-2018-04-ST2-044). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, The National Institute for Health Research or the Department of Health and Social Care.

This report is independent research funded by the National Institute for Health Research, Yorkshire and Humber Applied Research Collaborations. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care.

Availability of data and materials

This study uses data from NHS Digital for research purposes and therefore data will not be available following completion of the study in accordance with the data sharing agreement. Findings from the research will be published in a peer-reviewed, open-access journal and disseminated at relevant conferences. The tool will be presented to appropriate stakeholders for real-world prospective evaluation.

Declarations

Ethics approval and consent to participate

This research study has received ethical approval from the NHS Health Research Authority. It received ethical approval from the Yorkshire and Humber Research Ethics Committee (REC) on 11 November 2019 (ref: 19/YH/0360). As this study uses data without the participants' consent, it has also undergone approval from the Confidentiality Advisory Group (CAG). This was approved on 14 July 2020 (ref: 20/CAG/0035).

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

¹CURE Group, School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent Street, Sheffield S1 4DA, UK. ²Yorkshire Ambulance Service, Brindley Way, Wakefield WF2 0XQ, UK. ³School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent Street, Sheffield S1 4DA, UK

Received: 6 May 2021 Accepted: 25 October 2021

Published online: 08 November 2021

References

- National Audit Office. NHS Ambulance services. 2017.
- Coster JE, Turner JK, Bradbury D, Cantrell A. Why do people choose emergency and urgent care services? A rapid review utilizing a systematic literature search and narrative synthesis. *Acad Emerg Med*. 2017;24 [cited 2020 Sep 16]. p. 1137–49. Available from: <https://doi.org/onlineibrary.wiley>.
- NHS England [online]. Statistics > Urgent and emergency care daily situation reports. [cited 2021 Feb 15]. Available from: <https://www.england.nhs.uk/statistics/statistical-work-areas/uec-sitre/>
- Andrew E, Nehme Z, Cameron P, Smith K. Drivers of increasing emergency ambulance demand. *Prehospital Emerg Care*. 2020; [cited 2020 Dec 4];24(3):385. Available from: <https://www.tandfonline.com/action/journalInformation?journalCode=ipecc20>.
- O' Cathain A, Knowles E, Long J, Connell J, Bishop-Edwards L, Simpson R, et al. Drivers of 'clinically unnecessary' use of emergency and urgent care: the DEUCE mixed-methods study. *Heal Serv Deliv Res*. 2020;8(15):1–256. <https://doi.org/10.3310/hsdr08150>.
- O'Hara R, Johnson M, Hirst E, Weyman A, Shaw D, Mortimer P, et al. A qualitative study of decision-making and safety in ambulance service transitions. *Heal Serv Deliv Res*. 2014;2(56):1–138. Available from: <https://www.journalslibrary.nihr.ac.uk/hsdr/hsdr02560/>. <https://doi.org/10.3310/hsdr02560>.
- Burrell L, Noble A, Ridsdale L. Decision-making by ambulance clinicians in London when managing patients with epilepsy: a qualitative study. *Emerg Med J*. 2013;30(3):236–40. <https://doi.org/10.1136/emmermed-2011-200388>.
- Halter M, Vemon S, Snooks H, Porter A, Close J, Moore F, et al. Complexity of the decision-making process of ambulance staff for assessment and referral of older people who have fallen: a qualitative study. *Emerg Med J*. 2011;28(1):44–50. <https://doi.org/10.1136/emj.2009.079566>.
- Simpson P, Thomas R, Bendall J, Lord B, Lord S, Close J. 'Popping nana back into bed' - a qualitative exploration of paramedic decision making when caring for older people who have fallen. *BMC Health Serv Res*. 2017;17(1):1–14. <https://doi.org/10.1186/s12913-017-2243-y>.
- Hoikka M, Silfvast T, Ala-Kokko TI. A high proportion of prehospital emergency patients are not transported by ambulance: a retrospective cohort study in Northern Finland. *Acta Anaesthesiol Scand*. 2017;61(5):549–56. <https://doi.org/10.1111/aas.12889>.
- Brydges M, Spearen C, Birze A, Tavares W. A culture in transition: paramedic experiences with community referral programs. *Can J Emerg Med*. 2015; 17(6):631–8. <https://doi.org/10.1017/cem.2015.6>.
- Patton GG, Thakore S. Reducing inappropriate emergency department attendances - a review of ambulance service attendances at a regional teaching hospital in Scotland. *Emerg Med J*. 2013;30(6):459–61. <https://doi.org/10.1136/emmermed-2012-201116>.
- O'Keeffe C, Mason S, Jacques R, Nicholl J. Characterising non-urgent users of the emergency department (ED): a retrospective analysis of routine ED data. *PLoS One*. 2018;13(2):1–14. <https://doi.org/10.1371/journal.pone.0192855>.
- Miles J. 17 Exploring ambulance conveyances to the emergency department: a descriptive analysis of non-urgent transports. *Emerg Med J*. 2017; Available from: <http://europepmc.org/abstract/med/29170314>.
- Miles J, Coster J, Jacques R. Using vignettes to assess the accuracy and rationale of paramedic decisions on conveyance to the emergency department. *Br Paramed J*. 2019;4(1):6–13. <https://doi.org/10.29045/14784726.2019.06.4.1.6>.
- North West Ambulance Service. Paramedic Pathfinder and Community Care Pathways. 2014;(September):52. Available from: <https://www.nwas.nhs.uk/DownloadFile.ashx?id=286&page=16586>
- Newton M, Tunn E, Moses J, Ratcliffe D, MacKway-Jones K. Clinical navigation for beginners: The clinical utility and safety of the Paramedic Pathfinder. *Emerg Med J*. 2013;31(e1):e29–34. <https://doi.org/10.1136/emmermed-2012-202033>.
- Miles J, Turner J, Jacques R, Williams J, Mason SM. Using machine-learning risk prediction models to triage the acuity of undifferentiated patients entering the emergency care system: a systematic review. *BMC Diagnostic Progn Res*. 2020; [cited 2020 Oct 2];4(1):16. Available from: <https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-020-00084-1>.
- Meisel ZF, Pollack CV, Mechem CC, Pines JM. Derivation and internal validation of a rule to predict hospital admission in prehospital patients. *Prehospital Emerg Care*. 2008;12(3):314–9. <https://doi.org/10.1080/10903120802096647>.
- Li J, Guo L, Handy N. Hospital admission prediction using pre-hospital variables. 2009 IEEE Int Conf Bioinforma Biomed BIBM 2009. 2009;283–6.
- Seymour CW, Kahn JM, Cooke CR, Watkins TR, Heckbert SR, Rea TD. Prediction of critical illness during out-of-hospital emergency care. *JAMA*. 2010;304(7):747–54. <https://doi.org/10.1001/jama.2010.1140>.
- van Rein EAJ, van der Sluijs R, Voskens FJ, Lansink KWW, Houwert RM, Lichtveld RA, et al. Development and validation of a prediction model for prehospital triage of trauma patients. *JAMA Surg*. 2019;154(5):421–9. Available from: <http://search.ebscohost.com/login.aspx?direct=true&db=cin20&AN=136501962&site=ehost-live>. <https://doi.org/10.1001/jamasurg.2018.4752>.
- Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann Intern Med*. 2015 Jan 6;162(1):W1–73. <https://doi.org/10.7326/M14-0698>.
- NHS Digital. Non-urgent A&E attendances. 2020 [cited 2020 Sep 16]. Available from: <https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/innovative-uses-of-data/demand-on-healthcare/unnecessary-a-and-e-attendances>
- Raita Y, Goto T, Faridi MK, Brown DFMM, Camargo CAJ, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit Care*. 2019;23(1):1–13. Available from: <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=prem&NEWS=N&N=30795786>. <https://doi.org/10.1186/s13054-019-2351-7>.
- Goto T, Camargo CAJ, Faridi MK, Freishtat RJ, Hasegawa K. Machine learning-based prediction of clinical outcomes for children during emergency department triage. *JAMA Netw Open*. 2019;2(1):e186937 Available from: <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=prem&NEWS=N&AN=30646206>.
- Ensor J, Martin EC, Riley RD. Package "prmsampsiz": calculates the minimum sample size required for developing a multivariable prediction model. 2020

- [cited 2020 Sep 10]. Available from: <https://cran.r-project.org/web/packages/pmsamplesize/pmsamplesize.pdf>
28. Riley RD, Snell KJ, Ensor J, Burke DL, Jr FEH, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med*. 2019; [cited 2021 Aug 26]; 38(7):1276–96. Available from: <https://onlinelibrary.wiley.com/doi/full/10.1002/sim.7992>.
 29. Riley RD, Debray TPA, Collins GS, Archer L, Ensor J, Smeden M van, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat Med*. 2021 [cited 2021 Aug 26];40(19):4230–51. Available from: <https://onlinelibrary.wiley.com/doi/full/10.1002/sim.9025>
 30. Riley RD, Snell KJ, Ensor J, Burke DL, Harrell Jr FE, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II-binary and time-to-event outcomes. 2018
 31. NHS Digital. Linked datasets supporting health and care delivery and research. 2018;(April):1–14. Available from: <https://digital.nhs.uk/services/data-access-request-service-dars/linked-datasets-supporting-health-and-care-delivery-and-research>
 32. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. [cited 2021 Aug 26]; Available from: <https://github.com/dmlc/xgboost>
 33. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer New York; 2009. (Statistics for Biology and Health). Available from: <http://link.springer.com/10.1007/978-0-387-77244-8>
 34. Riley RD, Moons KGM, Snell KIE, Ensor J, Hooft L, Altman DG, et al. A guide to systematic review and meta-analysis of prognostic factor studies. [cited 2020 Oct 19]; Available from: <https://doi.org/10.1136/bmj.k4597> <http://www.bmj.com/>.
 35. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. [cited 2021 Sep 2]; Available from: <https://doi.org/10.1136/bmj.i3140> <http://www.bmj.com/>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



Chapter 7

The Protocol Expansion

7.1 Introduction

In the previous chapter, the protocol was published in manuscript form with succinct paragraphs detailing what is being done in this thesis. However, some elements of the manuscript need expanding to justify why these methods were chosen. This chapter provides justification where necessary and includes changes to the protocol that have happened since its publication.

7.2 Study design and setting

7.2.1 Design

An avoidable conveyance to the ED was a binary outcome as it was either avoidable or not. The positive class of the binary outcome is labelled (1) and indicates to support a decision not to convey the patient to the ED. The negative class is labelled (0) and supports a decision to convey the patient. No study in the previous literature has built a model using only a single candidate variable and therefore a multivariable prediction model was created. In the Transparent Reporting of multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines they describe the difference between a diagnostic prediction model and a prognostic one.¹⁴⁶ The description of a diagnostic model fulfils the primary aim of this thesis:

In the diagnostic setting, the probability that a particular disease is present can be used, for example, to inform the referral of patients for further testing, to initiate treatment directly, or to reassure patients that a serious cause for their symptoms is unlikely.¹⁴⁶

This is further reiterated in a more recent statistical note, which highlights the difference between the two is a 'temporal relationship between the moment of prediction and the outcome of interest'.¹⁴⁷

There are different study designs in prediction modelling, including retrospective, prospective, case-control and registry. In retrospective studies, the data is already collected at the point the study starts. A benefit to this design is that large samples can often be generated at low cost. However, the limitation is that only what has been collected can be used. This means if the ideal candidate predictors are not available in the data, or the outcome is not collected appropriately, it becomes a threat to the project. It is also difficult to achieve samples from different areas or information-governed by different systems due to issues with consent and data protection. Conversely, the prospective study design allows for customisable data collection with purposeful sampling. This extends to the case-control design, where two cohorts can be defined, collected and compared. These overcome the limitation of the retrospective design but inherit the problem of sample size. The prospective nature of recruitment renders the study design limited to smaller geographies or fewer patients in the sample. A registry design allows for larger sample sizes to be considered as they can cover large geographies and share the benefits of the prospective design in their data collection. They are more useful in longitudinal studies for following patients at different time points.¹²⁸ A consideration in this thesis is the quantity of patients required to create an accurate model. Due to the outcome prevalence being so low, the sample size needed to be greater in order to capture enough events to build a model. Furthermore, the sample needed to capture rural, urban and coastal events in order to be generalisable. In this respect, the two study designs that benefitted the situation were the retrospective or registry design. Fortunately, NHS Digital were able to offer a halfway house in that they collect data from all acute trusts in England and store the data centrally. This meant a retrospective study design, using the scale of data found in registry studies.

7.2.2 Setting

This study was set within the boundaries of Yorkshire Ambulance Service (YAS). YAS serves a population of over five million people and covers 6000 miles of varied terrain from the isolated Yorkshire Dales and North York Moors to urban

areas including Bradford, Hull, Leeds, Sheffield, Wakefield, and York. There are sixteen type 1 EDs in Yorkshire. This includes three Major Trauma Centres (MTCs) located in Sheffield, Leeds, and Hull. A type 1 ED has been described by NHS England as:

“A consultant led 24-hour service with full resuscitation facilities and designated accommodation for the reception of accident and emergency patients.”¹⁴⁸

There are two other types of ED possible, a type 2 and a type 3. These only offer a limited service and would not routinely manage the critical care of a patient and could lack on-site specialist services. Due to this, they are excluded from this study. A full list of the Emergency Departments used in this study can be found in appendix E. There was a single ED located outside of Yorkshire, which was the James Cook University Hospital in Middlesbrough. This is because it is the nearest hospital for a small geography in north Yorkshire and YAS transported several patients there.

7.2.3 Inclusion criteria

Ambulance services often have specific conveyance policies around the transportation of child patients. As such, only adult patients were included in this study. One such example of a policy was any patient under the age of 5 had to be transported to the ED. The data collection period was bound by two events. The start was from the point at which Yorkshire Ambulance Service was using an electronic Patient Care Record (ePCR) within the entirety of its footprint. This was a transition away from paper-based records into an electronic format which records the whole patient journey whilst they are with the ambulance service. The end was at the point there was a significant increase in COVID-19 infections in Yorkshire that could confound the dataset. As increasing non-conveyance was likely to be a benefit of the model, the non-conveyed patients from this period

were extracted for validation and form a separate cohort. This led to the following inclusion and exclusion criteria:

- | | |
|--|--|
| Inclusion 1 (for Dataset 1 (DS1) ED information) | <ul style="list-style-type: none">• Age 18 years old or older.• Transported to ED by Yorkshire Ambulance Service between July 1st 2019 and 29th February 2020.• Have an ED Care record of the event. |
| Inclusion 2 (for DS1 Ambulance information) | <ul style="list-style-type: none">• Age 18 years or older.• Assessed by a qualified ambulance clinician ((either paramedic (of any level) or technician grade II)).• Had an electronic patient care record completed.• Transported to an ED between July 1st 2019 and 29th February 2020.• Were handed over and booked in as a patient to the ED |
| Inclusion 3 (for Dataset 2 (DS2)) | <ul style="list-style-type: none">• Age 18 years or older.• Assessed by a qualified ambulance clinician (either paramedic or technician grade II).• Had an electronic patient care record completed.• Discharged on scene and not transported between July 1st 2019 and 29th February 2020. |
| Exclusion 1 (for DS1 ED) | <ul style="list-style-type: none">• Patient cases where they were less than 18 years old at time of episode.• Patient cases where there were five or more attendances within the data collection period. |
| Exclusion 2 (for DS1 Ambulance Service) | <ul style="list-style-type: none">• Patient cases where they were less than 18 years old at time of episode.• Patient cases where they had five or more patient contacts within the data collection period. |

- Exclusion 3
(for DS2
Ambulance
Service)
- Patient cases where they were less than 18 years old at time of episode.
 - Patient cases where they had five or more patient contacts within the data collection period.
 - Patient cases that were transported by the ambulance crew on scene.

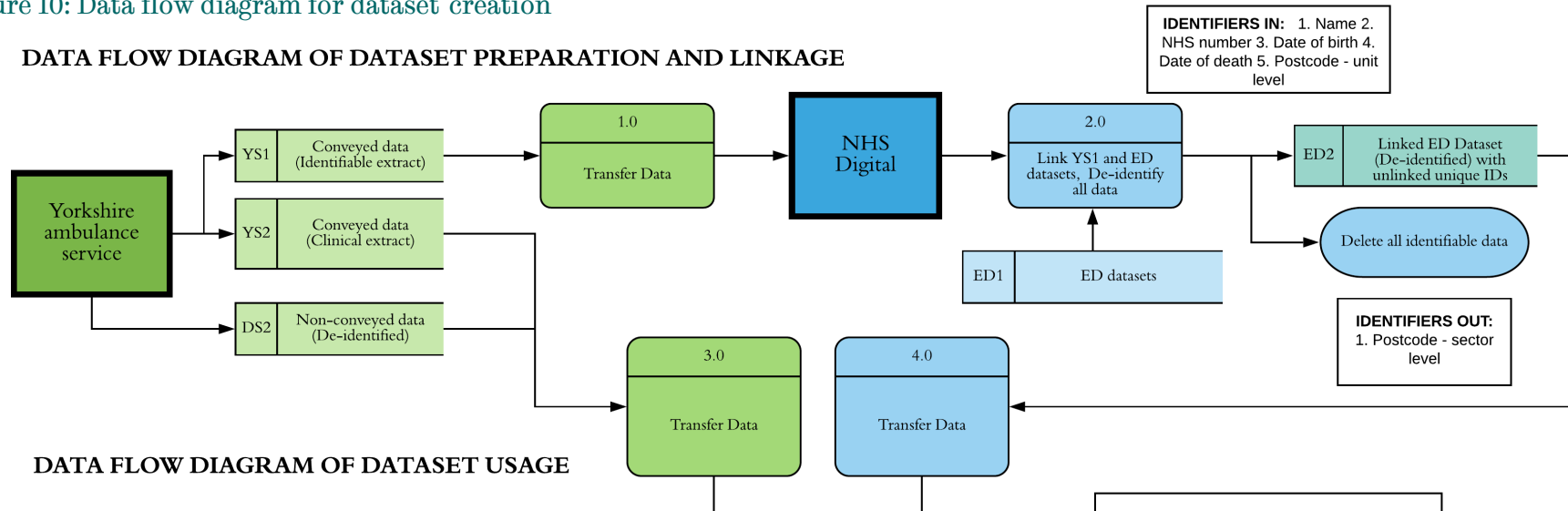
7.4 Dataset creation

At present, ambulance data and emergency department data are routinely collected but stored separately. Every patient that is seen by a paramedic face-to-face has an electronic Patient Care Record (ePCR) generated for that episode of care, which is stored locally in the ambulance services data warehouse. Similarly, every patient that attends the ED would also have a record of their ED attendance. This is stored locally within the trust, but also sent centrally to NHS digital where it is processed into a data product along with data submitted from other trusts. With linkage of these two datasets, it is possible to map a patient's journey from when they called an ambulance, to when they left the ED. Figure 10 below illustrates the process of creating the linked dataset used for this study. All ePCRs that met the ambulance service inclusion criteria were extracted from the YAS data warehouse. A unique random number was assigned to each record as a 'study ID'. The extract was then split into clinical information (YS1) and patient identifiable information (YS2). This was for the purpose of data security, so that no one outside of YAS would see the completely identifiable patient episode. The clinical extract with the study ID attached was sent to the University of Sheffield, whilst the identifiers and study ID were sent to NHS Digital, who hold the ED data. NHS Digital then linked ED records to the identifiers using an eight level deterministic matching algorithm.¹⁴⁹ This used a combination of NHS number (unique to every patient), date of birth, sex, and patient postcode. Once the records were linked, all identifiers except the study ID were destroyed and the ED data was then securely transferred to the University of Sheffield. The ambulance service ePCR data extract was then linked to the ED data extract using the study

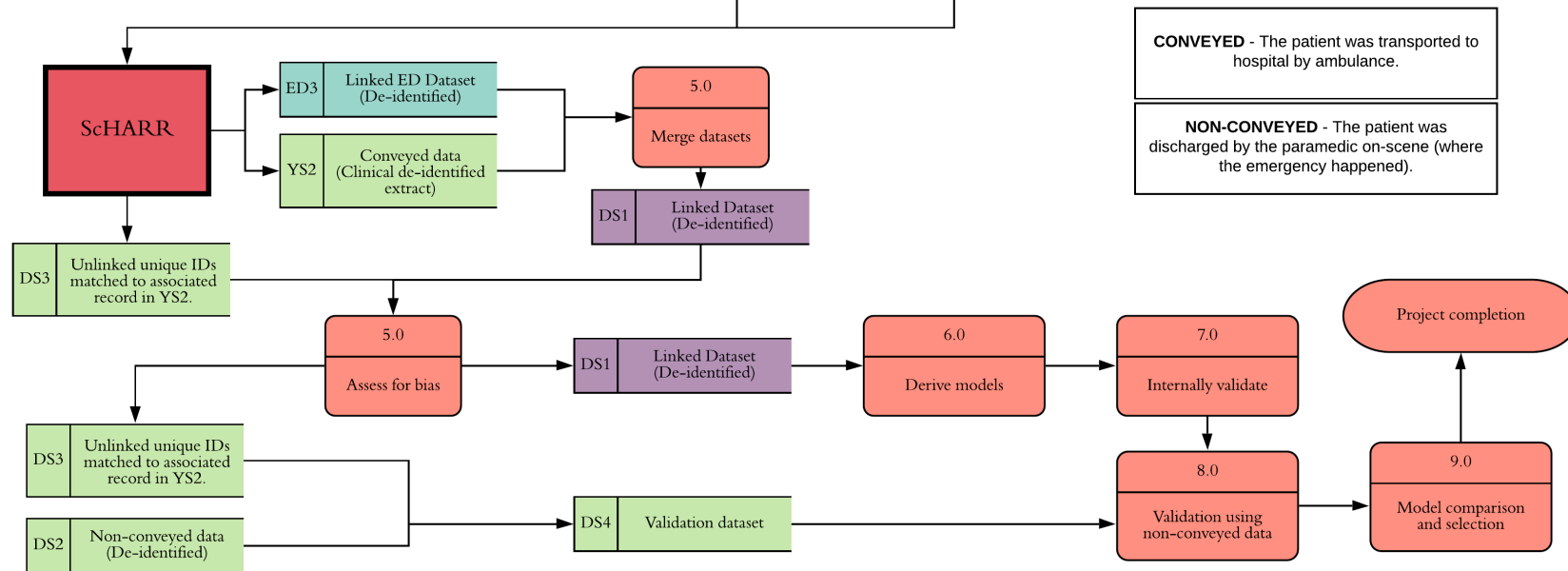
ID, which was present in both records. This creates DS1 whilst the non-conveyed clinical extract is DS2.

Figure 10: Data flow diagram for dataset creation

DATA FLOW DIAGRAM OF DATASET PREPARATION AND LINKAGE



DATA FLOW DIAGRAM OF DATASET USAGE



7.5 Outcome variable

The outcome variable for this study was an avoidable conveyance (by ambulance) to the ED. It was considered that if a patient attended the ED and had a non-urgent experience, then it would fulfil the definition of the outcome. However, it is important to note that this could be considered a conservative outcome measure as it will not capture accurately all patients who would have found a more appropriate care setting elsewhere. In a sense, it is bias towards the extreme end of low acuity. The reason that the experience threshold of non-urgent was selected, was because there was a clear definition that could be applied to the dataset.

It is, however, recognised that using a data driven definition as the gold standard has its limitations. For example, a 2019 study examining diagnostic error in the ED found that nationally in England between 2013 and 2015, there were 5,412 diagnostic errors in the ED. Of these, 2,288 resulted in a patient safety incident.¹⁵⁰ The gold standard is not perfect, and as mentioned in chapter 2, section 2.3 provider induced demand will limit the accuracy of the outcome measure. This is because the outcome measure assumes all investigations and treatments were needed, whereas the evidence referenced above shows that is not always the case.

7.5.1 The non-urgent definition

Lowy et al. used a random sample of 6439 patients to develop a process-based definition. Their definition contained five features outlined below:

- Registered with a GP
- Not investigated in the ED
- Not treated in the ED except for a prescription, bandage, sling, dressing or steristrips
- Did not come from a road traffic accident or an accident at work, school, or public place or a sporting event, and were:

- Discharged completely from care in the ED or referred to their GP.

*Adapted from Lowy et al.*¹⁵¹

O’Keeffe et al. updated this definition using a larger sample of 3,667,601 first time attendances to the Emergency Department.⁵ The definition was simplified to contain the following:

“First attendance with some recorded treatments or investigations, all of which may have reasonably been provided in a non-emergency care setting, followed by discharge home or to GP care”⁵

7.5.2 Instrument of measurement

For the purpose of sentinel surveillance, NHS Digital uses a definition of a non-urgent attendance empirically defined using clinical coding. By doing so, the data can be captured locally and analysed nationally. Their choice of definition was by O’Keeffe et al. and was initially operationalised into the dictionary of codes known as the Commissioning Data Set (CDS) type 10 codes but later translated into CDS type 11 - the Emergency Care Data Set (ECDS). The corresponding codes to the definition can be found in table 3:

Table 3: NHS Definition of low acuity attendance

| Criteria | ECDS codes |
|---|---|
| <i>Department type</i> | |
| Type-1 Emergency Department | 01 |
| <i>Attendance disposal (discharge status)</i> | |
| Discharged – follow-up treatment to be provided by general practitioner | 1077021000000100 |
| Discharged – did not require any follow-up treatment | 182992009 |
| Left department before being treated | 1066321000000107 |
| <i>Investigations</i> | |
| Urinalysis | 27171005 |
| Pregnancy test | 167252002 / 67900009 |
| Dental Investigation | 53115007 |
| None | 1088291000000101or blank |
| <i>Treatment</i> | |
| Guidance/advice only - written | 413334001 |
| Guidance/advice only - verbal | Not applicable |
| Recording vital signs | Not applicable |
| Dental treatment | 81733005 |
| Prescription/medicines prepared to take away | 266712008 |
| None (consider guidance/advice option) | 183964008 or blank |
| Prescriptions (retired code but still present in some records) | Not applicable |
| <i>Attendance category</i> | |
| First Accident and Emergency attendance | 01 |
| <i>Arrival mode</i> | |
| Non-ambulance arrivals | 1048071000000103, 1048061000000105, 1047991000000102, 1048001000000106 |

Adapted from O’Keeffe and NHS Digital^{5,152}

However, this definition falls short of describing avoidable ambulance conveyances. The ambulance service can provide a healthcare professional on scene who arrives in a mobile clinic. This means by the time the patient arrives at the ED they have already been triaged, investigated, treated and a care plan made by a fellow healthcare service. Transporting the patient to ED is a clinically decided upgrade in care to a more acute healthcare system. Furthermore, the definition by O’Keeffe et al. is bound to the premise that the level of care the patient requires falls below the threshold of emergency care. It is not necessarily what the ED can perform for the patient, but what level of care the patient needs. A definition of avoidable conveyance equally should not be bound to what the ambulance service can provide on scene but should stay true to the premise that the level of care required for their patient could be met in the community or primary care. This requires a modification to the NHS Digital definition.

7.5.3 Modifying the NHS Digital definition

In order to clearly define the parameters for this outcome measure, the definition used by NHS Digital needed modifying to be more stipulative. The first modification needed to be the arrival mode. NHS Digital only include non-ambulance arrivals in their definition, however, in the O’Keeffe study, there were 8.5% of ambulance attendances that qualified as a non-urgent attendance in ED.⁵ NHS Digital make an assumption that by virtue of arriving by ambulance, it characterises the patient as an emergency care patient. The inverse definition of arrival mode was applied, with only those arriving by road ambulance included. It was assumed that an ambulance conveyance via a helicopter or with a medical escort would still qualify as an emergency.

The second modification required was a robust translation from the CDS type 10 to the CDS type 11 (ECDS). NHS Digital performed a direct translation between the two coding types; however, there appear to be new codes that naturally fit with their original definition. The third modification is updating the definition to

include the routine clinical investigations and treatments that are widely offered by ambulance services in the UK.

7.5.3.1 Investigations

Investigations are largely translated correctly by NHS Digital into the ECDS however there is an extra code relating to urine testing ‘*Urine sent for culture – 168338000*’, which appeared rational to include. This was also included in the CDS coding under a code share with urinalysis. Table 4 shows all investigation codes.

Table 4: Investigation codes

| ECDS Description (Investigations) | ECDS code |
|--|--------------------------|
| <i>Clinical investigation not indicated*</i> | <i>1088291000000101*</i> |
| Dementia test | 165320004 |
| <i>Diagnostic dental procedure*</i> | <i>53115007*</i> |
| Glucose measurement, blood, test strip | 104686004 |
| <i>Human chorionic gonadotropin measurement*</i> | <i>67900009*</i> |
| Peak expiratory flow measurement | 29893006 |
| <i>Urinalysis*</i> | <i>27171005*</i> |
| <i>Urine pregnancy test*</i> | <i>167252002*</i> |
| Urine sent for culture | 168338000 |

* Validated definition by O’Keeffe et al. and officially adopted by NHS Digital.⁵

7.5.3.2 Treatments

In the latest version of ECDS there are extra codes, which could reasonably be provided in the community and do not necessarily require the expertise of the ED. These included social care treatments such as assessing a patient’s activities of daily living or mobility. Table 5 shows all treatment codes that could be provided in the community.

Table 5: Treatment codes

| ECDS description (Treatments) | ECDS Code |
|---|-------------------|
| Activities of daily living assessment | 304492001 |
| Application of dressing, minor | 15631002 |
| Assessment of mobility | 430481008 |
| Closure of skin wound by tape | 71810007 |
| <i>Dental surgical procedure*</i> | <i>81733005*</i> |
| Gluing | 284182000 |
| Mobility/transfers education, guidance, and counselling | 410267000 |
| <i>New medication commenced*</i> | <i>266712008*</i> |
| <i>Patient given written advice*</i> | <i>413334001*</i> |
| Psychosocial assessment | 371585000 |
| Review of medication | 182836005 |
| Social assessment | 406551008 |
| <i>Treatment not indicated*</i> | <i>183964008*</i> |

* Validated definition by O’Keeffe et al. and officially adopted by NHS

Digital.⁵

7.5.3.3 Discharge status

ECDS introduced streaming codes into the discharge status section. This was to solve the problem of inviting financial tariffs for patients whose care did not occur in the department itself.¹⁵³ Prior to these codes, a patient was recorded as normal (no investigations or treatments) if they were in fact streamed elsewhere and would attract the tariff of a ‘normal’ patient being seen in the Emergency Department. These codes are designed to be recorded when a patient is streamed immediately after initial assessment. They act as both a clinical and financial clarification of an individual patient’s experience. Table 6 below reproduced from Walker et al. states the exact ECDS codes that qualify as streaming codes:

Table 6: Discharge status codes

| ECDS Description (Discharge status) | ECDS Code |
|---------------------------------------|------------------|
| Streamed to primary care service / GP | 1077021000000100 |
| Streamed to Urgent Care Centre | 1077031000000103 |
| Streamed to falls service | 1077091000000102 |
| Streamed to frailty service | 1077101000000105 |
| Streamed to mental health service | 1077041000000107 |
| Streamed to pharmacy service | 1077071000000101 |
| Streamed to dental service | 1077051000000105 |
| Streamed to ophthalmology service | 1077061000000108 |

*Adapted from Walker et al.*¹⁵³

For patients who are triaged in the ED and then streamed to a lower level of care, it is a signal that they do not require the ED and can be classed as an avoidable conveyance. Whilst it is recognised these services may be co-located and thus the patient needs to be there, the destination of the ED is potentially avoidable as is transportation by ambulance. This is supported by clinical evidence and national policy which encourage the use of decision support in the ambulance service to stream patients to the most appropriate clinical area.^{95,154,155}

Using the modifications above, the experience-based definition for an avoidable conveyance to the ED in this thesis needed to meet all the criteria found in table 7. The ECDS data was still not ubiquitously collected and as such a CDS translation has also been included. A limitation in the translation is that CDS code shares with other investigations. For example, CDS code 05 could be glucose measurement, but it could also be any other type of biochemistry. It has these broader categories, which made a direct translation with ECDS difficult. The ECDS dataset is more granular and there are no codeshares as each code is unique in its definition. For example, the code for glucose measurement is 104686004, and each other biochemistry investigation also has its own unique code. The ideal situation was for all the data to be collected in ECDS and only a single outcome definition used. However, due to some trusts in the data collection only submitting CDS10 to NHS Digital, both the CDS and ECDS definitions must be

used. Missing values have been included in the definition below. On the balance of probability this is because there were no investigations/treatments etc. To elaborate, in the datasets there are up to twelve investigation ‘slots’ to record the first twelve investigations a patient had. If they only had two investigations, the remaining ten would record blank values. If an instance had all investigation or treatment fields missing, it was considered unable to create the outcome measure. It was deemed that this event was rare, due to the compulsory coding of these variables by clinical staff, and the presence of values that represent that there were no investigations or no treatments indicated.

Table 7: Full definition for this study

| Criteria | CDS | ECDS |
|---|-----|------------------|
| <i>Had the following attendance category:</i> | | |
| Unplanned First Emergency Care Attendance for a new clinical condition (or deterioration of a chronic condition). | 1 | - |
| <i>Had attended the following department type:</i> | | |
| Type 1: General Emergency Department (24 hour). | 1 | - |
| <i>Arrived only by:</i> | | |
| Arrival by emergency road ambulance | - | 1048031000000100 |
| Arrival by non-emergency road ambulance | - | 1048021000000102 |
| <i>Only had one or more of these Investigations:</i> | | |
| Clinical investigation not indicated | 24 | 1088291000000101 |
| Dementia test | 99 | 165320004 |
| Diagnostic dental procedure | 22 | 53115007 |
| Glucose measurement, blood, test strip | CS | 104686004 |
| Human chorionic gonadotropin measurement | 21 | 67900009 |
| Peak expiratory flow measurement | 99 | 29893006 |
| Urinalysis | 6 | 27171005 |
| Urine pregnancy test | 21 | 167252002 |
| Urine sent for culture | CS | 168338000 |
| visual acuity testing | CS | 16830007 |
| NA | - | - |

Only had one or more of these treatments:

| | | |
|--|-----|-----------|
| Activities of daily living assessment | 521 | 304492001 |
| Application of a dressing, minor | 11 | 15631002 |
| Assessment of mobility | 91 | 430481008 |
| Closure of skin wound by tape | CS | 71810007 |
| Dental surgical procedure | 56 | 81733005 |
| Gluing of wound | CS | 284182000 |
| Mobility/transfers education, guidance and counselling | 522 | 410267000 |
| New medication commenced | 57 | 266712008 |
| Patient given written advice | 221 | 413334001 |
| Psychosocial assessment | CS | 371585000 |
| Review of medication | CS | 182836005 |
| Social assessment | 54 | 406551008 |
| Treatment not indicated | 99 | 183964008 |
| Physiotherapy: Falls prevention | 92 | 391027005 |
| Observation/ cardiac monitor, pulse oximetry/ head injury / trends | 21 | 88140007 |
| NA | - | - |

Left the department in the following way:

| | | |
|--|----|------------------|
| Discharged – follow-up treatment to be provided by GP | 2 | - |
| Discharged – did not require any follow-up treatment | 3 | - |
| Transferred to other healthcare provider | 7 | - |
| Left department before being treated | 12 | - |
| Left department having refused treatment | 13 | - |
| Left care setting after initial assessment | - | 1066311000000101 |
| Left care setting before initial assessment | - | 1066301000000103 |
| Left care setting before treatment completed | - | 1066321000000107 |
| Streamed from ED to dental service following initial assessment | - | 1077051000000105 |
| Streamed from ED to falls service following initial assessment | - | 1077091000000102 |
| Streamed from ED to frailty service following initial assessment | - | 1077101000000105 |
| Streamed from ED to GP following initial assessment | - | 1077021000000100 |
| Streamed from ED to mental health following initial assessment | - | 1077041000000107 |
| Streamed from ED to ophthalmology following initial assessment | - | 1077061000000108 |
| Streamed from ED to pharmacy service following initial assessment | - | 1077071000000101 |
| Streamed from ED to urgent care service following initial assessment | - | 1077031000000103 |
| Treatment completed | - | 182992009 |

Discharge destination (for ECDS)

| | | |
|---|---|-----------|
| Home | - | 306689006 |
| Residential care facility without 24-hour nursing care (e.g., residential home) | - | 306691003 |
| Residential care facility with 24-hour nursing care (e.g., nursing home) | - | 306694006 |
| Police | - | 306705005 |
| Custodial services e.g., prison / detention centre | - | 50861005 |

7.5.4 Limitations with the outcome variable

The gold standard (reference standard) in this study is a data driven definition based on ED experience. However, it has not been considered or accounted for that the ED itself is susceptible to error. One in ten diagnoses in the ED is likely to be incorrect, and the rates of medical harm as a result of the incorrect diagnosis is higher amongst the ED population than the inpatient.¹⁵⁶ One Iranian study found that on direct observation of 202 patients, there was an average of 3.5 errors per patient.¹⁵⁷ It has been proposed that errors in the ED are attributed to interruptions during clinical procedures, multitasking, fatigue, and working memory capacity.¹⁵⁸ This evidences that the reference standard being used assumes an errorless environment, however this is unlikely to be the case. It is hoped that the sample size being large would mitigate this limitation.

By having a data driven definition, it also reduces information down into binary variables that are less representative of the real world. Elaborating on the discussion in chapter 5, section 5.2, there is a fluidity to both the urgent and emergency care system, the patients, and the interaction between the two. However, in a data driven definition, the fluidity has to be fixed to be able to describe whether a patient is low acuity. A more realistic approach would be to represent the fluidity of an unformed illness, yet to be given a clinical 'label' within the data. However, this is very difficult to capture, even if all possible resources were available. An expansion of the limitations of the outcome variable can be found in chapter 10, section 10.3.3.

7.6 Candidate predictors

A preliminary analysis of 100,000 YAS ePCR's was used to identify candidate predictors. These were non-conveyed patients that were attended by a YAS clinician between the 1st of July 2019 and the 29th February 2020. The cohort (known as DS2 on the figure 10 data flow diagram) had a total of 503 variables for consideration. Prior to selecting candidate variables, certain procedures were undertaken to assess missing data in DS2. The way the ePCR is completed by the clinician leaves many interventional fields blank. These are only completed if the intervention happened. For example, if the patient had a tube inserted down their airway into their lungs to help them breathe (a procedure known as intubation), then the variable intubation would have a positive value. Conversely, if the intervention did not happen, the variable would be left blank. As such, the missing values in this field are not missing, they are indeed the negative class. All variables that resemble this structure with only a positive value, had the negative class imputed. There was a decision to eliminate any variable with more than 35% missing data (except those pertinent to transportation). This was to reduce dimensionality within the data but also for practical reasons. If the variable was only completed in 65% of patients, then to include them would limit the decision support tools benefit in practice. The threshold of 35% is based on reasoning as opposed to evidence as there have not been studies to date that have explored candidate elimination based on different thresholds (and its impact on final model development). In DS2 there were initially 503 variables, but after applying missing data rules, there were only 60 candidate predictors for consideration. Candidate variables were sub-categorised into demographic, physiological, interventional, and social categories. Previous modelling studies identified in the systematic review were used as a theoretical justification for inclusion in the model. Although most of the studies in the review were predicting high-acuity outcomes, the selected variables may have the inverse relationship for predicting low-acuity patients and were worthy of inclusion. Outcome measures in the systematic review such as admission prediction were also appraised for inclusion as the outcome measure is clinically

a close neighbour to acuity. In the absence of evidence for selecting a candidate predictor based on urgency; admission was used. Where there was not a theoretical justification, a logical decision would be made for inclusion or exclusion. However, the decision support tool would have a greater impact on patient care if it could be applied earlier on in the patient journey and this was a key factor in deciding variable selection. Justifying candidate variables prior to feature selection safeguards against poor quality information being used to develop a model.

7.6.1 Demographic factors

The decision to include demographic variables as candidate predictors raises an interesting philosophical and ethical dilemma. Would the model being 'aware' of these characteristics become prejudiced, which could in-turn harm patients? Or would inclusion account for their association with the outcome and therefore optimise their care? There is no clear consensus on how to handle demographic variables, but there has been a discourse on the creation of 'fair algorithms' relating to demographic variables that are protected characteristics. From a legal basis in the UK, a protected characteristic includes: age, disability, gender reassignment, marriage status, pregnancy and maternity, race, religion or belief, sex and sexual orientation.¹⁵⁹ It is against UK law to discriminate someone because of these characteristics according to the Equalities act 2010.¹⁵⁹ This extends to the Data Protection Act (2018). which devised provisions to protect data subjects' 'fundamental rights and freedoms' and aimed to ensure that the processing of personal data does not lead to discrimination.¹⁶⁰ A useful starting point in the debate whether to include or exclude protected characteristics in a risk prediction model was an article authored by Reuben Binns and Valeria Gallo.¹⁶¹ It was highlighted that there are two main reasons that machine learning algorithms could inadvertently penalise protected characteristics. The first reason is imbalanced training data. This is where there is an imbalance in the protected characteristic (i.e., more males than females) during model development. To paraphrase their example, if the training data for a bank loan

repayment prediction model contained more males than females, then the model would perceive females to be less important. This could lead to a systemic reduction in predicted loan repayment rates for women, even if the training data would suggest otherwise. The second reason is that the training data is reflecting inherent discrimination at the data collection phase. Using the same example above, if there were more women's loan applications rejected compared to males due to historic gender discrimination, then to include this data will derive a prejudiced model.

It is important to emphasise the context of the risk prediction model being developed in this thesis. Unlike the bank loan example, this is not a 'decision making' model, but a 'decision support' model. Its intentional use as an eventual decision-support tool is by a healthcare professional. Therefore, there is an override mechanism by the end user, safeguarding against actual discrimination. Furthermore, the aim of the tool is to optimise the care of patients who may not require the ED. By omitting protected characteristics that have an association with the health outcome, it creates a model that could potentially be unfair. For example, studies have shown that age is associated with acuity. Younger patients are more likely to be low acuity compared to older patients. To include age could help younger patients navigate care when needed without penalising older patients. If age was omitted from the model through anti-classification, this information would be missing, and it could potentially lead to patients of all ages in a sub-optimal care settings for their need. Building on the work of Binns et al, a three-part strategy that ensures characteristics remain protected in this thesis has been developed.¹⁶¹⁻¹⁶⁴ Firstly, ensuring in the pre-processing that all demographics and protected characteristics are justified as candidate predictors. Secondly, during model development not to force these into the model if there is no association (identified through feature selection). Thirdly, through an assessment of outcome and error parity of all available protected characteristic variables post validation (even if they were eliminated in model development).

The systematic review was able to provide an insight into variables that are important in predicting patient acuity in the emergency care system. These variables have been categorised into demographic, social, clinical, and interventional. Tables 8-11 list each variable along with the justification for inclusion. Some variables that have been identified in the systematic review were unable to be used in this study, and more detail of these are found below.

7.6.1.1 Time and day of arrival

Time of arrival has been shown to have a relationship with the acuity of patients. Time of attendance can be split into 'in hours' and 'out of hours'. O'Keeffe defined in hours as between 8am and 6pm, Monday to Friday.⁵ Weekends and time-periods that fell outside of the in hours definition were classed as out of hours. Non-urgent attendances increased out of hours, with 62.4% of all non-urgent attendances occurring in this period. Compared with in hours the OR of a non-urgent attendance out of hours was 1.19 (95% CI 1.18-1.20). The weekend was identified as having considerable increases in the number of non-urgent attendances with the peak occurring at 3am on a Sunday. Other studies explored time and day of arrival when predicting admission and critical care, but their association varied and there was an inconsistent reference category for generating odds ratios.^{133,165-168} A limitation of using temporal variables in a prediction model of acuity is the variability in community services that have time-restricted access. For example, if there is a minor injury unit that is only open during the day, then even if the model classified them as low-acuity during the night, it is redundant information. It would therefore be inappropriate to include in the model development as a candidate predictor.

7.6.1.2 Past Medical History (PMH)

Prehospital modelling studies have previously used comorbidities as candidate variables for predicting admission to hospital. One study used only the most common comorbidities (diabetes, hypertension, asthma, seizures, cancer, end-stage renal disease and Chronic Obstructive Pulmonary Disease (COPD)) as

candidate variables in their model. The only two that made it into the final model were cancer and diabetes.¹⁰⁷ Raita et al. also studied comorbidities with congestive heart failure being the most significant in both predicting critical care need and hospitalisation. Compared to all the other variables in their model, comorbidities still did not rank that highly.¹⁶⁹ In prehospital clinical practice, patient comorbidities are found primarily through inquisition, when eliciting a past medical history (PMH). Alternative methods available are examining a patient's summary care record, which is a primary care document showing the PMH as well as the outcome of recent medical consultations. In DS2, the comorbidities are collected in a free-text format and cannot be extracted in an efficient manner. In the ECDS dataset, comorbidities are listed as a variable, but are rarely captured at present. This means that, as useful as comorbidities might be, this study is unable to use them.

7.6.1.3 Chief Complaint (CC)

In urgent and emergency care, patients can present with all types of physical, mental, and social health complaints. These complaints are the crux of the attendance and are highly predictive of outcome. For example, Meisel et al. included three chief complaints in their final model predicting admission. Shortness of breath had an OR of 6.8 (95%CI 2.9-6) compared to not having the complaint, chest pain 5.2 (95%CI 2.2-12.3), dizziness, weakness or syncope 3.5 (95%CI 1.8-6.5).¹⁰⁷ This is an important study to mention as it was using prehospital chief complaints, which are often arrived at using less information than primary care of the ED. Even when codes are grouped together, they are significant predictors. Zlotnik et al. grouped the Manchester Triage Codes (MTS) complaint codes together into five risk groups according to their risk of admission. In the final modelling all five groups of chief complaints had odds ratios where the confidence intervals did not overlap but rose drastically with each ascending group.¹⁷⁰ In the DS2, there was no chief complaint code, however there is a surrogate known as 'clinical impression'. The difference being that the chief complaint is often a coded version of what the patient describes as being the

clinical problem. Whereas the clinical impression is what the paramedic on scene believes to be the clinical problem once they have assessed them. There are 99 codes that represent clinical impression.

Table 8: Demographic candidate variables

| Variable name (units) | Values | Justification | Parameters | Cumulative |
|-------------------------------------|--|--|------------|------------|
| Age (Years) | 18,19,20 years etc. | Age was the most significant predictive variable identified in the literature. Studies predicting higher-acuity outcomes such as critical care or hospitalisation have found that as patients become older the risk is greater. ^{107,109-113,133,165,166,168,169,171-175} | 1 | 1 |
| Gender | Male, Female, Transgender, Unknown | Studies are inconclusive as to whether gender is a predictor of acuity. Therefore there is a benefit to including it in this model. ^{172 167 170,171} | 4 | 5 |
| Ethnicity | Black, Asian, Mixed, White, Other | It has been shown to be associated with decisions to admit patients into the hospital. ^{168 167} | 5 | 10 |
| Previous attendance within 24 hours | 1,0 | Fewer attendances within the last 12 months were a predictor of admission. ¹⁷⁰ | 1 | 11 |
| Incident location | Care home, Domestic address, Not selected, Public place, School, Work, Other | Two studies have identified nursing home residency as a potential predictor for admission. ^{109 168} | 7 | 18 |
| Social Deprivation (IMD) | 1.1, 9.2, 13.4 ... | The relationship between health and wealth is axiomatic with those more deprived having worse healthcare outcomes. | 1 | 19 |

Table 9: Social candidate variables

| Variable name (units) | Values | Justification | Parameters | Cumulative |
|-----------------------|---|---|------------|------------|
| GP address recorded | 1,0 | Social variables have rarely been used in prediction modelling. One example set in the USA used insurance type as a candidate predictor. ¹⁶⁷ As alluded to in the background, patients present with complex physical, mental and social needs. To include social variables can support safe non-conveyance by ensuring there is an appropriate support network available. In the data, there are flags for if a GP, Next of Kin, parent, guardian or social worker is named in the ePCR. | 1 | 20 |
| NOK named | 1,0 | | 1 | 21 |
| Parent named | 1,0 | | 1 | 22 |
| Guardian named | 1,0 | | 1 | 23 |
| Referral to service | Coroner, Police, Safeguarding adult, Safeguarding child ... | | 5 | 28 |
| Social worker named | 1,0 | | 1 | 29 |

Table 10: Clinical candidate variables

| Variable name (units) | Values | Justification | Parameters | Cumulative |
|---|--|--|------------|------------|
| Primary survey: Catastrophic haemorrhage | 1,0 | This is the rapid assessment an ambulance clinician will do as they enter the scene to check if any life-threatening problems are present. It is reasonable to include these variables, as they are the earliest triaging that occurs on scene. | 1 | 30 |
| Primary survey: Cervical spine tenderness | 1,0 | | 1 | 31 |
| Primary survey: Airway | Clear, Noisy, Occluded | | 3 | 34 |
| Primary survey: Breathing | Normal, Abnormal, Not breathing | | 3 | 37 |
| Primary survey: Pulse | Radial, Carotid, No palpable pulse | | 3 | 40 |
| Primary survey: Level of response | Alert, Confusion, Verbal, Pain, Unresponsive | | 5 | 45 |
| Mental capacity | 1,0 | This appears rational to include as mental capacity can have many different causes, and from a clinical perspective feature heavily in deciding if a patient can be left at home. | 1 | 46 |
| Clinical impression | Shortness of breath, Abdominal pain, Hypoglycaemia ... | As described above when discussing chief complaint, studies have identified an association between clinical impression and acuity. | 99 | 145 |
| Initial pulse rate (bpm) | 60,61,62 ... | Many studies have used physiological observations as candidate variables for predicting acuity and they have shown great significance in final models. ^{110,112,113,167,169,171,176-178} These variables are often limited to pulse rate, respiratory rate, blood | 1 | 146 |
| Initial respiratory rate (rpm) | 16,17,18 ... | | 1 | 147 |
| Initial SpO2 (%) | 96%,97%,98% ... | | 1 | 148 |

| Variable name (units) | Values | Justification | Parameters | Cumulative |
|--|---------------------------------|--|------------|------------|
| Initial temperature (°C) | 36.2, 37.1, 37.5 ... | pressure, temperature, oxygen saturations, blood glucose and level of consciousness. The common principle in using them to predict critical care is that observational values that are extreme or deviated from the norm become highly predictive. Therefore, for predicting low acuity, it would be logical to include them as candidate predictors and expect the inverse relationship. That a normal physiological observation is predictive of a low acuity patient. | 1 | 149 |
| Initial Systolic BP (mmHg) | 120,121,122 ... | | 1 | 150 |
| Initial diastolic BP (mmHg) | 80,81,82 ... | | 1 | 151 |
| Blood glucose (mmol/L) | 4.1, 5, 5.2, 10.1 | | 1 | 152 |
| Initial Glasgow Coma Scale (GCS) score | 15,14,13,12,11,10,9,8,7,6,5,4,3 | Level of consciousness is usually measured in adults using the Glasgow Coma Scale (GCS). In one study, the GCS was abbreviated to a simplified consciousness score (SCS). In regards to variable importance, it was ranked 1 st across all learners. ¹¹² | 13 | 165 |
| Initial GCS: Eye component | 4,3,2,1 | | 4 | 169 |
| Initial GCS: Verbal component | 5,4,3,2,1 | | 5 | 174 |
| Initial GCS: Motor component | 6,5,4,3,2,1 | | 6 | 180 |
| Initial NEWS score | 1,2,3 ... | The National Early Warning Score (NEWS) is a composite scoring system based on respiratory rate, the saturation of oxygen in the blood (SpO ₂), pulse rate, systolic blood pressure, body temperature and level of alertness. A copy of the latest version (NEWS2) can be found in appendix F. ^{179,180} Studies have shown there is a clear relationship between the NEWS score and patient acuity. ¹³³ | 1 | 181 |
| Initial pain score | 1,2, 3... | Pain has been a significant predictor for admission of patients from the ED. A study using natural language processing of free text fields in hospital documentation found that the most frequently used terms for admission were: pain, soreness and ache. ¹⁶⁷ | 1 | 182 |

| Variable name (units) | Values | Justification | Parameters | Cumulative |
|---------------------------------|----------------------|---|------------|------------|
| Hypercapnic respiratory failure | 1,0 | There are observations that are present in DS2, but not necessarily captured in previous studies. This presents itself as an opportunity to explore potential new candidates and further contribute new knowledge in this area. Pupil size is a neurological observation that includes the size of the person's pupil in the eye, and the reactivity of it. Peak flow is a measurement of force during exhalation and is used before and after treating a patient with respiratory conditions in order to assess effectiveness of treatment. A related variable is whether a person has hypercapnic respiratory failure. This is a relatively new field and is assessing whether a person has an abnormal respiratory physiology as part of their medical history, which would alter what their normal SpO ₂ should be. An example would be a patient with COPD who may live with reduced SpO ₂ , and therefore the target % oxygen should also be lower. | 1 | 183 |
| initial pupil size left | 1,2,3 ... | | 1 | 184 |
| initial pupil size right | 1,2,3 ... | | 1 | 185 |
| Initial pupil reaction left | 1,0 | | 1 | 186 |
| initial pupil reaction right | 1,0 | | 1 | 187 |
| Subsequent pulse rate | 60,61,62 ... | The subsequent observations were included for two reasons. The first is that abnormal observations in subsequent recordings is important information. But also, the subsequent observations can be used to create observation intervals (the difference) which can be used as a predictor of deterioration. | 1 | 188 |
| Subsequent respiratory rate | 16,17,18 ... | | 1 | 189 |
| Subsequent SpO ₂ | 96%,97%,98% ... | | 1 | 190 |
| Subsequent temperature | 36.2, 37.1, 37.5 ... | | 1 | 191 |
| Subsequent Svstolic BP | 120,121,122 ... | | 1 | 192 |
| Subsequent diastolic BP | 80,81,82 ... | | 1 | 193 |

| Variable name (units) | Values | Justification | Parameters | Cumulative |
|---|--|--|------------|------------|
| Subsequent Responsiveness | Alert, Confusion, Verbal, Pain, Unresponsive | | 5 | 198 |
| Subsequent Glasgow Coma Scale (GCS) score | 15,14,13,12,11,10,9,8,7,6,5,4,3 | | 13 | 211 |
| Subsequent GCS: Eye component | 4,3,2,1 | | 4 | 215 |
| Subsequent GCS: Verbal component | 5,4,3,2,1 | | 5 | 220 |
| Subsequent GCS: Motor component | 6,5,4,3,2,1 | | 6 | 226 |
| Subsequent NEWS score | -1,0,1 ... | | 1 | 227 |
| Subsequent peak flow | 300.301.302 ... | | 1 | 228 |
| subsequent pupil reaction left | 1,0 | | 1 | 229 |
| subsequent pupil reaction right | 1,0 | | 1 | 230 |
| subsequent pupil size left | 1,2,3 ... | | 1 | 231 |
| subsequent pupil size right | 1,2,3 ... | | 1 | 232 |
| Subsequent pain score | 1,2, 3... | | 1 | 233 |
| Difference pulse rate | -1,0,1 ... | As discussed above, these are useful in identifying a deteriorating patient. | 1 | 234 |
| Difference respiratory rate | -1,0,1 ... | | 1 | 235 |
| Difference SpO2 | -1,0,1 ... | | 1 | 236 |

| Variable name (units) | Values | Justification | Parameters | Cumulative |
|---|--|---|------------|------------|
| Difference temperature | -1,0,1 ... | | 1 | 237 |
| Difference Systolic BP | -1,0,1 ... | | 1 | 238 |
| Difference diastolic BP | -1,0,1 ... | | 1 | 239 |
| Difference Responsiveness | -1.0.1 ... | | 1 | 240 |
| Difference Glasgow Coma Scale (GCS) score | -1,0,1 ... | | 1 | 241 |
| Difference NEWS score | -1.0.1 ... | | 1 | 242 |
| Difference pain score | 1,2, 3... | | 1 | 243 |
| Difference peak flow | 300.301.302 ... | | 1 | 244 |
| difference pupil reaction left | 1,0 | | 1 | 245 |
| difference pupil reaction right | 1.0 | | 1 | 246 |
| difference pupil size left | 1,2,3 ... | | 1 | 247 |
| difference pupil size right | 1,2,3 ... | | 1 | 248 |
| Abnormal ECG on primary | Left Bundle Branch Block, Right BBB, STEMI | No studies have used the initial abnormal ECG finding as a predictor variable in the past. The variable is found in the YAS ePCR and is an early indicator of something wrong with the patient's heart. It is a categorical variable that only accepts ST elevation MI (STEMI) and bundle branch blocks, which are also an electrical problem with the heart. | 3 | 251 |

Table 11: Interventional candidate variables

| Variable name (units) | Values | Justification | Parameters | Cumulative |
|-----------------------|--|--|------------|------------|
| ECG monitored | 1,0 | | 1 | 252 |
| Supplemental oxygen | 1,0 | | 1 | 253 |
| ICN type | Intravenous, Intraosseous, None | | 3 | 256 |
| Drug 1 | Adrenaline 1:1000, Co-codamol 30/500, Diazepam ... | | 100 | 356 |
| Drug 2 | Adrenaline 1:1000, Co-codamol 30/500, Diazepam ... | Interventional candidate predictors utilise the massive benefit of using electronic healthcare records at the granular level. In DS2, there is a plethora of interventions captured in the data including investigations like an electrocardiogram (ECG). There are also fields detailing equipment used such as an airway device or immobilisation equipment. Treatments are also captured including which drug has been given. Being able to include these into the model as candidates has a tangible benefit as it could reveal which interventions make a difference to patient acuity. | 100 | 456 |
| Drug 3 | Adrenaline 1:1000, Co-codamol 30/500, Diazepam ... | | 54 | 510 |
| Drug 4 | Adrenaline 1:1000, Co-codamol 30/500, Diazepam ... | | 25 | 535 |
| Drug 5 | Adrenaline 1:1000, Co-codamol 30/500, Diazepam ... | | 17 | 552 |
| Drug 6 | Adrenaline 1:1000, Co-codamol 30/500, Diazepam ... | | 16 | 568 |
| Drug 7 | Adrenaline 1:1000, Co-codamol 30/500, Diazepam ... | | 14 | 582 |
| Drug 8 | Adrenaline 1:1000, Co-codamol 30/500, Diazepam ... | | 12 | 594 |
| Airway type | ETT, LMA, OPA, NPA... | | | 10 |
| Immobilisation | Scoop, Cervical collar, Extrinsic board ... | | 9 | 613 |
| Advice given | Wound care, bereavement, head injury | | 8 | 621 |
| Mobility | Stretcher, walked, hoist | | 15 | 636 |
| CPR | 1,0 | | 1 | 637 |

7.7 Sample size

The sample size calculations for risk prediction models with a binary outcome have only recently been described through a series of simulation studies.^{114,181–186} However, this study has the limitation of being bound by two time-points. The first time-point is when the moment in which the ePCR in YAS was rolled out across the region and the end point is when COVID-19 became prevalent and would thus confound the cohort. By this premise, the sample size is a convenient sample, and any calculations performed in line with the literature would serve a different purpose. Traditionally, the calculation of a sample size for a multivariable risk prediction model is completed using a ‘rule-of-thumb’ of between 10 and 50 events per variable.¹⁸¹ The requirement of a sample size is mainly to prevent the concept of overfitting. In order to calculate a sample size that prevents overfitting, the estimated R^2 needs to be defined. The R^2 is the expected variance explained by the model. There are many different R^2 statistics, the one used in the sample size calculations is the Cox-Snell R^2 . This is used to estimate the overall model fit in the developed model with a binary outcome. It is also known as the likelihood ratio R^2 .¹⁸⁴ Unfortunately, explained variance is not always reported in the results of prediction model studies, however the C-statistic is mostly reported. It is possible to obtain a Cox-Snell R^2 statistic from a known C-statistic.¹⁸⁶ For the estimated parameters, DS2 was analysed, and it was found that there were 60 variables once the missing data rules were applied. In the estimation, degrees of freedom were calculated for each variable. Continuous and binary variables had a single degree of freedom, and categorical variables had n degrees (where n = number of categories). The degrees of freedom calculations were based on one-hot encoding in subsequent data preparation instead of dummy variables being created. The difference is that in the creation of dummy variables for a categorical variable, there are $n-1$ degrees of freedom. Conversely, in one-hot encoding, the result is equal to n degrees of freedom. The degrees of freedom can be seen in the above tables 8-11.

7.8 Internal-external cross-validation (IECV)

Traditionally, validation has two components: internal and external validity. In internal validation, the data is either split into a training set and a test set, or resampling methods such as cross validation and bootstrapping are performed. By doing so, optimism can be corrected using the same dataset. External validation is applying the model to a different dataset, preferably in a different location, at a different time, by different researchers.^{114,128} This fully independent approach is not always feasible to achieve. A modern approach, identified by Steyerberg et al. in 2016 is the idea of internal-external validation, which was used in this thesis.

This is a modern method designed to overcome the limitations of data availability required for external validation. The idea behind the external validation element in internal-external validation is not so much ‘does the model work in different settings?’, rather ‘If the model was developed using different data, would the results be the same?’ Internal-external validation uses cross-validation as the framework to its approach.

Cross validation is an extension of the sample splitting procedure, except it is repeated many times. The whole sample is randomly split into k folds. A single fold is left out and becomes the test set. The remaining folds are the training sets. A model is built on the training set and evaluated on the test set. The process is then repeated with the next fold left out instead as the test set. The number of folds is user defined, but commonly five or ten folds. If the number of folds equals the number of instances in the sample, it is known as the ‘jack-knife procedure’. The method of cross-validation overcomes the limitations of apparent validity by using unseen test sets. It also overcomes the weaknesses of split-sample validation as it allows all instances to appear in a test set at least once. It can also average out evaluation results across all test sets, which gives a more realistic measure of model performance.¹²⁸

In the description of cross validation above, it was noted that the data is randomly partitioned into folds. In Internal-External Cross-Validation (IECV), the data is non-randomly split into folds. Each cluster in the dataset represents a single fold. The procedure used in this thesis was as follows:

The first cluster (in this case ED) was removed from the dataset to form the test set. The remaining data formed the training dataset. The training dataset was then split into k random folds and an internally validated model was developed on the training data. It was then tested on the test set. This was repeated for every cluster. The procedure is known as nested cross-validation as there is an inner-loop nested into an outer-loop.¹⁸⁷ In the context of this study, the outer-loop consisted of cross-validation for each ED. There were seventeen EDs included in this study, and therefore the outer-loop was 17-fold cross validation. The final step in IECV was to use a meta-analysis to pool all the clustered results to update the final model performance.^{114,187,188} The theory in the meta-analysis had been modified and applied for the purposes of IECV. Traditionally in a meta-analysis, the results from each study are listed and then a random effects meta-analysis pools them altogether to produce a summary statistic of the results with a confidence interval. In the IECV meta-analysis, the cluster results were all pooled together instead.¹⁸⁹ An illustrative example of this procedure can be found in chapter 6 of this thesis (the protocol manuscript). Figure 1 in the manuscript shows the whole model procedure for this study and steps 4-6 illustrate the internal-external cross-validation procedure. IECV has proven successful at testing the reproducibility and generalisability of models that have been developed on large, clustered datasets. It is the superior method of building a model that can internally- and externally-validate a model using the existing dataset.^{114,128}

7.9 Evaluation of model performance

In the systematic review from chapter 3 (section 3.4), it was found that those studies authored by computer scientists would sometimes have different methods of evaluating a model compared to those authored by statisticians. The lexicon between the two areas can be different, despite describing the same terms. A good example of this is the positive predictive value (PPV) and the sensitivity. These are terms common in statistics to describe a model's performance. The PPV is the proportion at which a positive prediction is correct, whereas sensitivity is reporting the proportion of all positive instances have been identified by the model. As an aside, computer science adopts different terminology for identical calculations, with PPV known as precision, and sensitivity known as recall. For clarity, the statistics terminology has been adopted here.

The simplest way to evaluate the performance of a classification model is to transform the results into a confusion matrix such as figure 11. This assigns every prediction to one of four categories and arranges them in a 2x2 matrix. Figure 11 has been created for this thesis to illustrate the basic components of a confusion matrix, which contains the elements needed to calculate PPV and sensitivity.

Figure 11: Confusion Matrix

| | | Predicted Class | |
|--------------|--------------------------|---------------------|---------------------|
| | | Positive (PP) | Negative (PN) |
| Actual Class | Total population = P + N | | |
| | Positive (P) | True Positive (TP) | False Negative (FN) |
| | Negative (N) | False Positive (FP) | True Negative (TN) |

Using the matrix in figure 11, PPV (precision) and sensitivity (recall) can be calculated in the following way:

$$PPV = \frac{TP}{(TP + FP)}$$

$$sensitivity = \frac{TP}{(TP + FN)}$$

In the equations above, the true positives (TP) are those where the predicted and actual class was positive. The false positive (FP) is where the actual class was negative, but the model predicted positive. The false negative (FN) is where the model predicted negative, but the actual class was positive. There have been defined above (TP, FP, FN), with the fourth being a true negative (TN). This is where both the predicted class and the actual class are negative. The confusion matrix can be used to calculate many statistics to evaluate model performance.¹⁹⁰ The most popular are sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and overall accuracy. These can all be found in the equations below:

$$sensitivity = \frac{TP}{(TP + FN)}$$

$$specificity = \frac{TN}{(TN + FP)}$$

$$PPV = \frac{TP}{(TP + FP)}$$

$$NPV = \frac{TN}{(TN + FN)}$$

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

In clinical practice, sensitivity and specificity have greater utility than cruder metrics such as accuracy. A pre-specified threshold can be established to determine the success of a model. For example, in the context of this study the aim was aligned more to predicting health than disease. Therefore, a false positive would be a patient requiring the skills of the ED but being identified as

non-urgent. This is an unsafe situation and thus the model needs to have a low false positive rate (in effect high positive predictive value). Traditional predictive modelling often aims to predict disease and so the inverse ideal threshold is true. There are limitations with statistics that are derived from the confusion matrix such as they lack information about the model itself, only the classification results. Also, due to the nature of the results being aggregated into four classes first, if there is a class imbalance the results become less valuable. For example, if the outcome is only prevalent 7% of the time, an extreme model of assigning every instance to the negative class will still result in 93% accuracy. The ideal structure for an analysis of a clinical prediction model is outlined in the Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) and the Prediction model Risk Of Bias ASSEssment Tool (PROBAST) guidelines.^{146,191} Their recommendations stand on the shoulders of the latest recommendations in prediction modelling research. Proper evaluation of a model lies in two specific areas. The first is examining whether the model predicts accurately across all predictions, this is known as calibration. The second is assessing whether the newly developed classifier can differentiate between a random pair of instances (one with and one without the outcome), this is known as discrimination. There are also measures of evaluation that report how well a model fits the data and are composites of both calibration and discrimination. These have their limitations and are discussed briefly below before calibration and discrimination are separated out.

7.9.1 Calibration

Calibration is producing a statistic that is an assessment of whether the predicted probabilities for each instance in a test set match with the observed probabilities. The simplest for binary outcomes is to take the ratio of observed and expected (O:E). The observed (O) is the prevalence in the dataset, therefore it is number of events divided by the whole sample. The expected (E) is the sum of predicted probabilities created by the model for each instance. In a perfect model, the O:E would be 1. If the model underpredicts, there will be a O:E ratio of greater than 1

as there will be more predicted outcomes than observed. Similarly, if the model overpredicts the O:E ratio will be less than 1. The benefit of using the O:E ratio is its intuitive nature. It is also the statistic that is recommended for meta-analysing the different clusters.¹⁹² In this thesis, all cluster results were pooled into a random effects meta-analysis and a summary O:E was calculated with a 95% confidence interval (95% CI).

A limitation with the O:E is that it informs of whether there is over-prediction or under-prediction, but it does not determine whether the model is miscalibrated or not. Conversely, the Spiegelhalter's z-test statistic does, and can be fine-tuned to ensure any miscalibration is adjusted. For this reason, the Spiegelhalter's z-test statistic was used for this purpose.¹⁹³ It can be calculated as follows:

$$z = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)(1 - 2\hat{y}_i)}{\sqrt{\sum_{i=1}^n (1 - 2\hat{y}_i)^2 \hat{y}_i (1 - \hat{y}_i)}}$$

In the equation, y_i is the actual value of the outcome, and \hat{y}_i is the predicted outcome. The test follows a standard normal distribution asymptotically. The null hypothesis associated with Spiegelhalter's z-test is that the model is well calibrated. Therefore, if the z-test is less than -1.96, or greater than 1.96, the test is statistically significant, and the related p-value will reflect this by being less than the conventional alpha of 0.05. A model with a Spiegelhalter's z-test that is well calibrated will have values that fall within the -1.96 to 1.96 window and will have a p-value > 0.05, i.e., it is not statistically significant, and the null hypothesis is rejected.

The final way of reporting calibration in a transparent way is to assess the calibration slope and intercept. In this thesis, it was reported visually on a calibration plot of predicted probabilities on the x-axis, and actual probabilities on the y-axis. Perfect calibration in this case lies on a straight line drawn at 45°. The intercept of the line would be 0, and the slope would be 1.

7.9.2 Discrimination

In the modelling of a binary outcome, it is useful to start with the predicted probabilities belonging to the positive class and then identifying a threshold with which to dichotomise patients into the two classes. Depending on the threshold, there will be different frequencies in each of the four areas in a contingency square. Discrimination is almost universally reported as the C-statistic for binary classification problems. A Receiver Operator Characteristic (ROC) curve plots 1-specificity on the x-axis and sensitivity on the y-axis at different thresholds. This allows the ideal threshold to be selected, but also allows for a summary statistic of how good the model is at differentiating between the two classes. The area underneath the curve is this statistic and is also known as the C-statistic. A model which has the predictive ability no better than tossing a coin would have a C-statistic of 0.5. It would not look like a curve visually but would appear on a 45° line. A perfectly discriminate model will have a C-statistic of 1, and the curve would reach far into the upper-left corner of the plot, effectively forming a (near) right-angle.¹²⁸ The C-statistic can be meta-analysed using each cluster, and further visualised using a forest plot of each individual C-statistic with confidence intervals. A note on the C-statistic is that it is not always possible to achieve a result of 1. Depending on the classification problem, a good result may be limited. In the systematic review, it was reported that the C-statistics found in previous modelling studies had a mean value of 0.80.¹⁰⁶

7.10 Further data preparation

One of the main features of the XGBoost algorithm is the speed at which it can function, and it achieves this through operating on sparse matrices. The dataset needed to be transformed into a sparse matrix through one-hot coding of all categorical variables. Continuous and binary variables stayed in their original format.

7.10.1 Hyperparameter tuning

The beauty of a machine learning algorithm lies in its ability to automatically detect patterns within a dataset. In XGBoost, the user does not have to manually build every decision tree, calculate errors and adjust weights for future trees accordingly. It is done by the algorithm itself. But this raises the question as to whether the final model is overfitting, or whether the bias variance trade-off needs trading off. This is the role of a hyperparameter.

Hyperparameters provide the algorithm with a set of rules that can be tweaked in order to optimise the results. Each rule (hyperparameter) controls how the algorithm operates. For example, some hyperparameters aim to increase or decrease model complexity (adjusting the bias) and others can be used to introduce randomness to prevent overfitting. The value of each hyperparameter is set prior to the model building process. The challenge is finding the best value of each hyperparameter. This is exacerbated by the fact that sometimes the hyperparameters influence each other, so the tuning of one can affect the performance of another. XGBoost has around 53 hyperparameters that can be customised. The limitations with algorithms that have an extensive list of tuneable hyperparameters is knowing those that are important and will affect the final model but also adjusting the computational expense to find the right value for each one.^{194–196} To address the first limitation in this thesis, the documentation for the XGBoost algorithm was studied and a decision was made to tune only hyperparameters that concerned themselves with the bias variance trade-off, model complexity or the class imbalance. For the second limitation, a restricted grid search approach was undertaken. These are explained below.

7.10.2 General parameters used for modelling

The hyperparameter 'booster' is to control the design of the XGBoost algorithm. This was set to the value 'gbtree' as it was the design of the chosen algorithm to build boosted trees. It was possible to select gblinear, or dart. Dart is short for 'Dropouts meet Multiple Additive Regression Trees' and is a tree boosting

algorithm that incorporates the random drop out of trees. This can create an unstable model with a slower training rate.¹⁹⁷ The `num_round` and `early_stopping` also needed specifying. These both relate to each other, and it is best practice to set `num_round` quite high and then tune the `early_stopping`. `Num_round` is the maximum number of trees to be created. `Early_stopping` controls when to stop building trees and uses a threshold of no improvement after X rounds. `Num_round` is set to a high value in case it takes a lot of trees to maximise an accuracy metric. In this thesis, `num_rounds` was set to 1000, and `early_stopping` was set to 10, which meant the algorithm kept building trees up to 1000 of them. The 1000 for `num_rounds` was an arbitrary selection, but the plan was to increase this if there was no early stopping. The 10 for `early_stopping` was used as the commonly used value in the XGBoost documentation.¹⁹⁸ To clarify how the two hyperparameters interact, if after building 360 trees there was no improvement in the next 10 trees, the algorithm would stop, and the model would only have 360 trees. The `eval_metric` must be specified as this is how performance is evaluated in `early_stopping`, and within the wider algorithm itself. It is chosen depending on the objective of the algorithm. As the objective in this thesis was binary classification, the evaluation metric was the C-statistic (AUC), as explained above. Most other non-booster hyperparameters were related to parallelisation of computing (which affects speed of modelling, as opposed to the results of the model), hence were left at default values.^{198,199}

7.10.3 Booster parameters

There are 23 hyperparameters that control the actual boosting procedure within the algorithm. These are discussed below.

Eta (η)

The hyperparameter eta is also known as the learning rate and accepts values between 0 and 1 and has a default value of 0.3. It controls the shrinkage of the feature weights after each boosting step. A larger eta value makes the boosting process more conservative. It is recommended to keep this high when tuning

other hyperparameters and then fine tune it to a lower value within 2 decimal places.¹⁹⁹

Max_depth

This is one of the most important hyperparameters for the XGBoost algorithm. It controls how deep the trees are that are built: therefore, the value of max_depth controls the model complexity of each tree. Higher values will allow the algorithm to build deeper trees, but risk overfitting. The values of max_depth can range from 1 to ∞ and is defaulted to 6. The values are more realistically between 1-10 to begin with.¹⁹⁹

Min_child_weight

The min_child_weight is the minimum hessian weight needed in a child node to qualify that split. If the sum of instances falls below the minimum specified, then the building process will give up further splits. For example, if the min_child_weight was set to 100, then the model would keep splitting until there were less than 100 instances left in the prospective child nodes. The higher the value, the more conservative the algorithm. The range can be 0 to ∞ , with the default set at 0. Similar to max_depth, the realistic values may be within a smaller range of 1-10. The hyperparameter is also used to control model complexity.¹⁹⁹

Subsample

Subsample is a ratio of the training instances. It ranges from 0 to 1 and has a default value of 1. The value sets the ratio of randomly sampled training data prior to growing trees. For example, if the subsample was set to 0.5, it would randomly sample half of the training data each time it developed a tree. This process prevents overfitting by introducing randomness.¹⁹⁹

Colsample_bytree

Colsample_bytree is part of a group of hyperparameters non as colsample_by*. These all have a range between 0 and 1 have a default of 1. They all control whether a fraction of columns is selected for splitting (much like the random forest modelling). It is possible in the XGBoost algorithm to control the fraction of randomly sampled columns for each tree, at each level and by each node. They work cumulatively because to specify a fraction for each one would significantly limit the number of variables to choose from at each node. However, the ability to randomly sample columns is a good way to prevent overfitting. As such, only the highest level (colsample_bytree) was included for tuning. The others (colsample_bylevel, and colsample_bynode) were left at their default values.¹⁹⁹

Gamma (γ)

This is also known as the minimum split loss. It specifies the minimum reduction in loss required to make a further split in the tree. The range can be from 0 to ∞ with a default value of 0. The larger values of gamma, the more conservative the algorithm, hence it controls model complexity.¹⁹⁹

Scale_pos_weight

This controls the weighted scaling of the positive class and helps with managing a class imbalance in the dataset. Adjusting the value of scale_pos_weight helps to calibrate the model in this situation. The values range from 1 to ∞ and the default value is 1. This hyperparameter was used as the method of recalibration during model development.¹⁹⁹

Alpha (α)

This is the regularisation penalty that can be tuned in the XGBoost model. Unlike in regularisation, where this represents the elastic net hyperparameter, in XGBoost it represents the L1 regularisation term, which is the LASSO. The L2 penalty (ridge regularisation) was available, as well; however, it was decided to

only tune the LASSO, as it would lead to a more parsimonious model if extra features were eliminated, even after recursive feature elimination.¹⁹⁹

7.10.4 Restricted grid search

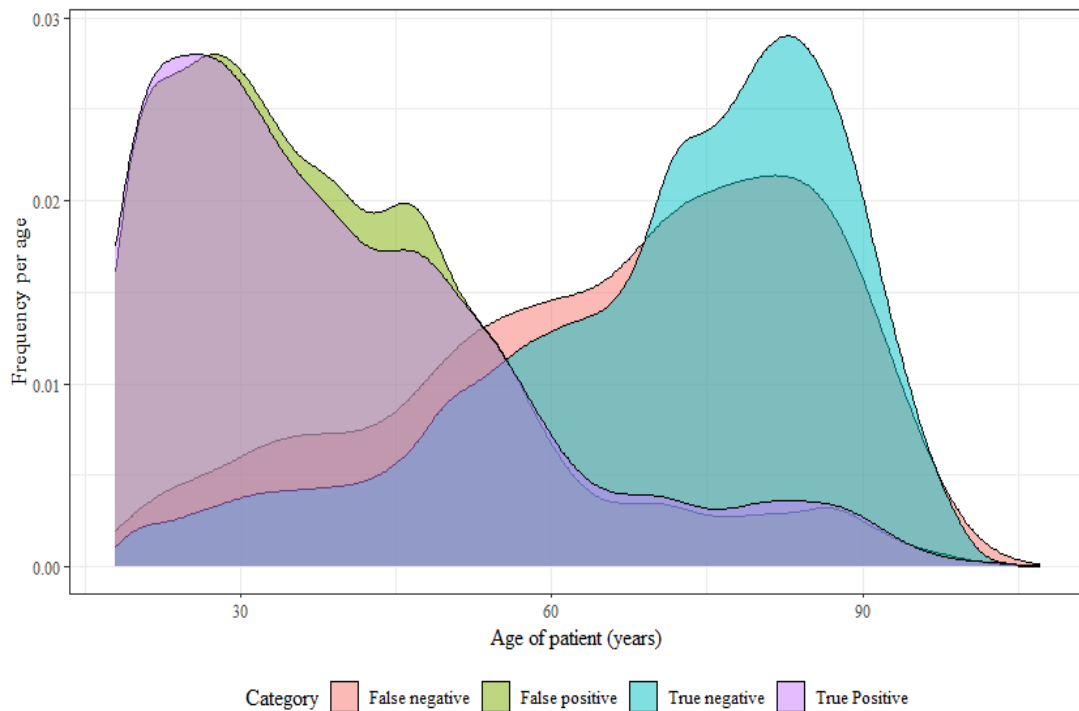
As demonstrated above, there are hyperparameters that need tuning but accept many different values. Even those are restricted between 0 and 1, have an infinite number of decimal places that can be incorporated into a search strategy. The most appropriate method of finding the best value for each hyperparameter is to use a grid-based approach because it is thorough and more robust for finding optimum hyperparameters (compared with stochastic methods).¹⁹⁶ In this strategy, each hyperparameter has a vector containing a set of values. These are then placed into a grid with all the other hyperparameters that are intended to be tuned. For every possible combination in the grid, the model is built and evaluated. However, the computational expense is tremendous. As an example, if six hyperparameters were given just ten values each, the number of possible combinations in the grid search would be 3600. This is 3600 models that need to develop through the process of nested CV for the best combination to be evaluated. There are seven tuneable hyperparameters in this thesis and so it was decided to create a restricted grid. The eighth hyperparameter (scale_pos_weight) was tuned separately, and the default was updated to the optimum value. It was then retuned if there any recalibration was necessary. The process of using a restricted grid has two parts. The first part (hyperparameter optimisation) is identifying the best performing hyperparameter values (the fewer the better). These are then entered into a grid, to form a restricted grid space. For the second part (using the restricted grid space), each time a model is developed, the ideal hyperparameters are selected from running a grid search on the restricted grid. In this thesis, the hyperparameters were sequentially tuned and then the three best performing values were chosen to be in a restricted grid. This way, the computational expense was mitigated as there had to be a grid search for each model developed in the IECV. In the sequential grid searching, some hyperparameters were tuned in tandem, as they interact with each other.

This included `max_depth` with `min_child_weight`, and `subsample` with `colsample_bytree`. The sequential nature of tuning was undertaken so that the default of each was updated to the best performing value after the search. The search space for each hyperparameter was defined using small values between 1 and 10 initially. If the best performing value was 10, the space was expanded to larger values. Equally, if the best performing value was 1, the space was changed to between 0 and 1, with increments of 0.1. If a hyperparameter had only two values in the top ten performing iterations in the optimisation process, then only those two values went through to the restricted grid to further restrict the grid.

7.11 Protocol deviations

The initial model findings included age, ethnicity and deciles of deprivation (according to the Indices of Deprivation) as predictor variables. During recursive feature elimination some ethnic categories and deciles of deprivation were omitted from the variable list. As such, it was decided to remove the two variables from the candidate list to ensure no group was penalised over another. Age remained in and it was the most significant predictor in the model by far. When variables were ranked according to their information gain in the model, age had twice the gain compared to the next variable. Because it held so much weight, it introduced a bias into the model that would penalise younger ages into a higher probability of an avoidable conveyance. The discussion in chapter 5 highlighted that this could be a benefit of the model as it could optimise the care of younger patients. However, the problem was that, in the misclassification analysis (shown in figure 12), it showed that most of the errors in prediction stemmed from age carrying so much weight. The variable was removed and the model re-run. The accuracy of the model barely changed, but the bias was removed. The results presented below do not have age as a predictor included.

Figure 12: Frequency distribution by confusion matrix category



7.12 Ethical considerations

The main ethical consideration for the wider context of the thesis was using a large volume of patient data without consent. The study underwent extensive review to ensure that there was a clear legal basis for doing so, and patients had every possible opportunity to opt out. The NHS Research Ethics Committee (REC) reviewed the study and gave a favourable outcome which can be found in appendix G. This is required before going for additional ethics review by the Confidentiality Advisory Group (CAG). This review is specifically for using patient data without consent. The initial CAG outcome asked for a more detailed plan on how patients who did not want their data to be used in this way could opt out. As such, the following strategy was developed.

The first method was to first run patient NHS numbers through the national opt out scheme and remove any instances where patients had preselected to opt out.²⁰⁰ Then, a public notice was placed on the YAS and University of Sheffield

websites. This notice gave the details of the study, the dates of data collection and details those individuals could contact and have their instances removed. A limitation was that each individual could not be identified and contacted as they would not have consented for their data to be processed this way. Furthermore, the cost and timings to do this for every patient in the sample would have been infeasible. The vast geography of the included sample meant that placing physical literature would not be an option either. Once this strategy was implemented, CAG gave their ethical approval. The letter from the Health Research Authority (HRA), the REC committee and the CAG committee can be found in appendix G .

As part of the application to NHS Digital for the data, a Data Access Request Service (DARS) application had to be completed. This extensively described the legal basis and appropriateness of the project for NHS Digital to send data. As part of this, NHS Digital requested that the NHS REC and CAG approval be in place, but they also undertook their own ethical review. The legal basis and ethical considerations passed this extra step.

One of the ethical considerations that was a focus of the ethics applications was the difference between decision-making and decision-support. The philosophy of this study is to help paramedics make decisions regarding the clinical benefit of transporting a patient to the ED. Therefore, the greatest utility is adding the model to existing clinical knowledge. A decision-making tool would go against this and requires significant more methodology prior to implementation, including observational and interventional studies. The firm belief is that the real tangible benefit of a clinical prediction model in this context was a decision-support tool only. A final ethical consideration was the inclusion of demographic features in the model; however, this has been covered in a previous section.

7.13 Patient and public involvement

This project was heavily influenced by the public. The Sheffield Emergency Care Forum (SECF) is a Public Involvement and Engagement forum that is situated at the University of Sheffield. This study had a standing agenda at their quarterly meetings. Through an iterative process, they steered the study. Changes made included the terminology used, which felt very impersonal at the start but with their support, reflected that this was at the heart of patient care. They also helped with the mechanics of involving patients, and their suggestions for how to involve the public were developed and revised by the group. They helped provide reassurance that using patient data anonymously to create a tool would be a benefit to patients.

A small grant from the NIHR Research Design Service (RDS) was awarded to conduct three Public Involvement events. These were designed to fulfil the following objectives:

- To elicit how much the public agreed with the feedback of the SECF and if they had any different thoughts and feelings.
- To discover how the public would like clinicians to communicate care plans in emergency situations.
- To recruit public members to form a panel for the duration of the project (and hopefully into post-doctoral research).

A total of twenty-two public members were involved who captured diversity in socio-economic status, gender, age, disability, and ethnicity. The events led to a multi-sided discourse; however, underlying themes caused changes in the application. For example, the public felt a tool could not be created in isolation and would require advice and guidance for patients. This led to the qualitative phase of the project embedding public perceptions around managing patient expectation. Unfortunately, the qualitative phase of the project was removed as a request from the funder. From a reflective lens of undertaking the events with

such a diverse membership, there were personal development areas highlighted such as conveying information that can be understood.

The members of the public who helped in the design of the research asked whether they would like to be invited to a WhatsApp group. This is a digital messaging application that is secure with end-to-end encryption.²⁰¹ Those who had access to the application but never used it were helped. Those who did not have access but wanted to participate were invited to face-to-face meetings and met on a one-to-one basis. The purpose of the WhatsApp group was to have a longitudinal conversation with the public about the project. As changes were being made, the public were asked in advance what their thoughts were. They were also invited to comment on reports to ensure the language can be understood.

There were annual meetings with members of the group to consolidate conversations in the WhatsApp group, and to present and discuss the progress of the project.

7.14 Conclusion

This chapter expanded on the protocol chapter found in chapter 6. This study is set within the geographical footprint of Yorkshire Ambulance Service NHS Trust and includes all type 1 ED's within. The methodology will use an XGBoost algorithm on a linked dataset of patient episodes that started in the ambulance and ended at the ED. The methods detailed in section 7.8 include using internal-external cross-validation. This allows for spatial validation between geographical areas to be explored. In the next chapter, the results of the study are presented in manuscript form.

Chapter 8

The Results

8.1 Introduction

This chapter presents the results of the study as a manuscript that is being prepared for publication. Due to this, there is some duplication with other chapters which is unavoidable. The accepted best practice is to frame the manuscript around accepted reporting guidelines. In the case of this study, the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement was considered most appropriate.¹⁴⁶

8.1.2 Background

In the emergency care system, pressure is rising amidst the growing quantity of patients accessing front door services such as the ambulance service, Emergency Department (ED) and General Practice (GP). This demand is rising at around 5% per annum.^{14,19} For the ambulance service, this means that patients who are transported to hospital may be held in a queue of other ambulances waiting to hand their patients over. In 2019/2020 in England alone, there were 137,009 delays in ambulance handover of between 30 and 60 minutes.²⁶ When these delays occur and ambulances are queueing, it has the potential to cause harm to those in the queue. A recent report from the Association of Ambulance Chief Executives (AACE) found that 80% of ambulance patients that queued for more than an hour experienced some level of harm.²⁰² There are also potential consequences for prehospital patients still waiting to be assessed in the community.

The case mix of these patients is not always life-threatening emergencies. Previous reports have demonstrated that the majority of prehospital patients have no immediate life-threatening care need and their actual need could be managed in the community.^{13,22} However, some of these patients are still transported to the ED and this can lead to an avoidable ED attendance.

When paramedics make decisions on-scene to transport a patient to hospital, it is often the most complex decision they make.⁷⁶ As such, the decision is not always accurate. Studies have found that there are between 9 and 32% of ambulance transports to ED that could have been avoided.^{5,13,49,54}

Existing transport decision support tools that are in practice have all been designed not to miss a higher acuity patient, which has led to significant over-triage of patient acuity. They have also failed to demonstrate significant benefit over clinician decision making. A vignette-based survey by Miles et al. found that conveyance decisions had a sensitivity of 0.89 (95% CI 0.86 – 0.92) and a specificity of 0.51 (95% CI 0.46 – 0.56).⁶ This is comparable to existing decision support tools such as the paramedic pathfinder.^{8,94} A systematic review into whether machine learning computerised decision support could offer an improvement on triage found that certain methods such as decision trees, neural networks and logistic regression were all able to provide accurate discrimination between different acuity levels. A limitation of the included studies was that they were often predicting high acuity.²⁰³

If current clinical judgement is already sensitive to identifying high-acuity patients, the benefit of a decision support tool is on triaging the mid- and low-acuity. If accuracy is improved at this level of triage, the benefit would be a reduction in the avoidable transportation of patients to an ED.

8.1.3 Objectives

Primary research question

In adult patients attending the ED by ambulance, can prehospital information predict an avoidable attendance?

Primary objectives

- 1) Extract prehospital variables from ambulance service electronic patient care records
- 2) Link the data with ED electronic patient care records
- 3) Identify low acuity patients in the dataset using the ED information
- 4) Build a predictive model using prehospital variables
- 5) Measure the success of the model in predicting an avoidable attendance using prehospital variables.

Secondary research questions

What is the simulated transportability of the model derived from the primary outcome?

Secondary objectives

6) Test spatial validation

7) Test model discrimination of protected characteristics

8.2 Methods

8.2.1 Source of data

This retrospective cohort study analysed a sample of ambulance service attendances between the 1st of July 2019 and the 29th February 2020. Each face-to-face attendance had an electronic Patient Care Record (ePCR) created which contained all demographic and clinical information. A similar record was also created at the Emergency Department (ED) if the patient was conveyed there. These records were linked so that the outcome could be generated in the ED data, and the candidate variables in the ambulance data.

8.2.2 Participants

In this study, all patients who were over the age of 18 and had a face-to-face ambulance in Yorkshire with a completed ePCR were eligible for inclusion. The patients were not selected by any specific demographic or disease in order to develop a model which could be applied to all patients. Children were excluded from the model as ambulance policy can dictate transport decisions to the clinicians on scene. For example, mandatory transportation of under 5s.

8.2.3 Outcome

The outcome is a avoidable conveyance attendance at the ED, which is an experience-based definition initially described by O’Keeffe et al. as “first attendance with some recorded treatments or investigations all of which may have reasonably been provided in a non-emergency care setting, followed by

discharge home or to GP care”.⁵ This was operationalised into a data-driven definition and can be found in the protocol publication.²⁰³

8.2.4 Predictors

All candidate variables were measured whilst the ambulance crew was with the patient prospectively. Data were retrieved after the data collection period, and no ambulance crew was aware of the study during data collection. Variables can be broadly categorised into demographic, clinical, social and interventional. The only demographic variable included was incident location as a categorical variable. This variable is user inputted by the ambulance crew depending on whether the patient is at a domestic address, public place, care home, work or other. Age was also initially included: however, after initial model building, it was found to introduce a bias and was removed. Clinical variables formed most of the candidate variables. When a paramedic arrives on scene, they will first undertake a primary survey. This records whether the patient has a catastrophic haemorrhage, whether their airway is clear, whether they are breathing normally, whether there are any obvious circulation issues. These are all recorded as categorical variables. The patient will then have physiological variables recorded in order to assess how serious their medical complaint may be. Pulse rate is measured in beats per minute (bpm) and is the frequency at which the heart beats in a minute. Traditionally this is measured by palpation of the pulse, however modern technology allows this to be measured using medical equipment. Respiratory rate is measured as respirations per minute (rpm) and is a manual count of the number of breaths the patient takes in one minute. Temperature is a continuous variable measured in °C using a tympanic thermometer. The peripheral capillary oxygen saturation in the blood (SpO₂) is measured using medical equipment as a percentage. Blood sugar levels are also recorded using a machine that takes a small blood sample. The results are recorded as mmol per litre. Blood pressure is recorded using millimetres of mercury (mmHg). Two measurements are recorded, the systolic blood pressure and the diastolic blood pressure. The level of consciousness is calculated using a four-scale system (AVPU

= Alert, Voice, Pain, Unresponsive) in the primary survey and the Glasgow Coma Scale (GCS) in the physiological observations. GCS is a composite score of labelled scales. The minimum score is three and maximum fifteen.²⁰⁴ Baseline oxygen demands, and current oxygen demands are recorded as binary variables. All the physiological variables are combined to calculate a National Early Warning 2 score (NEWS2).^{179,180} The NEWS2 score has been included as a candidate predictor and treated as categorical. Other clinical variables include pain scores out of ten, subsequent measurements of observations and feature engineered intervals between primary measurements and subsequent ones. All clinical interventions (e.g., cannulation, intubation etc.) were included as binary variables. The patient's mobility was recorded depending on what resource they required, i.e., self-mobile, stretcher needed, carry chair needed etc. The final clinical impression was also included as a categorical variable with 99 different values to possibly select. Examples include 'head injury', 'shortness of breath', and 'abdominal pain'. Social variables were included as binary variables. These include network variables such as GP details recorded, social worker recorded etc. It also included referral variables if the patient was referred to a service such as falls, safeguarding or diabetes clinic etc. A full list of candidate variables, their units and variable types can be found in chapter 7, tables 8-11.²⁰³

8.2.5 Sample size

The sample size was calculated using the 'pmsampsize v1.1.0' for R v3.6.1 for windows.²⁰⁵ Two studies by Riley et al. also informed the sample size calculation.^{183,206} Previous studies have found a conservative estimate of the outcome prevalence to be 0.9.⁵ A meta-analysis found that the average C-statistic was 0.8.²⁰⁷ A preliminary analysis of a separate dataset found that there were potentially 637 parameters in the ambulance service dataset. This gave an estimated sample size of 52,958 with an anticipated 4,767 event and an events per parameter (EPP) of 7.48. A full list of the parameters can be found in chapter 7, tables 8-11.²⁰³

8.2.6 Missing data

The strategy for handling missing data was to first elicit if missing values in each variable were the negative class. For example, the clinical procedure of intravenous cannulation is only recorded in the ePCR if the patient was cannulated. Therefore, it is logical, in the absence of a positive recording to assume the patient was not cannulated and the missing data can be transformed into the negative class. Once this had been completed, any variable with more than 30% missing data was excluded from the analysis. The rationale for this was that it may not be routinely, or accurately completed in the ePCR and to include them could lead to model failure in practice.

8.2.7 Statistical analysis methods

The full statistical analysis plan has been published in the study protocol.²⁰³ In this study, an XGBoost algorithm was used for model development. Recursive feature elimination was used to subset the candidate variables into only the most important that provided the most accurate prediction model. Then the algorithms hyperparameters were tuned in order to prevent model overfitting. The model was first evaluated for its calibration using Spiegelhalter's Z-test. Then, model discrimination was assessed using the C-statistic (area under the ROC curve). The optimal threshold was identified by finding the closest top left point of the ROC curve. This was then used to assess accuracy statistics. Once the full model was completely developed, symmetrical procedures were undertaken using different Emergency Departments as held-out test sets with all remaining data as the training data. This in effect created a full model and seventeen other models which could then be meta-analysed. The summary statistics generated in a random effects meta-analysis were then used to update the final model for its performance. In the protocol paper, the full procedures are outlined in detail.²⁰³ This study is a development study with internal-external validation using a meta-analysis of ED clusters. There is no external validation.

8.3 Results

8.3.1 Participants

There were 101,522 individual patient episodes included in the analysis. Of these, 7228 (7.12%) were defined as having an avoidable conveyance attendance at the ED. Table 12 provides key demographic information between those with, and without, the outcome. It also shows physiological observations as a surrogate for comparative patient acuity. In appendix H, the table is extended to show the clinical impression fields.

Table 12: Characteristics of participants

| | Unavoidable (N=94294) | Avoidable (N=7228) | Overall (N=101522) |
|--------------------------------------|--------------------------|-----------------------|-----------------------|
| Gender | | | |
| Female | 52620 (93%) | 4120 (7%) | 56740 |
| Male | 41572 (93%) | 3100 (7%) | 44672.00 |
| Transgender | 7 (88%) | 1 (13%) | 8 |
| Unknown | 95 (93%) | 7 (7%) | 102 |
| Age | | | |
| Mean (SD) | 66.8 (20.3) | 50.9 (22.6) | 65.7 (20.9) |
| Median [Min, Max] | 72.0 [18.0, 107] | 48.0 [18.0, 107] | 71.0 [18.0, 107] |
| Ethnicity | | | |
| African (Black or Black British) | 269 (89%) | 33 (11%) | 302 |
| Caribbean (Black or Black British) | 380 (91%) | 36 (9%) | 416 |
| Any other Black background | 164 (86%) | 26 (14%) | 190 |
| Bangladeshi (Asian or Asian British) | 124 (84%) | 23 (16%) | 147 |
| Chinese (Asian or Asian British) | 59 (86%) | 10 (14%) | 69 |
| Indian (Asian or Asian British) | 521 (90%) | 55 (10%) | 576 |
| Pakistani (Asian or Asian British) | 2894 (87%) | 415 (13%) | 3309 |
| Any other Asian background | 382 (85%) | 67 (15%) | 449 |
| British (White) | 78401 (94%) | 5420 (6%) | 83821 |
| Irish (White) | 361 (93%) | 29 (7%) | 390 |
| Any other White | 2464 (90%) | 263 (10%) | 2727 |
| White and Asian (Mixed) | 76 (85%) | 13 (15%) | 89 |
| White and Black African (Mixed) | 35 (81%) | 8 (19%) | 43 |
| White and Black Caribbean (Mixed) | 113 (92%) | 10 (8%) | 123 |
| Any other Mixed background | 150 (90%) | 17 (10%) | 167 |
| Any other ethnic group | 761 (84%) | 142 (16%) | 903 |
| Unknown | 2554 (90%) | 281 (10%) | 2835 |
| Not stated | 4586 (92%) | 380 (8%) | 4966 |
| Incident location | | | |
| Care Home | 7614 (95%) | 372 (5%) | 7986 |
| Domestic Address | 68004 (93%) | 5281 (7%) | 73285 |
| Not Selected | 27 (96%) | 1 (4%) | 28 |
| Other | 4449 (93%) | 320 (7%) | 4769 |
| Public Place | 2710 (89%) | 335 (11%) | 3045 |
| School | 30 (70%) | 13 (30%) | 43 |
| Work | 473 (88%) | 65 (12%) | 538 |
| Missing | 10987 (93%) | 841 (7%) | 11828 |

| Transported ED | | | |
|--|-------------------|-------------------|-------------------|
| Airedale General Hospital | 3058 (93%) | 240 (7%) | 3298 |
| Barnsley District General | 5810 (95%) | 323 (5%) | 6133 |
| Bradford Royal Infirmary | 6705 (87%) | 1004 (13%) | 7709 |
| Calderdale Royal Hospital | 3865 (94%) | 242 (6%) | 4107 |
| Dewsbury District Hospital | 827 (86%) | 137 (14%) | 964 |
| Doncaster Royal Infirmary | 6258 (94%) | 420 (6%) | 6678 |
| Harrogate District Hospital | 2598 (94%) | 163 (6%) | 2761 |
| Huddersfield Royal Infirmary | 4392 (94%) | 283 (6%) | 4675 |
| Hull Royal Infirmary | 10099 (94%) | 612 (6%) | 10711 |
| James Cook University Hospital | 749 (93%) | 55 (7%) | 804 |
| Leeds General Infirmary | 4839 (95%) | 263 (5%) | 5102 |
| Northern General Hospital | 9793 (91%) | 929 (9%) | 10722 |
| Pinderfields General Hospital | 9481 (93%) | 764 (7%) | 10245 |
| Rotherham District General Hospital | 5618 (94%) | 352 (6%) | 5970 |
| Scarborough District General Hospital | 4374 (97%) | 120 (3%) | 4494 |
| St James University Hospital | 8078 (91%) | 824 (9%) | 8902 |
| York District Hospital | 5719 (94%) | 382 (6%) | 6101 |
| Missing | 2031 (95%) | 115 (5%) | 2146 |
| Indices of Deprivation | | | |
| 1 | 22882 (91%) | 2331 (9%) | 25213 |
| 2 | 12177 (92%) | 1054 (8%) | 13231 |
| 3 | 9934 (92%) | 817 (8%) | 10751 |
| 4 | 7439 (93%) | 518 (7%) | 7957 |
| 5 | 7560 (94%) | 484 (6%) | 8044 |
| 6 | 8025 (94%) | 504 (6%) | 8529 |
| 7 | 7801 (94%) | 459 (6%) | 8260 |
| 8 | 7199 (94%) | 432 (6%) | 7631 |
| 9 | 5959 (95%) | 346 (5%) | 6305 |
| 10 | 5199 (95%) | 272 (5%) | 5471 |
| Missing | 119 (92%) | 11 (8%) | 130 |
| Initial Pulse rate (bpm) | | | |
| Mean (SD) | 89.2 (22.3) | 88.1 (18.2) | 89.1 (22.0) |
| Median [Min, Max] | 86.0 [5.00, 220] | 87.0 [6.00, 220] | 86.0 [5.00, 220] |
| Missing | 2186 (2.3%) | 309 (4.3%) | 2495 (2.5%) |
| Initial Respiratory rate (rpm) | | | |
| Mean (SD) | 20.7 (6.30) | 18.7 (4.45) | 20.5 (6.21) |
| Median [Min, Max] | 18.0 [0, 99.0] | 18.0 [0, 96.0] | 18.0 [0, 99.0] |
| Missing | 1820 (1.9%) | 188 (2.6%) | 2008 (2.0%) |
| Initial Systolic Blood Pressure (mmHg) | | | |
| Mean (SD) | 143 (28.3) | 143 (24.4) | 143 (28.1) |
| Median [Min, Max] | 142 [0, 265] | 140 [1.00, 288] | 142 [0, 288] |
| Missing | 2991 (3.2%) | 388 (5.4%) | 3379 (3.3%) |
| Initial Diastolic Blood Pressure (mmHg) | | | |
| Mean (SD) | 82.9 (17.7) | 86.4 (15.6) | 83.2 (17.6) |
| Median [Min, Max] | 83.0 [0, 200] | 86.0 [4.00, 182] | 83.0 [0, 200] |
| Missing | 3114 (3.3%) | 397 (5.5%) | 3511 (3.5%) |
| Initial Oxygen saturations (%) | | | |
| Mean (SD) | 95.3 (5.41) | 97.1 (2.84) | 95.4 (5.29) |
| Median [Min, Max] | 97.0 [11.0, 100] | 98.0 [18.0, 100] | 97.0 [11.0, 100] |
| Missing | 2543 (2.7%) | 329 (4.6%) | 2872 (2.8%) |
| Initial temperature (Celsius) | | | |
| Mean (SD) | 37.0 (0.965) | 36.8 (0.735) | 37.0 (0.952) |
| Median [Min, Max] | 36.9 [31.7, 42.1] | 36.8 [33.0, 40.7] | 36.9 [31.7, 42.1] |
| Missing | 5935 (6.3%) | 796 (11.0%) | 6731 (6.6%) |
| Initial Pain Score | | | |
| Mean (SD) | 3.10 (3.58) | 2.94 (3.50) | 3.09 (3.57) |
| Median [Min, Max] | 1.00 [0, 10.0] | 0 [0, 10.0] | 1.00 [0, 10.0] |

| | | | |
|----------------------------|---------------|--------------|---------------|
| Missing | 26475 (28.1%) | 2073 (28.7%) | 28548 (28.1%) |
| Self Mobile | | | |
| Yes | 26079 (87%) | 3792 (13%) | 29871 |
| No | 68215 (95%) | 3436 (5%) | 71651 |
| Initial NEWS2 score | | | |
| 0 | 20807 (90%) | 2194 (10%) | 23001 |
| 1 | 16801 (90%) | 1779 (10%) | 18580 |
| 2 | 10527 (92%) | 928 (8%) | 11455 |
| 3 | 9910 (94%) | 610 (6%) | 10520 |
| 4 | 6899 (95%) | 330 (5%) | 7229 |
| 5 | 5172 (97%) | 186 (3%) | 5358 |
| 6 | 4564 (97%) | 128 (3%) | 4692 |
| 7 | 3313 (98%) | 60 (2%) | 3373 |
| 8 | 2696 (99%) | 40 (1%) | 2736 |
| 9 | 2033 (99%) | 16 (1%) | 2049 |
| 10 | 1475 (99%) | 10 (1%) | 1485 |
| 11 | 969 (99%) | 9 (1%) | 978 |
| 12 | 595 (100%) | 2 (0%) | 597 |
| 13 | 441 (100%) | 2 (0%) | 443 |
| 14 | 241 (99%) | 3 (1%) | 244 |
| 15 | 138 (100%) | 0 (0%) | 138 |
| 16 | 58 (100%) | 0 (0%) | 58 |
| 17 | 34 (100%) | 0 (0%) | 34 |
| 18 | 13 (100%) | 0 (0%) | 13 |
| 19 | 2 (100%) | 0 (0%) | 2 |
| Missing | 7606 (89%) | 931 (11%) | 8537 |

8.3.2 Model Development

8.3.3 Dataset preparation

During the preparation of the dataset there were 215 possible candidate variables for inclusion which comprised of 190 categorical variables (including 169 binary variables), and 25 continuous variables. After one hot encoding there were 452 candidate predictors in the final dataset. During recursive feature elimination, the ideal set of variables was found to be only 90 of the total candidate variables. These were condensed down into 19 variables, comprising of 14 clinical variables, 3 interventional and 2 demographics. A full list of included candidate variables can be found in chapter 7, tables 8-11.

8.3.4 Model performance

In an XGBoost algorithm, the hyperparameters that control how the model is built prevents the model overfitting the training data. Therefore the apparent validity can be perceived as less optimistic from the outset.¹⁴² Table 13 is a brief summary of the performance measures being used to evaluate the model.

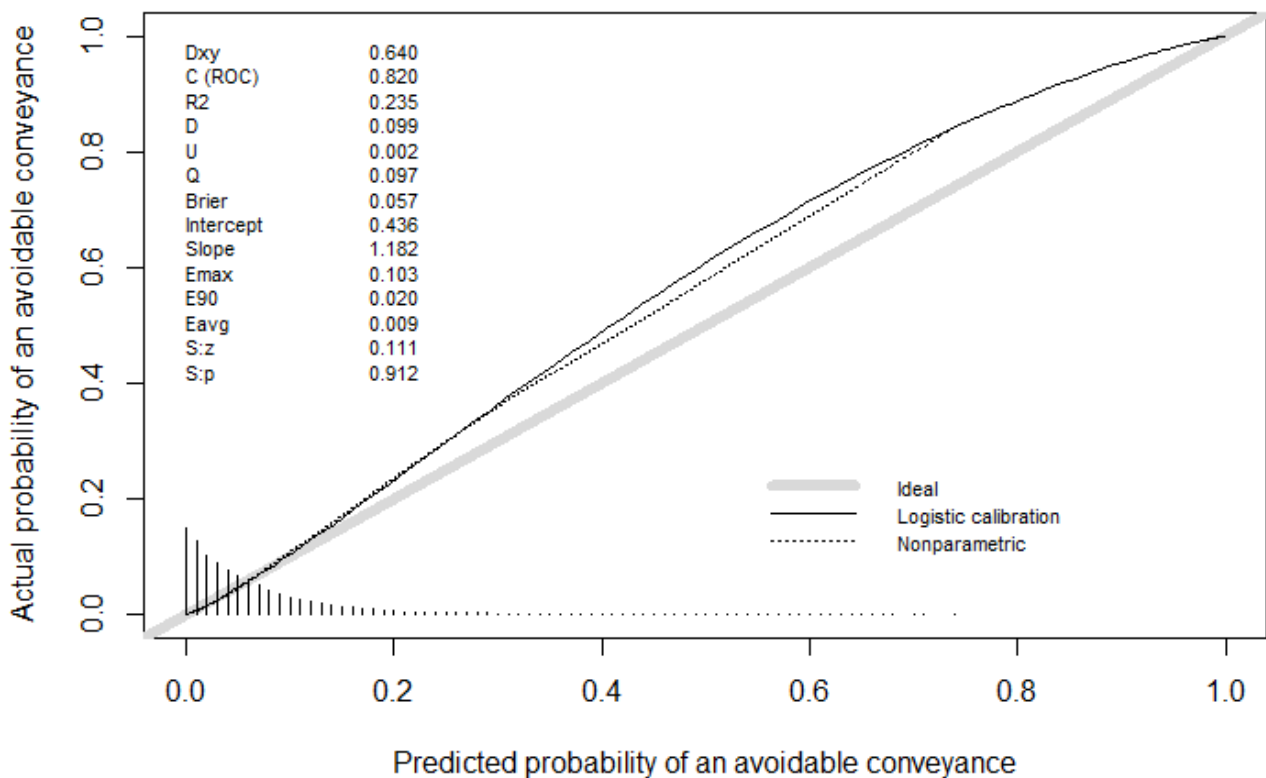
Table 13: Model performance measures

| Test | Description | Statistic | Interpretation |
|----------------|---|------------------------|---|
| Calibration | Assessment of whether the predicted probabilities match with the observed probabilities. | O:E ratio | A perfect O:E ratio would be 1. If the model is over-trianging, the O:E ratio will be greater than 1 as it would predict more than observed, and vice versa. |
| | | Spiegelhalter's z-test | A Spiegelhalter's z-test that falls outside the interval of -1.96 – 1.96 will have a p-value greater than 0.05 and it means the model is miscalibrated. |
| Discrimination | Discrimination is assessing whether the model can take two random instances (one with and without the outcome) and tell them apart. | C-statistic | The C-statistic of 0.5 means the model is no better than chance at telling apart the two random instances. A C-statistic of 1 means the model will tell the two random instances apart every time. A good C-statistic achieved in prior studies for this clinical problem is 0.8. |

8.3.5 Calibration

Calibration was assessed using Spiegelhalter's Z-test and calculated using the Rmisc package v1.5.²⁰⁸ The interpretation of this Z-test is such that a statistically significant test result means the model is miscalibrated as the null hypothesis is that it is a well calibrated model. The initial model was miscalibrated with a Spiegelhalter's Z-test of -3.668 ($p = 0.001$). Therefore, the weighting of the positive class was tuned to two decimal places to yield the smallest Z-test with no statistical significance. The optimum value for scale_pos_weight was 0.95 which gave a Spiegelhalter's Z-test of 0.111 ($p = 0.912$). The ratio of the observed and expected (O:E) was 1.042 (95% CI 1.02 – 1.07). The full calibration plot with intercept and slope can be found in figure 13.

Figure 13: Full model calibration plot

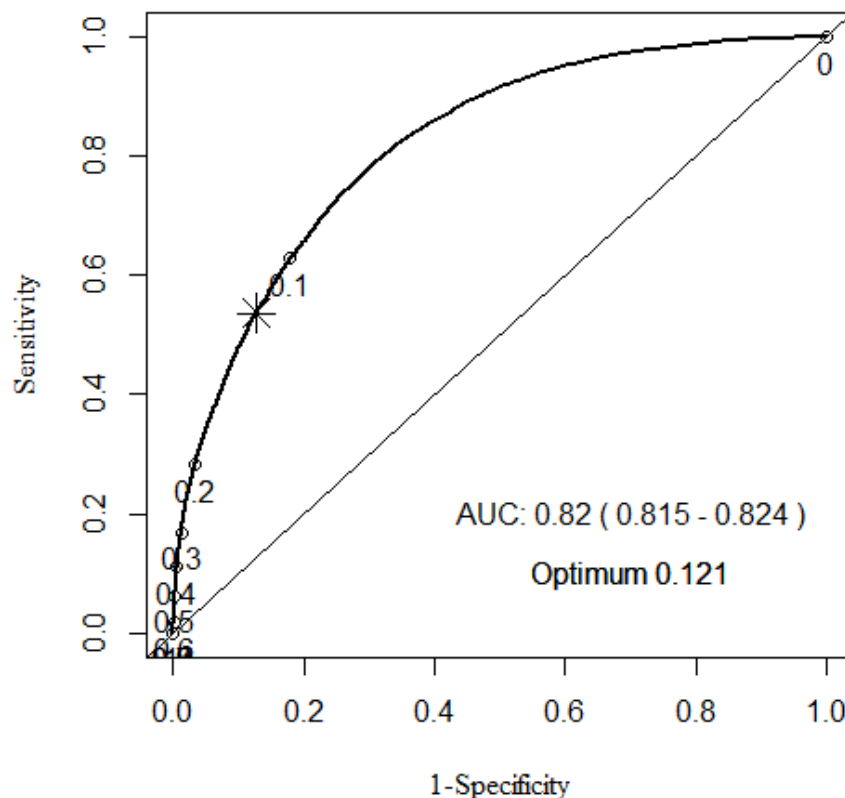


8.3.6 Discrimination

The C-statistic for the full model was 0.815 (95% CI 0.820-0.824). The optimum cut point was 0.121, which gave a specificity of 0.87 and a sensitivity of 0.54. The ROC curve with different thresholds including the optimal threshold (marked with a star) can be found in figure 14. The threshold was chosen as the 'closest top left' point mathematically. Experiments were performed by maximising specificity, but the model was unstable, and the sensitivity decreased by such a significant amount that it would miss-classify far more often than it would classify.

Using the optimal cut point, the full model had an accuracy of 0.85 (95% CI 0.847 – 0.852). The model had a preference towards specificity as it was predicting health and not disease. The positive predictive value (PPV) was 0.25 (95% CI 0.24 - 0.25) and the negative predictive value was 0.96 (95% CI 0.96 - 0.963).

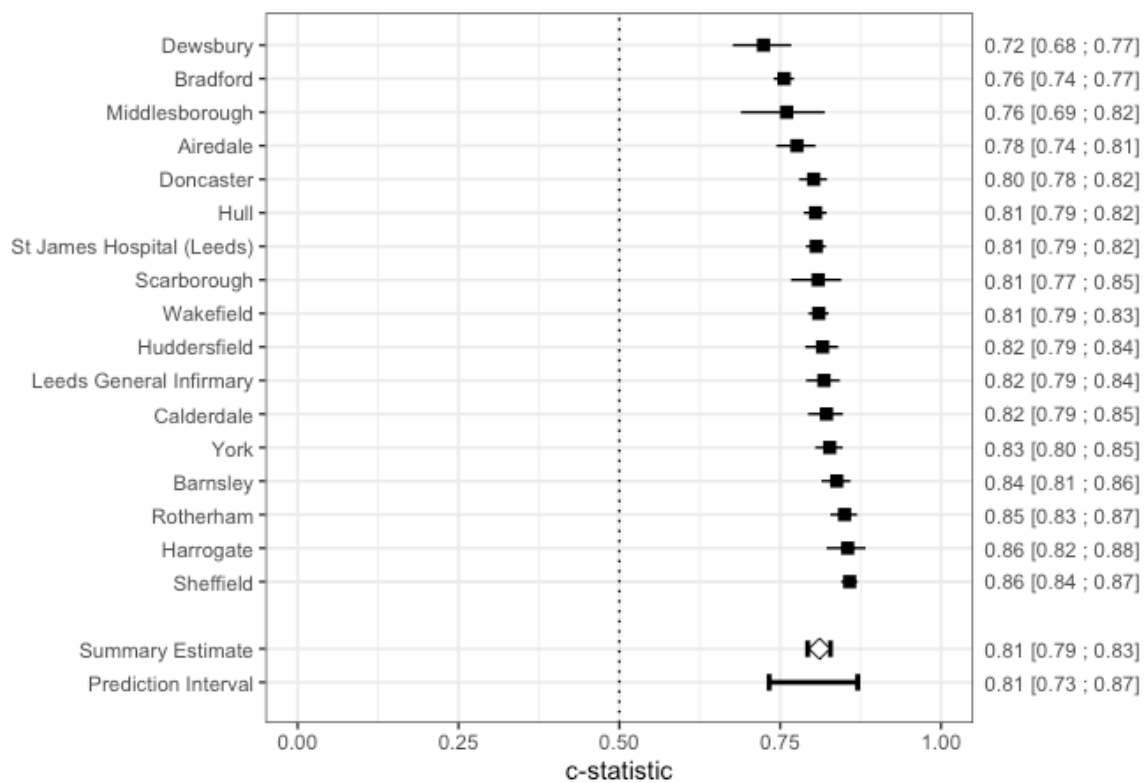
Figure 14: ROC curve of the full model



8.3.7 Model updating

The meta-analysis was undertaken using the framework by Debray et al. and used the metamisc package v0.2.5.^{189,209} In the meta-analysis of clusters, the C-statistic was found to be 0.81 (95%CI 0.79-0.83). The prediction interval was between 0.73 and 0.87. Figure 15 shows the forest plot of C-statistic results for each cluster. The meta-analysed O:E ratio was 0.995 (95% CI 0.97 – 1.03) with a prediction interval between 0.93 and 1.06. In appendix I, there is further information on the hyperparameters chosen in each cluster.

Figure 15: Meta-analysis of cluster discrimination



8.3.8 Fair Machine Learning analysis

In the analysis of fair machine learning, each demographic was assessed on two criteria. The first was comparing the probability distribution of each category within the variable and the second was examining how many were misclassified in each category. If age is left in as a candidate variable, the model becomes more

accurate but introduces a bias towards younger patients. When excluded, the model slightly decreases in performance but removes the bias. There were no significant differences in the mean probabilities, distributions or misclassification for any of the demographic variables assessed. This included ethnicity, gender, and social deprivation.

8.3.9 Misclassification analysis

There were (3880 (3.8%)) true positive predictions where the model correctly identified an avoidable ambulance conveyance and (82,340 (81.1%)) true negatives where it identified an unavoidable conveyance. There were (11,954 (11.8%)) false positives and (3348 (3.3%)) false negatives. This gave a misclassification rate of (0.151).

8.3.10 Variable importance

Variable importance can be broken down into three features - frequency (weight), coverage and gain. Frequency represents how many times a particular feature appears in the trees of the full model as a percentage of all the frequencies. Coverage is the number of instances that are contained within a feature when it is used as a split. Gain is the relative contribution of each feature to the whole model. Figures 16,17 and 18 show the frequency, coverage and gain for the model.

Figure 17: Top 20 variables used in the full model by frequency

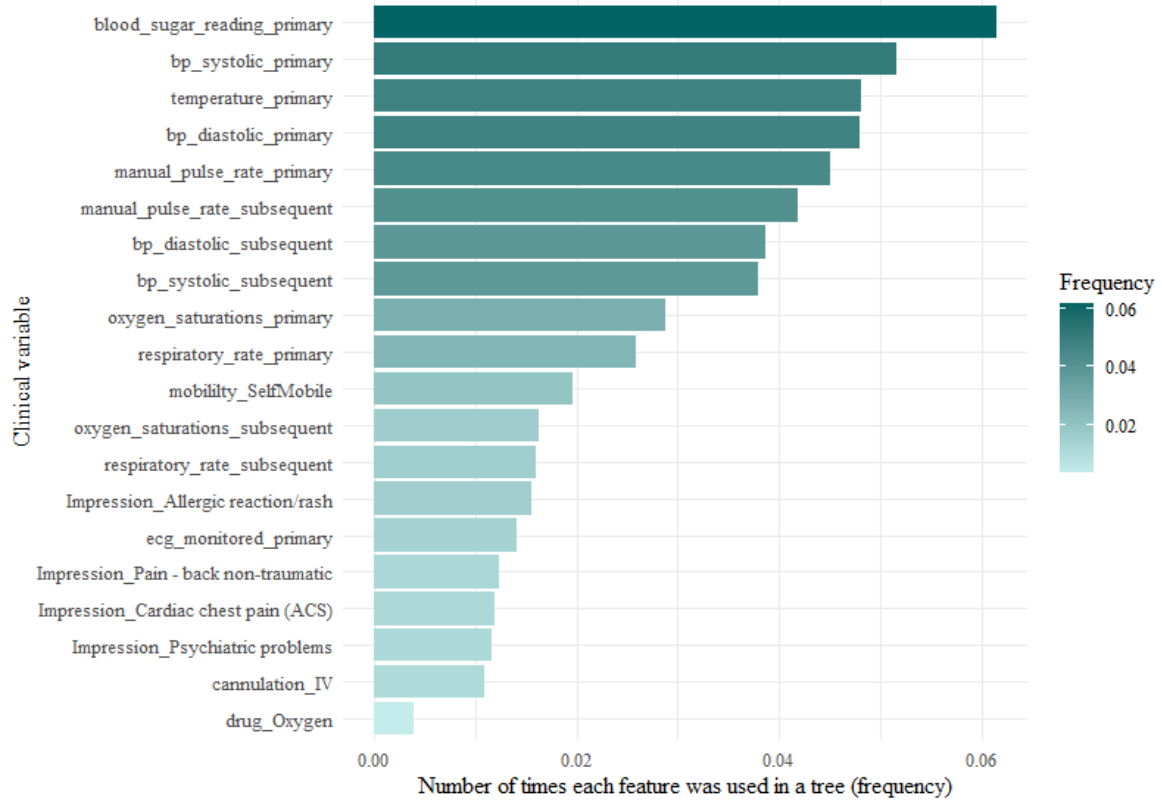


Figure 16: Top 20 variables with the highest relative contribution to the full model (gain)

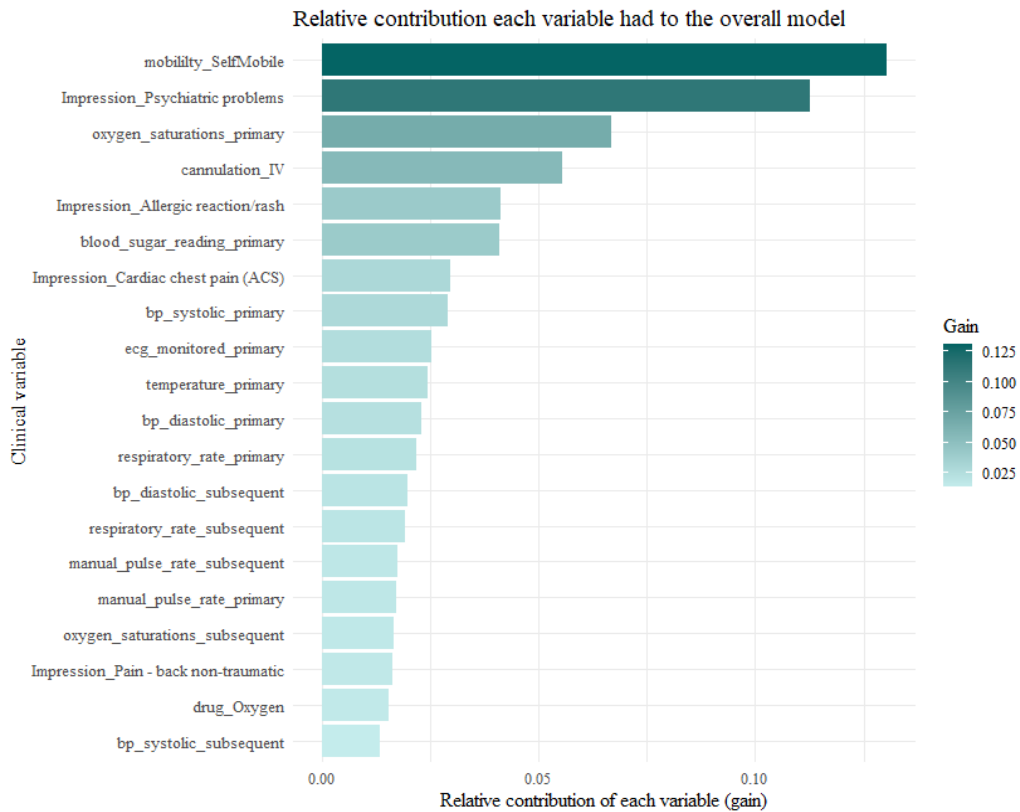
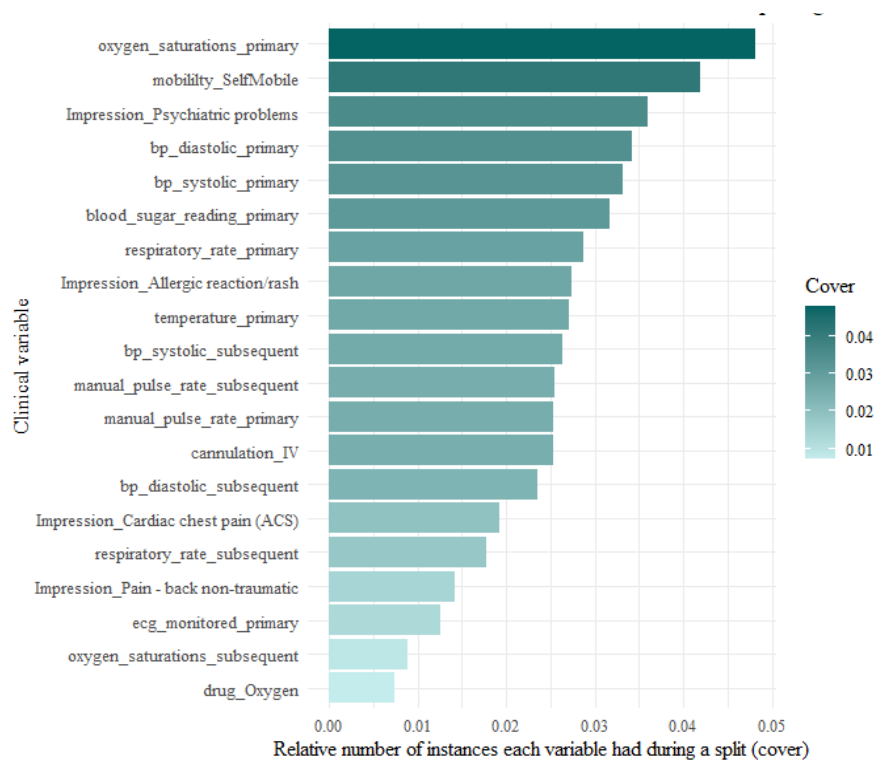


Figure 18: Top 20 variables with the greatest number of instances when splitting (cover)



8.4 Discussion

This study used a large sample of conveyed ambulance patients linked to their ED record to derive a risk prediction model. Using feature selection, it was found that only a limited number of features contributed to identifying an avoidable conveyance. Most of these related to physiological observations, a patient's mobility and clinical impression.

8.4.1 Feature importance

A novel finding from this study was the identification of six clinical impressions that were important in predicting avoidable conveyances. There were six clinical impressions that featured in the top twenty. The most important was patients presenting with psychiatric problems. This could be a reflection on the experience of mental health presentations at the ED. They rarely require investigations or treatments that physical health presentations may require. The main purpose of the ED for these patients is to offer a place of safety and access to a mental health practitioner who can better meet their care need. Other clinical

impressions were allergic reactions, cardiac chest pain, head injury, non-traumatic back pain, minor cuts and bruising. These have been previously identified in observational studies as being associated with a non-urgent ambulance conveyance.^{53,54} All physiological observations appeared in the top twenty, however the NEWS2 score did not. Only three NEWS2 scores were included in the full model. A NEWS2 score of 0 appeared as the 31st variable, a score of 1 as the 58th, a score of 2 as the 78th and a score of 5 as the 93rd. This may mean that low NEWS2 scores are not strong predictors of an avoidable conveyance attendance. This is an interesting finding, as the decision tree should have associated higher NEWS2 scores with information gain from ruling out an avoidable conveyance. Conversely, it omitted most NEWS2 scores during recursive feature elimination. In the full model and all the clusters, the frequency, cover and gain did not change rank order, which shows the stability of their importance. When predicting high acuity, it is often easier to find significant variables, as physiological observations, such as pulse rate and respiration rate, will change when patients are acutely unwell. However, when predicting avoidable conveyance patient episodes, physiological observations will often be normal. Interestingly, there were clinical variables more important than physiological observations that have featured in other triage models as main candidate predictors.^{110,171} In the development of decision tree models, splits are made based on the information gained. This can be either gain, in deciding what an avoidable conveyance patient is, or gain in deciding what an avoidable conveyance patient is not. As such, variables associated with higher acuity appear high in variable importance as they rule out necessary attendances. The algorithm has identified signals of higher acuity patients with high prevalence of completion within the ePCR. For example, delivering advanced life support to someone in cardiac arrest does not often happen in the overall case-mix of ambulance patients. Therefore, the skills and procedures associated with undertaking ALS were rarely captured and were not identified as important. However, far more patients had the clinical procedure of intravenous cannulation or were monitored by ECG, and these appeared as the fifth and

eighth most important variables respectively. This theory can be extended to the patient's mobility. In the model, a patient's mobility status is important, as being stretcher bound, self-mobile or needing a carry chair all featured in the top twenty.

8.4.2 Model performance

The model was well calibrated with a meta-analysed O:E ratio of 0.99 (95% CI 0.96 – 1.02). This means that the model is making accurate predictions across all values. The model is also successful in distinguishing between an avoidable ambulance conveyance and one that needed transport to hospital with a C-statistic of 0.81 (95% CI 0.79-0.83). The optimal threshold for classification was 0.125 which appears low, but so is the proportion of avoidable ambulance conveyances and this reflects the class imbalance. The model provided many false negatives with a sensitivity of 0.58, meaning that 42% of patients who were classified as needing ED care were avoidable conveyances. The choice of threshold is a point of discussion. It could be adjusted to a higher or lower value, but this would influence the sensitivity and specificity. To illustrate, the ROC curve in figure 14 shows the thresholds above 0.2 have limited effect on the specificity but a large effect on sensitivity. If the threshold was changed to 0.2 for example, the sensitivity drops dramatically to 0.28. The optimum threshold was chosen to be the highest specificity with the highest sensitivity, which is also known as a balanced approach. It was also possible to take the Youden index, which would place the threshold at the nearest point to the top left corner, but this placed too much of a penalty on specificity to create a functioning tool.

The meta-analysis of clusters revealed that there were no significant performance differences between test sets in urban areas, rural areas or coastal areas. There were significant differences in the calibration slopes; however, this was at the latter part of the plots where predicted outcome was rare. They all produced O:E ratios that were acceptable except for two smaller test sets (Dewsbury hospital and James Cook University Hospital), which had significant under-triage.

There was only a prevalence of 7% for avoidable conveyance attendances in the study sample. This is fewer cases than the literature had previously reported (9-13%).^{5,13,49,54} This may appear low; however, the quantity of high acuity patients is similar, indicating that to predict avoidable conveyance and high-acuity would be predicting the two tails of a normal distribution. Future studies should examine the mid-acuity patients and begin to unpick differences between these patients to improve on the outcome definition of a patient who is unlikely to gain a clinical benefit from being transported to a higher-acuity clinical setting than community care.

8.4.3 Limitations

This study has its limitations. It was a retrospective, observational study using routine data. A strength of using routine data is the ability to use large volumes of patient episodes, which can produce accurate models. A limitation, however, is that it is not feasible to tailor data collection to the project. It is only possible to use what is routinely collected. This extends to not being able to control how a variable is collected. For example, ethnicity may not have been explicitly stated by the patient every time but could have been assumed by the clinician. Another limitation is the computational expense of selecting an algorithm with many hyperparameters. It would take a significant amount of time to be able to scan all combinations of hyperparameters through a grid search every time a model was developed. As such, the grid was restricted. The anticipated impact of the restricted grid search is expected to be minimal as the differences in AUC performance (the evaluation metric of choice) had a narrow interval of between 0.7 and 0.85. The validation does not benefit from true external validation, and it would be a sensible conclusion to revisit the definition of an avoidable ambulance conveyance, or indeed the taxonomy of how prehospital care systems classify their patients, based on their need before further validation of the SINEPOST model.

There is a limitation in using a risk (or probability) based approach to answering this research question when later transforming the model into a tool. The process is very analytical and does not successfully answer how staff would use this information or whether they would trust the outcome of the tool. This was particularly highlighted in an article by Kappen et al. who evaluated the impact of prediction models. It was highlighted that a prediction model is, in effect, a complex intervention. Due to this, it is difficult to ascertain the exact benefit if used in practice.²¹⁰

8.4.5 Interpretation

This study can conclude that it is possible, with good accuracy, to predict an avoidable ambulance conveyance to the ED using prehospital clinical data. The XGBoost model developed here, known as the SINEPOST model, can discriminate between those with non-urgent needs and those without. It can also accurately provide what the probability of an avoidable conveyance is. The model does not bias different ages, ethnicities, genders, or Indices of Deprivation. It is robust to all different prehospital settings. However, to maximise its potential, if it was to be transformed into a computerised clinical decision support tool, there needs to be a more robust definition of what an avoidable conveyance should be. It is recommended to revise the taxonomy of prehospital patients according to the care setting they need, as opposed to the paradigm of describing patient acuity.

8.5 Conclusion

This chapter has presented the findings of this thesis in the form of a manuscript. The model was successfully created as demonstrated by its calibration, discrimination, and accuracy. The manuscript was written for a clinical audience, and as a result there were details omitted. The next chapter will elaborate on these results including the hyperparameter optimisation and the results for each cluster.

Chapter 9

The Further Results

9.1 Introduction

This chapter presents more detailed results that were not included in the manuscript but submitted as a supplementary file. The main results were aimed at a clinical audience and therefore some of the detail on model development that may distract from the clinical context were left out. In this chapter, the hyperparameters and recursive feature elimination are discussed as well as the age variable in further detail. The results from all of the clusters are also displayed.

9.2 Data modelling

9.2.1 Full model hyperparameter optimisation

The procedure for sequential tuning was to set an initial list of default hyperparameter values but with eta initially to a high value as this was the learning rate. Then `max_depth` and `min_child_weight` were optimised first in tandem, and the list updated. This was followed by `subsample` and `colsample_bytree` in tandem and list updated, then `gamma`, then `alpha` and then eta was the last to be tuned. Once the hyperparameters were tuned, the best performing values were taken forward into a restricted grid. This restricted grid is then used for hyperparameter optimisation for each model. Table 14 shows the restricted grid based on the sequential tuning of each hyperparameter. Where hyperparameters only presented two values within the top ten combinations, only two values were selected to reduce computational expense.

Table 14: Hyperparameter grid

| Hyperparameter | 1 st value | 2 nd value | 3 rd value |
|-------------------------------|-----------------------|-----------------------|-----------------------|
| <code>max_depth</code> | 3 | 4 | - |
| <code>min_child_weight</code> | 2 | 4 | - |
| <code>subsample</code> | 0.7 | 0.9 | 1 |
| <code>colsample_bytree</code> | 0.9 | 0.6 | - |
| <code>gamma</code> | 0 | 0.5 | 1 |
| <code>alpha</code> | 0.6 | 0.7 | 0.8 |
| <code>eta</code> | 0.08 | 0.06 | - |

The full model grid search found the following optimal hyperparameter values: `max_depth = 3`, `min_child_weight = 2`, `subsample = 0.9`, `colsample_by_tree = 0.6`, `gamma = 0.5`, `alpha = 0.8`, `eta = 0.06`. The hyperparameter values of each cluster model can be found below.

9.2.2 Model recalibration

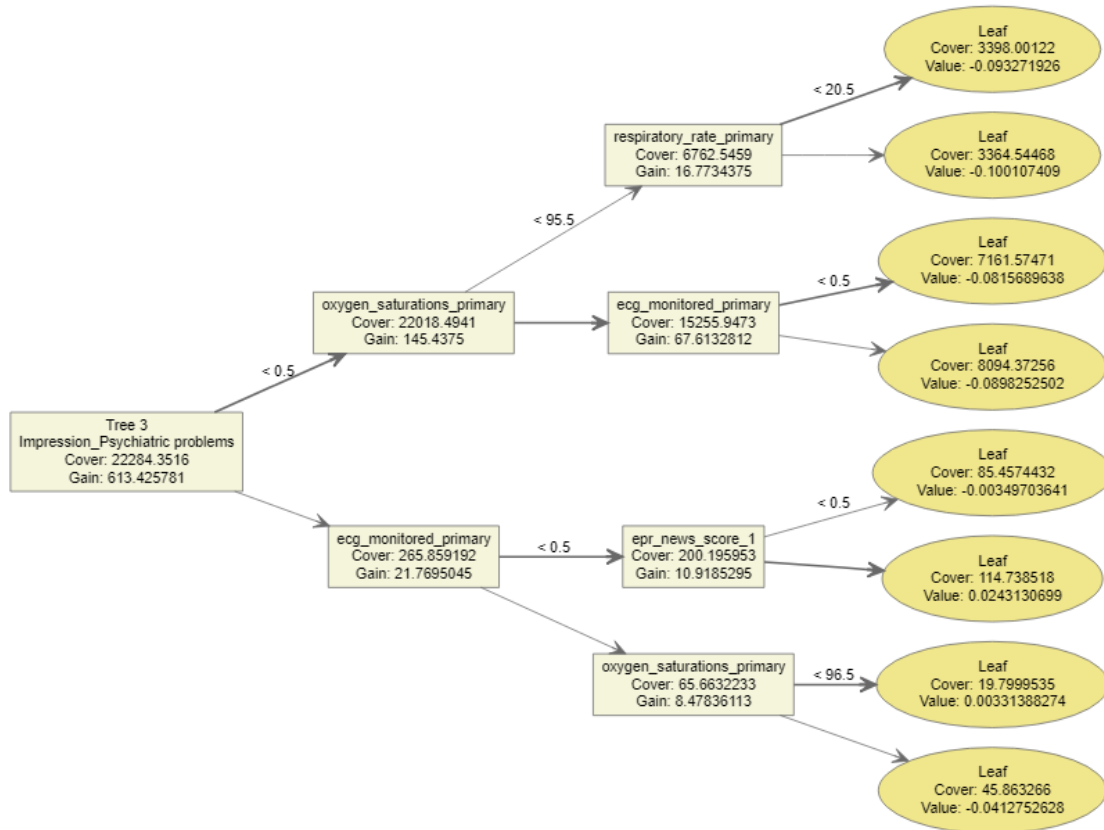
The initial model appeared to be miscalibrated with a Spiegelhalter's Z-test of -3.668. Recalibration was performed by manually fine tuning the hyperparameter `scale_pos_weight`, which adjusts the weight of the positive class. Tuning this hyperparameter also increase the discrimination; however, it comes at the expense of predicting the right probability.¹⁴² The value of `scale_pos_weight` changed from 1, to 0.95 in the final model.

9.2.3 Trees

There was a total of 557 trees developed in the SINEPOST model. Figure 19 is an example of one of the trees in the final model. Despite decision trees not being as explanatory as logistic regression, it is technically possible to print every tree and go through each one to calculate a predicted probability. The tree in figure 19 was the fourth tree to be developed. It was chosen as it had a good mix of continuous and categorical information. In the tree, the cover refers to the sum of second order gradients (Hessian) classified to the leaf. The deeper the tree node, the lower will be the metric. The second order is calculated by taking the number of observations and multiplying it by the probability multiplied by one minus the probability or $n \times (p = 1) \times (1 - (p = 1))$. Gain refers to the amount of information gained at the split. The value in the leaf is summed with all the values of the other trees to derive the predicted probability. For example, using the tree in figure 19, if a patient had a clinical impression of psychiatric problems, was not monitored by electrocardiogram and had a NEWS score of 1, the value is 0.0243. This is a positive integer, and so the tree is adding to the probability of an avoidable ambulance conveyance. Conversely, if the patient did not have a

psychiatric problem, had oxygen saturations less than 95.5%, and a respiratory rate of more than 20.5 breaths per minute, their value is -0.1. This is moving the probability in a negative direction.

Figure 19: The fourth tree in the full XGBoost model



9.3 Cluster results

In appendix J, there are the ROC curves and calibration curves for each cluster. Figure 20 shows the ROC curves grouped together by geographical region.

9.3.1 The Airedale model

Airedale NHS Foundation Trust is a 350-bed hospital that serves approximately 200,000 people in a 700 square mile radius. It is situated on the border of

Yorkshire and Lancashire, north-west of Bradford. Its ED treats 55,000 patients per annum.²¹¹ There were 3298 patients transported to Airedale hospital with 240 (7.3%) avoidable conveyances. This gave a training sample of 98,244 with 6988 events. The hyperparameters that were selected using the restricted grid search were: $\eta = 0.06$, $\gamma = 0.5$, $\alpha = 0.6$, $\text{max_depth} = 4$, $\text{min_child_weight} = 4$, $\text{subsample} = 0.9$, $\text{colsample_bytree} = 0.6$, $\text{scale_pos_weight} = 1$. The optimum number of trees was 408. The model calibration had a Spiegelhalter's Z-test statistic of 0.733 (p-value 0.463). The ratio of observed vs expected (O:E) was 0.95 (95% CI 0.84 – 1.07). The optimum cut point was 0.116, which gave a C-statistic of 0.776 (95% CI 0.747–0.805). Model accuracy was 0.8 (95% CI 0.79 – 0.82), with a sensitivity of 0.55 and specificity of 0.83.

9.3.2 The Barnsley model

Barnsley NHS Foundation Trust is a district general hospital in South Yorkshire, situated between two large cities: it lies north of Sheffield and south of Leeds. It sees 84,000 patients in its ED every year.²¹² There were 6133 patients transported to Barnsley hospital with 323 (5.3%) defined as avoidable conveyances. Selected model hyperparameters were $\eta = 0.06$, $\gamma = 1$, $\alpha = 0.7$, $\text{max_depth} = 3$, $\text{min_child_weight} = 4$, $\text{subsample} = 1$, $\text{colsample_bytree} = 0.6$, $\text{scale_pos_weight} = 0.67$. The optimum number of trees was 630. Spiegelhalter's Z-test statistic was 0.014 (p-value = 0.989). The O:E ratio was 1.04 (95% CI 0.93 – 1.15), and there was slight over triage at higher predicted probabilities. The optimum cut point was 0.094 which gave a C-statistic of 0.838 (95%CI 0.817 – 0.86). Model accuracy was 0.86 (95% CI 0.85 – 0.87), sensitivity was 0.62 and specificity 0.88.

9.3.3 The Bradford model

Bradford Royal Infirmary is a large trust which sees nearly 400 daily attendances.²¹³ It is a large city, north of Huddersfield and west of Leeds. There were 7709 patients transported to Bradford Royal Infirmary ED with 1004 (13%) avoidable conveyances. Selected hyperparameters were $\eta = 0.06$, $\gamma = 0.5$, $\alpha = 0.6$,

max_depth = 3, min_child_weight = 2, subsample = 0.7, colsample_bytree = 0.9, scale_pos_weight = 2.1. The optimum number of trees was 395. Spiegelhalter's Z-test statistic was -0.241 (p-value 0.81). The O:E ratio was 0.91 (95% CI 0.86 – 0.96). There was slight under triage at higher probabilities. Optimum cut point was 0.27, giving a C-statistic of 0.756 (95% CI 0.74 – 0.772). Accuracy was 0.81 (95% CI 0.8-0.82), sensitivity was 0.45, specificity was 0.86.

9.3.4 The Calderdale model

The Calderdale Royal Hospital is situated in Halifax, which is a large town between Huddersfield and Bradford. The hospital is joined with Huddersfield Royal Infirmary to form the Calderdale and Huddersfield NHS Trust. Between them they see around 125,000 attendances in ED every year.²¹⁴ There were 4107 patients transported to Calderdale Royal ED with 242 (5.9%) avoidable conveyances. Selected hyperparameters were $\eta = 0.06$, $\gamma = 0.5$, $\alpha = 0.8$, max_depth = 3, min_child_weight = 2, subsample = 0.9, colsample_bytree = 0.6, scale_pos_weight = 0.68. The optimum number of trees was 477. Spiegelhalter's Z-test statistic was -0.181 (p value = 0.856). The O:E ratio was 1 (95% CI 0.88 – 1.12). The optimum cut point was 0.127, C-statistic 0.822 (95% CI 0.796 – 0.848). Accuracy was 0.87 (95% CI 0.86 – 0.88), sensitivity was 0.53, specificity was 0.89.

9.3.5 The Dewsbury model

Dewsbury and District hospital couples with the larger Pinderfields Hospital in Wakefield to form the Mid Yorkshire NHS Trust. Between them they have 120,000 ED attendances per year.²¹⁵ There were 964 patients transported to Dewsbury and District ED with 137 (14.2%) avoidable conveyances. Selected hyperparameters were $\eta = 0.06$, $\gamma = 1$, $\alpha = 0.6$, max_depth = 3, min_child_weight = 4, subsample = 0.9, colsample_bytree = 0.6, scale_pos_weight = 1.49. The optimum number of trees was 463. Spiegelhalter's Z-test statistic was -0.035 (p value = 0.972) with under triage at later probabilities. The O:E ratio was 0.89 (95% CI 0.75 – 1.02).

The optimum cut point was 0.249, with a C-statistic 0.724. (95% CI 0.682 – 0.765). Accuracy was 0.77 (95% CI 0.75 – 0.8), sensitivity was 0.44, specificity 0.83.

9.3.6 The Doncaster model

Doncaster Royal Infirmary is a large acute hospital with over 500 beds.²¹⁶ It is situated northeast of Sheffield and Rotherham in South Yorkshire. There were 6678 patients transported to Doncaster ED with 420 (6.3%) avoidable conveyances. Selected hyperparameters were $\eta = 0.06$, $\gamma = 1$, $\alpha = 0.7$, $\text{max_depth} = 3$, $\text{min_child_weight} = 2$, $\text{subsample} = 0.9$, $\text{colsample_bytree} = 0.6$, $\text{scale_pos_weight} = 1.01$. The optimum number of trees was 453. Spiegelhalter's Z-test statistic was -0.1 (p-value = 0.921) with near perfect calibration across all predicted probabilities. The O:E ratio was 0.98 (95% CI 0.88 – 1.07). The optimum cut point 0.115, with a C-statistic of 0.802 (95% CI 0.782 – 0.823). Accuracy was 0.84 (95% CI 0.83 – 0.85), sensitivity was 0.52 and specificity was 0.86.

9.3.7 The Harrogate model

Harrogate District Hospital is situated north of Leeds and west of York. It treats around 52,000 patients a year in the ED.²¹⁷ There were 2761 patients transported to Harrogate ED with 163 (5.9%) avoidable conveyances. Selected hyperparameters were $\eta = 0.08$, $\gamma = 0$, $\alpha = 0.7$, $\text{max_depth} = 3$, $\text{min_child_weight} = 4$, $\text{subsample} = 0.9$, $\text{colsample_bytree} = 0.6$, $\text{scale_pos_weight} = 0.85$. The optimum number of trees was 467. Spiegelhalter's Z-test statistic was 0.080 (p-value = 0.936) with over triage at higher predicted probabilities. The O:E ratio was 1.06 (95% CI 0.90 – 1.21). The optimum cut point 0.094, with a C-statistic of 0.855 (95% CI 0.828 – 0.883). Accuracy was 0.86 (95% CI 0.85 – 0.87), sensitivity was 0.6, specificity was 0.88.

9.3.8 The Huddersfield model

Huddersfield Royal Infirmary is situated west of the city of Wakefield. As previously mentioned, it forms part of the Calderdale and Huddersfield NHS

Foundation Trust which sees around 125,000 ED attendances between them a year.²¹⁴ There were 4675 patients transported to Huddersfield ED with 283 (6.1%) avoidable conveyances. Selected hyperparameters were $\eta = 0.06$, $\gamma = 0$, $\alpha = 0.7$, $\text{max_depth} = 3$, $\text{min_child_weight} = 2$, $\text{subsample} = 0.9$, $\text{colsample_bytree} = 0.9$, $\text{scale_pos_weight} = 0.83$. The optimum number of trees was 516. Spiegelhalter's Z-test statistic was -0.065 (p-value = 0.948). The O:E ratio was 1 (95% CI 0.89 – 1.12). The optimum cut point 0.103, with a C-statistic 0.816 (95% CI 0.793 – 0.84). Accuracy was 0.83 (95% CI 0.82 – 0.84), sensitivity was 0.58 and specificity was 0.85.

9.3.9 The Hull model

The Hull Royal Hospital is based in the large city of Hull on the East Coast. They have a catchment area with approximately 600,00 people and see around 107,000 patients in the ED each year. It is the Major Trauma Centre for the east of Yorkshire, and forms one of four MTCs in this study.²¹⁸ There were 10,711 patients transported to Hull ED with 612 (5.7%) avoidable conveyances. Selected hyperparameters were $\eta = 0.06$, $\gamma = 1$, $\alpha = 0.7$, $\text{max_depth} = 3$, $\text{min_child_weight} = 4$, $\text{subsample} = 0.9$, $\text{colsample_bytree} = 0.6$, $\text{scale_pos_weight} = 0.9$. The optimum number of trees was 472. Spiegelhalter's Z-test statistic was 0.066 (p-value = 0.947). The O:E ratio was 1 (95% CI 0.92 – 1.08). The optimum cut point 0.106 with a C-statistic 0.805 (95% CI 0.788 – 0.822). Accuracy was 0.85 (95% CI 0.85 – 0.86), sensitivity was 0.53 and specificity was 0.87.

9.3.10 The Middlesbrough model

The James Cook University Hospital in Middlesbrough is situated just north of Yorkshire but serves the north Yorkshire community. It is a large trust and hosts an MTC. There were 804 patients transported to James Cook University ED by Yorkshire Ambulance Service with 55 (6.8%) avoidable conveyances. The figures are considerably lower for this trust than others due to Yorkshire only being a partial catchment area. Selected hyperparameters were $\eta = 0.08$, $\gamma = 0.5$, $\alpha = 0.8$,

max_depth = 4, min_child_weight = 2, subsample = 0.9, colsample_bytree = 0.9, scale_pos_weight = 1.37. The optimum number of trees was 261. Spiegelhalter's Z-test statistic was -0.077 (p-value = 0.939). The O:E ratio was 0.90 (95% CI 0.67 – 1.13). The optimum cut point was 0.126 with a C-statistic 0.76 (95% CI 0.694 – 0.825). Accuracy was 0.83 (95% CI 0.80 – 0.85), sensitivity was 0.53, specificity was 0.85.

9.3.11 The Leeds models

Leeds Teaching Hospitals NHS Trust comprises of two hospitals: Leeds General Infirmary (LGI) and St. James Hospital. Both have an ED, with LGI hosting the MTC for the people of west Yorkshire. They serve a population of around 770,000 people. In the study sample, there were 5102 patients transported to Leeds General Infirmary ED with 263 (5.2%) avoidable conveyances. Selected hyperparameters in the LGI model were $\eta = 0.06$, $\gamma = 0$, $\alpha = 0.7$, max_depth = 3, min_child_weight = 4, subsample = 0.9, colsample_bytree = 0.6, scale_pos_weight = 0.68. The optimum number of trees was 485. The Spiegelhalter's Z-test in the LGI model was 0.060 (p-value = 0.952), and the O:E ratio was 1 (95% CI 0.88 – 1.12). There was slight over triage at higher probabilities. The optimum cut point was 0.099 and the C-statistic was 0.818 (95% CI 0.795 – 0.841). The accuracy was 0.86 (95% CI 0.85 – 0.87), sensitivity was 0.53 and specificity was 0.88.

There were 8902 patients transported to St James Hospital University ED with 824 (9.3%) avoidable conveyances. Selected hyperparameters in the St. James model were $\eta = 0.06$, $\gamma = 0$, $\alpha = 0.7$, max_depth = 4, min_child_weight = 4, subsample = 0.9, colsample_bytree = 0.9, scale_pos_weight = 1.26. The optimum number of trees was 348. Spiegelhalter's Z-test statistic was 0.009 (p-value 0.993) and an O:E ratio of 1 (95% CI 0.93 – 1.06). There was almost perfect calibration across all predicted probabilities. The optimum cut point was 0.168 and a C-statistic of 0.806 (95% CI 0.792 – 0.821). Accuracy was 0.84 (95% CI 0.83 – 0.85), sensitivity was 0.53 and specificity was 0.87.

9.3.12 The Sheffield model

The Northern General Hospital is a large MTC in Sheffield, South Yorkshire. There were 10,722 patients transported to Northern General Hospital ED with 929 (8.7%) avoidable conveyances. Selected hyperparameters were $\eta = 0.06$, $\gamma = 0$, $\alpha = 0.6$, $\text{max_depth} = 3$, $\text{min_child_weight} = 4$, $\text{subsample} = 0.9$, $\text{colsample_bytree} = 0.6$, $\text{scale_pos_weight} = 1.13$. The optimum number of trees was 411. Spiegelhalter's Z-test statistic was 0.115 (p-value = 0.909), and an O:E ratio of 1.06 (95% CI 0.99 – 1.12). There was significant over triage at higher predicted probabilities. The optimum cut was 0.157. C-statistic 0.858 (95% CI 0.847 – 0.869). Accuracy was 0.87 (95% CI 0.86 – 0.87), sensitivity was 0.60, specificity was 0.89.

9.3.13 The Wakefield model

Pinderfields is a large acute hospital based in Wakefield, West Yorkshire and is the larger part of the Mid Yorkshire Hospitals NHS Trust. There were 10,245 patients transported to Pinderfields ED with 764 (7.5%) avoidable conveyances. Selected hyperparameters were $\eta = 0.06$, $\gamma = 1$, $\alpha = 0.7$, $\text{max_depth} = 4$, $\text{min_child_weight} = 2$, $\text{subsample} = 0.7$, $\text{colsample_bytree} = 0.6$, $\text{scale_pos_weight} = 1.06$. The optimum number of trees was 330. Spiegelhalter's Z-test statistic was -0.293 (p-value = 0.770) and O:E ratio was 1 (95% CI 0.93 – 1.07). There was almost perfect calibration across all probabilities (slope 1.021, intercept 0.039). The optimum cut point was 0.128 and the C-statistic 0.81 (95% CI 0.795 – 0.826). Accuracy was 0.83 (95% CI 0.82 – 0.85), sensitivity was 0.57 and specificity was .85.

9.3.14 The Rotherham model

Rotherham NHS Foundation Trust serves a catchment area of around 265,000 people in a large town of Rotherham, which is northeast of Sheffield. It has over 370 beds and sees 75,000 patients attend the ED each year. There were 5970 patients transported to Rotherham ED with 120 (2%) avoidable conveyances. Selected hyperparameters were $\eta = 0.06$, $\gamma = 0.5$, $\alpha = 0.8$, $\text{max_depth} = 3$, $\text{min_child_weight} = 4$, $\text{subsample} = 0.9$, $\text{colsample_bytree} = 0.6$, scale_pos_weight

= 0.81. The optimum number of trees was 524. Spiegelhalter's Z-test statistic was -0.023 (p-value = 0.982) with an O:E ratio of 1.06 (95% CI 0.96 – 1.17). There was over triage at higher predicted probabilities. The optimum cut point was 0.1 and a C-statistic of 0.85 (95% CI 0.831 – 0.868). Accuracy was 0.85 (95% CI 0.85 – 0.86), sensitivity was 0.6, specificity 0.87.

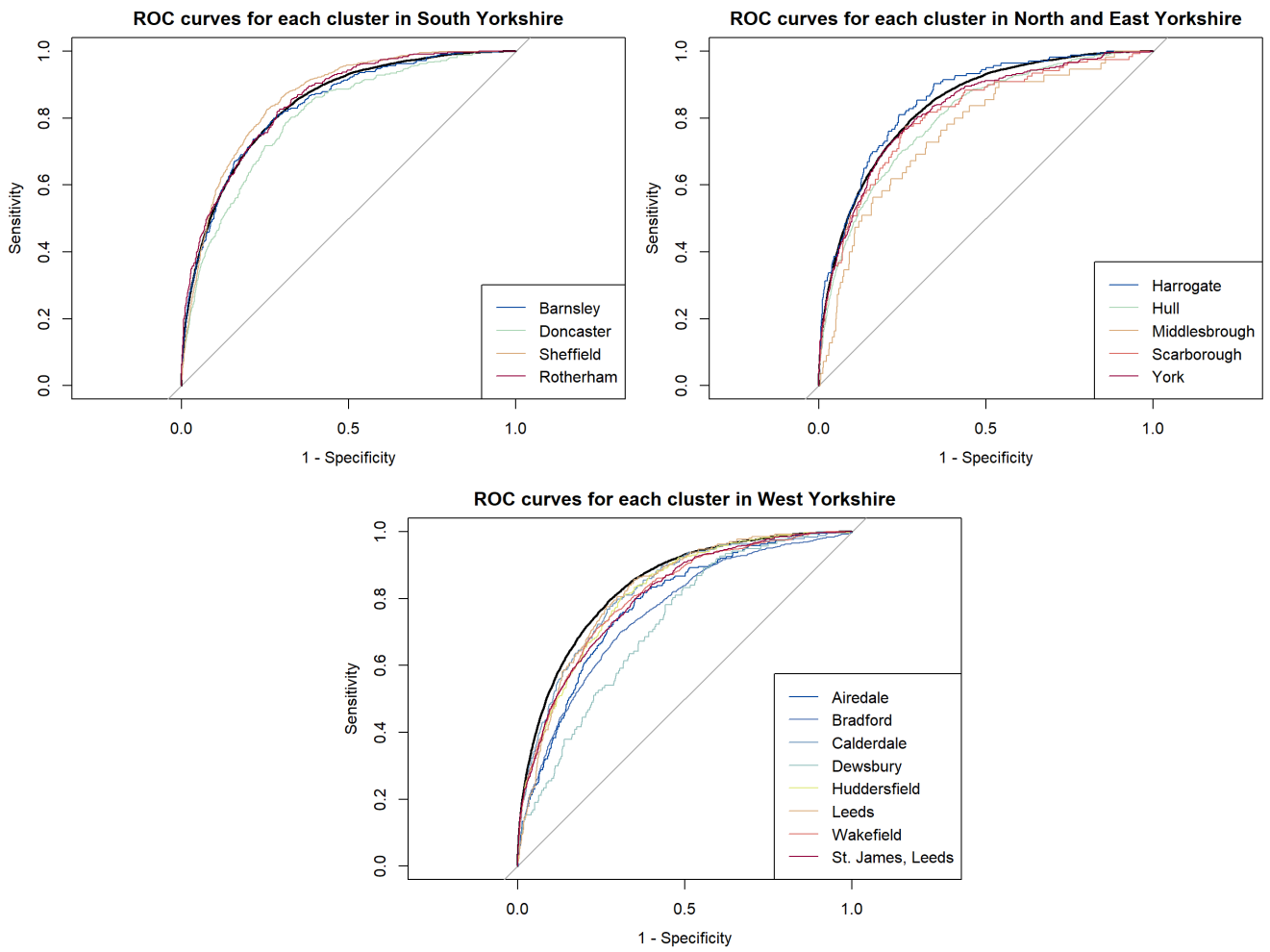
9.3.15 The Scarborough model

Scarborough hospital is the second largest hospital in the York and Scarborough Teaching Hospitals NHS Foundation Trust. It is situated on the East Coast and serves the population of northeast Yorkshire. There were 4494 patients transported to Scarborough ED with 120 (2.7%) avoidable conveyances. Selected hyperparameters were $\eta = 0.06$, $\gamma = 0.5$, $\alpha = 0.7$, $\text{max_depth} = 3$, $\text{min_child_weight} = 4$, $\text{subsample} = 0.9$, $\text{colsample_bytree} = 0.9$, $\text{scale_pos_weight} = 0.41$. The optimum number of trees was 577. Spiegelhalter's Z-test statistic was -0.058 (p-value 0.954) and the O:E ratio was 1 (95% CI 0.83 – 1.19). There was slight over triage at higher probabilities. The optimum cut point was 0.058 and the C-statistic was 0.809 (95% CI 0.768 – 0.851). Accuracy was 0.9 (95% CI 0.89 – 0.91), sensitivity was 0.48 and specificity was 0.91.

9.3.16 The York model

York is the largest hospital of the York and Scarborough Teaching Hospitals NHS Foundation Trust. It largely serves the population of North Yorkshire. There were 6101 patients transported to York ED with 382 (6.3%) avoidable conveyances. Selected hyperparameters were $\eta = 0.06$, $\gamma = 1$, $\alpha = 0.8$, $\text{max_depth} = 4$, $\text{min_child_weight} = 4$, $\text{subsample} = 0.9$, $\text{colsample_bytree} = 0.9$, $\text{scale_pos_weight} = 0.86$. The optimum number of trees was 367. Spiegelhalter's Z-test was -0.077 (p-value = 0.939) and the O:E ratio was 1.04 (95% CI 0.94 – 1.14). The optimum cut point was 0.105, and the C-statistic was 0.827 (95% CI 0.805 – 0.848). Accuracy was 0.85 (95% CI 0.84 – 0.86), sensitivity was 0.59 and specificity was 0.87.

Figure 20: ROC curves grouped by geographical area



9.4 Fair Machine Learning analysis

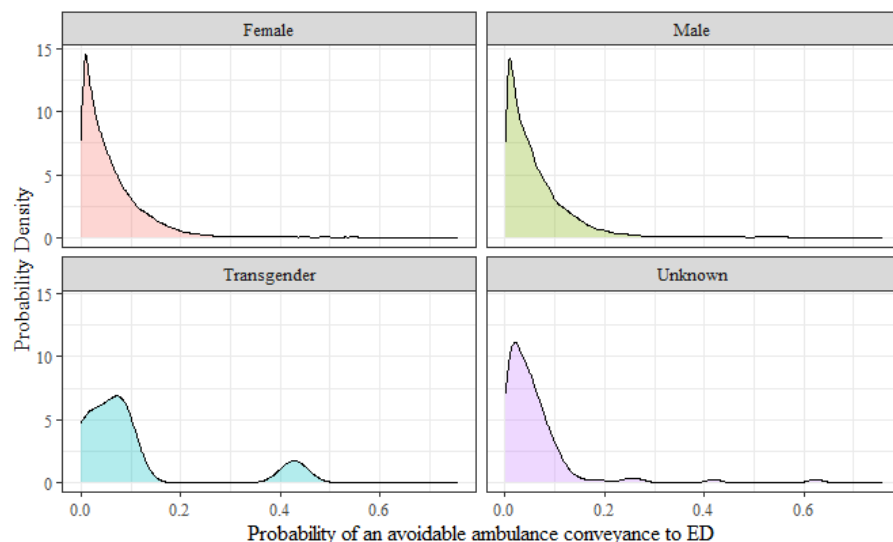
Fair machine learning is ensuring that any decision support (or making) prediction algorithm treats all individuals fairly and is not prejudiced. To evaluate whether the SINEPOST model is fair in a post-analysis, each individual group within a characteristic had their probability density plotted for comparison. If the model is fair, the distributions should look the same. In the analysis of fair machine learning, each category had the probability distributions mapped out along with the mean probability for the group. This was undertaken for age, gender, ethnicity and decile of deprivation.

9.4.1 Gender

Gender did demonstrate a difference in the probability distribution of transgender patients. However, there were only 8 transgender patients (1 classed as an avoidable ambulance conveyance) in the whole dataset, and it is more likely that this is just a result of low sampling as opposed to a bias within the model.

Figure 21: Fair machine learning: Gender

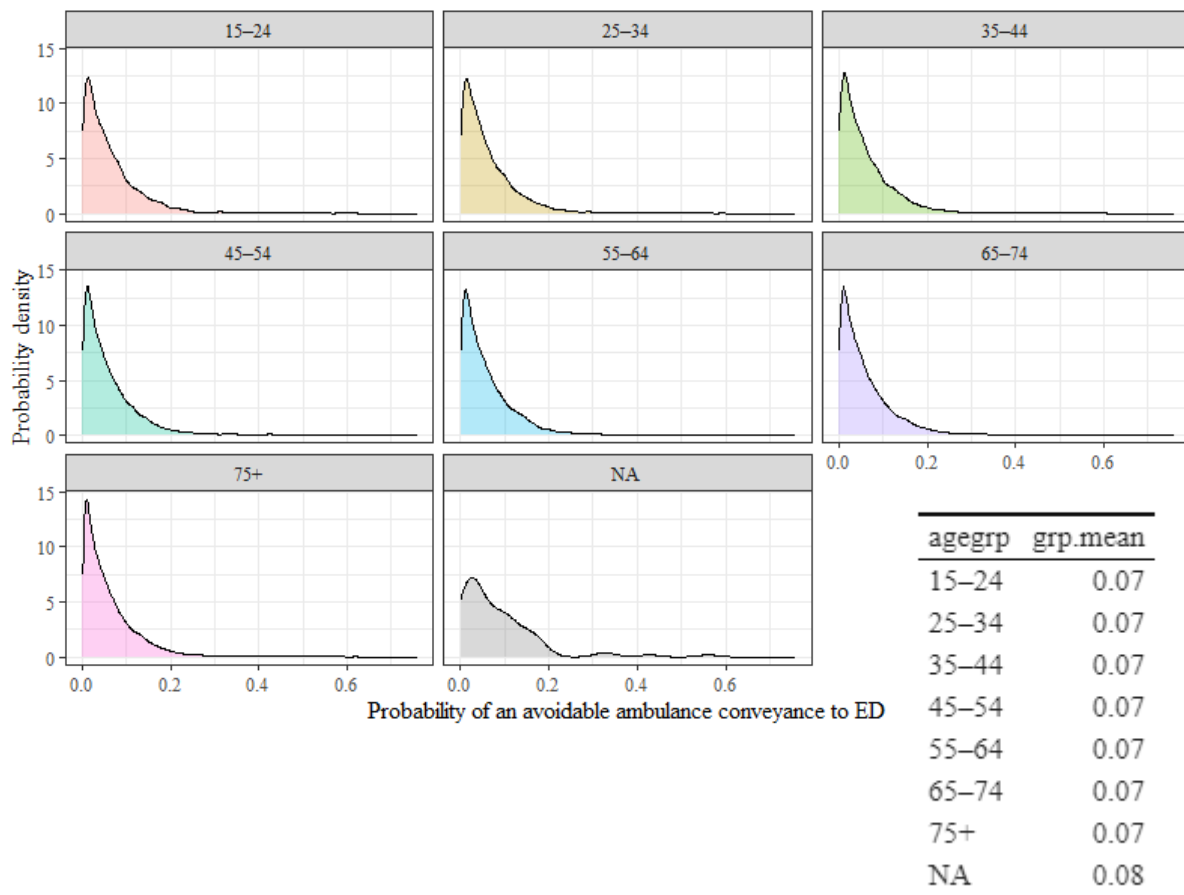
| Gender | grp.mean |
|-------------|----------|
| Female | 0.07 |
| Male | 0.07 |
| Transgender | 0.10 |
| Unknown | 0.06 |



9.4.2 Age

For the purposes of making the analysis interpretable, age was categorised into groups. When age was initially left in as a candidate predictor, the distributions per age category differed significantly from each other and the younger age categories had higher predicted probabilities. This was not the case when age was removed as seen by figure 22. The SINEPOST model does not discriminate based on age.

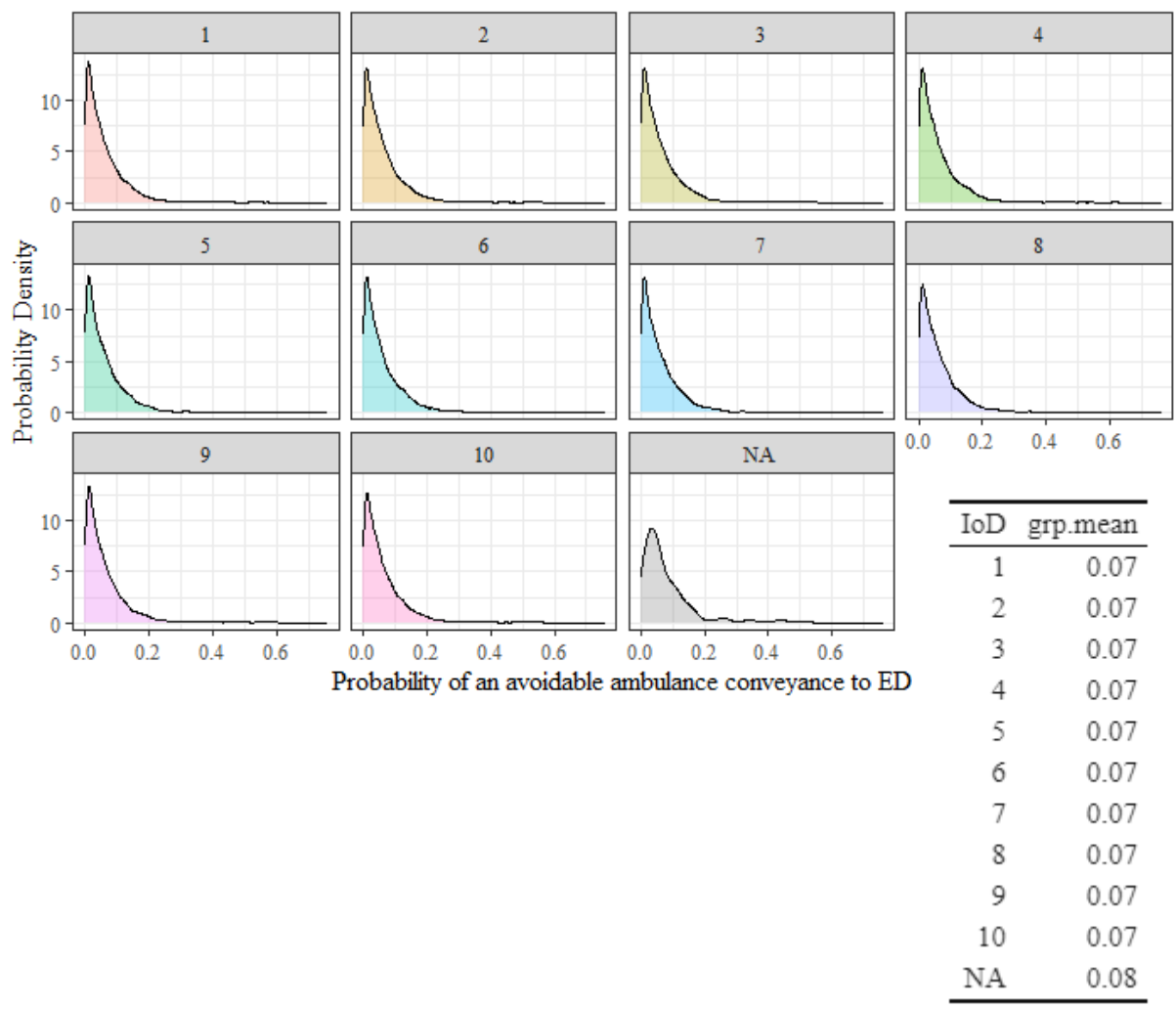
Figure 22: Fair machine learning: Age



9.4.4 Deciles of the Indices of Deprivation

The deciles of deprivation do not show any bias or discrimination between deciles. The probability distributions all appear similar with the only exception being the 'NA' category. Like the transgender category above, the 'NA' only had 180 instances which is small in comparison with the rest.

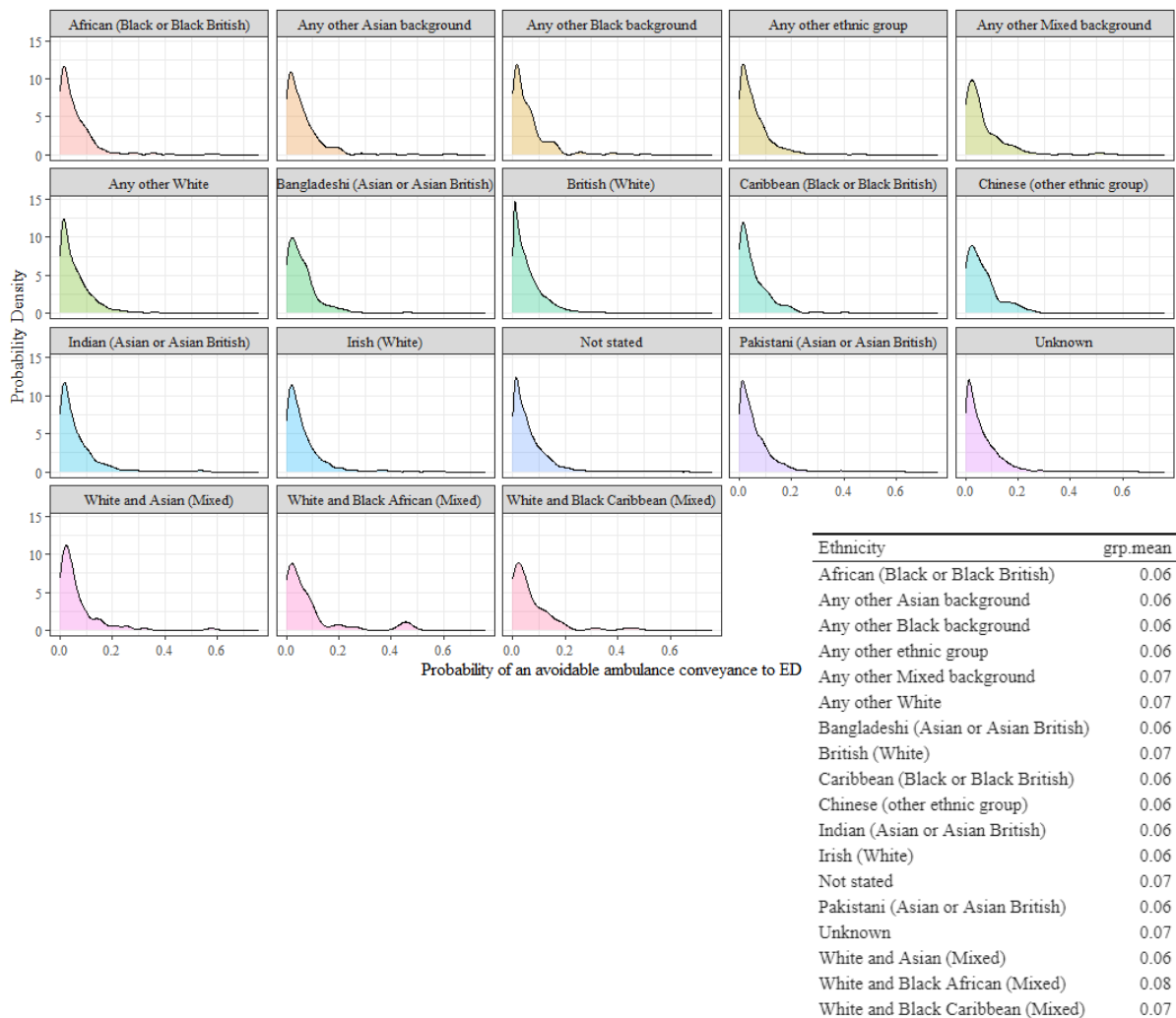
Figure 23: Fair machine learning: Indices of Deprivation



9.4.5 Ethnicity

On initial modelling, the recursive feature elimination removed around two thirds of the ethnic categories. Due to this, it was decided to completely remove ethnicity to ensure the model was fair. On examining the distributions of the full model (figure 24) it appears that this was the right decision as there are no differences in the probabilities per ethnicity.

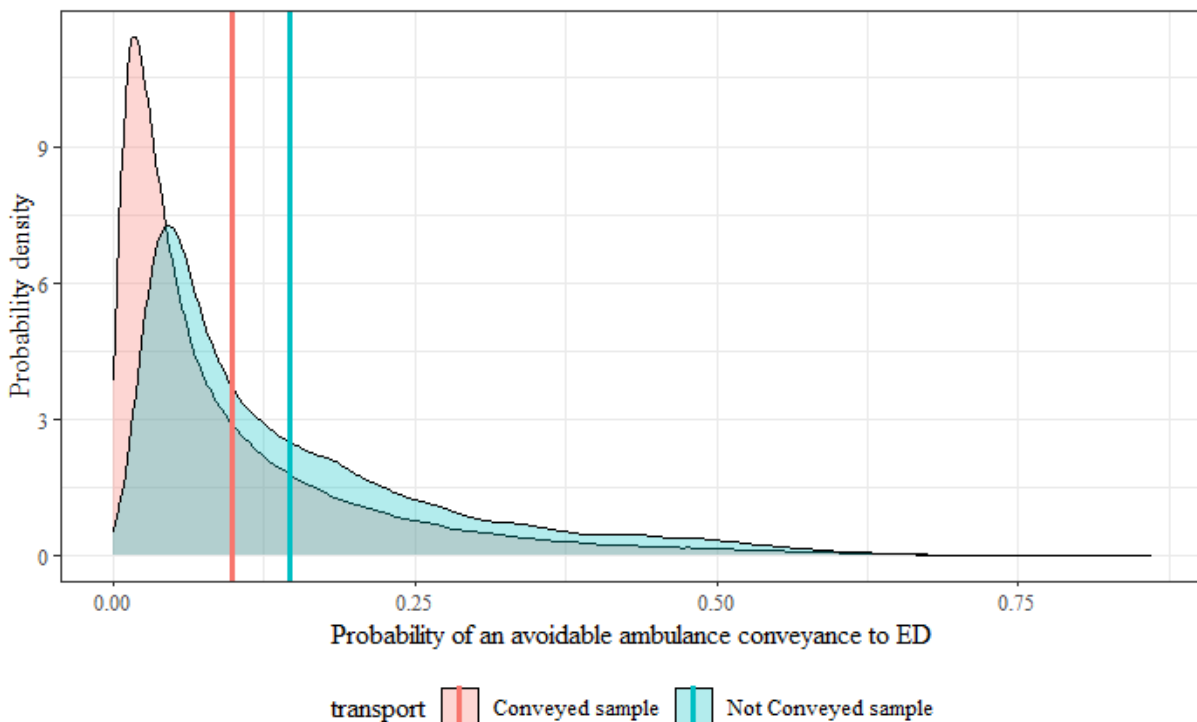
Figure 24: Fair machine learning: Ethnicity



9.5 Evaluation on two different cohorts

The model was applied to a sample of 195078 conveyed but unlinked patients and 112067 non-conveyed patients from the same dataset. These were excluded from the training data as they were unlabelled. The application of the model to these significant sample sizes shows a great overlap, but with the people who were non-conveyed having a different mean than the conveyed (0.15 vs 0.10). The SINEPOST model predicted that there were (45,370 (23%)) avoidable ambulance conveyances in the sample of conveyed patients. When this is multiplied by the false positive rate, only 6805 (3.4%) were likely to be actually avoidable. This model was not predicting non-conveyance, but reassuringly has identified more low-acuity patients who were non-conveyed rather than conveyed.

Figure 25: Probability densities of conveyed vs non conveyed



9.6 Conclusion

This chapter has expanded on the results and discussed the chosen hyperparameters, both for the restricted grid search and for the final model. The results of each cluster were presented, with graphical representation found in appendix J. The fair machine learning analysis was also reported, with further details on what this is in chapter 7, section 7.6.1. The next chapter will discuss the findings and place them in the wider context of clinical and scientific research.

Chapter 10

The Discussion

10.1 Introduction

This chapter elaborates on the principal findings of the study and relates them back to the original research question, aims and objectives. The model developed in this research is known as the SINEPOST model (Safety INDEX of Prehospital On Scene Triage) and is joined with the existing body of evidence on predicting low acuity situations such as an avoidable conveyance to the ED. It is also hypothetically described as a decision support tool, with the potential benefits to clinical decision making and wider impact in the urgent and emergency care system. This study did have limitations, and these are thoroughly discussed with the next steps identified following. Once the chapter has concluded, a personal reflection has been written at the end of the thesis.

10.2 Principal findings of the study

10.2.1 The SINEPOST model can predict an avoidable conveyance to the ED using prehospital patient information

The primary research question in this study was asking whether ambulance service clinical data can predict an avoidable attendance at the ED in adults. To answer this question, a unique and bespoke dataset was created, containing 101,522 ambulance ePCRs collected over an 8-month period linked to ED records. The outcome had a prevalence of 7.12% which is lower than previously reported in the literature. The prediction modelling was assessed in a robust way by deconstructing the results into calibration and discrimination. Calibration was used as assessment of whether the predicted probabilities for each instance in a test set matched with the observed probabilities. The SINEPOST model had good calibration, with a meta-analysed O:E ratio of 0.995 (95%CI 0.97-1.03) and a Spiegelhalter's z-test of -0.031 (p-value 0.975). However, the calibration curve showed some miscalibration at higher predicted probabilities. This is likely to be due to the small number of patients that had high predicted probabilities. This means that even though the calibration curve appears to show miscalibration, it

does not actually have a great impact on prediction. Discrimination is assessing whether the model can take two random instances (one with and without the outcome) and tell them apart. It is measured using the C-statistic whereby a result of 0.5 is no better than chance, and 1 is perfect. The result of the SINEPOST model was 0.81 (95%CI 0.79-0.83). This can be interpreted as good discrimination. The optimum threshold for decision support was statistically calculated as 0.125. Those below this threshold would be appropriate for conveyance to ED, whereas those above could be considered for an alternative care plan. The threshold appears to be low, but again this reflects the distribution of probabilities and the prevalence within the population. Most instances result in a conveyance to ED, and this is why the lower predicted probabilities contain most of the instances. With good calibration and discrimination, it can be concluded that the SINEPOST model can predict an avoidable conveyance to the ED using prehospital clinical information.

10.2.2 The predictions of the SINEPOST model are stable across all different geographies

The second research question was asking what the simulated transportability of the newly developed SINEPOST model. This was tested by using a process known as Internal-External Cross Validation (IECV) and more information on this can be found in chapter 7, section 7.8. The reason it is described as the simulated transportability is that the study did not conduct external validation and so this is the best estimate of true transportability using the available data. It is important to remember that the reason the model was not developed and then applied to each cluster as a test set was because this would be testing the model in the data it was developed on. The question being asked in simulated transportability is not 'does my model work in different geographies', but 'if I did not have my data, but instead had different data, could I still build an accurate model?'. In a sense, the cluster results could be indexed to be SINEPOST1, SINEPOST2, SINEPOST3...SINEPOST17. The aim is to detect any heterogeneity in the different models, but also to meta-analyse the results to update and shrink

any optimism in the original model. The modelling procedures were conducted for 17 clusters with unique test sets for each ED included in the study. A qualitative comparison of the calibration curves demonstrated that the model would act differently in different areas. Some models, such as Calderdale, Doncaster, Huddersfield, Hull, St. James's Hospital (Leeds), and Wakefield had near perfect calibration curves along all predictions. Other models, such as Barnsley, Harrogate, LGI (Leeds), Sheffield, Rotherham, Scarborough, and York acted in the same way as the full model, having higher predicted probabilities than actual probabilities in the tail of the distribution. Airedale, Bradford, Dewsbury, and Middlesbrough had the opposite effect to the full model and would have lower predicted probabilities than actual in the tail. Discrimination varied and was not associated with sample size, indicating that the model would act slightly differently between geographies. For example, Bradford had a C-statistic of 0.76 (95% CI 0.74-0.77), whereas Sheffield had a C-statistic of 0.86 (95%CI 0.84-0.87). Both had large sample sizes, and this can be seen through the small confidence intervals for each. However, despite the small variations, all models had good discrimination and calibration. This means that it is possible to transport this model to different geographies within the catchment of a type 1 ED in Yorkshire. This is an important finding because there are so many different areas in Yorkshire. The county covers 9% of the land mass of England. It has rural areas such as the Yorkshire Dales, and North Yorkshire Moors, as well as large urban areas. The hospitals that receive patients from rural areas to their EDs include Airedale, Harrogate, York, and Middlesbrough. These had avoidable conveyances ranging from 5.9% in Harrogate to 7.3% in Airedale. All rural models performed well, but without any trends in calibration or discrimination which demonstrates that rural patients are not penalised by developing a model to predict an avoidable conveyance. Yorkshire also has coastal areas to the east, with Scarborough and Hull providing EDs to these patients. Scarborough only had 2.7% avoidable conveyances, whereas Hull had 5.7%. It is important to note that Hull serves a larger population and is more urban than Scarborough. There could be a rurality-urbanity interaction, but this study cannot determine this.

There was good calibration and good discrimination for these two hospitals showing the model can be developed in a coastal population. Yorkshire also has urban areas with large inland cities such as Bradford, Leeds, and Sheffield. Urban areas had more avoidable conveyances than rural and coastal with Bradford having 13% avoidable, Leeds 7.76%, and Sheffield 8.7%. The models all performed well in these urban populations. The summary of all different geographies is that there is a difference in prevalence, with avoidable conveyances more common in urban areas. However, each model performed well and there were no trends that were associated with a specific geography other than the urbanity of the population. This means that the answer to the second research question is that the SINEPOST model is spatially validated to be transported across different geographies. A conclusion which provides an advantage over triage tool such as paramedic pathfinder that over triaged in rural areas.⁹⁵ One element of spatial validation the model struggles to answer is why there is variation demonstrated between different sites. This is something which falls outside of the realms of this study as the models are not implemented in this thesis. Future studies should consider using more accurate geographical boundaries and collecting data about the population it serves to try and reduce the unknown reasons associated with the heterogeneity.

10.2.3 The SINEPOST model is fair to age, gender, ethnicity, and deprivation.

Fair machine learning is an important and emerging necessity within the practice of developing decision-support and decision-making models. The limitation with using retrospective data is that any biases that are present in the data collection stage will influence the model. In the initial model it was decided to include protected characteristics such as age, ethnicity, and indices of deprivation in order to account for them in model development. However, recursive feature elimination identified that some ethnicities were more predictive than others and kept them in as candidate predictors whilst removing others. This was also the case with the deciles of deprivation. It is reasonable to think that a model

where not every ethnicity or decile is included as a candidate predictor would lead to introducing a bias. In many other algorithms, a categorical variable would be used in its entirety and not be deconstructed into binary variables for all the categories within. It was decided after recursive feature elimination to remove all ethnicities and social deprivation as candidate predictors, a process known as anti-classification. Age was included in the initial model development, but at the evaluation stage it was the single most important variable. It stood out and had the highest gain, cover, and frequency in the XGBoost model. It improved the accuracy of the model by leaving it in, but it also created a bias. If a new patient was young, the model would steer towards an avoidable ambulance conveyance, and an analysis of misclassification shown in figure 11 illustrates this. The distribution of the false positives matched the true positives peaking at the younger age group, whereas the false negatives were older patients. The trade-off between accuracy of the model, and the bias towards a demographic is a fine balance. When age was removed from the model, the accuracy only dropped by 0.01, but the distribution of age became balanced. Therefore, it can be said that the final SINEPOST model is not biased towards any protected characteristic. The philosophy of anti-classification in the development of the SINEPOST model stems from whether to account, or not, for demographic variables in the model. Binns et al. created the argument of 'minimising harm to the least advantaged' in their stance on fairness.¹⁶³ An example from the criminal justice system is gender in recidivism. Females are much less likely to reoffend than males. But a model that predicts reoffending and ignores gender as a variable is likely to over predict females as reoffending, which then places them at a disadvantage because of the model. There is a clear association between the protected characteristic and the outcome.^{163,219} If their argument is applied here, age had a clear association with the outcome. But here, the risk is needing the ED but being discharged on scene by the ambulance service. A model which included age is likely to proliferate this risk. Furthermore, there is not enough evidence in this study, or the wider literature, to suggest that different ethnicities have barriers accessing an ambulance or ED in the UK. Nor is there evidence, beyond small sample

descriptive statistics, that demonstrate different ethnicities and deciles of deprivation are more or less likely to have a non-urgent attendance at the ED.^{167,168} This means that there is no known discrimination that needs accounting for by including these variables in the model. Anti-classification was an appropriate strategy to mitigate the initial bias, and the results show that irrespective of ethnicity, age, gender or deprivation the model will treat patients the same.

10.2.4 Potential impact of applying the model in practice

Two different systematic reviews concluded that the most effective clinical decision support should be computer-based, providing support as part of the natural workflow, offering practical advice and being available at the time of decision making. Computerised Clinical Decision Support (CCDS) in the prehospital system increasingly plays an important role in delivering efficient care that can meet the needs of its users. In an environment where information is difficult to obtain but decisions are crucial and time limited, CCDS tools appear to offer a potential solution. The implementation of such tools also aligns with current policy in the UK, which is focusing attention on improving information flow throughout the ambulance service. In a Department for Health and Social Care review of operational productivity of ambulance services in England, the first recommendation for future contracting was for ambulance services to have ‘technology, processes and systems in place to support clinical decision making’.⁵⁸ In this section, a distinction is made between the SINEPOST model and the SINEPOST tool. The model has been developed and validated in this thesis, but the tool is a hypothetical implementation of the model in practice. The conceptual idea for implementation would be the model embedded into the ePCR. As the record is filled in, the probability of an avoidable outcome would be calculated and presented to the clinician as a figure on the screen. It is anticipated the tool would show a predicted probability of the outcome as opposed to a classification. This is following the argument in chapter 5, section 5.3.1 that concluded it is the probability that a patient belongs to a certain class

being displayed. Further research is needed to implement the model as a CCDS but conceptualising the model as a tool allows a discussion on the potential placement within the context of computerised decision support.

In prehospital care, CCDS has woven itself into every aspect of service delivery. In the emergency call centre, CCDS structures have long been established in the Emergency Operations Centre (EOC). These are the centres that were traditionally the emergency call centres but provide a wider service of care now, including more advanced triage, hear and treat, and access to a directory of services where patients can be referred. CCDS systems in EOC provide support for ambulance prioritisation and initial triage of patients. In the UK, two computerised algorithms that incorporate decision tree structures are used: the Advanced Medical Priority Dispatch System (AMPDS) and NHS Pathways.⁷⁶ When a patient calls for an ambulance, it is one of these algorithms which helps grade their acuity and predict what resources are needed to meet their required level of care. The decision making needs to be accurate as the effect of poor decisions means that high acuity patients are at risk of not receiving appropriate care and, in certain circumstances, of dying. The results in developing the SINEPOST model demonstrate that prehospital variables available at the time the patient calls the EOC could be used to increase accuracy in prediction of patient acuity. A patient's mobility status was one of the most important variables and can be captured prior to a face-to-face assessment. Even for clinical impression there are surrogates for this in both AMPDS and NHS Pathways and therefore there is a future opportunity to explore an up-stream version of the SINEPOST tool.

Unlike the EOC, Computerised decision support is relatively novel to clinicians on scene with a patient. This is owing to the requirement of electronic patient care records. In Yorkshire Ambulance Service, ePCRs were only fully launched in July 2019, and this formed a barrier to data availability. However, evidence is mounting about the benefits of on-scene CCDS, and the results in this study could have the greatest benefit if a prospective tool is used on scene with the patient.

One of the more neoteric advancements of on scene CCDS is predicting end diagnosis to expedite specialist care or to instigate earlier treatment. As an example, The Japanese Urgent Stroke Triage Score using Machine Learning (JUST-ML) predicted a major neurological event such as a large vessel occlusion, subarachnoid haemorrhage, intracranial haemorrhage or cerebral infarction better than any other available model.²²⁰ The benefit of predicting a major neurological event in the pre-hospital phase of care is that it can steer transport destination decisions to ensure the right patients go to a stroke unit for specialist care. Predicting a downstream outcome has been seen in many clinical conditions including Acute Coronary Syndrome (ACS) and major trauma.^{221–223} The results of this study cannot extend to predicting an end diagnosis, however, they support the idea of modifying a care plan according to the outcome of a CCDS tool. The model has demonstrated it can predict avoidable ambulance conveyances and contributes evidence that computerised decision support can not only predict a high acuity outcome, but also low.

In chapter 2, section 2.9, the decision making of paramedics was explored, and electronic decision support was examined within the context of prehospital decision making. In the SAFER1 trial, the computerised decision support tool was embedded into the ePCR.²²⁴ However, the application required manual selection and usage of the tool. In the qualitative evaluation, it was found that the paramedics who had access to the tool were twice more likely to refer patients to a falls service than those without. However, the paramedics only applied the tool in 12% of eligible patients. One of the barriers to implementation identified in the qualitative element to the study included the labour involved in accessing and using the tool. This resonates with the work of Kawamoto et al.²²⁵ In their systematic review, they were aiming to identify key features of success in the implementation of clinical decision support systems. The most important feature was automation and ensuring that the effort on the end user was minimised. The reason that machine learning algorithms were considered for developing the

SINEPOST model was their potential accuracy and ability to be embedded in an electronic healthcare system. Whilst the Occam's razor approach of making the model as simple as possible was the intended philosophy of the SINEPOST model, machine learning algorithms can be complicated, if needed, and still provide automated prediction. An implementation of the SINEPOST tool would automatically calculate and remove all barriers of labour for the end user because of its design. A limitation, however, is that, by automating the process, the tool becomes somewhat nebulous in how it made the decision. Most machine learning algorithms are not explanatory, they are only predictive. This is true, to an extent, of the SINEPOST model, and its future development into a tool would be dependent on whether clinicians had an appetite to trust a decision support tool where there is no explanation. It is technically possible to print all 557 trees of the SINEPOST model out, map a patient through each tree and sum the predicted probability that way. To do so, however, would take an inordinate length of time and would be impractical.

10.2.5 The SINEPOST model will improve paramedic decision making on conveying patients to the ED

This thesis had a purpose of developing a model that would support paramedics with a specific decision of whether to transport low acuity patients to the ED or not, to reduce avoidable ambulance conveyances. Decision support systems that are already in place for triaging patients include the paramedic pathfinder and the Manchester Triage System (MTS).^{8,9,95} The outcomes of these tools are different, and it so it would be inappropriate to compare performance between them. The intended use of these tools was to risk stratify patients to support non conveyance decisions.

Of course, it was entirely feasible within this study to take the non-conveyed sample and the conveyed to create a prediction model predicting non-conveyance using just the ambulance data. However, the gold standard used would be paramedic decision making, and therefore the model would only be as good as

what is already out there. This is a limitation in both the paramedic pathfinder and the MTS. The strength in this study was taking information that the ambulance crew would not know and predicting that information for them to use whilst they were on scene. The results of this study have demonstrated that using the prehospital variables, it is entirely possible to predict the experience they may have if they were transported to ED. This brings with it a benefit to paramedic decision making.

The idea that this model will improve paramedic decisions on non-conveyance is hypothetical and a future research question needs to be asked that leads to an implementation study. However, it is conceivable that, by bringing the knowledge of this study and the knowledge of presenting the likelihood of the prospective ED experience to the clinician on scene, a decision would be made with more information.

In the study by Miles et al. I explored paramedic decision making using a mixed methodology⁸⁶. In the qualitative part, it was found that paramedics either framed a decision around the scene, or the ED. When they framed the decision around the scene, their language would often be why it is not safe to be left at home, or that the patient requires a GP appointment (for example). When it was framed around the ED, the justifications would be anchored to the patient either receiving a certain benefit from attending, or that the ED would probably not find anything abnormal.⁸⁶ The findings from this study have the opportunity to support those who use the ED to frame their decisions. By knowing what the predicted probability is, it provides new information to them that would not have been available for decision making. However, perhaps the largest benefit to transport decision making on scene from this study is the revealing of clinically important variables that should be accounted for in making such a decision.

This is one of the first studies ever to use prehospital clinical variables linked to an ED outcome and use this data for predictive analytics in the UK. In chapter 5,

the justification for each candidate variable had to be taken from a surrogate outcome such as admission prediction as this was available in existing literature. The results in this thesis generate new knowledge on which prehospital variables matter the most when predicting an avoidable ambulance conveyance and can contribute to areas of patient improvement, such as safe non-conveyance.

The most important variable in the model was age, and so much so that it had to be removed as it was creating a bias towards younger patients. However, the knowledge that age is so important can be used to target policies and interventions specifically designed for younger aged populations. It comes with a caution and the knowledge identified in this thesis should be used as a warning label for future prediction models in this area.

In the final model, there were 19 variables that could be categorised, 1 social, 1 demographic, 14 clinical and 3 interventional. Of the clinical variables, ten were physiological observations (and the composite score of NEWS).

As mentioned in chapter 5, social variables have rarely been used in clinical prediction modelling, except in private healthcare systems where the insurance cover has been used.¹⁶⁷ The rationale to include data flags for if the patient had a named GP, next of kin, parent, guardian or social worker was because they acted as markers for the social support of the patient. The only variable that remained in the model was next of kin and this was not regarded by the final model as a significant variable according to gain, cover and frequency. It is likely that these social variables do matter, but the way they were presented to the model did not allow for their potential to be realised. It would be more useful to know how often they access these networks, and whether they are actively reaching out to the networks at the time of the incident.

The included demographic variable was not related to a protected characteristic but was the incident location. The variable itself had seven choices (care home,

domestic address, not selected, public place, and school, work, other) but school, work and not selected were eliminated during feature selection. Domestic address was the most important of these variables, followed by care home, other and then public place. Two studies identified an association between nursing home residency and admission to hospital. However, they argued against each other and so no conclusions were drawn.^{109,168} This study has shown that the incident location has a relationship with an avoidable conveyance, but more information is needed to define what this relationship is. Furthermore, the variable importance did not consider incident location to be a high-ranking variable, this may hint at a weaker association. In chapter 5, the only other variable that was mentioned in the demographic category that has not yet been discussed is the previous attendance. In the results of this study, a previous attendance within 24 hours was not associated with an avoidable conveyance to the ED and was removed during recursive feature elimination.

Physiological observations were the most frequently used variables in the final model. The top five frequently used variables were the patient's blood sugar, systolic blood pressure, temperature, diastolic blood pressure and pulse rate. When examining the actual trees, it became apparent that these observations were naturally split along values perceived in clinical practice to be abnormal. For example, the XGBoost algorithm identified a threshold for respiratory rate to be 20.5 breaths per minute. A faster rate placed a negative value on the leaf node, whereas a slower rate had a positive. In clinical practice, it is perceived that a respiratory rate between 16 and 20 is normal. Subsequent observations also appeared more frequently, and this makes logical sense. The model is blinded to the construct of primary and secondary observations, which means that it will treat them equally as predictor variables. A way round this was the engineering of interval variables. The difference between the primary and secondary observations were calculated and used as predictor variables. These intervals feature in the final model, but not as high as the original variables. Primary observations also featured heavily in the top 20 variables according to cover and

gain. This was more so in cover, which means that these variables were used higher up in each tree. This study can conclude that physiological variables are predictive of an avoidable conveyance to the ED.

Other than physiological observations, the other clinical variables included in the final model include clinical impression, the patient's mobility, if they had an underlying oxygen requirement and whether the patient had a catastrophic haemorrhage. The latter variable forms part of the primary survey and is one of the first things that paramedics assess when examining a patient. It is a life-threatening condition, and so it seems logical that it appears in the final model. However, it did not feature highly on the three criteria for variable importance, which means it probably does not contain enough information to be predictive. Similarly, a patient's baseline oxygen requirement was included in the full model, but not in any of the top 20 variables. The two that did though, were clinical impression and mobility.

Clinical impression contains a list of 99 potential illnesses or injuries that a patient might be suffering from and the ambulance clinician on scene can only select one. Of the 99 conditions, 38 made it into the full model which represented 42% of all the SINEPOST model variables. Clinical impression was heavily associated with the outcome, and this is a significant finding of the study. In the top 20 variables according to gain, psychiatric problems were the second most important, followed by allergic reaction/ rash in fifth, cardiac chest pain in seventh, and back pain (non-traumatic) in eighteenth. These four also appeared in the top 20 for cover, and frequency, so it can now be said that they associated with an avoidable conveyance to the ED.

Mobility was also a variable that featured highly in all aspects of variable importance. Initially, all mobility options were included in the model, but after the initial modelling, self-mobile and stretcher were the highest important variables even though they are mutually exclusive. It was decided that only one

mobility category is needed to reduce model noise, and so all mobility categories were removed except for self-mobile. A patient being self-mobile and not requiring any assistance was the single most important variable in the SINEPOST model. This variable was the highest-ranking variable in the full model and had the second highest cover. However, it ranked 11th for frequency, but was the first categorical variable in the frequency rankings. With continuous variables (such as pulse rate etc.) they can be used more than once per tree, which is a likely contributor to all the physiological observations appearing as high frequency variables.

Interventional variables that remained in the full model include IV cannulation, whether a drug was given and if a patient was monitored by electrocardiogram. Not all drugs were included in the final model, and only one drug appeared in all three measures of variable importance and that was oxygen. Other drugs included in the final model were aspirin, glyceryl trinitrate (GTN), morphine, Entonox, chlorphenamine, adrenaline, activated charcoal, ondansetron, and salbutamol.

All three interventional variables indicate that a patient is of higher acuity, and this can be typified by the example tree in figure 23, when a patient was monitored by electrocardiogram, it resulted in a negative leaf value. The drugs in the included model are often given in high acuity situations. For example, morphine is given for severe pain, and can be given alongside aspirin and GTN for suspected cardiac chest pain.

The variable importance has generated significant new knowledge and can help future research, and future clinical decisions become more discerning even if the prediction model derived in this thesis does not translate into practice. Self-mobility and clinical impression had the greatest impact, and perhaps future policy in prehospital care should explore the self-mobile cohort in more detail to

determine whether there is a subgroup that may not need to be transported by ambulance at all. This would help with the wider topic of non-conveyance.

10.2.6 The SINEPOST model will impact the wider Urgent and Emergency Care System

Using the predictive model developed in this study and applying it to a large, conveyed cohort from the same time-period and geography, it was found that 45,370 (23%) cases of conveyance were potentially avoidable. The model was validated across geographies and any implementation could be reasonably done on a national level. If this figure was applied to national level data in England, the predictive model could support 85,560 conveyance decisions per month to change to non-conveyance. This is based on the latest NHS England Ambulance Quality Indicators which identified 372,002 ambulance transports to the ED in November 2021.¹⁶ First and foremost, this impacts the patients who avoided being transported to a higher level of care that was not needed. It also benefits patients waiting to be seen by an ambulance, as the crew can be dispatched as soon as they have finished the current patient episode. In theory, this will have an effect downstream and reduce the compounding effect on ED crowding by minimising ambulance queuing to just those who need to be there. Upstream it can also have an effect as theoretically, there would be increased fleet availability to respond to the next emergency.

10.3 Limitations

10.3.1 Data availability

This study was one of the first studies to link prehospital clinical data to corresponding ED data. Previous studies have managed to link the Computer Aided Dispatch (CAD) data, which is collected in the Emergency Operations Centre (EOC) and contains information such as the predicted triage, and demographic information. It also contains response times and journey times for the attending resources. The CAD data has been used successfully in the past for

reforming how ambulances respond to emergencies. The Ambulance Response Program (ARP) identified that the response time targets in national policy did not reflect the case mix of the modern ambulance service. Through a robust program of research, they were able to change the national policy and improve clinical care.¹⁵ The limitation with CAD data is that there is very little on-scene information. The only source of prehospital clinical information comes from the electronic patient care record. A strength of the study was being able to access this data, but a weakness was this data was bound by its implementation date only being July 2019. Being so juvenile in its use is a limitation to this study as it is less understood how it is used. For example, the ethnicity of a patient may be assumed by the clinician instead of through a discussion. The clinical impression code is represented as 99 different options, which is arguably too few to accurately represent the incoming case-mix of ambulance patients. As a result, clinicians may choose the ‘nearest neighbour’ to their actual clinical impression. Furthermore, the context of the clinical impression could be a limitation to the study. Labelling a condition provides a framework for the subsequent clinical care of a patient and can be helpful to both the clinician and the patient.²²⁶ For example, there are certain conditions such as cancer, which have specific diagnostic markers, signs and symptoms. However, when signs and symptoms can only be described by the patient, an accurate clinical label becomes more challenging and there may be harms associated with doing so. For example, if someone is given a mental health clinical label, there is a risk of harm when the patient takes the label into the wider context of their life, such as their home or workplace, and the label results in stigma or discrimination.²²⁶ Premature condition labelling could potentially affect prehospital and emergency care more than a normal clinical environment as the condition is not completely formed due to the unscheduled nature of access. There is an argument that less emphasis should be placed on forming a clinical label, and more on the probability of future events happening given a current clinical state.²²⁷ This issue relates to this thesis as the accuracy of the data may not reflect the actual care need of the patient and may not accurately describe whether the care need was met. It assumes all

clinical impressions are exact states and not temporary labels almost used in the context of a placeholder. This limitation should temper the importance of clinical impressions as candidate variables in the final model.

A more important weakness is the data availability of ED data. Despite this being collected by NHS digital nationally, the quality of the data omits any granular level detail. For example, it is possible to see how many patients had a blood test, even a specific one like troponin levels in the blood, but not the actual test result. The data is collected primarily for the purpose of commissioning health care services and therefore the result of a test is not as concerning as the cost. This study also used retrospective data to achieve the necessary sample size but using already collected data has the limitation in that it cannot be changed. It is whatever it is, and so if the quality is not there it is difficult to improve.

10.3.2 Outcome measure used

In chapter 7, section 7.5.4 there was a discussion surrounding the limitations of using the data-driven outcome measure. This section expands on this limitation. The data derived definition was generated using values found within certain variables. Due to NHS Digital collecting the ED data in different coding languages, it meant the outcome measure had to be translated into two coded definitions. This is a limitation of the study as, ideally, every patient would be measured with an identical outcome measure. The outcome being data driven also meant it was a conservative description of the patient episode. Patients excluded from the definition may not have needed the skills and expertise of the ED but may have received them anyway by virtue of being there. For example, if a patient presents with minor signs of anaemia (such as fatigue) that they have had for more than a month, they may have a blood test in the ED which would show low levels of haemoglobin, or serum iron. This is a primary care condition but because the patient had a blood test, they were considered a necessary attendance at the ED by the outcome definition. Ideally, there would be variable completed

after the patient episode that describes what clinical level the care needed to be. This would be inputted by the clinician responsible for their care.

This idea that patients should be classified according to their care need as opposed to their acuity is one which should transcend into future research. Currently, acuity is triaged at multiple steps in the urgent and emergency care system and has the single benefit of ensuring high acuity patients have an expedited pathway of care. Labelling mid- and low-acuity patients yields little benefit, and this study demonstrates this: it is not informative enough. The SINEPOST model can predict those who may have an avoidable attendance at ED, but it would be better if it could predict the most appropriate care setting required. Instead of a metaphorical signpost, it would be a compass. Able to separate prehospital patients into distinct and mutually exclusive groups that can be used to develop a multi-class prediction model.

10.3.3 Concept drift

It is already likely that the SINEPOST model is degrading in validity and, even more so, the SINEPOST tool, which has not even been conceived yet. This is because of concept drift, which is the notion that the input data and output data change over time and this causes prediction models to degrade.^{228,229} In the SINEPOST model, it was established above that the outcome measure was not ideal, but was the best available. In the future, the outcome measure may change in definition and even in context. Furthermore, the ambulance services are only at the beginning of their use of electronic patient care records, and these are constantly updating and changing. Variables could be defined or collected differently, even omitted from the record.

10.3.4 The empiricist approach

As outlined in chapter 5, section 5.2, this thesis adopted an empiricist approach to answering the research question. The argument for doing so is outlined in that

section, but the limitations of doing such are expanded here. The true reality of ambulance decision making on whether they need to transport a patient to hospital or not, is one of the most complicated to make.^{6,76} Paramedics felt a fear of litigation and lack of managerial support if a decision was wrong, and therefore had a preference to over-convey. There are also differences between how clinicians make decisions, and some will look at the patient in the context of their wider health, whereas others will focus on the specific complaint.⁶ This provides evidence for the complexity of decision making and highlights that there would be merit in adopting a rationalist approach to studies of this type. By using the empiricist approach, the subjective nature of decision making has been ignored, and so the research cannot reasonably be extended into a tool. Had a rationalist approach been undertaken, the qualitative aspects could have been accounted for and the SINEPOST model could have got to the stage of a SINEPOST tool. This was initially proposed when submitting this study to the funding body, however it was recommended to remove this.

10.4 Future research

The immediate next steps would be to mobilise the knowledge from this study into practice. This would be through dissemination events and meetings with key stakeholders. The knowledge that can be easily translated is that an avoidable conveyance is predictable on scene, and that clinical impression, physiological observations and a patient's mobility are all important for this.

An implementation study and strategy would need to be developed to test the SINEPOST model as a tool embedded into the ePCR. This could be done using multiple trusts as the model was internally-externally validated, with no signs that it cannot be transported across ambulance services. Strict external validation may need to be undertaken in a different region to be thoroughly sure that it is transportable across different regions. An interesting external validation would be in another ambulance service healthcare setting, such as Australia or the USA. It is important to reiterate that the ideal SINEPOST tool

would be used as decision support (as opposed to decision making) as part of an arsenal of tools that will help ambulance clinicians make discernible transport decisions in the future.

Future studies aiming to improve patient care by getting ambulance service patients to the right place, first time, should consider how to differentiate patients according to the care setting for their need. By accurately defining the cohort, clinical prediction models with greater utility can be developed.

10.5 Conclusion

This section has placed the findings within the wider context of the urgent and emergency care system. There has been a critical argument on how the model being transformed into a tool could lead to significant benefits for both patients and the healthcare services. The model's ability to support decision making has been explored, as has the political implications of upscaling the model to a national level. In the next chapter, there is a personal reflection from the author around undertaking this study.

Chapter 11

The Personal Reflection

11.1 Personal reflection

There have been many victories and challenges in undertaking this study, and this reflection aims to give my subjective thoughts on what they have been. It also provides an insight on how I felt undertaking the PhD. There were more elements to the project than just the research study. More words written, documents created, approvals requested, and contracts drafted than made it into this thesis. Managing a research project brings its challenges, and I learnt from the experience of managing such a project what those challenges are. Obtaining a linked dataset is difficult and became the biggest threat to the study. I underestimated the time and labour required from starting the project to receiving the data in the analysis format and it took around 30 months for this to happen. This was despite the help of the data experts at YAS and the data architects at the University of Sheffield all helping to make the process as expeditious as possible.

I think the biggest personal trial I faced in undertaking this study was finding the right balance between clinical and academic practice. The lexicons, methodologies and skills have rare overlap, which made it quite challenging to balance the two. In addition, there were times when the study could be likened to taking a stroll in the fog. My underpinning knowledge and identity is that of a paramedic and not of a statistician or computer scientist. To play in these arenas felt like an imposter at times. One example of this was deciding if logistic regression was machine learning or not. It appeared to be a vocal debate between statisticians and computer scientists, and I felt by drawing my own conclusion to the debate it would validate my existence being in their spaces. I spent a long time trying to draw a conclusion one way or the other almost like Odysseus trying to navigate the two sea monsters Scylla and Charybdis. In the end, I decided the debate was not helpful to me at all, but a distraction. I taught myself to code in R, revisited my knowledge in statistics and undertook multiple courses in both machine learning and predictive modelling. I felt undertaking the PhD broadened my knowledge base, and I now feel I boundary span them all. I believe

the future of clinical research in prehospital care will feature digital systems and I would like to be competent to lead such research and remain spanned across these clinical academic boundaries. My current thoughts are the art of clinical academia is to feel comfortable being an expert in neither as the peak lies somewhere between the two.

The journey of completing this study has perhaps changed my own beliefs on how we should categorise patients and the labels that are commonly used. At the beginning, it felt logical to categorise patients according to their acuity. This was the traditional model, and the accepted paradigm of classifying prehospital patients. From the moment someone calls for an ambulance, their acuity is triaged. But as the study developed and after building the SINEPOST model, I felt that this classification system lacked clear boundaries when it came to on scene decision making. All high acuity patients largely end up in the same place. Some specific conditions might lend themselves to a specialist centre such as strokes or heart attacks, but in general, high acuity patients get transported to the ED. These patients may have abnormal physiological observations, or a specific sign or symptom like breathlessness or chest pain that makes them easier to classify as high acuity. But the other acuity categories I think contain more diversity and have more complex care needs within their case mix. They may not have abnormal observations, or a clear sign or symptom. This makes describing them according to their acuity unhelpful, and harder to identify the right care need. I think future clinical practice and research needs to navigate a way of describing patients in a way that creates meaningful groups.

Chapter 12

The Conclusion

12.1 Introduction

The flow of patients through the urgent and emergency care system is dependent on ensuring that they get to the right level of care, first time. Chapter 2, section 2.6 outlined that currently this is not the case and there are a cohort of patients that are transported to the ED by ambulance that do not need the clinical expertise of emergency medicine.

Transporting these patients contributes to a demand-induced phenomenon known as offload delay at the handover gate, and ambulance ramping (see chapter 2, section 2.5). These situations have an increased risk of causing harm to current patients in the system, but also to prospective emergency care and ambulance patients.

If ambulance staff (i.e., on-scene decision makers such as paramedics) were able to identify avoidable conveyances whilst on scene, it would have an impact on these problems. This is exactly what this study set out to achieve. It aimed to build a prediction model using prehospital data and an ED outcome (for more information on the methods, please see chapters 6 and 7). If successful, it would provide information on scene that otherwise would not be available knowledge to clinicians making transport decisions. The research questions asked in this thesis (as identified in chapter 4) were the following:

In adult patients attending the ED by ambulance, can prehospital information predict an avoidable attendance?

Can the model derived from the primary outcome be spatially transported?

In this chapter, section 12.2 summarises the findings from the thesis and demonstrates how the research questions have been answered. In section 12.3, the contributions of new knowledge to clinical practice and to the scientific

community are outlined, with recommendations to policy and future research described in section 12.4 and 12.5. The chapter concludes in section 12.6.

12.2 Summary of findings

A linked dataset was created especially for this thesis. It consisted of 101,522 patient episodes that started in the ambulance service and ended at the ED. In this sample, 7228 (7.12%) had an avoidable conveyance to the ED according to the definition used. The candidate (predictor) variables were all in the ambulance service data, whilst the outcome variable was from ED data. This meant that the prediction model would be telling paramedics new information on scene that is not currently available. Machine learning methods were explored for their feasibility of building a predictive model, with a gradient boosted decision tree known as an XGBoost algorithm being the chosen method. Through a process of tuning hyperparameters and utilising novel prediction methodology such as internal-external cross validation and meta-analysing the results, it was indeed possible to build a model that successfully predicted an avoidable conveyance to the ED. The final model had nineteen variables, fourteen of which were clinical, three were interventional and two were demographic. The model demonstrated success in the form of its calibration, discrimination, and accuracy statistics. For calibration the Spiegelhalter's z-test was mainly used, with the O:E for the meta-analysis. For more information on these tests, please see chapter 7, section 7.9.1. The final model had a Spiegelhalter's z-test of 0.111 ($p=0.912$), and a meta-analysed O:E ratio of 0.995 (95% CI 0.97-1.03). For discrimination, the C-statistic was used and more information on this can be found in chapter 7, section 7.9.2. This was meta-analysed as 0.81 (95% CI 0.79 – 0.83). A fair machine learning analysis was performed to ensure the model would not discriminate against any protected characteristic. It was found the model was fair to all age, ethnicity, gender and indices of multiple deprivation.

12.3 Contributions to knowledge

The novel contributions in this thesis can be broadly categorised into clinical knowledge and scientific knowledge. Clinically, this thesis has produced new knowledge that it is possible to predict an avoidable conveyance to the ED, whilst the ambulance is still on scene with the patient as evidenced by chapters 8 and 9. The variables that contribute to this prediction have been revealed, and by knowing this information new clinical pathways can be developed to target interventions at these groups. The important variables include the patient's mobility, their vital signs, and their clinical impression. Certain clinical impressions such as psychiatric problems, and allergic reactions were very important. Self-mobility was a key variable, and this could lead to accounting for a patient's mobility as a consideration for prospective transport decisions. If the model was transformed into a tool, it could potentially prevent 85,560 ambulance transportations a month in England. This is a significant reduction in ambulance conveyances.

From a scientific perspective, this thesis has contributed to an emerging method of deriving and validating a risk prediction model. Ideally, a model will first be derived and internally validated, then externally validated in a separate study.²³⁰ There are challenges to this method, including time and funding. The method used in this thesis uses the same data in a novel way, to simulate the external validation of a model. It is an efficient method, that can be easily visualised. It is also the first study in the context of urgent and emergency care to use an XGBoost algorithm to develop a risk prediction model.

12.4 Recommendations for policy

The main recommendation for policy would be to revise the taxonomy of prehospital patients and move away from ordering them by acuity. This should not negate the need to identify life threatening emergencies, but to identify a more accurate way of determining immediate care need at the point of call. The

SINEPOST model developed here could be seen as the start of the journey, as it is largely possible to elicit the important variables found in the model whilst the call handler is talking to the patient. A different strategy would be to take the important variables and target pathway interventions for these groups. A pathway intervention would be to embed a different solution to managing the care needs of certain patients instead of transporting them to the ED. A good example would be patients with mental health needs. If there was a system of support that could accept emergency referrals, these patients may find themselves resolving their care need sooner.

12.5 Recommendations for future research

Future studies need to focus beyond defining a patient by their acuity and more on the place for their care need. The ambulance service is already transitioning into a treatment service as opposed to just transport, and decision making by clinicians on scene need to reflect this. Studies should now concentrate on redefining the taxonomy of prehospital patients before proceeding to develop new predictive tools.

12.6 Conclusion

This thesis has demonstrated that it is possible to predict an avoidable conveyance to the ED whilst still on scene with paramedics. The methods used to demonstrate this include a novel application of the XGBoost algorithm, which allows the model to be developed under different circumstances. The SINEPOST model is accurate and does work in rural, urban, and coastal areas. However, the idea of classifying patients according to their acuity holds marginal value once the high acuity is classified. Future studies should consider the taxonomy of patients entering the urgent and emergency care system, and shape solutions according to their care need. By doing so, patients will receive timely care, and in the right place.

References

1. World Health Organization. WHO Emergency Care Systems Framework.

- 2015 [cited 2022 Apr 26]; Available from:
<https://www.who.int/publications/i/item/who-emergency-care-system-framework>
2. Copyright, Designs and Patents Act. The Stationary Office; 1988.
 3. Nissen A. A Right to Access to Emergency Health Care: The European Court of Human Rights Pushes the Envelope. *Med Law Rev* [Internet]. 2018 Nov 1 [cited 2022 Feb 2];26(4):693–702. Available from:
<https://academic.oup.com/medlaw/article/26/4/693/4731564>
 4. Anderson PD, Suter RE, Mulligan T, Bodiwala G, Razzak JA, Mock C. World Health Assembly Resolution 60.22 and Its Importance as a Health Care Policy Tool for Improving Emergency Care Access and Availability Globally. *Ann Emerg Med*. 2012;60:35–44.
 5. O’Keeffe C, Mason S, Jacques R, Nicholl J. Characterising non-urgent users of the emergency department (ED): A retrospective analysis of routine ED data. *PLoS One*. 2018;13(2):1–14.
 6. Miles J. Using vignettes to assess the accuracy and rationale of paramedic decisions on conveyance to the emergency department. *Br Paramed J* [Internet]. 2019; Available from:
<http://doi.org/10.29045/14784726.2019.06.4.1.6>
 7. Newton M, Tunn E, Moses I, Ratcliffe D, MacKway-Jones K. Clinical navigation for beginners: The clinical utility and safety of the Paramedic Pathfinder. *Emerg Med J*. 2013;31(e1):e29–34.
 8. North West Ambulance Service. Paramedic Pathfinder and Community Care Pathways. 2014;(September):52. Available from:
<https://www.nwas.nhs.uk/DownloadFile.ashx?id=286&page=16586>
 9. Scholes S. Can Paramedics use the Manchester Triage System to triage and refer patients into clinical pathways of care from scene?
 10. Thind A, Hsia R, Mabweijano J, Hicks ER, Zakariah A, Mock CN. Prehospital and Emergency Care. In: *Disease Control Priorities, Third Edition (Volume 1): Essential Surgery* [Internet]. The World Bank; 2015 [cited 2020 Nov 6]. p. 245–62. Available from: https://elibrary.worldbank.org/doi/abs/10.1596/978-1-4648-0346-8_ch14
 11. NHS England. Transforming urgent and emergency care services in England Urgent and Emergency Care Review End of Phase 1 Report Appendix 1 – Revised Evidence Base from the Urgent and Emergency Care Review High quality care for all , now and for future generations. *NHS Engl*. 2013;1–87.
 12. NHS England. Five Year Forward View. 2014;(October).
 13. Andrew E, Nehme Z, Cameron P, Smith K. Drivers of Increasing Emergency Ambulance Demand. *Prehospital Emerg Care* [Internet]. 2020 [cited 2020 Dec 4];24(3):385. Available from:
<https://www.tandfonline.com/action/journalInformation?journalCode=ipeec20>
 14. National Audit Office. NHS Ambulance services. 2017.
 15. Turner J, Jacques R, Crum A, Coster J, Stone T, Nicholl J. Ambulance Response Programme Evaluation of Phase 1 and Phase 2 Final Report. 2017.
 16. NHS England [online]. Statistics » Ambulance Quality Indicators [Internet].

- NHS England. 2021 [cited 2020 Dec 17]. Available from: <https://www.england.nhs.uk/statistics/statistical-work-areas/ambulance-quality-indicators/>
17. NHS England [online]. Statistics » COVID-19 Daily Deaths [Internet]. NHS England. 2020 [cited 2020 Dec 17]. Available from: <https://www.england.nhs.uk/statistics/statistical-work-areas/covid-19-daily-deaths/>
 18. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *J Chronic Dis*. 1987 Jan 1;40(5):373–83.
 19. Coster JE, Turner JK, Bradbury D, Cantrell A. Why Do People Choose Emergency and Urgent Care Services? A Rapid Review Utilizing a Systematic Literature Search and Narrative Synthesis [Internet]. Vol. 24, *Academic Emergency Medicine*. 2017 [cited 2020 Dec 4]. p. 1137–49. Available from: <https://doi.org/onlinelibrary.wile>
 20. Seyedin H, Afshari M, Isfahani P, Hasanzadeh E, Radinmanesh M, Corani Bahador R. The main factors of supplier-induced demand in health care: A qualitative study. 2021 [cited 2022 Jun 7]; Available from: www.jehp.net
 21. Hussain F, Cooper A, Carson-Stevens A, Donaldson L, Hibbert P, Hughes T, et al. Diagnostic error in the emergency department: Learning from national patient safety incident report analysis. *BMC Emerg Med* [Internet]. 2019 [cited 2022 Jun 7];19(1). Available from: <https://doi.org/10.1186/s12873-019-0289-3>
 22. O’Cathain A, Knowles E, Long J, Connell J, Bishop-Edwards L, Simpson R, et al. Drivers of ‘clinically unnecessary’ use of emergency and urgent care: the DEUCE mixed-methods study. *Heal Serv Deliv Res*. 2020;8(15):1–256.
 23. Li M, Vanberkel P, Carter AJE. A review on ambulance offload delay literature. *Health Care Manag Sci* [Internet]. 2019;22(4):658–75. Available from: <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medc&NEWS=N&AN=29982911>
 24. Segal E, Verter V, Colacone A, Afilalo M. The in-hospital interval: A description of EMT time spent in the emergency department. In: *Prehospital Emergency Care* [Internet]. Taylor & Francis; 2006 [cited 2021 Feb 12]. p. 378–82. Available from: <https://www.tandfonline.com/doi/abs/10.1080/10903120600725884>
 25. Silvestri S. Evaluation of Patients in Delayed Emergency Medical Services Unit Off-load Status. *Acad Emerg Med*. 2006 May 1;13(5Supplement 1):S70–S70.
 26. NHS England [online]. Statistics » Urgent and Emergency Care Daily Situation Reports [Internet]. [cited 2021 Feb 15]. Available from: <https://www.england.nhs.uk/statistics/statistical-work-areas/uec-sitrep/>
 27. Lee YJ, Shin S Do, Lee EJ, Cho JS, Cha WC. Emergency department overcrowding and ambulance turnaround time. *PLoS One* [Internet]. 2015;10(6):1–9. Available from: <http://dx.doi.org/10.1371/journal.pone.0130758>
 28. Asplin BR, Magid DJ, Rhodes K V., Solberg LI, Lurie N, Camargo CA. A

- conceptual model of emergency department crowding. *Ann Emerg Med* [Internet]. 2003 Aug 1 [cited 2021 Jan 25];42(2):173–80. Available from: <http://www.annemergmed.com/article/S019606440300444X/fulltext>
29. Eckstein M, Chan LS. The Effect of Emergency Department Crowding on Paramedic Ambulance Availability. *Ann Emerg Med* [Internet]. 2004 [cited 2021 Feb 15];43(1):100–5. Available from: <https://pubmed.ncbi.nlm.nih.gov/14707948/>
 30. RCEM, AACE, Improvement N. Ambulance handover: tactical advice to hospitals and ambulance services. 2017.
 31. Pham JC, Patel R, Millin MG, Kirsch TD, Chanmugam A. The Effects of Ambulance Diversion: A Comprehensive Review. *Acad Emerg Med* [Internet]. 2006 Nov 1 [cited 2021 Feb 15];13(11):1220–7. Available from: <http://doi.wiley.com/10.1197/j.aem.2006.05.024>
 32. Knowles E, Shephard N, Stone T, Bishop-Edwards L, Hirst E, Abouzeid L, et al. Closing five Emergency Departments in England between 2009 and 2011: the closed controlled interrupted time-series analysis. *Heal Serv Deliv Res*. 2018;6(27):1–234.
 33. Cooney DR, Millin MG, Carter A, Lawner BJ, Nable JV, Wallus HJ. Ambulance diversion and emergency department offload delay: Resource document for the national association of ems physicians position statement. *Prehospital Emerg Care*. 2011;15(4):555–61.
 34. Majedi M. A Queueing Model To Study Ambulance Of f load Delays [Internet]. University of Waterloo; 2008 [cited 2021 Feb 15]. Available from: <https://uwspace.uwaterloo.ca/handle/10012/4019>
 35. Kingswell C, Shaban RZ, Crilly J. The lived experiences of patients and ambulance ramping in a regional Australian emergency department: An interpretive phenomenology study. *Australas Emerg Nurs J* [Internet]. 2015;18(4):182–9. Available from: <http://dx.doi.org/10.1016/j.aenj.2015.08.003>
 36. Hammond E, Holzhauser K, Shaban R, Melton N. The effects of ambulance ramping on Emergency Department length of stay and in-patient mortality. *Australas Emerg Nurs J*. 2009;12(4):170.
 37. Kingswell C, Shaban RZ, Crilly J. Concepts, antecedents and consequences of ambulance ramping in the emergency department: A scoping review. *Australas Emerg Nurs J*. 2017;20(4):153–60.
 38. Crilly J, Keijzers G, Tippett V, O'Dwyer J, Lind J, Bost N, et al. Improved outcomes for emergency department patients whose ambulance off-stretcher time is not delayed. *EMA - Emerg Med Australas*. 2015 Jun 1;27(3):216–24.
 39. Olausson A, Abetz JW, Smith K, Bernard S, Gaddam R, Banerjee A, et al. Paramedic streaming upon arrival in emergency department: A prospective study. *EMA - Emerg Med Australas*. 2021 Apr 1;33(2):286–91.
 40. Kue R, Ramstrom E, Weisberg S, Restuccia M. Evaluation of an emergency medical services based social services referral program for elderly patients. *Prehospital Emerg Care* [Internet]. 2009 [cited 2021 Jun 15];13(3):273–9. Available from: <https://pubmed.ncbi.nlm.nih.gov/19499461/>
 41. Department of Health. Taking Healthcare to the Patient: transforming NHS

- ambulance services. 2005;(June 2005):1–63.
42. NHS England [online]. About urgent and emergency care [Internet]. [cited 2020 Sep 17]. Available from: <https://www.england.nhs.uk/urgent-emergency-care/about-uec/>
 43. Pope C, McKenna G, Turnbull J, Prichard J, Rogers A. Navigating and making sense of urgent and emergency care processes and provision. *Heal Expect* [Internet]. 2019 Jun 10 [cited 2020 Oct 22];22(3):435–43. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1111/hex.12866>
 44. Egan M, Murar F, Lawrence J, Burd H. Identifying the predictors of avoidable emergency department attendance after contact with the NHS 111 phone service: Analysis of 16.6 million calls to 111 in England in 2015-2017. *BMJ Open* [Internet]. 2020 Mar 9 [cited 2020 Oct 22];10(3):32043. Available from: <https://europepmc.org/articles/PMC7066618>
 45. Parkinson B, Meacock R, Checkland K, Sutton M. Clarifying the concept of avoidable emergency department attendance. *J Health Serv Res Policy*. 2020;0(0):1–6.
 46. Hannay DR. The iceberg’ of illness and “trivial” consultations. *J R Coll Gen Pract*. 1980;
 47. Snooks H, Wrigley H, George S, Thomas E, Smith H, Glasper A. Appropriateness of use of emergency ambulances. *21 Accid Emerg Med* [Internet]. 1998 [cited 2021 Jun 17];15:212–8. Available from: <http://emj.bmj.com/>
 48. Department of Health. Ambulance services, England: 1998-9. 1999.
 49. Patton GG, Thakore S. Reducing inappropriate emergency department attendances - A review of ambulance service attendances at a regional teaching hospital in Scotland. *Emerg Med J*. 2013;30(6):459–61.
 50. Morris T, Mason SM, Moulton C, O’keeffe C. Calculating the proportion of avoidable attendances at UK emergency departments: analysis of the Royal College of Emergency Medicine’s Sentinel Site Survey data. *Emerg Med J* [Internet]. 2018 [cited 2020 Sep 16];35:114–9. Available from: <http://dx.doi.org/10.1136/>
 51. Mchale P, Wood S, Hughes K, Bellis MA, Demnitz U, Wyke S. Who uses emergency departments inappropriately and when-a national cross-sectional study using a monitoring data system [Internet]. 2013 [cited 2020 Sep 16]. Available from: <http://www.biomedcentral.com/1741-7015/11/258>
 52. Hjalte L, Suserud BO, Herlitz J, Karlberg I. Why are people without medical needs transported by ambulance? A study of indications for pre-hospital care. *Eur J Emerg Med*. 2007 Jun;14(3):151–6.
 53. Miles J. 17 Exploring ambulance conveyances to the emergency department: a descriptive analysis of non-urgent transports. *Emerg Med J* [Internet]. 2017 Dec; Available from: <http://europepmc.org/abstract/med/29170314>
 54. Miles J. 59 Ambulance over-conveyance to the emergency department: a large data analysis of ambulance journeys. *BMJ Open* [Internet]. 2018; Available from: http://bmjopen.bmj.com/content/8/Suppl_1/A22.3
 55. Snooks H, Williams S, Crouch R, Foster T, Hartley-Sharpe C, Dale J. NHS emergency response to 999 calls: Alternatives for cases that are neither life

- threatening nor serious. Vol. 325, British Medical Journal. SAS Institute; 2002. p. 330–3.
56. Turner J, Coster J, Chambers D, Cantrell A, Phung V-H, Knowles E, et al. What evidence is there on the effectiveness of different models of delivering urgent care? A rapid review. *Heal Serv Deliv Res.* 2015;3(43):1–134.
 57. UEC Review Team, ECIST. Transforming urgent and emergency care services in England: Safer, faster, better: good practice in delivering urgent and emergency care A guide for local health and social care communities. 2015.
 58. Lord Carter of Coles. Operational productivity and performance in English NHS acute hospitals: Unwarranted variations. *Dep Heal [Internet].* 2018 [cited 2020 Sep 16];(February):87. Available from: <https://improvement.nhs.uk/about-us/corporate-publications/publications/lord-carters-review-unwarranted-variation-nhs-ambulance-trusts/>
 59. NHS England. High quality care for all , now and for future generations : Transforming urgent and emergency care services in England - The Evidence Base from the Urgent and Emergency Care Review. 2013;1–79. Available from: <http://www.england.nhs.uk/wp-content/uploads/2013/06/urg-emerg-care-ev-bse.pdf>
 60. ECIST, UEC Review Team. Transforming urgent and emergency care services in England. 2015;1–57. Available from: <https://www.england.nhs.uk/wp-content/uploads/2015/06/trans-uec.pdf>
 61. Snooks HA, Kingston MR, Anthony RE, Russell IT. New models of emergency prehospital care that avoid unnecessary conveyance to emergency department: Translation of research evidence into practice? *Sci World J.* 2013;2013.
 62. Agressie KM. Ambulances in-zicht 2015.
 63. Mikolaizak AS, Simpson PM, Tiedemann A, Lord SR, Close JC. Systematic review of non-transportation rates and outcomes for older people who have fallen after ambulance service call-out. *Australas J Ageing.* 2013;32(3):147–57.
 64. Ebben RHA, Vloet LCM, Speijers RF, Tönjes NW, Loef J, Pelgrim T, et al. A patient-safety and professional perspective on non-conveyance in ambulance care: A systematic review. *Scand J Trauma Resusc Emerg Med.* 2017;25(1):1–20.
 65. Fraess-Phillips AJ. Can Paramedics Safely Refuse Transport of Non-Urgent Patients? *Prehosp Disaster Med.* 2016;31(6):667–74.
 66. Tohira H, Williams TA, Jacobs I, Bremner A, Finn J. The impact of new prehospital practitioners on ambulance transportation to the emergency department: a systematic review and meta-analysis. *Emerg Med J.* 2014;31(e1):e88–94.
 67. Turner J. Building the evidence base in pre-hospital urgent and emergency care. DoH. 2010;
 68. NHS England. Ambulance Quality Indicators. [online] [Internet]. [cited 2019 Dec 6]. Available from: <https://www.england.nhs.uk/statistics/statistical-work-areas/ambulance-quality-indicators/>

69. Thompson MP, Kaplan CM, Cao Y, Bazzoli GJ, Waters TM. Reliability of 30-Day Readmission Measures Used in the Hospital Readmission Reduction Program. *Health Serv Res.* 2016;51(6):2095–114.
70. Tohira H, Fatovich D, Williams TA, Bremner AP, Arendts G, Rogers IR, et al. Is it Appropriate for Patients to be Discharged at the Scene by Paramedics? *Prehospital Emerg Care* [Internet]. 2016 Jul 3 [cited 2021 Feb 12];20(4):539–49. Available from: <https://www.tandfonline.com/doi/abs/10.3109/10903127.2015.1128028>
71. Hipskind JE, Gren JM, Barr DJ. Patients who refuse transportation by ambulance: A case series. *Prehosp Disaster Med* [Internet]. 1997 [cited 2021 Feb 12];12(4):45–50. Available from: <https://www.cambridge.org/core/journals/prehospital-and-disaster-medicine/article/abs/patients-who-refuse-transportation-by-ambulance-a-case-series/82B23E0C19292C685A89F4FD94A9A7EA>
72. Staudenmayer K, Hsia R, Wang E, Sporer K, Ghilarducci D, Spain D, et al. The forgotten trauma patient: Outcomes for injured patients evaluated by emergency medical services but not transported to the hospital. *J Trauma Acute Care Surg* [Internet]. 2012 Mar [cited 2021 Feb 12];72(3):594–600. Available from: [/pmc/articles/PMC3489913/](https://pubmed.ncbi.nlm.nih.gov/22489113/)
73. Coster J, O’Cathain A, Jacques R, Crum A, Siriwardena AN, Turner J. Outcomes for Patients Who Contact the Emergency Ambulance Service and Are Not Transported to the Emergency Department: A Data Linkage Study. *Prehospital Emerg Care.* 2019 Jul 4;23(4):566–77.
74. O’Cathain A, Jacques R, Stone T, Turner J. Why do ambulance services have different non-transport rates? A national cross sectional study. *PLoS One.* 2018 Sep 1;13(9).
75. National Institute for Health and Care Excellence. Emergency and acute medical care in over 16s Quality standard [QS174]. 2018.
76. O’Hara R, Johnson M, Hirst E, Weyman A, Shaw D, Mortimer P, et al. A qualitative study of decision-making and safety in ambulance service transitions. *Heal Serv Deliv Res* [Internet]. 2014;2(56):1–138. Available from: <https://www.journalslibrary.nihr.ac.uk/hsdr/hsdr02560/>
77. Burrell L, Noble A, Ridsdale L. Decision-making by ambulance clinicians in London when managing patients with epilepsy: A qualitative study. *Emerg Med J.* 2013;30(3):236–40.
78. Halter M, Vernon S, Snooks H, Porter A, Close J, Moore F, et al. Complexity of the decision-making process of ambulance staff for assessment and referral of older people who have fallen: A qualitative study. *Emerg Med J.* 2011;28(1):44–50.
79. Simpson P, Thomas R, Bendall J, Lord B, Lord S, Close J. ‘Popping nana back into bed’ - a qualitative exploration of paramedic decision making when caring for older people who have fallen. *BMC Health Serv Res.* 2017;17(1):1–14.
80. Al-Sulaiti M, Snooks H, Porter A. Non-conveyance of 999 callers: Early findings related subsequent health service callers. *Emerg Med J.* 2009;26(10):7.

81. Hoikka M, Silfvast T, Ala-Kokko TI. A high proportion of prehospital emergency patients are not transported by ambulance: a retrospective cohort study in Northern Finland. *Acta Anaesthesiol Scand*. 2017;61(5):549–56.
82. Brydges M, Spearen C, Birze A, Tavares W. A culture in transition: Paramedic experiences with community referral programs. *Can J Emerg Med*. 2015;17(6):631–8.
83. Snooks HA, Dale J, Hartley-Sharpe C. On-scene alternatives for emergency ambulance crews attending patients who do not need to travel to the accident and emergency department: a review of the literature. *Emerg Med J [Internet]*. 2004;21:212–5. Available from: www.emjonline.com
84. Silvestri S, Rothrock SG, Kennedy D, Ladde J, Bryant M, Pagane J. Can paramedics accurately identify patients who do not require emergency department care? *Prehospital Emerg Care [Internet]*. 2002 [cited 2021 Jun 17];6(4):387–90. Available from: <https://pubmed.ncbi.nlm.nih.gov/12385603/>
85. Hauswald M. Can paramedics safely decide which patients do not need ambulance transport or emergency department care? *Prehospital Emerg Care [Internet]*. 2002 Oct 1 [cited 2021 Jun 17];6(4):383–6. Available from: <https://europepmc.org/article/med/12385602>
86. Miles J, Coster J, Jacques R. Using vignettes to assess the accuracy and rationale of paramedic decisions on conveyance to the emergency department. *Br Paramed J*. 2019;4(1):6–13.
87. Standards Committee Council. National Triage Scale. *Emerg Med*. 1994;145–6.
88. Eitel DR, Travers DA, Rosenau AM, Gilboy N, Wuerz RC. The Emergency Severity Index triage algorithm version 2 is reliable and valid. *Acad Emerg Med*. 2003 Oct 1;10(10):1070–80.
89. Tanabe P, Gimbel R, Yarnold PR, Kyriacou DN, Adams JG. Reliability and Validity of Scores on the Emergency Severity Index Version 3. *Acad Emerg Med*. 2004;11(1):59–65.
90. Wuerz RC, Milne LW, Eitel DR, Travers D, Gilboy N. Reliability and validity of a new five-level triage instrument. *Acad Emerg Med [Internet]*. 2000 Mar 1 [cited 2021 Jun 18];7(3):236–42. Available from: <https://onlinelibrary.wiley.com/doi/full/10.1111/j.1553-2712.2000.tb01066.x>
91. Wuerz RC, Travers D, Gilboy N, Eitel DR, Rosenau A, Yazhari R. Implementation and refinement of the Emergency severity index. *Acad Emerg Med [Internet]*. 2001 Feb 1 [cited 2021 Jun 18];8(2):170–6. Available from: <https://onlinelibrary.wiley.com/doi/full/10.1111/j.1553-2712.2001.tb01283.x>
92. Ebrahimi M, Heydari A, Mazlom R, Mirhaghi A. The reliability of the Australasian Triage Scale: a meta-analysis. *World J Emerg Med*. 2015;6(2):94–9.
93. Manchester Triage Group, Mackway-Jones K. *Emergency triage*. London: BMJ Publishing Group; 1997.
94. Newton M, Tunn E, Moses I, Ratcliffe D, MacKway-Jones K. *Clinical navigation for beginners: The clinical utility and safety of the Paramedic*

- Pathfinder. *Emerg Med J*. 2013;31(e1):e29–34.
95. Newton M, Tunn E, Moses I, Ratcliffe D, MacKway-Jones K. Clinical navigation for beginners: The clinical utility and safety of the Paramedic Pathfinder. *Emerg Med J*. 2013;31(e1):e29–34.
 96. Patel R, Nugawela MD, Edwards HB, Richards A, Le Roux H, Pullyblank A, et al. Can early warning scores identify deteriorating patients in pre-hospital settings? A systematic review [Internet]. Vol. 132, *Resuscitation*. Elsevier Ireland Ltd; 2018 [cited 2020 Oct 20]. p. 101–11. Available from: <https://pubmed.ncbi.nlm.nih.gov/30171976/>
 97. Snooks HA, Carter B, Dale J, Foster T, Humphreys I, Logan PA, et al. Support and assessment for fall emergency referrals (SAFER 1): Cluster randomised trial of computerised clinical decision support for paramedics. *PLoS One* [Internet]. 2014;9(9). Available from: <https://www.gov.uk/government/organisations/department-of-health>.
 98. Snooks H, Cheung WY, Close J, Dale J, Gaze S, Humphreys I, et al. Support and Assessment for Fall Emergency Referrals (SAFER 1) trial protocol. Computerised on-scene decision support for emergency ambulance staff to assess and plan care for older people who have fallen: Evaluation of costs and benefits using a pragmatic. *BMC Emerg Med*. 2010;10:1–8.
 99. Blomberg SN, Folke F, Ersbøll AK, Christensen HC, Torp-Pedersen C, Sayre MR, et al. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. *Resuscitation*. 2019 May;138:322–9.
 100. Blomberg SN, Christensen HC, Lippert F, Ersbøll AK, Torp-Petersen C, Sayre MR, et al. Effect of Machine Learning on Dispatcher Recognition of Out-of-Hospital Cardiac Arrest during Calls to Emergency Medical Services: A Randomized Clinical Trial. *JAMA Netw Open* [Internet]. 2021 Jan 1 [cited 2021 Jun 22];4(1):2032320. Available from: <https://jamanetwork.com/>
 101. Snooks H, Bailey-Jones K, Burge-Jones D, Dale J, Davies J, Evans B, et al. Predictive risk stratification model: a randomised stepped-wedge trial in primary care (PRISMATIC). *Heal Serv Deliv Res*. 2018;6(1):1–164.
 102. Porter A, Dale J, Foster T, Logan P, Wells B, Snooks H. Implementation and use of computerised clinical decision support (CCDS) in emergency pre-hospital care: A qualitative study of paramedic views and experience using Strong Structuration Theory. *Implement Sci* [Internet]. 2018 [cited 2021 Jun 11];13(1). Available from: <https://doi.org/10.1186/s13012-018-0786-x>
 103. Levin S, Toerper M, Hamrock E, Hinson JS, Barnes S, Gardner H, et al. Machine-Learning-Based Electronic Triage More Accurately Differentiates Patients With Respect to Clinical Outcomes Compared With the Emergency Severity Index. *Ann Emerg Med* [Internet]. 2018;71(5):565-574.e2. Available from: <https://doi.org/10.1016/j.annemergmed.2017.08.005>
 104. Raita Y, Goto T, Faridi MK, Brown DFMM, Camargo CAJ, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit Care* [Internet]. 2019;23(1):1–13. Available from: <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=prem&NEWS=N&AN=30795786>

105. Grant K, McParland A, Mehta S, Ackery AD. Artificial Intelligence in Emergency Medicine: Surmountable Barriers With Revolutionary Potential [Internet]. Vol. 75, *Annals of Emergency Medicine*. 2020 [cited 2021 Jun 22]. p. 721–6. Available from: <https://reader.elsevier.com/reader/sd/pii/S0196064419314659?token=DEC53C32D979F9AFC249FA34EC3688BA58EEE05A5857B75FA31D0852A542719158F44279BE380D34DA69F5F4757384B3&originRegion=eu-west-1&originCreation=20210622080254>
106. Miles J, Turner J, Jacques R, Williams J, Mason SM. Using machine-learning risk prediction models to triage the acuity of undifferentiated patients entering the emergency care system: a systematic review. *BMC Diagnostic Progn Res* [Internet]. 2020 Dec 2 [cited 2020 Oct 2];4(1):16. Available from: <https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-020-00084-1>
107. Meisel ZF, Pollack C V., Mechem CC, Pines JM. Derivation and internal validation of a rule to predict hospital admission in prehospital patients. *Prehospital Emerg Care*. 2008;12(3):314–9.
108. Li J, Guo L, Handly N. Hospital admission prediction using pre-hospital variables. 2009 *IEEE Int Conf Bioinforma Biomed BIBM 2009*. 2009;283–6.
109. Seymour CW, Kahn JM, Cooke CR, Watkins TR, Heckbert SR, Rea TD. Prediction of critical illness during out-of-hospital emergency care. *JAMA*. 2010 Aug;304(7):747–54.
110. van Rein EAJ, van der Sluijs R, Voskens FJ, Lansink KWW, Houwert RM, Lichtveld RA, et al. Development and Validation of a Prediction Model for Prehospital Triage of Trauma Patients. *JAMA Surg* [Internet]. 2019 May;154(5):421–9. Available from: <http://search.ebscohost.com/login.aspx?direct=true&db=cin20&AN=136501962&site=ehost-live>
111. Newgard CD, Hsia RY, Mann NC, Schmidt T, Sahni R, Bulger EM, et al. The trade-offs in field trauma triage. *J Trauma Acute Care Surg* [Internet]. 2013 May;74(5):1298–306. Available from: <http://insights.ovid.com/crossref?an=01586154-201305000-00017>
112. Kim D, You S, So S, Lee J, Yook S, Jang DP, et al. A data-driven artificial intelligence model for remote triage in the prehospital environment. *PLoS One*. 2018;13(10):1–14.
113. Goto T, Camargo CAJ, Faridi MK, Freishtat RJ, Hasegawa K. Machine Learning-Based Prediction of Clinical Outcomes for Children During Emergency Department Triage. *JAMA Netw open* [Internet]. 2019;2(1):e186937. Available from: <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=pem&NEWS=N&AN=30646206>
114. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation HHS Public Access. *J Clin Epidemiol*. 2016;69:245–7.
115. Meyers RG. Understanding empiricism. 2006;183.
116. Humphreys P. *Extending Ourselves: Computational Science, Empiricism,*

- and Scientific Method. Extending Ourselves Comput Sci Empiricism, Sci Method. 2006 Feb 1;1–182.
117. Chalmers DJ, Manley D, Wasserman R. *Metametaphysics: new essays on the foundations of ontology*. 2009.
 118. Jacquette D. *Ontology*. Oxford: Routledge; 2002.
 119. Harari YN. *Sapiens : a brief history of humankind*. London: Penguin Random House; 2015. 443 p.
 120. Bodenreider O, Smith B, Kumar A, Burgun A. Investigating subsumption in SNOMED CT: An exploration into large description logic-based biomedical terminologies. *Artif Intell Med [Internet]*. 2007 [cited 2022 May 27];39(3):183–95. Available from: <http://www.snomed.org/>
 121. Chiang MF, Hwang JC, Yu AC, Casper DS, Cimino JJ, Starren JB. Reliability of SNOMED-CT coding by three physicians using two terminology browsers. *AMIA Annu Symp Proc [Internet]*. 2006 [cited 2022 May 27];2006:131–5. Available from: </pmc/articles/PMC1839418/>
 122. Honderich T, editor. *The Oxford Companion to Philosophy*. The Oxford Companion to Philosophy. Oxford University Press; 2005.
 123. Borchert DM, editor. *Macmillan Encyclopedia of Philosophy*. In: 2nd Editio. New York: Macmillan Library Reference; 2005 [cited 2022 May 30]. Available from: <https://philpapers.org/rec/MONMEO-3>
 124. *Stanford Encyclopedia of Philosophy*. *Stanford Encyclopedia of Philosophy - Value Pluralism [Internet]*. <https://plato.stanford.edu/entries/value-pluralism/>. 2018 [cited 2022 May 30]. Available from: <https://plato.stanford.edu/entries/value-pluralism/>
 125. Murphy KP. *Machine Learning: A probabilistic perspective*. 1st ed. Malden, Massachusetts: the MIT press; 2012.
 126. Badillo S, Banfai B, Birzele F, Davydov II, Hutchinson L, Kam-Thong T, et al. An Introduction to Machine Learning. *Clin Pharmacol Ther*. 2020;107(4):871–85.
 127. Lever J, Krzywinski M, Altman N. Points of Significance: Model selection and overfitting. Vol. 13, *Nature Methods*. 2016. p. 703–4.
 128. Steyerberg EW. *Clinical Prediction Models: A practical approach to Development, Validation, and Updating [Internet]*. New York, NY: Springer New York; 2009. (Statistics for Biology and Health). Available from: <http://link.springer.com/10.1007/978-0-387-77244-8>
 129. Bellman RE. *Adaptive Control Processes*. *Adapt Control Process*. 1961 Dec 31;
 130. Chen L. Curse of Dimensionality. In: *Encyclopedia of Database Systems [Internet]*. Boston, MA: Springer US; 2009. p. 545–6. Available from: http://link.springer.com/10.1007/978-0-387-39940-9_133
 131. Dunn R. *The Science of Conjecture: Evidence and Probability before Pascal (review)*. *Parergon*. 2002;19(2):196–8.
 132. GeeksforGeeks. *Advantages and Disadvantages of Logistic Regression [Internet]*. 2021 [cited 2021 Dec 10]. Available from: <https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/>

133. Cameron A, Rodgers K, Ireland A, Jamdar R, McKay GA. A simple tool to predict admission at the time of triage. *Emerg Med J* [Internet]. 2015;32(3):174–9. Available from: <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med11&NEWS=N&AN=24421344>
134. Steyerberg EW, Kievit J, de Mol Van Otterloo JC, van Bockel JH, Eijkemans MJ, Habbema JD. Perioperative mortality of elective abdominal aortic aneurysm surgery. A clinical prediction rule based on literature and individual patient data. *Arch Intern Med* [Internet]. 1995 Oct 9 [cited 2020 Mar 9];155(18):1998–2004. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/7575054>
135. Fritsch S, Guenther F, Wright MN, Suling M, Mueller SM. Package “neuralnet” Title Training of neural networks. 2015 [cited 2021 Dec 13]; Available from: <http://www.dfg.de>
136. De Veaux RD, Ungar LH. Multicollinearity: A tale of two nonparametric regressions. In 1994. p. 393–402.
137. Smieja M, Struski Ł, Tabor J, Zielinski B, Spurek P. Processing of missing data by neural networks. In: *Advances in Neural Information Processing Systems*. 2018. p. 2719–29.
138. Arik SO, Pfister T. TabNet: Attentive Interpretable Tabular Learning. 2019 [cited 2021 Dec 13]; Available from: www.aaai.org
139. Therneau T, Atkinson B, Ripley B. Package “rpart” v4.1-15 [Internet]. 2019 [cited 2021 Dec 21]. Available from: <https://cran.r-project.org/package=rpart>
140. Kelleher JD, Mac Namee B, D’Arcy A. *Fundamentals of machine learning for predictive data analytics : algorithms, worked examples, and case studies*. Cambridge, Massachusetts: The MIT Press; 2015.
141. Breiman L. Bagging predictors. *Mach Learn*. 1996;24(2):123–40.
142. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. [cited 2021 Aug 26]; Available from: <https://github.com/dmlc/xgboost>
143. Ramírez-Hernández JA, Fernandez E. Control of a re-entrant line manufacturing model with a reinforcement learning approach. *Proc - 6th Int Conf Mach Learn Appl ICMLA 2007*. 2007;330–5.
144. Darst BF, Malecki KC, Engelman CD. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genet* [Internet]. 2018 Sep 17 [cited 2021 Dec 21];19(1):1–6. Available from: <https://bmcbgenomdata.biomedcentral.com/articles/10.1186/s12863-018-0633-8>
145. Yin Z, Wang Y, Liu L, Zhang W, Zhang J. Cross-subject EEG feature selection for emotion recognition using transfer recursive feature elimination. *Front Neurobot*. 2017 Apr 10;11(APR):19.
146. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann Intern Med*. 2015 Jan 6;162(1):W1–73.
147. van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: diagnosis versus prognosis. *J Clin Epidemiol*. 2021 Apr

- 1;132:142–5.
148. NHS England. A&E Attendances and Emergency Admissions Monthly Return Definitions. 2015; Available from: <https://www.england.nhs.uk/statistics/statistical-work-areas/ae-waiting->
 149. NHS Digital. Linked datasets supporting health and care delivery and research. 2018;(April):1–14. Available from: <https://digital.nhs.uk/services/data-access-request-service-dars/linked-datasets-supporting-health-and-care-delivery-and-research>
 150. Hussain F, Cooper A, Carson-Stevens A, Donaldson L, Hibbert P, Hughes T, et al. Diagnostic error in the emergency department: Learning from national patient safety incident report analysis. *BMC Emerg Med*. 2019;19(1).
 151. Lowy A, Kohler B, Nicholl J. Attendance at accident and emergency departments: unnecessary or inappropriate? *J Public Health (Bangkok)* [Internet]. 1994;16(2):134–40. Available from: <https://academic.oup.com/jpubhealth/article/1545251/Attendance>
 152. NHS Digital. Non-urgent A&E attendances [Internet]. 2020 [cited 2020 Sep 16]. Available from: <https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/innovative-uses-of-data/demand-on-healthcare/unnecessary-a-and-e-attendances>
 153. Walker C. Emergency Care Data Set (ECDS) Technical User Guidance Provisional Reviewer nam. 2017.
 154. NHS England. Transforming Urgent and Emergency Care Services in England Improving referral pathways between urgent and emergency services in England Advice for Urgent and Emergency Care Networks.
 155. England N. Transforming urgent and emergency care services in England Clinical models for ambulance services. 2015.
 156. Hussain F, Cooper A, Carson-Stevens A, Donaldson L, Hibbert P, Hughes T, et al. Diagnostic error in the emergency department: Learning from national patient safety incident report analysis. *BMC Emerg Med* [Internet]. 2019 Dec 4 [cited 2022 Jun 21];19(1):1–9. Available from: <https://bmccemergmed.biomedcentral.com/articles/10.1186/s12873-019-0289-3>
 157. Vazin A, Zamani Z, Hatam N. Frequency of medication errors in an emergency department of a large teaching hospital in southern Iran. *Drug Healthc Patient Saf* [Internet]. 2014;6:179–84. Available from: <http://dx.doi.org/10.2147/DHPS.S75223>
 158. Westbrook JI, Raban MZ, Walter SR, Douglas H. Task errors by emergency physicians are associated with interruptions, multitasking, fatigue and working memory capacity: a prospective, direct observation study. *BMJ Qual Saf* [Internet]. 2018 [cited 2022 Jun 21];0:1–9. Available from: <http://qualitysafety.bmj.com/>
 159. Government U. Equality Act 2010. Response [Internet]. 2010 [cited 2021 Sep 16];2010(April):1–216. Available from: <https://www.legislation.gov.uk/ukpga/2010/15/contents>
 160. UK Government. Data Protection Act 2018. 2018 [cited 2021 Sep 16];2018(January):338. Available from: <http://www.legislation.gov.uk/ukpga/2018/12/part/2/chapter/2/enacted%0A>

- <https://ico.org.uk/for-organisations/data-protection-act-2018/>
161. Binns R, Gallo V. Human bias and discrimination in AI systems. 2021 [cited 2021 Jul 30]; Available from: <https://ico.org.uk/about-the-ico/news-and-events/ai-blog-human-bias-and-discrimination-in-ai-systems/>
 162. Veale M, Binns R. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data Soc* [Internet]. 2017 [cited 2021 Jul 29];4(2). Available from: <https://us.sagepub.com/en-us/nam/open-access>
 163. Binns R. Fairness in Machine Learning: Lessons from Political Philosophy. *Proc Mach Learn Res* [Internet]. 2017 [cited 2021 Jul 29];81:1–11. Available from: <http://arxiv.org/abs/1712.03586>
 164. Binns R, Gallo V. Fully automated decision making AI systems: the right to human intervention and other safeguards. 2021 [cited 2021 Sep 16]; Available from: <https://ico.org.uk/about-the-ico/news-and-events/ai-blog-fully-automated-decision-making-ai-systems-the-right-to-human-intervention-and-other-safeguards/>
 165. Dinh MM, Russell SB, Bein KJ, Rogers K, Muscatello D, Paoloni R, et al. The Sydney Triage to Admission Risk Tool (START) to predict Emergency Department Disposition: A derivation and internal validation study using retrospective state-wide data from New South Wales, Australia. *BMC Emerg Med* [Internet]. 2016;16(1):1–7. Available from: <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med12&NEWS=N&AN=27912757>
 166. Kim SW, Li JY, Hakendorf P, Teubner DJJO, Ben-Tovim DI, Thompson CH. Predicting admission of patients by their presentation to the emergency department. *EMA - Emerg Med Australas*. 2014 Aug;26(4):361–7.
 167. Zhang X, Kim J, Patzer RE, Pitts SR, Patzer A, Schrage JD. Prediction of Emergency Department Hospital Admission Based on Natural Language Processing and Neural Networks. *Methods Inf Med* [Internet]. 2017 Jan 24;56(05):377–89. Available from: <http://www.thieme-connect.de/DOI/DOI?10.3414/ME17-01-0024>
 168. Rendell K, Koprinska I, Kyme A, Ebker-White AA, Dinh MM. The Sydney Triage to Admission Risk Tool (START2) using machine learning techniques to support disposition decision-making. *EMA - Emerg Med Australas*. 2019;31(3):429–35.
 169. Raita Y, Goto T, Faridi MK, Brown DFM, Camargo CA, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit Care* [Internet]. 2019 Dec 22;23(1):64. Available from: <https://ccforum.biomedcentral.com/articles/10.1186/s13054-019-2351-7>
 170. Zlotnik A, Alfaro MC, Pérez MCP, Gallardo-Antolín A, Martínez JMM. Building a Decision Support System for Inpatient Admission Prediction With the Manchester Triage System and Administrative Check-in Variables. *CIN Comput Informatics, Nurs* [Internet]. 2016 May;34(5):224–30. Available from: <http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&>

- an=00024665-201605000-00006
171. Dugas AF, Kirsch TD, Toerper M, Korley F, Yenokyan G, France D, et al. An Electronic Emergency Triage System to Improve Patient Distribution by Critical Outcomes. *J Emerg Med*. 2016;50(6):910–8.
 172. Golmohammadi D. Predicting hospital admissions to reduce emergency department boarding. *Int J Prod Econ*. 2016;182(September):535–44.
 173. Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency department triage using machine learning. *PLoS One* [Internet]. 2018;13(7):1–13. Available from: <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=30028888>
 174. Kwon J myoung, Lee YY, Lee YY, Lee S, Park H, Park J. Validation of deep-learning-based triage and acuity score using a large national dataset. *PLoS One* [Internet]. 2018;13(10):1–10. Available from: <http://dx.doi.org/10.1371/journal.pone.0205836>
 175. Zlotnik A, Alfaro MC, Perez MCP, Gallardo-Antolin A, Martinez JMM. Building a Decision Support System for Inpatient Admission Prediction With the Manchester Triage System and Administrative Check-in Variables. *Comput Inform Nurs*. 2016 May;34(5):224–30.
 176. Levin S, Toerper M, Hamrock E, Hinson JS, Barnes S, Gardner H, et al. Machine-Learning-Based Electronic Triage More Accurately Differentiates Patients With Respect to Clinical Outcomes Compared With the Emergency Severity Index. *Ann Emerg Med*. 2018;71(5):565-574.e2.
 177. Kwon J, Jeon K-H, Lee M, Kim K-H, Park J, Oh B-H. Deep Learning Algorithm to Predict Need for Critical Care in Pediatric Emergency Departments. *Pediatr Emerg Care* [Internet]. 2019 Jul;1. Available from: <http://insights.ovid.com/crossref?an=00006565-900000000-98117>
 178. Seymour CW, Kahn JM, Cooke CR, Watkins TR, Rea TD. Prediction of Critical Illness During Out-of-Hospital Emergency Care. 2010;304(7):747–54.
 179. RCP London. Royal College of Physicians. “National early warning score (NEWS) 2.” [Internet]. 2012. 1–30 p. Available from: <https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2>
 180. Royal College of Physicians. National Early Warning Score (NEWS) 2: Standardising the assessment of acute-illness severity in the NHS. Vol. 17, Updated report of a working party. 2017. 318–318 p.
 181. Van Smeden M, Moons KG, Ah De Groot J, Collins GS, Altman DG, Eijkemans MJ, et al. Sample size for binary logistic prediction models: Beyond events per variable criteria.
 182. Floares AG, Ferisgan M, Onita D, Ciuparu A, Calin GA, Manolache FB. The Smallest Sample Size for the Desired Diagnosis Accuracy. [cited 2021 Aug 26]; Available from: <http://cancergenome.nih.gov/>
 183. Riley RD, Debray TPA, Collins GS, Archer L, Ensor J, Smeden M van, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat Med* [Internet]. 2021 Aug 30 [cited 2021 Aug 26];40(19):4230–51. Available from:

- <https://onlinelibrary.wiley.com/doi/full/10.1002/sim.9025>
184. Riley RD, Snell KI, Ensor J, Burke DL, Harrell Jr FE, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II-binary and time-to-event outcomes. 2018;
 185. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ* [Internet]. 2020 Mar 18 [cited 2020 Sep 9];368. Available from: <http://www.bmj.com/permissionsSubscribe:http://www.bmj.com/subscribe> BMJ2020;368:m441doi:10.1136/bmj.m441
 186. Riley RD, Van Calster B, Collins GS. A note on estimating the Cox-Snell R² from a reported C statistic (AUROC) to inform sample size calculations for developing a prediction model with a binary outcome. *Stat Med*. 2021 Feb 20;40(4):859–64.
 187. van der Sluijs R, Debray TPA, Poeze M, Leenen LPH, van Heijl M. Development and validation of a novel prediction model to identify patients in need of specialized trauma care during field triage: design and rationale of the GOAT study. *Diagnostic Progn Res*. 2019;3(1):1–8.
 188. Takada T, Nijman S, Denaxas S, Snell KIE, Uijl A, Nguyen T-L, et al. Internal-external cross-validation helped to evaluate the generalizability of prediction models in large clustered datasets. *J Clin Epidemiol* [Internet]. 2021;137:83–91. Available from: <https://doi.org/10.1016/j.jclinepi.2021.03.025>
 189. Debray TPA, Damen JAAG, Riley RD, Snell K, Reitsma JB, Hooft L, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. *Stat Methods Med Res* [Internet]. 2019 [cited 2021 Nov 26];28(9):2768–86. Available from: <https://orcid.org/0000-0002-1790-2719>
 190. Tharwat A. Classification assessment methods. *Appl Comput Informatics* [Internet]. 2018 [cited 2021 Dec 20];17(1):168–92. Available from: <https://www.emerald.com/insight/2210-8327.htm>
 191. Moons KGM, Wolff RF, Riley RD, Penny ;, Whiting F, Westwood M, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration *Annals of Internal Medicine* RESEARCH AND REPORTING METHODS. *Ann Intern Med* [Internet]. 2019 [cited 2020 Mar 8];170:1–33. Available from: www.probast.org
 192. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. [cited 2021 Sep 2]; Available from: <http://dx.doi.org/10.1136/bmj.i3140><http://www.bmj.com/>
 193. Lindhiem O, Petersen IT, Mentch LK, Youngstrom EA. The Importance of Calibration in Clinical Psychology. *Assessment*. 2020;27(4):840–54.
 194. Probst P, Boulesteix A-L, Bischl B. Tunability: Importance of Hyperparameters of Machine Learning Algorithms. 2018;
 195. Hutter F, Hoos H, Leyton-Brown K. An Efficient Approach for Assessing Hyperparameter Importance.
 196. Claesen M, De Moor B. Hyperparameter Search in Machine Learning. 2015 [cited 2021 Dec 23]; Available from:

- <https://www.codalab.org/competitions/2321>.
197. Harrou F, Dairi A, Kadri F, Sun Y. Forecasting emergency department overcrowding: A deep learning framework. *Chaos Solitons Fractals* [Internet]. 2020;139:110247. Available from: <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=prem6&NEWS=N&AN=32982079>
 198. Chen T, He T, Benesty M, Khotilovitch V. Package “xgboost” v1.5. 2021 [cited 2021 Dec 24]; Available from: <https://github.com/dmlc/xgboost/issues>
 199. Chen T, He T, Benesty M, Khotilovitch V. XGBoost Parameters — xgboost 1.5.1 documentation [Internet]. [cited 2021 Dec 24]. Available from: <https://xgboost.readthedocs.io/en/stable/parameter.html>
 200. NHS Digital. National data opt-out - NHS Digital. NHS Digital. 2020.
 201. WhatsApp LLC. WhatsApp [Internet]. 2021 [cited 2021 Dec 27]. Available from: <https://www.whatsapp.com/>
 202. Association of Ambulance Chief Executives. Delayed hospital handovers: Impact assessment of patient harm. London; 2021.
 203. Miles J, Jacques R, Turner J, Mason S. The Safety INdEx of Prehospital On Scene Triage (SINEPOST) study: the development and validation of a risk prediction model to support ambulance clinical transport decisions on-scene—a protocol. *Diagnostic Progn Res* [Internet]. 2021 Dec 8 [cited 2021 Nov 8];5(1):18. Available from: <https://doi.org/10.1186/s41512-021-00108-4>
 204. Allan D. Glasgow coma scale. *Nurs Mirror* [Internet]. 1984 [cited 2021 Oct 1];158(23):32–4. Available from: <https://www.glasgowcomascale.org/>
 205. Ensor J, Martin EC, Riley RD. Package “pmsampsize”: Calculates the Minimum Sample Size Required for Developing a Multivariable Prediction Model [Internet]. 2020 [cited 2020 Sep 10]. Available from: <https://cran.r-project.org/web/packages/pmsampsize/pmsampsize.pdf>
 206. Riley RD, Snell KI, Ensor J, Burke DL, Jr FEH, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med* [Internet]. 2019 Mar 30 [cited 2021 Aug 26];38(7):1276–96. Available from: <https://onlinelibrary.wiley.com/doi/full/10.1002/sim.7992>
 207. Miles J, Turner J, Jacques R, Williams J, Mason SM. Using machine-learning risk prediction models to triage the acuity of undifferentiated patients entering the emergency care system: a systematic review. *BMC Diagnostic Progn Res*. 2020;
 208. Hope RM. Package “Rmisc” v.1.5. 2016;
 209. Debray T, de Jong V. Package “metamisc” Title Meta-Analysis of Diagnosis and Prognosis Research Studies. 2021 [cited 2021 Nov 26]; Available from: <https://orcid.org/0000-0002-1790-2719>
 210. Kappen TH, van Klei WA, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, Moons KGM. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. *Diagnostic Progn Res* 2018 21 [Internet]. 2018 Jun 12 [cited 2022 Jun 15];2(1):1–11. Available from: <https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-018-0033-6>
 211. Airedale NHS Foundation Trust. Annual Report Summary 2020/2021. 2020.

212. Barnsley Hospital NHS Foundation Trust. Changing lives: Trust strategy 2018-2021. 2018.
213. Bradford Teaching Hospitals NHS Foundation Trust. Annual Report and Accounts 2018/2019 [Internet]. Bradford; 2019 [cited 2021 Nov 27]. Available from: <https://www.bradfordhospitals.nhs.uk/wp-content/uploads/2019/06/Annual-Report-Quality-and-AccountsFinal18.6.19-reduced.pdf>
214. Calderdale and Huddersfield Foundation Trust. About us - CHFT [Internet]. [cited 2021 Nov 27]. Available from: <https://www.cht.nhs.uk/about-us>
215. The Mid Yorkshire Hospitals NHS Trust. Annual Report and Financial Accounts 2020/2021 [Internet]. 2021 [cited 2021 Nov 27]. Available from: <https://www.midyorks.nhs.uk/download.cfm?doc=docm93jijm4n540.pdf&ver=170>
216. Doncaster and Bassetlaw Teaching Hospitals NHS Foundation Trust. Annual Report and Accounts 20/21 [Internet]. 2021 [cited 2021 Nov 28]. Available from: https://oesn11hpbml2xaq003wx02ib-wpengine.netdna-ssl.com/wp-content/uploads/2021/11/dbth_annualreport2021.pdf
217. Harrogate and District NHS Foundation Trust. Annual Summary 2019/20 [Internet]. [cited 2021 Nov 28]. Available from: <https://www.flipsnack.com/harrogatenhsft/annual-summary-2019-20.html>
218. Hull University Teaching Hospitals NHS Trust. Quality Accounts 2021 [Internet]. 2021 [cited 2021 Nov 28]. Available from: [https://www.hey.nhs.uk/downloads/quality-account/?ind=1626707737894&filename=Quality Accounts 2020-2021.pdf&wpdmdl=9713&refresh=61a34f5ca5a9b1638092636](https://www.hey.nhs.uk/downloads/quality-account/?ind=1626707737894&filename=Quality%20Accounts%202021.pdf&wpdmdl=9713&refresh=61a34f5ca5a9b1638092636)
219. Corbett-Davies S, Goel S, Chohlas-Wood A, Chouldechova A, Feller A, Huq A, et al. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning *. 2018;
220. Uchida K, Kouno J, Yoshimura S, Kinjo N, Sakakibara F, Araki H, et al. Development of Machine Learning Models to Predict Probabilities and Types of Stroke at Prehospital Stage: the Japan Urgent Stroke Triage Score Using Machine Learning (JUST-ML). [cited 2022 Jan 3];1:3. Available from: <https://doi.org/10.1007/s12975-021-00937-x>
221. Dixon J, Burkholder T, Pigoga J, Lee M, Moodley K, de Vries S, et al. Using the South African Triage Scale for prehospital triage: a qualitative study. *BMC Emerg Med.* 2021;21(1):1–10.
222. Wibring K, Lingman M, Herlitz J, Ashfaq A, Bång A. Development of a prehospital prediction model for risk stratification of patients with chest pain ☆. 2021 [cited 2022 Jan 3]; Available from: <https://doi.org/10.1016/j.ajem.2021.09.079>
223. Knoery CR, Heaton J, Polson R, Bond R, Iftikhar A, Rjoob K, et al. Systematic Review of Clinical Decision Support Systems for Prehospital Acute Coronary Syndrome Identification. *Crit Pathw Cardiol* [Internet]. 2020 Sep 1 [cited 2022 Jan 4];19(3):119–25. Available from: https://journals.lww.com/critpathcardio/Fulltext/2020/09000/Systematic_Review_of_Clinical_Decision_Support.4.aspx

224. Snooks HA, Carter B, Dale J, Foster T, Humphreys I, Logan PA, et al. Support and assessment for fall emergency referrals (SAFER 1): Cluster randomised trial of computerised clinical decision support for paramedics. *PLoS One*. 2014;9(9).
225. Kawamoto K, Houlihan CA, Balas A, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. [cited 2022 Jan 4]; Available from: <http://www.bmj.com/>
226. Sims R, Kazda L, Michaleff ZA, Glasziou P, Thomas R. Consequences of health condition labelling: protocol for a systematic scoping review. Available from: <http://bmjopen.bmj.com/>
227. Croft P, Altman DG, Deeks JJ, Dunn KM, Hay AD, Hemingway H, et al. The science of clinical practice: Disease diagnosis or patient prognosis? Evidence about “what is likely to happen” should shape clinical practice. *BMC Med*. 2015 Jan 30;13(1).
228. Tsymbal A. The problem of concept drift: definitions and related work. 2004;
229. Zliobaitis, Pechenizkiy M, Gama J. An overview of concept drift applications.
230. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLoS Med*. 2013;10(2).
231. Yancey CC, O'Rourke MC. Emergency Department Triage [Internet]. StatPearls. 2021. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/32491515>
232. Harrell FE. Regression modeling strategies : with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001. (Springer series in statistics).
233. Royston P, Sauerbrei W. Multivariable Model-Building: A Pragmatic Approach to Regression Analysis based on Fractional Polynomials for Modelling Continuous Variables [Internet]. Multivariable Model-Building: A Pragmatic Approach to Regression Analysis based on Fractional Polynomials for Modelling Continuous Variables. Wiley; 2008 [cited 2021 Dec 10]. 1–303 p. Available from: <https://www.wiley.com/en-us/Multivariable+Model+Building%3A+A+Pragmatic+Approach+to+Regression+Analysis+based+on+Fractional+Polynomials+for+Modelling+Continuous+Variables-p-9780470028421>
234. Hastie T, Tibshirani R, Tibshirani R, Tibshirani R, Tibshirani R, Tibshirani R. Statistical Learning with Sparsity The Lasso and Generalizations Statistical Learning with Sparsity.
235. Zou H, Hastie T. Regularization and variable selection via the elastic net. Vol. 67, *J. R. Statist. Soc. B*. 2005.
236. Tibshirani R. Regression shrinkage and selection via the lasso: A retrospective. *J R Stat Soc Ser B Stat Methodol*. 2011 Jun;73(3):273–82.
237. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Stat Methodol* [Internet]. 2005 Apr 1 [cited 2021 Dec 10];67(2):301–20. Available from:

- <https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9868.2005.00503.x>
238. Van Houwelingen JC. Shrinkage and penalized likelihood as methods to improve predictive accuracy. Vol. 55, *Statistica Neerlandica*. Blackwell Publishing Ltd; 2001. p. 17–34.
 239. Kartalopoulos S V. Artificial Neural Networks: Concepts. In: *Understanding Neural Networks and Fuzzy Logic*. IEEE; 2010.
 240. Rebala G, Ravi A, Churiwala S. *An Introduction to Machine Learning* [Internet]. Cham: Springer International Publishing; 2019. Available from: <http://link.springer.com/10.1007/978-3-030-15729-6>
 241. Aggarwal CC. *Data classification : algorithms and applications*. Bosa Roca, FL: Bosa Roca, FL : CRC Press, 2015; 2015.
 242. Breiman L. Arcing Classifiers. *Ann Stat*. 1998;26(3):801–24.
 243. Donges N. Gradient Descent: A Quick, Simple Introduction | Built In [Internet]. 2021 [cited 2021 Dec 15]. Available from: <https://builtin.com/data-science/gradient-descent>
 244. Jerome H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *Ann Stat* [Internet]. 2001 [cited 2021 Dec 15];29(5):1189–232. Available from: <https://www.jstor.org/stable/2699986>
 245. Mason L, Baxter J, Bartlett P, Frean M. Boosting algorithms as gradient descent. *Adv Neural Inf Process Syst*. 2000;512–8.
 246. Hastie, Trevor, Tibshirani, Robert, Friedman J. *The Elements of Statistical Learning The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition* [Internet]. Springer series in statistics. 2009 [cited 2021 Dec 15]. 282 p. Available from: <http://www.worldcat.org/oclc/405547558%5CnHastie, Tibshirani et al - The elements of statistical learning.pdf%5Cnhttp://www.springer.com.libproxyl.nus.edu.sg/statistics/statistical+theory+and+methods/book/978-0-387-84857-0%5Cnhttp://statweb.stanford.edu/~>
 247. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. In: *ACM International Conference Proceeding Series* [Internet]. 2006 [cited 2021 Dec 15]. p. 161–8. Available from: www.cs.cornell.edu
 248. NHS. NHS England » About NHS England [Internet]. <https://www.england.nhs.uk/about/>. 2019 [cited 2022 Apr 26]. Available from: <https://www.england.nhs.uk/urgent-emergency-care/nhs-111/accessing-nhs-111/>
 249. Sedgwick P. Statistical question: Incidence rate ratio [Internet]. Vol. 341, *BMJ (Online)*. British Medical Journal Publishing Group; 2010 [cited 2022 Apr 28]. p. 561. Available from: <https://www.bmj.com/content/341/bmj.c4804>

Chapter 14

The Appendices

Appendix A: Licenses and permissions

A1: WHO License

WORLD HEALTH ORGANIZATION (WHO)

Non-exclusive licence to use selected WHO published materials

You submitted a request, through WHO's online platform, for permission to reproduce certain WHO copyrighted material (the "Licensed Materials"). This is a legal agreement (the "Agreement") between you and WHO, granting you a licence to use the Licensed Materials subject to the terms and conditions herein.

Read this Agreement in its entirety before using the Licensed Materials.

By using the Licensed Materials, you enter into, and agree to be bound by, this Agreement. This licence is granted only for original materials belonging to WHO.

If you enter into this Agreement on behalf of an organization, by using the Licensed Materials you confirm (represent and warrant) that you are authorized by your organization to enter into this Agreement on the organization's behalf. In such a case, the terms "you" and "your" in this Agreement refer to, and this Agreement applies to, the organization.

WHO grants this licence to you based on the representations and warranties you made in the licence request you submitted through WHO's online platform. If any of those representations and/or warranties are or become false or inaccurate, this licence agreement shall automatically terminate with immediate effect, without prejudice to any other remedies which WHO may have.

If you have questions regarding this Agreement, please contact permissions@who.int.

1. Licence. Subject to the terms and Conditions of this Agreement, WHO grants to you a worldwide, royalty free, non-transferable, non-sublicensable, non-exclusive licence to use, reproduce, publish, and display the Licensed Materials in the manner indicated in the Permissions Request Form you submitted to WHO (the "Licensed Use"). This licence is limited to the current edition of the Licensed Materials. Future editions or a different use of the Licensed Materials will require additional permission from WHO.

2. Retained Rights. Copyright in the Licensed Materials remains vested in WHO, and WHO retains all rights not specifically granted under this Agreement. Furthermore, the rights granted under this permission shall not be transferred or assigned.

3. Mandatory Acknowledgement. You must make suitable acknowledgement of WHO, either as a footnote or in a reference, as follows:

"Reproduced with permission of the World Health Organization + url link of the material"

Translations of the Licensed Materials should be attributed as follows:

"Translated with permission of the World Health Organization, from url link of the material"

4. Altering or Modifying the Licensed Materials. The Licensed Materials shall be faithfully reproduced in their entirety including the logo of WHO because it is an integral part of the Work. If Licensee finds it necessary to translate the Material, WHO logo cannot be used. If Licensee so wishes, the PDF of the translation can be provided to WHO, with permission 1) to make the PDF available on WHO web site and 2) to allow the use of the PDF by third parties for non-commercial purposes.

5. Appropriate and Prohibited Uses. You must use the Licensed Materials in a factual and appropriate context. You may not use the Licensed Materials in association with any product marketing, promotional, or commercial activities, including, without limitation, in advertisements, product brochures, company-sponsored web sites, annual reports, or other non-educational publications or distributions.

6. No WHO endorsement. You shall not state or imply that WHO endorses or is affiliated with your publication or the Licensed Use, or that WHO endorses any entity, organization, company, or product.
7. No use of the WHO logo. In no case shall you use the WHO name or emblem, or any abbreviation thereof. Notwithstanding the foregoing, if the WHO name and/or emblem appear as an integral part of the Licensed Materials (e.g. on the poster or checklist) you may use the name and/or emblem in your use of the Licensed Materials, provided the name and/or logo is not used separately from the Licensed Materials. If Licensee finds it necessary to adapt the Material to its own facility's needs, WHO logo cannot be used.
8. No Warranties by WHO. All reasonable precautions have been taken by WHO to verify the information contained in the Licensed Materials. However, WHO provides the Licensed Materials to you without warranty of any kind, either expressed or implied, and you are entirely responsible for your use of the Licensed Materials. In no event shall WHO be liable for damages arising from your use of the Licensed Materials.
9. Your Indemnification of WHO. You agree to indemnify WHO for, and hold WHO harmless against, any claim for damages, losses, and/or any costs, including attorneys' fees, arising in any manner whatsoever from your use of the Licensed Materials or for your breach of any of the terms of this Agreement.
10. Termination. The licence and the rights granted under this Agreement shall terminate automatically upon any breach by you of the terms of this Agreement. Further, WHO may terminate this licence at any time with immediate effect for any reason by written notice to you.
11. Entire Agreement, Amendment. This Agreement is the entire agreement between you and WHO with respect to its subject matter. WHO is not bound by any additional terms that may appear in any communication from you. This Agreement may only be amended by mutual written agreement of you and WHO.
12. Headings. Paragraph headings in this Agreement are for reference only.
13. Dispute resolution. Any dispute relating to the interpretation or application of this Agreement shall, unless amicably settled, be subject to conciliation. In the

event of failure of the latter, the dispute shall be settled by arbitration. The arbitration shall be conducted in accordance with the modalities to be agreed upon by the parties or, in the absence of agreement, with the rules of arbitration of the International Chamber of Commerce. The parties shall accept the arbitral award as final.

14. Privileges and immunities. Nothing in or relating to this Agreement shall be deemed a waiver of any of the privileges and immunities enjoyed by WHO under national or international law and/or as submitting WHO to any national court jurisdiction.

Online form originally submitted for WHO licensing

From: permissions@who.int <permissions@who.int>

Sent: Wednesday, November 4, 2020 12:55 PM

To: jamie.miles@nhs.net

Cc: permissions <permissions@who.int>

Subject: ID: 366020 Permission request for WHO copyrighted material

Dear Mr Miles

Thank you for your request for permission to reproduce, reprint or translate certain WHO copyrighted material.

Your request ID: 366020 is under review

Please be assured that we are working on your request and will get back to you as soon as we possibly can.

Kind regards,

WHO Permissions team

DataCol Web: Form for requesting permission to reproduce, reprint or translate
WHO copyrighted material

=====

ID: 366020

Section: Contact details

* Title

* Mr

* First name

* Jamie

* Family name

* Miles

* Organization/affiliation

* Yorkshire Ambulance Service / University of Sheffield

* Web site address

* <https://www.sheffield.ac.uk/scharr/people/pgr-students/jamie-miles>

* Type of organization

* University/Academic

* If other, please specify

*

* If STM signatory, please select

*

* Position

* NIHR Clinical Doctoral Research Fellow (Paramedic)

* Telephone

* +44 7557955748

* Address

* Yorkshire Ambulance Service NHS Trust

Springhill 1

Brindley way

Wakefield

WF2 0XQ

* Country

* United Kingdom of Great Britain and Northern Ireland

* Email

* jamie.miles@nhs.net

Section: Information about WHO material to be reproduced

* Full title of WHO material requested

* WHO Emergency Care System Framework Infographic

* Website URL where WHO material is published

* https://www.who.int/emergencycare/emergencycare_infographic/en/

* ISBN / WHO Reference Number

*

* Please select the item(s) to be reproduced

* Entire Document, Posters/Infographics

* Type of reuse

* Dissertation or thesis

* No of item(s) to be reproduced

* 5 items or less

* For each item selected, provide a reference and page number. If entire document, please state "Entire document".

* Entire document

Section: Information about the reuse

* Please provide information on where WHO's material will be used

* It will be used in the PhD thesis only.

* Publishing format

* Print, PDF, Ebook

* Will you be translating?

* No

* If yes, please indicate languages

*

* If web please provide URL / If other, please specify

* Number of copies (if applicable)

* How are you planning to distribute your material and to whom?

* The thesis will be printed and bound for personal copies and final submission.
An electronic form of the thesis will be published on the White Rose repository
and at the University of Sheffield.

* What is your planned publication or distribution date?

* March 2022

* Are you selling your material?

* No

* If yes, please provide additional information

*

* Is the material sponsored or funded by an organisation other than your own?

* No

* If yes, please provide additional information

*

* Will there be any advertising associated with the material?

* No

* If yes, please provide additional information

*

* Subject of interest that most correspond to your request

* Emergency and trauma care

* Additional information about your request

* There will be personal copies of the thesis printed and bound, and also e-copies uploaded to the White Rose repository and the University of Sheffield.

The entire image will be split into three separate images to illustrate each process. Specifically:

WHO Edit 1 - on scene - The picture has been cropped to include only the scene processes. The WHO logo has been moved to the bottom. The key has been moved above the picture. The transport banner has been resized and included at the end of the illustration of the road. No text has been changed or added, no images have been modified beyond movement and resizing.

WHO Edit 2 - transport - The picture has been cropped to include only the transport processes. The WHO logo has been added at the bottom. The key has been added to the top of the picture. The equipment illustrations have been moved down the page. The scene banner has been resized and added to the start of the illustration of the road. No text has been changed or added, no images have been modified beyond movement and resizing.

WHO Edit 3 - Emergency Department - The picture has been cropped to include only the facility processes. The key has been moved to the side of the picture. The inpatient image has been removed. The WHO logo has been added to the bottom. The transport banner has been added to the start of the road illustration. The disposition has been resized (enlarged). No text has been changed or added, no images have been modified beyond movement and resizing.

* Copy of Subject(s) of interest that most correspond to your request

* Approval

* To review

* Latest approval modification

* WHO Department

* ACP, ACT

* Correct WHO URL

* https://www.who.int/emergencycare/emergencycare_infographic/en/

Section: Terms and conditions

* By submitting this request you confirm that you will abide by the terms and conditions if WHO grants you permission.

* I have read and agree with the terms and conditions

Click the following link to access a format view of this record:

http://apps.who.int/datacol/survey.asp?survey_id=258&respondent_id=366020

This email was automatically sent to you by the WHO Intranet Data Collector.

The DataCol can send emails to accounts specified by the Form focalpoint.

A2: Badillo et al. permission

RE: Permission to use image

2 messages

Jamie Miles <j.miles@sheffield.ac.uk>
To: solveig.badillo@roche.com

16 May 2022 at 16:06

Good afternoon Solveig,

I am finalising my PhD thesis, where I used an XGBoost algorithm to train a model to predict hospital need for prehospital patients.

I would like to use an unedited copy of figure three in your article titled 'An introduction to machine learning' to help explain the bias-variance trade-off. I have referenced your work in the text and on the figure, but I am asking for permission to use it. The thesis is likely to be published on the white rose repository. I have attached the image below.

Yours Sincerely

Jamie Miles**Clinical Research Fellow
ACP and Paramedic
Specialty Lead for Health Services Research (NIHR CRN Yorkshire and Humber)****NIHR** | Applied Research Collaboration
Yorkshire and HumberRoom 3030
School of Health and Related Research (SCHARR)
University of Sheffield
S1 4DA**E-mail:** j.miles@sheffield.ac.uk | jamie.miles@nhs.net

* Please note I work clinically on a Monday and may not respond straight away *

Badillo, Solveig <solveig.badillo@roche.com>
To: Jamie Miles <j.miles@sheffield.ac.uk>

18 May 2022 at 10:19

Dear Jamie,

Sure, as long as you reference the figure and the corresponding paper, you can use this image.

Best Regards,
Solveig Badillo
[Quoted text hidden]

--

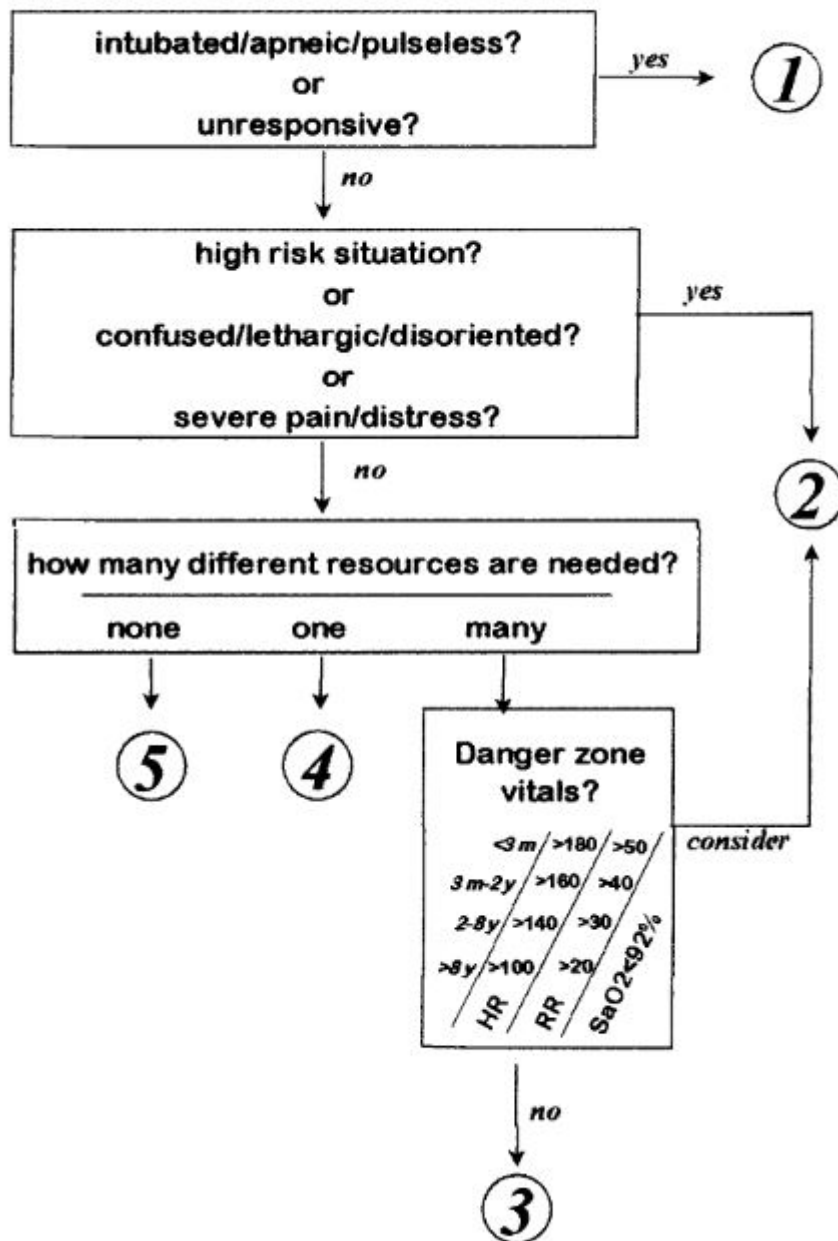
Solveig Badillo, PhD (brain imaging), Senior ScientistRoche Pharmaceutical Research and Early Development (pRED)
Pharmaceutical Sciences -
Predictive Modeling and Data Analytics (PMDA)

Appendix B: International triage scales

*Australasian triage scale*²³¹

| Australasian Triage Scale Category | Response Time | Category Description | Clinical Indicators |
|------------------------------------|-------------------------|---|--|
| Category 1 (RED) | Seen Immediately | Life Threatening Conditions | Cardiac/Respiratory Arrest Immediate risk of airway, respiratory rate < 10/min, Extreme Respiratory Distress. BP less than 80 in adult. Severe shock in child/infant GCS scale less than 9 Prolonged seizure IV overdose Severe behavioral disorder |
| Category 2 (ORANGE) | Seen within 10 minutes | Imminently life threatening, time sensitive treatment needed, or Severe pain. | Airway risk (stridor) Circulatory Compromise (HR less than 50 or greater than 150, Hypotension, severe blood loss, poor perfusion). Chest pain likely cardiac related Suspected sepsis, Febrile Neutropenia, Fever with lethargy Acute Stroke GCS less than 13 Suspected Testicular Torsion High Risk History (toxic ingestion, venomous bite, pain suggesting PE, AAA, ectopic pregnancy. |
| Category 3 (GREEN) | Seen within 30 minutes | Potentially life threatening, situational urgency, or severe pain | Severe Hypertension, Moderate blood loss Moderate Shortness of breath Vomiting Dehydration Seizure (post ictal), Head Injury with LOC (now alert) Physiologically stable suspected sepsis Severe pain Limb injury consisting of limb deformity or severe laceration, altered sensation, absent pulse. Potential child abuse Behavioral/Psychiatric patient very distressed, risk of self-harm, potentially aggressive. |
| Category 4 (BLUE) | Seen within 60 minutes | Potentially serious condition, situational urgency or complex case | Mild Hemorrhage Foreign Body Aspiration without respiratory distress Chest injury without rib pain or respiratory distress Minor head injury without LOC Moderate pain Vomiting or diarrhea without dehydration Inflammation or foreign body in eye without vision changes Minor limb trauma (ankle sprain, fracture, uncomplicated laceration with normal vital signs) Swollen, erythematous joint Semi Urgent mental health problems with no immediate risk to personnel. |
| Category 5 (white) | Seen within 120 minutes | Less urgent or Clinical-Administrative problems | Minimal pain with no risk factors Low risk history Minor symptoms of illness Minor symptoms of low risk condition Abrasions or minor laceration Scheduled revisit Immunizations Patient with chronic psychiatric symptoms in social crisis. |

ESI Triage Algorithm



Appendix C: Supplementary information for the systematic review

Search term identification

| Key terms | Alternative terms | Subject headings |
|---------------------------|---|--|
| Machine Learning | "Machine Learn*" ML "Artificial Intelligence" AI Deriv* Valid* | Supervised Machine Learning/ Unsupervised Machine Learning/ Machine Learning/ Algorithms/ Logistic Models/ |
| Clinical triage | "Clinic* Triag*" Triag* "Clinic* Classif*" Classif* "Clinic* sort*" Sort" | Clinical Triage/ |
| Patient acuity | "Patient severity" Prognos* Predict* Rule* | Patient Acuity/ Ingui filter* Haynes broad filter** |
| Urgent and Emergency care | "*Emergenc* care*" Emergenc* "Urgent Care" Urgent Prehospital Pre-hospital Pre hospital "Emergency Department" ED "Accident and Emergency" "Accident & Emergency" | Emergency Medical Services/ Emergency Medicine/ Emergency Treatment/ Emergencies/ Ambulatory care/ Ambulances/ Emergency Medical Tags/ Emergency Medical Technicians/ |

| | | |
|----------------------|---|---|
| | A&E Ambulanc* “Ambulanc* Serv* EMS “Emergency Medical Service” | Emergency Responders/ Emergency Service, Hospital/ |
| *Ingui filer | (Validat* OR Predict*.ti. OR Rule*) OR (Predict* AND (Outcome* OR Risk* OR Model*)) OR ((History OR Variable* OR Criteria OR Scor* OR Characteristic* OR Finding* OR Factor*) AND (Predict* OR Model* OR Decision* OR Identif* OR Prognos*)) OR (Decision* AND (Model* OR Clinical* OR Logistic Models/)) OR (Prognostic AND (History OR Variable* OR Criteria OR Scor* OR Characteristic* OR Finding* OR Factor* OR Model*)) | |
| *Haynes broad filter | (Predict*[tiab] OR Predictive value of tests[mh] OR Scor*[tiab] OR Observ*[tiab] OR Observer variation[mh]) | |

Search strategy

Last 10 years – Clinical contexts and computer capabilities are both rapidly changing industries and thus older studies have a higher risk of being void or outdated. Can review this.

English only – Don’t have access to interpretation services and therefore including foreign language studies could lead to misinterpretation.

MEDLINE via Ovid 257 results

(Machine Learn* OR ML OR Artificial Intelligence OR AI OR Big data OR Gaussian process OR Cross-validation OR Cross validation OR Crossvalidation OR Regularized logistic OR Linear discriminant analysis OR LDA OR Random forest OR Na#ve Bayes* OR Least Absolute selection shrinkage operator OR elastic net OR LASSO OR RVM OR relevance vector machine OR pattern recognition OR Computational Intelligence OR Computational Intelligences OR Machine Intelligence OR Knowledge Representation OR Knowledge Representations OR

support vector OR SVM OR pattern classification OR Supervised Machine Learning/ OR Unsupervised Machine Learning/ OR Machine Learning/ OR Algorithms/ OR Logistic Models/)

AND

(Clinic* Triag* OR Triag* OR Clinic* Classif* OR Classif* OR Clinic* sort* OR Sort* OR electro* triag* OR Digital triag* OR Clinical Triage/)

AND

(Patient severity OR Prognos* OR Predict* OR Rule* OR Patient Acuity/)

AND

(Emergenc* care* OR Urgent and Emergency Care OR Urgent & Emergency Care OR UEC OR Emergenc* OR Urgent Care OR Urgent OR Prehospital OR Pre-hospita OR Pre hospital OR Emergency Department OR ED OR Accident and Emergency OR Accident & Emergency OR A&E OR Ambulanc* OR Ambulanc* Serv* OR EMS OR Emergency Medical Service OR Emergency Medical Services/ OR Emergency Medicine/ OR Emergency Treatment/ OR Emergencies/ OR Ambulatory care/ OR Ambulances/ OR Emergency Medical Tags/ OR Emergency Medical Technicians/ OR Emergency Responders/ OR Emergency Service, Hospital/)

CINAHL via EBSCO 298 results

(Machine Learn* OR ML OR Artificial Intelligence OR AI OR Big data OR Gaussian process OR Cross-validation OR Cross validation OR Crossvalidation OR Regularized logistic OR Linear discriminant analysis OR LDA OR Random forest OR Na#ve Bayes* OR Least Absolute selection shrinkage operator OR elastic net OR LASSO OR RVM OR relevance vector machine OR pattern recognition OR Computational Intelligence OR Computational Intelligences OR Machine Intelligence OR Knowledge Representation OR Knowledge Representations OR support vector OR SVM OR pattern classification OR Supervised Machine Learning/ OR Unsupervised Machine Learning/ OR Machine Learning/ OR Algorithms/ OR Logistic Models/)

AND

(Clinic* Triag* OR Triag* OR Clinic* Classif* OR Classif* OR Clinic* sort* OR Sort* OR electro* triag* OR Digital triag* OR Clinical Triage/)

AND

(Patient severity OR Prognos* OR Predict* OR Rule* OR Patient Acuity/)

AND

(Emergenc* care* OR Urgent and Emergency Care OR Urgent & Emergency Care OR UEC OR Emergenc* OR Urgent Care OR Urgent OR Prehospital OR Pre-hospita OR Pre hospital OR Emergency Department OR “ED” OR Accident and Emergency OR Accident & Emergency OR “A&E” OR Ambulanc* OR Ambulanc* Serv* OR “EMS” OR Emergency Medical Service OR Emergency Medical Services/ OR Emergency Medicine/ OR Emergency Treatment/ OR Emergencies/ OR Ambulatory care/ OR Ambulances/ OR Emergency Medical Tags/ OR Emergency Medical Technicians/ OR Emergency Responders/ OR Emergency Service, Hospital/)

PubMed 150 results

(Machine Learn* OR ML OR Artificial Intelligence OR AI OR Big data OR Gaussian process OR Cross-validation OR Cross validation OR Crossvalidation OR Regularized logistic OR Linear discriminant analysis OR LDA OR Random forest OR Na#ve Bayes* OR Least Absolute selection shrinkage operator OR elastic net OR LASSO OR RVM OR relevance vector machine OR pattern recognition OR Computational Intelligence OR Computational Intelligences OR Machine Intelligence OR Knowledge Representation OR Knowledge Representations OR support vector OR SVM OR pattern classification OR Supervised Machine Learning[mh] OR Unsupervised Machine Learning[mh] OR Machine Learning[mh] OR Algorithms[mh] OR Logistic Models[mh])

AND

(Clinic* Triag* OR Triag* OR Clinic* Classif* OR Classif* OR Clinic* sort* OR Sort* OR electro* triag* OR Digital triag* OR Clinical Triage[mh])

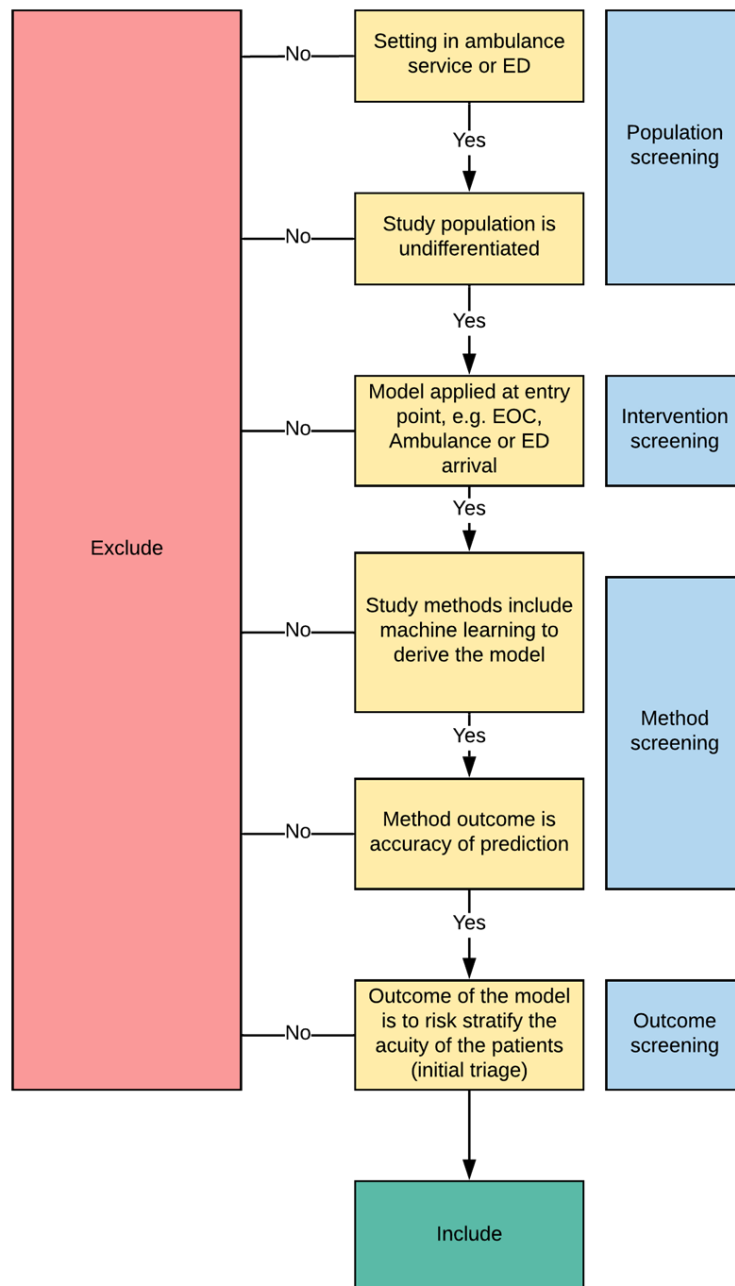
AND

(Patient severity OR Prognos* OR Predict* OR Rule* OR Patient Acuity[mh])

AND

(Emergenc* care* OR Urgent and Emergency Care OR Urgent & Emergency Care OR UEC OR Emergenc* OR Urgent Care OR Urgent OR Prehospital OR Pre-hospita OR Pre hospital OR Emergency Department OR “ED” OR Accident and Emergency OR Accident & Emergency OR “A&E” OR Ambulanc* OR Ambulanc* Serv* OR “EMS” OR Emergency Medical Service OR Emergency Medical Services[mh] OR Emergency Medicine[mh] OR Emergency Treatment[mh] OR Emergencies[mh] OR Ambulatory care[mh] OR Ambulances[mh] OR Emergency Medical Tags[mh] OR Emergency Medical Technicians[mh] OR Emergency Responders[mh] OR Emergency Service, Hospital[mh])
782 results after duplicates removed.

Visual schematic of inclusion criteria for systematic review



Appendix D: Technical elaboration on algorithm methods

Logistic regression

Simply, to make a prediction on a new individual (y_i), the dependent variable is regressed on independent predictors (x_i). These could be age, gender, heart rate etc. Each x_i is multiplied by its β_i coefficient that is derived from its linear association with the outcome. The interpretation of the β_i is that for each single unit increase in x_i , would lead to a β_i increase in y_i , hence it is the slope of a linear equation. The other coefficient is, β_0 which represents the value of y_i , when x_i is zero (for continuous variables) or the reference category in discrete variables. Hence, the β_0 is the intercept. In logistic regression, a logit link function is applied to scale the results of the equation between 0 and 1. The full equation can be found below. Because logistic regression is calculating the expected value of y , which can then be used to make a classification given an assumption of a cut off; the result of the equation is classifying y_i into 0 or 1.

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + (\beta_1x_1) + (\beta_2x_2) + \dots (\beta_ix_i)$$

The benefit of using logistic regression is it is an explanatory model as well as predictive. This means that the whole final model can be statistically represented, and each individual variable can have its association with the outcome. Further benefits of using this method is its flexibility and can incorporate different types of data, non-linear transformations (such as fractional polynomials) and interaction terms.¹²⁸ However, one of the drawbacks is that it assumes a constant (linearity) in the coefficients. The dependent variable will always have the same relationship with the independent variable. Fractional polynomials or restricted cubic splines can be used to model non-linearity; however, these are used with the variables and not the beta coefficients themselves.^{232,233} For example, in the equation below, a fractional polynomial to the power of two has been added to the value of x_1 .

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + (\beta_1x_1^2) + (\beta_2x_2) + \dots (\beta_ix_i)$$

In logistic regression, the model to fit must be specified, which means ‘cherry picking’ variables. Methods such as stepwise selection methods can be used for feature selection, however using regularisation within the regression equation can reduce model overfitting, help with any multicollinearity and incorporate feature selection.

Penalised regression

Regularisation aims to reduce variance by introducing a bias (in the form of a penalty term) and therefore reducing the Mean Squared Error (MSE).^{234,235} It operates by adding a penalty term to estimating equations during model development, which is different to other shrinkage techniques such as uniform or heuristic shrinkage. These add a penalty into the regression equation itself, e.g.

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + S((\beta_1x_1) + (\beta_2x_2) + \dots (\beta_ix_i))$$

Where S is a universal shrinkage factor. In regularisation, the penalty is added to the estimation of model fit. Less model error is likely to increase correct predictions in new individuals. In order to identify the optimum values of β_0 (the model intercept) and β_i (the model slope), the residual sum of squares (RSS) is calculated. This is the variance between the fitted regression line and the plotted points where individual values of the independent variables intersect with the values of the dependent. Therefore, the RSS can be calculated using the following equation:

$$RSS = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_{ij} x_j \right)^2$$

The predicted value of y_i is calculated by deducting the slope (β_{ij}), multiplied by the independent variable (x_j) from the intercept (β_0). This predicted value is then deducted from the actual value (y_i) to give the errors for an individual point. The results of all the errors are then added together and squared to give the RSS.

Regularisation penalty terms are added onto the calculation of the RSS. Ridge regression also known as the λ_1 penalty, shrinks each beta coefficient towards 0,

according to its relationship with the outcome by summing the absolute values of the beta 1 coefficients. This differs to the Least Absolute Shrinkage and Selector Operator (LASSO) or λ_2 , which will shrink variables to zero, effectively eliminating them from the model by summing the squares of the beta coefficients.²³⁵ Each penalty term can be used individually as a tuning parameter (hyper-parameter) and both penalties have been included in the below equations:^{236,237}

$$RSS(\beta_{L1}) = \sum_{i=1}^n \left(Y_i - \beta_0 - \sum_{j=1}^p \beta_{ij} x_j \right)^2 + \lambda_1 \overbrace{\sum_{j=1}^p |\beta_j|}^{\text{Ridge}}$$

$$RSS(\beta_{L2}) = \sum_{i=1}^n \left(Y_i - \beta_0 - \sum_{j=1}^p \beta_{ij} x_j \right)^2 + \lambda_2 \overbrace{\sum_{j=1}^p \beta_j^2}^{\text{LASSO}}$$

It is possible to keep both terms in the equation, however when used in combination the two parameters are dependent on each other, which makes optimisation through cross-validation difficult as in effect, there is a hyper-hyper-parameter. To counter this, an alpha term (α) is introduced into the equation, as:^{236,237}

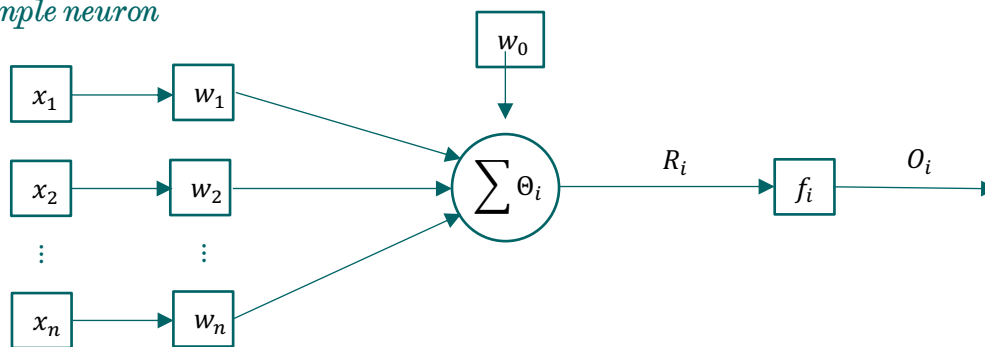
$$RSS(\beta_{L1\&2}) = \sum_{i=1}^n \left(Y_i - \beta_0 - \sum_{j=1}^p \beta_{ij} x_j \right)^2 + \lambda \left(\frac{(1-\alpha)}{2} \sum_{j=1}^p \beta_j^2 + \alpha \sum_{j=1}^p |\beta_j| \right)$$

The alpha is the ratio between the both penalty terms.²³⁴ When α is equal to 1, the LASSO penalty is being used, whereas when the α is 0, it is the ridge regression. Cross-validation can be used to identify the optimum α and λ . There are limitations with regularisation though. The lambda and alpha penalties are estimated with large uncertainty and final model predictors may not be stable.²³⁸ A limitation with the GLM is the model algorithm does not directly handle missing data, so this needs to be resolved in the data preparation phase. The options are limited but depending on the type of missing data, it would either be to delete columns or rows, or to impute the values.

The Generalised Non-Linear Model (GNLM)

A good example of the generalised non-linear model (and used with great success in the studies identified in the systematic review) is the artificial neural network (ANN) and can be described as an extension of the GLM.¹²⁸ The figure below is based on the basic components of a neuron.²³⁹ Many neurons are often needed to solve complex problems.

A simple neuron



In a neuron, each input variable (x_i) is weighted by multiplying the value by the associated w_i . A bias term (w_0) is then added to each weighted input before a sum function is applied. The weighted sum is then measured against a predetermined activation threshold (θ). If it meets the threshold, the value (R_i) is passed to an activation function (f_i). This is a non-linear transformation, often resembling a sigmoid curve. The non-linear function then provides an output between 0 and 1. Figure 8 in the main thesis shows an illustrative example of a neural network.

Hyperparameters within neural networks includes adjusting the number of hidden layers and the weights and biases of each node. These are often deduced through trial and error. Backpropagation can be used to re-evaluate the weights of the variables. However, too many iterations can lead to overtraining and should be avoided.^{125,128,240} The neural network can handle non-linearity effectively through the activation function, however the algorithm design struggles to overcome multicollinearity without undergoing procedures such as PCA in the data preparation stage. This means that the management of

collineated variables comes at the expense of model interpretability.¹³⁶ The algorithm by the nature of it being an extension of the GLM can inherit the same issues with missing data. Solutions that overcome this for the neural network include modelling the uncertainty of attributes with probability density functions.¹³⁷ Furthermore, a neural network can be computationally greedy, and it can take many hidden layers in order to create an accurate model. There are also limitations in how neural networks handle structured data. In classification problems that use unstructured data such as images or sounds, canonical architectures translate these forms into meaningful inputs for neural networks. However, this is a difficult task with structured, tabular data. As a result, decision trees have dominated competitions that have required prediction modelling using this type of data.¹³⁸

Tree-based methods

The basic components of a decision tree are nodes. These simple filters take an input and split it into two or more outputs based on the split criteria. Trees include a root node, which is the first or starting node. Figure 9 in the main thesis provides an illustrative example of a simple decision tree along with an explanation. The way a decision tree decides on which variables to use as the parent node, or any internal node (internal nodes are any node situated between the root and the leaf) is not random but calculated as a measure of entropy and information gain.

Claude Shannon's entropy model measures the amount of impurity of the elements in a dataset. In decision tree modelling, the objective is to create pure groups in the leaf nodes. As a hyperbole, a perfect binary classification decision tree can split a population into two classes with probabilities of each being pure (both being 1). In effect, a perfectly discriminative model. Any new participants that the tree is applied to will be perfectly classified (on the assumption there is no over fitting to the training data). The perfect classification tree rarely exists

when applied to real world problems, but the objective remains to reduce impurity as new child nodes and branches are created. In order to achieve this objective, the correct nodes and split criteria that will eliminate the most impurity in the resultant child nodes (or leaves) need to be identified. Shannon's model of entropy is the weighted sum of the logs of the probabilities of each possible outcome when making a random selection from a set of variables.¹⁴⁰ It is the logarithm of the probability because smaller values of probability need to be represented by large numbers and larger values of probability need to be represented by smaller numbers. Using the logarithm of the probabilities to the base of 2 achieves this goal. The large numbers will be negative; however, this can be rectified in the equation for Shannon's model of entropy by placing a minus sign at the start:¹⁴⁰

$$H(t) = - \sum_{i=1}^l (P(t = i) \times \log_2(P(t = i)))$$

Here, the entropy (H) of a variable (t) in the dataset is calculated by summing the probability that the value of the variable is i , multiplied by the \log_2 of this probability. As an imagined example, a dataset contains ten participants measured over a single variable (eye colour), two participants have blue eyes, three have green and five have brown. The entropy would be calculated as:

$$\begin{aligned} H(\text{Eye}) &= - \sum_{l \in (\text{blue}, \text{green}, \text{brown})} P(\text{Eye} = l) \times \log_2(P(\text{Eye} = l)) \\ &= - \left((P(\text{blue}) \times \log_2(P(\text{blue}))) + (P(\text{green}) \times \log_2(P(\text{green}))) + (P(\text{brown}) \times \log_2(P(\text{brown}))) \right) \\ &= - \left(\left(\frac{2}{10} \times \log_2\left(\frac{2}{10}\right) \right) + \left(\frac{3}{10} \times \log_2\left(\frac{3}{10}\right) \right) + \left(\frac{5}{10} \times \log_2\left(\frac{5}{10}\right) \right) \right) \\ &= - \left((0.2 \times -2.32) + (0.3 \times -1.37) + (0.5 \times -1) \right) \\ &= \mathbf{1.375} \end{aligned}$$

Entropy can be used to calculate the information gain for a variable. The variable in a dataset with the most information gain would lead to the most purity in the child nodes. It is therefore beneficial to use this variable as the root node in a simple decision tree to create a parsimonious model that minimises

computational expense. To calculate information gain, it is required to work out the entropy for the whole dataset, then remaining entropy after a split for each variable, and then deduct the latter from the former. The table below is a fictitious structured tabular dataset that closely resembles the study dataset in this thesis.

Fictitious data set for information gain example

| Abnormal observations | Intervention required | Urgency |
|-----------------------|-----------------------|-------------------|
| No | Yes | Non urgent (nurg) |
| Yes | No | Urgent (urg) |
| Yes | No | Urgent |
| No | No | Non urgent |
| No | No | Non urgent |
| Yes | Yes | Urgent |

Step one is to calculate the entropy for the whole dataset, note the (\mathcal{D}) represents the whole dataset:¹⁴⁰

$$\begin{aligned}
 H(t, \mathcal{D}) &= - \sum_{l \in \text{levels}(t)} P(t = l) \times \log_2(P(t = l)) \\
 &= H(\text{urg}, \mathcal{D}) = - \left((P(t = \text{urg}) \times \log_2(P(t = \text{urg}))) \right. \\
 &\quad \left. + (P(t = \text{nurg}) \times \log_2(P(t = \text{nurg}))) \right) \\
 &= H(\text{urg}, \mathcal{D}) = -((0.5 \times -1) + (0.5 \times -1))
 \end{aligned}$$

Dataset entropy = 1

Then the entropy that remains (*rem*) after partitioning on each variable is calculated. In the following equations, (d) represents a variable found within dataset (\mathcal{D})¹⁴⁰

$$rem(d, \mathcal{D}) = \sum_{l \in levels(d)} \overbrace{\frac{|\mathcal{D}_{d=l}|}{\mathcal{D}}}^{weighting} \times \overbrace{H(t, \mathcal{D}_{d=l})}^{entropy\ of\ partition\ \mathcal{D}_{d=l}}$$

For the variable ‘intervention required’ (shortened to ‘intereq’), this would be:

$$\begin{aligned} & rem(intereq, \mathcal{D}) \\ &= \left(\frac{|\mathcal{D}_{intereq=yes}|}{\mathcal{D}} \times H(t, \mathcal{D}_{intereq=yes}) \right) \\ &+ \left(\frac{|\mathcal{D}_{intereq=no}|}{\mathcal{D}} \times H(t, \mathcal{D}_{intereq=no}) \right) \\ &= \left(\frac{2}{6} \times \left(- \sum_{l \in (urg, nurg)} P(t=l) \times \log_2(P(t=l)) \right) \right) + \left(\frac{4}{6} \times \left(- \sum_{l \in (urg, nurg)} P(t=l) \times \log_2(P(t=l)) \right) \right) \\ &= \left(\frac{2}{6} \times \left(- \left(\left(\frac{1}{2} \times \log_2\left(\frac{1}{2}\right) \right) + \left(\frac{1}{2} \times \log_2\left(\frac{1}{2}\right) \right) \right) \right) \right) + \left(\frac{4}{6} \times \left(- \left(\left(\frac{2}{4} \times \log_2\left(\frac{2}{4}\right) \right) + \left(\frac{2}{4} \times \log_2\left(\frac{2}{4}\right) \right) \right) \right) \right) \end{aligned}$$

remaining entropy for intereq after splitting = 1

For the variable abnormal observations (shortened to abobs), it would be:

$$\begin{aligned} & rem(abobs, \mathcal{D}) \\ &= \left(\frac{|\mathcal{D}_{abobs=yes}|}{\mathcal{D}} \times H(t, \mathcal{D}_{abobs=yes}) \right) \\ &+ \left(\frac{|\mathcal{D}_{abobs=no}|}{\mathcal{D}} \times H(t, \mathcal{D}_{abobs=no}) \right) \\ &= \left(\frac{3}{6} \times \left(- \sum_{l \in (urg, nurg)} P(t=l) \times \log_2(P(t=l)) \right) \right) + \left(\frac{3}{6} \times \left(- \sum_{l \in (urg, nurg)} P(t=l) \times \log_2(P(t=l)) \right) \right) \\ &= \left(\frac{3}{6} \times \left(- \left(\left(\frac{3}{3} \times \log_2\left(\frac{3}{3}\right) \right) + \left(\frac{0}{3} \times \log_2\left(\frac{0}{3}\right) \right) \right) \right) \right) + \left(\frac{3}{6} \times \left(- \left(\left(\frac{0}{3} \times \log_2\left(\frac{0}{3}\right) \right) + \left(\frac{3}{3} \times \log_2\left(\frac{3}{3}\right) \right) \right) \right) \right) \end{aligned}$$

remaining entropy for abobs after splitting = 0

The final step in calculating information gain is to deduct the remaining entropy from the total in the data set. Therefore, the information gain for intereq is 1-1 = 0, and for abobs is 1-0 = 1. The variable ‘intervention required’ gained no information (with a score of 0), whereas the variable ‘abnormal observations’ was

able to match the entropy of the full dataset (with a score of 1) and therefore could diminish all impurity in the tree by splitting on this single variable.¹⁴⁰ These are both extreme examples and an information gain of 1 is not only rare in practice, but also invalidates the requirement of a decision tree model. Any variable with a gain of 1 could just be used as an individual predictor on its own. There are alternatives to using information gain as the split criterion. A popular alternative is using the Gini index, which is like information gain, but instead of taking the logarithm of the probability, it is squared as:¹⁴⁰

$$Gini(t, \mathcal{D}) = 1 - \sum_{l \in \text{levels}(t)} (t = l)^2$$

Different splitting criteria can result in different variables to be selected as the root node. There is no consensus on the correct splitting criteria, and it is recommended practice to try different approaches and assess model accuracy. Studies have shown that there is often no statistically significant differences in accuracy between techniques, however there is sometimes a difference in training time.^{140,241} Information gain is an important concept of a decision tree, and the calculations are intuitive for binary and categorical variables. However, a limitation with decision tree models is that they cannot handle continuous variables in their continuous form. Instead, a threshold needs to be defined within the variables that dichotomises them. The process in which this occurs is to first order all the values in the continuous variable from smallest to largest (can also be largest to smallest). The table below is another fictitious example where blood pressure has been ordered already. Then, adjacent values of the continuous values that result in a different outcome are identified. From figure 11, there are two pairs of values where the outcome changes. The first is 90 and 100, the second is 130 and 150. To identify the possible thresholds, take the average of the pair ($90+100/2=95$) and ($130+150/2=140$). The two thresholds for consideration in the dataset are therefore $\geq 95\text{mmHg}$ and $\geq 140\text{mmHg}$. The information gain for these thresholds can now be calculated to identify the best.

Fictitious example for transforming continuous variable

| Systolic Blood Pressure (mmHg) | Urgency |
|---|----------------|
| 70 | Urgent |
| 77 | Urgent |
| 90 | Urgent |
| 100 | Non urgent |
| 120 | Non urgent |
| 121 | Non urgent |
| 130 | Non urgent |
| 150 | Urgent |
| 155 | Urgent |
| 160 | Urgent |

This method is a crude heuristic and may not find the ideal threshold. Other proposed methods include an exact greedy approach, which scans every possible value and measures the gain. This will lead to the true optimal threshold, but can be extremely computationally expensive, especially if the continuous variable has a large range with small increments to scan.

In decision tree modelling, because of the recursive partitioning a categorical variable can only be used once; however, a continuous variable can be used multiple times providing its subsequent use operates at a different threshold to one that's already been used. Decision trees have the advantage of being able to overcome the weaknesses of logistic regression and neural networks such as non-linearity and multicollinearity. However, they have their own limitations. The way a decision tree handles continuous variables is to categorise them as detailed above. This leads to a loss of information and is discouraged in prediction modelling.¹²⁸ Simple decision trees also can generalise beyond the data. For example, if a binary variable is being split into two leaf nodes but all the sample are in the positive class, the tree will automatically assign an outcome for the negative class. This is decided by automatically choosing whichever outcome was the majority at the parent node. Decision trees can also be prone to overfitting to the training data. If there are no safeguards in place, the model will keep splitting until there is either no more variables left to select, or there is no more information gained from splitting. This makes a greedy algorithm, and as the tree becomes deeper with more splits, it will fit the training data with less error but will increase the error in the test set. There are methods to reduce overfitting that are effective and easy to implement. One method is tree pruning.

Pruning is the removal of subbranches and replacement with leaf nodes. This can be achieved either before the modelling (pre-pruning) or after the tree is developed (post-pruning). When pre-pruning, early stopping criteria are specified. The criteria can be based on a functional limiter or a statistical test. The limiters include identifying a minimum information gain to activate a split or using the number of instances in a partition. The statistical test that is commonly used is the χ_2 (chi-squared) test to determine whether each partition is important to the overall tree. Post-pruning measures often evaluate the error rate between a training set and test set when the sub-branch is included vs excluded.¹⁴⁰ Even with the addition of pruning, a single tree classifier is considered a weak learner on its own and can easily be overfit to the dataset used

to train it. Modern approaches use ensemble methods which constitute a whole forest of decision trees.

Ensemble decision tree models

Ensemble decision tree models were described in detail by Leo Breiman in the 1990's and are an extension of the simple recursive partitioning tree detailed above.¹⁴¹ Ensembles create many trees that are individually diverse in their decision making. These diverse models are aggregated in a voting system to make a final prediction. The two well-known techniques of creating a forest of trees are known as bagging and boosting.

Bagging is an abbreviation of bootstrap aggregation. The purpose of bootstrap aggregation is to prevent a forest of trees from group decision making. This would occur if all the trees were created using the same data. In bagging, each tree is technically built on a different dataset. A limitation with even a simple ensemble model using tree bagging is collinearity between strong predictors. Bagging will de-correlate to an extent by deriving the model on different training sets. However, strongly correlated predictors will consistently yield the highest entropy regardless of the sample space within the ensemble. This is because bootstrapping creates new datasets, but the distributions are largely the same over the variables themselves. If there is a noisy variable, it will be noisy in all the trees and will encourage the group decision making in the forest. To counter this, a common method is to build the bagging ensemble model but use a random subset of predictor variables in each classifier. This is known as random forest modelling. If a dataset had 100 variables, the architect of a random forest model could specify how many variables out of the 100 should be randomly selected each time for inclusion in the tree. This does not necessarily mean all randomly selected variables will be included in the tree, but it does mean that strongly correlated variables with the outcome will not be in every tree model.^{125,140,141}

In boosting, trees are iteratively created with subsequent trees adjusting a penalty on misclassification. First described by Freund and Schapire in 1996, boosting algorithms operate by weighting the initial dataset ($w_i \geq 0$) to 1 divided by the number of instances in the dataset. The weighting for each instance forms a distribution for resampling in the next tree. In the subsequent resampling, an instance may be replicated multiple times, and this is proportionate to the weighting of the previous tree. The decision on how to adjust the weights is statistical. Error (ϵ) is calculated by summing the weights of the misclassified samples. Then the weights for misclassified samples are increased:¹⁴⁰

$$w[i] \leftarrow w[i] \times \left(\frac{1}{2 \times \epsilon} \right)$$

Whilst the weights for correctly classified instances are decreased:¹⁴⁰

$$w[i] \leftarrow w[i] \times \left(\frac{1}{2 \times (1 - \epsilon)} \right)$$

Trees continue to be created in this iterative way until the maximum error has been removed from the model. In this respect, boosting is operating in the same way as logistic regression and neural networks in that its objective is to minimise a loss function. This can also be described as a form of gradient descent in function space.²⁴² Gradient descent can be likened to walking down into a steep valley. The objective in the analogy is to stand at the lowest point in the valley. The 'perfect' next step is calculated so that it moves the subject closer to the lowest point. The simplified equation for gradient descent can be found here:²⁴³

$$b = a - \gamma \Delta f(a)$$

In the equation b is where the subject needs to step next, and a is where they are currently standing. Gamma (γ) is the weighting factor and the rest ($\Delta f(a)$) is the gradient term which aims to identify the direction of the next step. The term is the negative of the functional derivative (gradient) of the cost function. The length of the step is known as the learning rate. If this is too high, the gradient descent equation will unlikely find the lowest point. This is because the graphical representation of the equation is convex. A learning rate that is too large has the chance of over-shooting the lowest point every time. Conversely, a learning rate

too small will find the lowest point, but it may take longer. Different boosting algorithms have been proposed depending on the loss function that is being minimised. For example, L2boost uses the square error loss whereas Adaboost uses the exponential.¹²⁵ A more advanced mechanism for gradient descent is to incorporate a weak learner into the gradient term. This negates the need for a different algorithm for each loss function as it generalises to them all.^{125,244,245} Boosting algorithms have been described as the best off-the-shelf classifiers in the world.²⁴⁶ This claim has been supported by high quality evidence.²⁴⁷

Appendix E: List of Emergency Departments included in this study

Barnsley Hospital NHS Foundation
Trust
Gawber Rd
Barnsley
S75 2EP
United Kingdom

Pinderfields Hospital
Aberford Rd
Wakefield
WF1 4DG
United Kingdom

St. James's University Hospital
Beckett Street
Leeds
LS9 7TF
United Kingdom

Leeds General Infirmary
Great George Street
Leeds
LS1 3EX
United Kingdom

Harrogate and District NHS
Foundation Trust
Lancaster Park Road
Harrogate
HG2 7SX
United Kingdom

Huddersfield Royal Infirmary
Acre Street Lindley
Huddersfield
HD3 3EA
United Kingdom

Calderdale Royal Hospital
Salterhebble
Halifax
HX3 0PW
United Kingdom

Hull Royal Infirmary
Anlaby Road
Hull
HU3 2JZ
United Kingdom

Rotherham NHS Foundation Trust
Moorgate Road
Rotherham
S60 2UD
United Kingdom

York Teaching Hospital NHS
Foundation Trust
Wigginton Road
York

YO31 8HE
United Kingdom

Airedale General Hospital
Skipton Road Steeton
Keighley
BD20 6TD
United Kingdom

Doncaster Royal Infirmary
Armthorpe Road
Doncaster
DN2 5LT
United Kingdom

Northern General Hospital
Herries Road
Sheffield
S5 7AU
United Kingdom

TS4 3BW

Bradford Royal Infirmary
Smith Lane
Bradford
BD9 6DA
United Kingdom

Dewsbury and District Hospital
Halifax Rd
Dewsbury
WF13 4HS
United Kingdom

Scarborough General Hospital
Woodlands Drive
Scarborough
YO12 6QL
United Kingdom

The James Cook University Hospital
Marton Rd
Middlesbrough

Appendix F: NEWS Score

Reproduced with permission? from the Royal College of Physicians¹⁸⁰

| Physiological parameter | Score | | | | | | |
|--------------------------------|-------|--------|-----------|---------------------|-----------------|-----------------|---------------|
| | 3 | 2 | 1 | 0 | 1 | 2 | 3 |
| Respiration rate (per minute) | ≤8 | | 9–11 | 12–20 | | 21–24 | ≥25 |
| SpO ₂ Scale 1 (%) | ≤91 | 92–93 | 94–95 | ≥96 | | | |
| SpO ₂ Scale 2 (%) | ≤83 | 84–85 | 86–87 | 88–92 ≥93 on air | 93–94 on oxygen | 95–96 on oxygen | ≥97 on oxygen |
| Air or oxygen? | | Oxygen | | Air | | | |
| Systolic blood pressure (mmHg) | ≤90 | 91–100 | 101–110 | 111–219 | | | ≥220 |
| Pulse (per minute) | ≤40 | | 41–50 | 51–90 | 91–110 | 111–130 | ≥131 |
| Consciousness | | | | Alert | | | CVPU |
| Temperature (°C) | ≤35.0 | | 35.1–36.0 | 36.1–38.0 | 38.1–39.0 | ≥39.1 | |

Appendix G: HRA, REC and CAG Approval



Professor Suzanne Mason
University of Sheffield
CURE, ScHARR
30 Regent Street
S1 4DA

Email: approvals@hra.nhs.uk
HCRW.approvals@wales.nhs.uk

06 August 2020

Dear Professor Mason

**HRA and Health and Care
Research Wales (HCRW)
Approval Letter**

Study title: Safety INdEx of Prehospital On Scene Triage (SINEPOST): The derivation and validation of a risk prediction model to support ambulance clinical transport decisions on scene.

IRAS project ID: 260505

Protocol number: YASRD109

REC reference: 19/YH/0360

Sponsor Yorkshire Ambulance Service

I am pleased to confirm that [HRA and Health and Care Research Wales \(HCRW\) Approval](#) has been given for the above referenced study, on the basis described in the application form, protocol, supporting documentation and any clarifications received. You should not expect to receive anything further relating to this application.

Please now work with participating NHS organisations to confirm capacity and capability, in line with the instructions provided in the "Information to support study set up" section towards the end of this letter.

How should I work with participating NHS/HSC organisations in Northern Ireland and Scotland?

HRA and HCRW Approval does not apply to NHS/HSC organisations within Northern Ireland and Scotland.

If you indicated in your IRAS form that you do have participating organisations in either of these devolved administrations, the final document set and the study wide governance report (including this letter) have been sent to the coordinating centre of each participating nation. The relevant national coordinating function/s will contact you as appropriate.

Please see [IRAS Help](#) for information on working with NHS/HSC organisations in Northern Ireland and Scotland.

How should I work with participating non-NHS organisations?

HRA and HCRW Approval does not apply to non-NHS organisations. You should work with your non-NHS organisations to [obtain local agreement](#) in accordance with their procedures.

What are my notification responsibilities during the study?

The standard conditions document "[After Ethical Review – guidance for sponsors and investigators](#)", issued with your REC favourable opinion, gives detailed guidance on reporting expectations for studies, including:

- Registration of research
- Notifying amendments
- Notifying the end of the study

The [HRA website](#) also provides guidance on these topics, and is updated in the light of changes in reporting expectations or procedures.

Who should I contact for further information?

Please do not hesitate to contact me for assistance with this application. My contact details are below.

Your IRAS project ID is **260505**. Please quote this on all correspondence.

Yours sincerely,
Hayley Henderson
Approvals Manager

Email: approvals@hra.nhs.uk

Copy to: Ms Jane Shewan, Yorkshire Ambulance Service, Sponsor Contact

List of Documents

The final document set assessed and approved by HRA and HCRW Approval is listed below.

| <i>Document</i> | <i>Version</i> | <i>Date</i> |
|--|----------------|-------------------|
| Contract/Study Agreement template | | 16 September 2019 |
| IRAS Application Form [IRAS_Form_19122019] | | 19 December 2019 |
| Letter from funder [NIHR Intent to fund letter] | | 11 December 2018 |
| Letter from sponsor | 1.0 | 16 September 2019 |
| Research protocol or project proposal [Research protocol] | 1.1 | 04 December 2019 |
| Summary CV for Chief Investigator (CI) [J Miles CV] | 1.0 | 13 September 2019 |
| Summary CV for student [Research CV] | version 1.0 | 13 September 2019 |
| Summary CV for supervisor (student research) | | |
| Summary CV for supervisor (student research) [S.Mason] | 1.0 | 18 September 2019 |
| Summary, synopsis or diagram (flowchart) of protocol in non technical language [SINEPOST data flow v1.0] | 1.0 | 01 July 2019 |



Health Research Authority

Yorkshire & The Humber - South Yorkshire Research Ethics Committee

NHSBT Newcastle Blood Donor Centre
Holland Drive
Newcastle upon Tyne
NE2 4NQ

Telephone: 0207 104 8079

Please note: This is the favourable opinion of the REC only and does not allow you to start your study at NHS sites in England until you receive HRA Approval

08 November 2019

Mr Jamie Miles
Clinical Doctoral Research Fellow
Yorkshire Ambulance Service
Springhill 1
Brindley Way
Wakefield
WF2 0XQ

Dear Mr Miles

Study title: Safety INdEx of Prehospital On Scene Triage (SINEPOST): The derivation and validation of a risk prediction model to support ambulance clinical transport decisions on scene.

REC reference: 19/YH/0360
Protocol number: YASRD109
IRAS project ID: 260505

The Research Ethics Committee reviewed the above application at the meeting held on 31 October 2019.

Ethical opinion

The members of the Committee present gave a favourable ethical opinion of the above research on the basis described in the application form, protocol and supporting documentation, subject to the conditions specified below.

Conditions of the favourable opinion

The REC favourable opinion is subject to the following conditions being met prior to the start of the study.

| Number | Condition |
|--------|---|
| 1. | The Committee agreed the student could not be the Chief Investigator or the Data Custodian and required these "roles" to be undertaken by one or two of the Academic Supervisors. |

A Research Ethics Committee established by the Health Research Authority

You should notify the REC once all conditions have been met (except for site approvals from host organisations) and provide copies of any revised documentation with updated version numbers. Revised documents should be submitted to the REC electronically from IRAS. The REC will acknowledge receipt and provide a final list of the approved documentation for the study, which you can make available to host organisations to facilitate their permission for the study. Failure to provide the final versions to the REC may cause delay in obtaining permissions.

Confirmation of Capacity and Capability (in England, Northern Ireland and Wales) or NHS management permission (in Scotland) should be sought from all NHS organisations involved in the study in accordance with NHS research governance arrangements. Each NHS organisation must confirm through the signing of agreements and/or other documents that it has given permission for the research to proceed (except where explicitly specified otherwise).

Guidance on applying for HRA and HCRW Approval (England and Wales)/ NHS permission for research is available in the Integrated Research Application System.

For non-NHS sites, site management permission should be obtained in accordance with the procedures of the relevant host organisation.

Sponsors are not required to notify the Committee of management permissions from host organisations.

Registration of Clinical Trials

It is a condition of the REC favourable opinion that **all clinical trials are registered** on a publicly accessible database. For this purpose, 'clinical trials' are defined as the first four project categories in IRAS project filter question 2. Registration is a legal requirement for clinical trials of investigational medicinal products (CTIMPs), except for phase I trials in healthy volunteers (these must still register as a condition of the REC favourable opinion).

Registration should take place as early as possible and within six weeks of recruiting the first research participant at the latest. Failure to register is a breach of these approval conditions, unless a deferral has been agreed by or on behalf of the Research Ethics Committee (see here for more information on requesting a deferral: <https://www.hra.nhs.uk/planning-and-improving-research/research-planning/research-registration-research-project-identifiers/>

As set out in the UK Policy Framework, research sponsors are responsible for making information about research publicly available before it starts e.g. by registering the research project on a publicly accessible register. Further guidance on registration is available at: <https://www.hra.nhs.uk/planning-and-improving-research/research-planning/transparency-responsibilities/>

You should notify the REC of the registration details. We routinely audit applications for compliance with these conditions.

Publication of Your Research Summary

We will publish your research summary for the above study on the research summaries section of our website, together with your contact details, no earlier than three months from the date of this favourable opinion letter. Should you wish to provide a substitute contact point, make a request to defer, or require further information, please visit: <https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/research-summaries/>

It is the responsibility of the sponsor to ensure that all the conditions are complied with before the start of the study or its initiation at a particular site (as applicable).

After ethical review: Reporting requirements

The attached document “After ethical review – guidance for researchers” gives detailed guidance on reporting requirements for studies with a favourable opinion, including:

- Notifying substantial amendments
- Adding new sites and investigators
- Notification of serious breaches of the protocol
- Progress and safety reports
- Notifying the end of the study, including early termination of the study
- Final report

The latest guidance on these topics can be found at <https://www.hra.nhs.uk/approvals-amendments/managing-your-approval/>.

Ethical review of research sites

NHS/HSC Sites

The favourable opinion applies to all NHS/HSC sites taking part in the study taking part in the study, subject to confirmation of Capacity and Capability (in England, Northern Ireland and Wales) or NHS management permission (in Scotland) being obtained from the NHS/HSC R&D office prior to the start of the study (see “Conditions of the favourable opinion” below).

Non-NHS/HSC sites

I am pleased to confirm that the favourable opinion applies to any non NHS/HSC sites listed in the application, subject to site management permission being obtained prior to the start of the study at the site.

Approved documents

The documents reviewed and approved at the meeting were:

| <i>Document</i> | <i>Version</i> | <i>Date</i> |
|--|----------------|-------------------|
| Contract/Study Agreement template | | 16 September 2019 |
| Initial Assessment for REC [IAL for REC] | 1 | 09 October 2019 |
| IRAS Application Form [IRAS_Form_01102019] | | 01 October 2019 |
| IRAS Checklist XML [Checklist_08102019] | | 08 October 2019 |
| Letter from funder [NIHR Intent to fund letter] | | 11 December 2018 |
| Letter from sponsor | 1.0 | 16 September 2019 |
| Research protocol or project proposal [Research protocol] | 1.0 | 16 September 2019 |
| Summary CV for Chief Investigator (CI) [J Miles CV] | 1.0 | 13 September 2019 |
| Summary CV for student [Research CV] | version 1.0 | 13 September 2019 |
| Summary CV for supervisor (student research) [S.Mason] | 1.0 | 18 September 2019 |
| Summary CV for supervisor (student research) | | |
| Summary, synopsis or diagram (flowchart) of protocol in non technical language [SINEPOST data flow v1.0] | 1.0 | 01 July 2019 |

Membership of the Committee

The members of the Ethics Committee who were present at the meeting are listed on the attached sheet.

The Committee is constituted in accordance with the Governance Arrangements for Research Ethics Committees and complies fully with the Standard Operating Procedures for Research Ethics Committees in the UK.

User Feedback

The Health Research Authority is continually striving to provide a high quality service to all applicants and sponsors. You are invited to give your view of the service you have received and the application procedure. If you wish to make your views known please use the feedback form available on the HRA website: <http://www.hra.nhs.uk/about-the-hra/governance/quality-assurance/>

HRA Learning

We are pleased to welcome researchers and research staff to our HRA Learning Events and online learning opportunities– see details at: <https://www.hra.nhs.uk/planning-and-improving-research/learning/>

| | |
|------------|--|
| 19/YH/0360 | Please quote this number on all correspondence |
|------------|--|

With the Committee's best wishes for the success of this project.

Yours sincerely
Pp



Dr Ian Woollands
Chair

E-mail: nrescommittee.yorkandhumber-southyorks@nhs.net

Enclosures: List of names and professions of members who were present at the meeting and those who submitted written comments

"After ethical review – guidance for researchers" [[SL-AR2 for other studies](#)]

Copy to: Ms Jane Shewan, Yorkshire Ambulance Service
Confidentiality Advise Team

Lead Nation England: HRA.Approval@nhs.net

Yorkshire & The Humber - South Yorkshire Research Ethics Committee

Attendance at Committee meeting on 31 October 2019

Committee Members:

| <i>Name</i> | <i>Profession</i> | <i>Present</i> | <i>Notes</i> |
|--------------------------|--|----------------|--------------|
| Ms Helen Barlow | Knowledge Service Manager | Yes | |
| Dr Geraldine Boyle | Senior Lecturer | Yes | |
| Dr David Broomhead | Therapy Consultant | No | |
| Dr Ruth Clark | Retired CT Senior Project Manager | Yes | |
| Mr John de Bartolome | Retired - Lay+ Member | Yes | |
| Mr Martin Edmunds | Editor | Yes | |
| Dr Max Huxham | Retired Scientist | Yes | |
| Dr Alison Patrick | Lecturer in Law and Ethics | Yes | |
| Mrs Carole Taylor | Deputy Chief Pharmacist | Yes | |
| Miss Sarah Varga | Staff Nurse/ Clinical Academic Nurse | Yes | |
| Dr Ian Woollands (Chair) | Retired Clinical Director, Occupational Health | Yes | |

Also in attendance:

| <i>Name</i> | <i>Position (or reason for attending)</i> |
|--------------------|---|
| Miss Donna Bennett | Approvals Administrator |
| Ms Rebecca Evans | Approvals Specialist |



**Health Research
Authority**

Yorkshire & The Humber - South Yorkshire Research Ethics Committee

NHSBT Newcastle Blood Donor Centre
Holland Drive
Newcastle upon Tyne
NE2 4NQ

Telephone: 0207 1048091

**Please note: This is an
acknowledgement letter from
the REC only and does not
allow you to start your study
at NHS sites in England until
you receive HRA Approval**

20 December 2019

Mr Jamie Miles
Yorkshire Ambulance Service HQ
Springhill 1, Brindley way
Wakefield
WF2 0XQ

Dear Mr Miles

Study title: Safety INdEx of Prehospital On Scene Triage
(SINEPOST): The derivation and validation of a risk
prediction model to support ambulance clinical
transport decisions on scene.

REC reference: 19/YH/0360
Protocol number: YASRD109
IRAS project ID: 260505

Thank you for your letter of 19 December 2019. I can confirm the REC has received the documents listed below and that these comply with the approval conditions detailed in our letter dated 11 November 2019.

Documents received

The documents received were as follows:

| <i>Document</i> | <i>Version</i> | <i>Date</i> |
|---|----------------|------------------|
| IRAS Application Form [IRAS_Form_19122019] | | 19 December 2019 |
| IRAS Checklist XML [Checklist_19122019] | | 19 December 2019 |
| Other [IRAS responses and clarification] | | |
| Research protocol or project proposal [Research protocol] | 1.1 | 04 December 2019 |

Approved documents

The final list of approved documentation for the study is therefore as follows:

| <i>Document</i> | <i>Version</i> | <i>Date</i> |
|--|----------------|-------------------|
| IRAS Application Form [IRAS_Form_19122019] | | 19 December 2019 |
| IRAS Checklist XML [Checklist_19122019] | | 19 December 2019 |
| Letter from funder [NIHR Intent to fund letter] | | 11 December 2018 |
| Letter from sponsor | 1.0 | 16 September 2019 |
| Other [IRAS responses and clarification] | | |
| Research protocol or project proposal [Research protocol] | 1.1 | 04 December 2019 |
| Summary CV for Chief Investigator (CI) [J Miles CV] | 1.0 | 13 September 2019 |
| Summary CV for student [Research CV] | version 1.0 | 13 September 2019 |
| Summary CV for supervisor (student research) [S.Mason] | 1.0 | 18 September 2019 |
| Summary CV for supervisor (student research) | | |
| Summary, synopsis or diagram (flowchart) of protocol in non technical language [SINEPOST data flow v1.0] | 1.0 | 01 July 2019 |

You should ensure that the sponsor has a copy of the final documentation for the study. It is the sponsor's responsibility to ensure that the documentation is made available to R&D offices at all participating sites.

| | |
|-------------------|---|
| 19/YH/0360 | Please quote this number on all correspondence |
|-------------------|---|

Yours sincerely



Helen Wilson
Approvals Officer

E-mail: nrescommittee.yorkandhumber-southyorks@nhs.net

Copy to: Ms Jane Shewan, Yorkshire Ambulance Service

Professor Suzanne Mason, Chief Investigator

Lead Nation England: HRA.Approval@nhs.net

A Research Ethics Committee established by the Health Research Authority



**Health Research
Authority**

Skipton House
80 London Road
London
SE1 6LH

Tel: 020 797 22557
Email: cag@hra.nhs.uk

14 July 2020

Professor Suzanne Mason
University of Sheffield
CURE, ScHARR
30 Regent street
Sheffield
S1 4DA

Dear Professor Suzanne Mason

Application title: Safety INdEx of Prehospital On Scene Triage (SINEPOST): The derivation and validation of a risk prediction model to support ambulance clinical transport decisions on scene.

CAG reference: 20/CAG/0035
IRAS project ID: 260505
REC reference: 19/YH/0360

Thank you for your research application, submitted for support under Regulation 5 of the Health Service (Control of Patient Information) Regulations 2002 to process confidential patient information without consent. Supported applications enable the data controller to provide specified information to the applicant for the purposes of the relevant activity, without being in breach of the common law duty of confidentiality, although other relevant legislative provisions will still be applicable.

The role of the Confidentiality Advisory Group (CAG) is to review applications submitted under these Regulations and to provide advice to the Health Research Authority on whether an application should be supported, and if so, any relevant conditions. This application was considered at the CAG meeting held on 05 March 2020.

This outcome should be read in conjunction with the provisional support letter dated 18 March 2020.

Health Research Authority decision

The Health Research Authority, having considered the advice from the Confidentiality Advisory Group as set out below, has determined the following:

1. The application, to allow the disclosure of confidential patient information from the Yorkshire Ambulance Service NHS Trust to NHS Digital, for linkage to Emergency Department Care data for the two participating organisations, is fully supported, subject to compliance with the standard and specific conditions of support.

Please note that the legal basis to allow access to the specified confidential patient information without consent is now in effect.

Context

Purpose of application

This application from the University of Sheffield set out the purpose of medical research which aims to determine whether ambulance service clinical data can predict an avoidable attendance at the Emergency Department in adults using classification models.

Paramedics have specialist knowledge and skills in helping people in emergencies. The bulk of ambulance service patients who call have problems that are described as 'urgent'. These are cases where the patient may need access to healthcare and medical help, but there is only a very small chance that the problem is life threatening. The care of urgent patients is complex and trying to find the right place for their care can be hard. In 2014 in Yorkshire, up to 16.9% of patients could have avoided being taken by ambulance to the Emergency Department (ED). This group of patients had no special tests or treatments and were sent home. This means they had a minor problem that could have been managed elsewhere. When the ED is busy, ambulances have to wait a long time to handover the care of their patients. This delay stops ambulances being free to respond to the next emergency. These problems mean paramedics need to make sure the ED is the right place for their patient before they take them there. This project aims to develop a tool to help with that decision by showing the paramedic the likelihood of treatment at an ED being of benefit to the patient.

The applicants are seeking support under s251 to access the Electronic Patient Care Record at Yorkshire Ambulance Service NHS Trust and to obtain Emergency Department Care data from NHS Digital. The first stage of the study requires that data from the Yorkshire Ambulance Service NHS Trust is linked to data from a large hospital via NHS Digital, in order to provide show us the complete patient journey from their call for help through to leaving the ED. An anonymised dataset will then be provided to the applicants at the University of Sheffield. This information will be used to create a tool that identifies patients who may not need to be taken to the ED. The public will be invited to face-to-face meetings to help the researcher produce a lay summary of this phase. In the second stage, data from the Yorkshire Ambulance Service NHS Trust will be linked to Emergency Department Care data from NHS Digital for a different hospital, to see if the test can work in different settings.

A recommendation for class 1, 2, 4, 5 and 6 support was requested to cover access to the relevant unconsented activities as described in the application.

Confidential patient information requested

The following sets out a summary of the specified cohort, listed data sources and key identifiers. Where applicable, full datasets and data flows are provided in the application

form and relevant supporting documentation as this letter represents only a summary of the full detail.

| | |
|---|---|
| Cohort | 467329 patients treated by Yorkshire Ambulance Service NHS Trust and by the two participating Trusts. |
| Data sources | 1. Electronic Patient Care Record at Yorkshire Ambulance Service NHS Trust 2. Emergency Department Care data provided by NHS Digital |
| Identifiers required for linkage purposes | 1. Name 2. NHS number 3. Date of birth 4. Date of death 5. Postcode -unit level |
| Identifiers required for analysis purposes | 1. Postcode – sector level |

Confidentiality Advisory Group advice

This letter summarises the outstanding elements set out in the provisional support letter, and the applicant response. The applicant response was considered by Chair's Action.

- 1. Provide the names of the two organisations whose Emergency Department Care data would be combined by NHS Digital with the Electronic Patient Care Record from Yorkshire Ambulance Service NHS Trust.**

The applicant advised that the study would use regional data from every acute trust in Yorkshire, however this will be done through NHS Digital using routinely collected data. The specific trusts data which will be included are Airedale General Hospital, Barnsley Hospital, Bradford Royal Infirmary, Calderdale Royal Hospital, Dewsbury and district Hospital, Doncaster Royal Infirmary, Harrogate district Hospital, Huddersfield Royal Infirmary, Hull Royal Infirmary, Leeds General Infirmary, Northern General Hospital, Pinderfields Hospital, Rotherham Hospital, Scarborough Hospital, St. James's University Hospital and York Hospital.

The CAG had requested that DSPTs for the two organisations were provided. However, due to the increase in the number of organisations, it was decided that individual DSPT submissions will not be required for the purpose of the application. Support is recommended on the basis that the applicant ensures the required security standards are in place at each site prior to any processing of confidential patient information with support under the Regulations.

- 2. Confirm that patient and public involvement will be undertaken around developing a patient notification strategy. Please provide a summary of the feedback received about the patient notification put in place within three months of support under s251 being confirmed.**

The applicant confirmed that further public involvement will be undertaken to develop a patient notification strategy. A report on this further patient and public involvement will be

submitted within 3 months of the s251 being confirmed. The CAG noted this and raised no further queries.

Confidentiality Advisory Group advice conclusion

The CAG agreed that the minimum criteria under the Regulations appeared to have been met, and therefore advised recommending support to the Health Research Authority, subject to compliance with the specific and standard conditions of support as set out below.

Specific conditions of support

1. Favourable opinion from a Research Ethics Committee. **Confirmed 08 November 2019.**
2. Confirmation provided from the IG Delivery Team at NHS Digital to the CAG that the relevant Data Security and Protection Toolkit (DSPT) submission(s) has achieved the 'Standards Met' threshold. See section below titled 'security assurance requirements' for further information. **(Confirmed: NHS Digital (by NHS Digital email 10 June 2019) and Yorkshire Ambulance Service NHS Trust (by NHS Digital email 07 January 2019) have confirmed 'Standards Met' grade on DSPT 2018/19).**

Due to the number of participating sites where confidential patient information will be accessed, individual DSPT submissions are not required for the purpose of the application. Support is recommended on the basis that the applicant ensures the required security standards are in place at each site prior to any processing of confidential patient information with support under the Regulations.

As the above conditions have been accepted and met, this letter provides confirmation of final support. I will arrange for the register of approved applications on the HRA website to be updated with this information.

Application maintenance

Annual review

Please note that this legal support is subject to submission of an annual review report, for the duration of support, to show that the minimal amount of patient information is being processed and support is still necessary, how you have met the conditions or report plans, any public benefits that have arisen and action towards meeting them. It is also your responsibility to submit this report every 12 months for the entire duration that confidential patient information is being processed without consent.

The next annual review should be provided no later than **14 July 2021** and preferably 4 weeks before this date. Reminders are not issued so please ensure this is provided annually to avoid jeopardising the status of the support. Submission of an annual review in line with this schedule remains necessary even where there has been a delay to the commencement of the supported activity, or a halt in data processing. Please ensure you review the HRA website to ensure you are completing the most up to date 'section 251' annual review form as these may change.

For an annual review to be valid, there must also be evidence that the relevant DSPT submission(s) for organisations processing confidential patient information without consent are in place and have been reviewed by NHS Digital. Please plan to contact NHS Digital in advance of the CAG annual review submission date to check they have reviewed the relevant DSPTs and have confirmed these are satisfactory.

Register of Approved Applications

All supported applications to process confidential patient information without consent are listed in the published 'Register of Approved Applications'. It is a statutory requirement for the Register to be published and it is available on the CAG section of the Health Research Authority website. It contains applicant contact details, a summary of the research and other pertinent points.

This Register is used by controllers to check whether support is in place.

Changes to the application

The application and relevant documents set out the scope of the support which is in place for the application activity and any relevant restrictions around this.

Any amendments which are made to the scope of this support, including but not limited to, purpose, data flows, data sources, items of confidential patient information and processors, require submission of a formal amendment to the application. Changes to processors will require evidence of satisfactory DSPT submission. The amendment form can be found in the Confidentiality Advisory Group pages on the Health Research Authority website.

Support for any submitted amendment would not come into effect until a positive outcome letter has been issued.

Changes to the controller

Amendments which involve a change to the named controller for the application activity require the submission of a new and signed CAG application form and supporting documentation to support the application amendment. This is necessary to ensure that the application held on file appropriately reflects the organisation taking responsibility for the manner and purpose of data processing within the application, and that the legal support in place is related to the correct legal entity.

Applicants are advised to make contact with the Confidentiality Advice Team to discuss a change in controllership for an existing application in sufficient time ahead of the transfer of project responsibility to discuss the submission process timings.

Further information and relevant forms to amend the support is available on the HRA website.

Reviewed documents

The documents reviewed at the meeting were:

| <i>Document</i> | <i>Version</i> | <i>Date</i> |
|--|----------------|-------------|
| CAG application from (signed/authorised) [JMiles | | |

| | | |
|---|-----|------------------|
| CAG_Form_ReadyForSubmission] | | |
| Written recommendation from Caldicott Guardian (or equivalent) of applicant's organisation [Written approval from Caldicott Guardian] | | 24 February 2020 |
| SINEPOST protocol v1.1 (edits highlighted) | 1.1 | 04 December 2019 |

Membership of the Committee

The members of the Confidentiality Advisory Group who were present at the consideration of this item are listed below.

There were no declarations of interest in relation to this item.

User Feedback

The Health Research Authority is continually striving to provide a high-quality service to all applicants and sponsors. You are invited to give your view of the service you have received and the application procedure. If you wish to make your views known please use the feedback form available on the HRA website: <http://www.hra.nhs.uk/about-the-hra/governance/quality-assurance/>

HRA Training

We are pleased to welcome researchers and R & D staff at our training days – see details at <http://www.hra.nhs.uk/hra-training/>

Please do not hesitate to contact me if you have any queries following this letter. I would be grateful if you could quote the above reference number in all future correspondence.

With the Group's best wishes for the success of this project.

Yours sincerely

Kathleen Cassidy
Confidentiality Advisor

On behalf of the Health Research Authority

Email: cag@hra.nhs.uk

Included: List of members who considered application
Standard conditions of support

Copy to: southyorks.rec@hra.nhs.uk
approvals@hra.nhs.uk

**Confidentiality Advisory Group meeting attendance
05 March 2020 held via teleconference**

Members present:

| <i>Name</i> | |
|---------------------|----------------------------|
| Dr Tony Calland MBE | CAG Chair |
| Dr Malcolm Booth | CAG member |
| Ms Sophie Brannan | CAG member |
| Dr Lorna Fraser | CAG member |
| Dr Katie Harron | CAG member |
| Dr Rachel Knowles | CAG member |
| Mr Andrew Melville | CAG member |
| Ms Clare Sanderson | CAG alternative vice-chair |
| Mr Marc Taylor | CAG member |
| Ms Gillian Wells | CAG member |

Also in attendance:

| <i>Name</i> | <i>Position (or reason for attending)</i> |
|-----------------|---|
| Ms Katy Cassidy | HRA Confidentiality Advisor |

Standard conditions of support

Support to process the specified confidential patient information without consent, given by the Health Research Authority, is subject to compliance with the following standard conditions of support.

The applicant and those processing the information under the terms of the support will ensure that:

1. The specified confidential patient information is only used for the purpose(s) set out in the application.
2. Confidentiality is preserved and there are no disclosures of information in aggregate or patient level form that may inferentially identify a person, nor will any attempt be made to identify individuals, households or organisations in the data.
3. Requirements of the Statistics and Registration Services Act 2007 are adhered to regarding publication when relevant, in addition to other national guidance.
4. All staff with access to confidential patient information have contractual obligations of confidentiality, enforceable through disciplinary procedures.
5. All staff with access to confidential patient information have received appropriate ongoing training to ensure they are aware of their responsibilities and are acting in compliance with the application detail.
6. Activities must be compliant with the General Data Protection Regulation and Data Protection Act 2018.
7. Audit of data processing by a designated agent is facilitated and supported.
8. The wishes of patients who have withheld or withdrawn their consent are respected.
9. Any significant changes (for example, people, purpose, data flows, data items, security arrangements) must be approved via formal amendment prior to changes coming into effect.
10. An annual review report is submitted to the CAG every 12 months from the date of the final support letter, for the duration of the support.
11. Any breaches of confidentiality around the supported flows of information should be reported to CAG within 10 working days of the incident, along with remedial actions taken/to be taken. This does not remove the need to follow national/legal requirements for reporting relevant security breaches.

Appendix H: Included variables in the model

| | Unavoidable (N=94294) | Avoidable (N=7228) | Overall (N=101522) |
|--|--------------------------|-----------------------|-----------------------|
| ED (not used as a candidate variable) | | | |
| AIREDALE GENERAL HOSPITAL | 3058 (3.2%) | 240 (3.3%) | 3298 (3.2%) |
| BARNSLEY DISTRICT GENERAL | 5810 (6.2%) | 323 (4.5%) | 6133 (6.0%) |
| BRADFORD ROYAL INFIRMARY | 6705 (7.1%) | 1004 (13.9%) | 7709 (7.6%) |
| CALDERDALE ROYAL HOSPITAL | 3865 (4.1%) | 242 (3.3%) | 4107 (4.0%) |
| DEWSBURY DISTRICT HOSPITAL | 827 (0.9%) | 137 (1.9%) | 964 (0.9%) |
| DONCASTER ROYAL INFIRMARY | 6258 (6.6%) | 420 (5.8%) | 6678 (6.6%) |
| HARROGATE DISTRICT HOSPITAL | 2598 (2.8%) | 163 (2.3%) | 2761 (2.7%) |
| HUDDERSFIELD ROYAL INFIRMARY | 4392 (4.7%) | 283 (3.9%) | 4675 (4.6%) |
| HULL ROYAL INFIRMARY | 10099 (10.7%) | 612 (8.5%) | 10711 (10.6%) |
| JAMES COOK UNIVERSITY HOSPITAL | 749 (0.8%) | 55 (0.8%) | 804 (0.8%) |
| LEEDS GENERAL INFIRMARY | 4839 (5.1%) | 263 (3.6%) | 5102 (5.0%) |
| NORTHERN GENERAL HOSPITAL | 9793 (10.4%) | 929 (12.9%) | 10722 (10.6%) |
| PINDERFIELDS GENERAL HOSPITAL | 9481 (10.1%) | 764 (10.6%) | 10245 (10.1%) |
| ROTHERHAM DISTRICT GENERAL HOS | 5618 (6.0%) | 352 (4.9%) | 5970 (5.9%) |
| SCARBOROUGH DISTRICT GENERAL HOSPITAL | 4374 (4.6%) | 120 (1.7%) | 4494 (4.4%) |
| ST JAMES UNIVERSITY HOSPITAL | 8078 (8.6%) | 824 (11.4%) | 8902 (8.8%) |
| YORK DISTRICT HOSPITAL | 5719 (6.1%) | 382 (5.3%) | 6101 (6.0%) |
| Missing | 2031 (2.2%) | 115 (6.1%) | 2146 (2.1%) |
| Impression_Psychiatric problems | | | |
| Did not Occur | 93634 (99.3%) | 6642 (91.9%) | 100276 (98.8%) |
| Occurred | 660 (0.7%) | 586 (8.1%) | 1246 (1.2%) |
| cannulation_IV | | | |
| Did not Occur | 79660 (84.5%) | 6940 (96.0%) | 86600 (85.3%) |
| Occurred | 14634 (15.5%) | 288 (4.0%) | 14922 (14.7%) |
| mobility_SelfMobile | | | |
| Did not Occur | 68215 (72.3%) | 3436 (47.5%) | 71651 (70.6%) |
| Occurred | 26079 (27.7%) | 3792 (52.5%) | 29871 (29.4%) |
| Impression_Allergic reaction/rash | | | |
| Did not Occur | 93881 (99.6%) | 6971 (96.4%) | 100852 (99.3%) |
| Occurred | 413 (0.4%) | 257 (3.6%) | 670 (0.7%) |
| Impression_Cardiac chest pain (ACS) | | | |
| Did not Occur | 88507 (93.9%) | 7094 (98.1%) | 95601 (94.2%) |
| Occurred | 5787 (6.1%) | 134 (1.9%) | 5921 (5.8%) |
| temperature_primary | | | |
| Mean (SD) | 37.0 (0.965) | 36.8 (0.735) | 37.0 (0.952) |
| Median [Min, Max] | 36.9 [31.7, 42.1] | 36.8 [33.0, 40.7] | 36.9 [31.7, 42.1] |

| | | | |
|---|--------------------|--------------------|--------------------|
| Missing | 5935 (6.3%) | 796 (11.0%) | 6731 (6.6%) |
| ecg_monitored_primary | | | |
| Did not Occur | 27983 (29.7%) | 2988 (41.3%) | 30971 (30.5%) |
| Occurred | 51293 (54.4%) | 2946 (40.8%) | 54239 (53.4%) |
| Missing | 15018 (15.9%) | 1294 (17.9%) | 16312 (16.1%) |
| oxygen_saturations_primary | | | |
| Mean (SD) | 95.3 (5.41) | 97.1 (2.84) | 95.4 (5.29) |
| Median [Min, Max] | 97.0 [11.0, 100] | 98.0 [18.0, 100] | 97.0 [11.0, 100] |
| Missing | 2543 (2.7%) | 329 (4.6%) | 2872 (2.8%) |
| respiratory_rate_primary | | | |
| Mean (SD) | 20.7 (6.30) | 18.7 (4.45) | 20.5 (6.21) |
| Median [Min, Max] | 18.0 [0, 99.0] | 18.0 [0, 96.0] | 18.0 [0, 99.0] |
| Missing | 1820 (1.9%) | 188 (2.6%) | 2008 (2.0%) |
| pain_score_primary | | | |
| Mean (SD) | 3.10 (3.58) | 2.94 (3.50) | 3.09 (3.57) |
| Median [Min, Max] | 1.00 [0, 10.0] | 0 [0, 10.0] | 1.00 [0, 10.0] |
| Missing | 26475 (28.1%) | 2073 (28.7%) | 28548 (28.1%) |
| blood_sugar_reading_primary | | | |
| Mean (SD) | 7.43 (3.43) | 6.72 (2.72) | 7.38 (3.39) |
| Median [Min, Max] | 6.50 [0.600, 35.0] | 6.00 [0.400, 33.0] | 6.40 [0.400, 35.0] |
| Missing | 23704 (25.1%) | 2593 (35.9%) | 26297 (25.9%) |
| manual_pulse_rate_primary | | | |
| Mean (SD) | 89.2 (22.3) | 88.1 (18.2) | 89.1 (22.0) |
| Median [Min, Max] | 86.0 [5.00, 220] | 87.0 [6.00, 220] | 86.0 [5.00, 220] |
| Missing | 2186 (2.3%) | 309 (4.3%) | 2495 (2.5%) |
| obs_supplimental_oxygen_subsequent | | | |
| Did not Occur | 56181 (59.6%) | 4030 (55.8%) | 60211 (59.3%) |
| Occurred | 12364 (13.1%) | 149 (2.1%) | 12513 (12.3%) |
| Missing | 25749 (27.3%) | 3049 (42.2%) | 28798 (28.4%) |
| Impression_Head injury | | | |
| Did not Occur | 92710 (98.3%) | 6937 (96.0%) | 99647 (98.2%) |
| Occurred | 1584 (1.7%) | 291 (4.0%) | 1875 (1.8%) |
| Impression_Pain - back non-traumatic | | | |
| Did not Occur | 92414 (98.0%) | 6882 (95.2%) | 99296 (97.8%) |
| Occurred | 1880 (2.0%) | 346 (4.8%) | 2226 (2.2%) |
| manual_pulse_rate_subsequent | | | |
| Mean (SD) | 87.8 (22.0) | 86.5 (17.3) | 87.7 (21.8) |
| Median [Min, Max] | 85.0 [9.00, 220] | 85.0 [37.0, 188] | 85.0 [9.00, 220] |
| Missing | 27444 (29.1%) | 3166 (43.8%) | 30610 (30.2%) |
| bp_diastolic_subsequent | | | |

| | | | |
|---|-------------------|--------------------|-------------------|
| Mean (SD) | 81.4 (17.3) | 85.2 (15.3) | 81.6 (17.2) |
| Median [Min, Max] | 81.0 [0, 200] | 85.0 [32.0, 168] | 81.0 [0, 200] |
| Missing | 29222 (31.0%) | 3263 (45.1%) | 32485 (32.0%) |
| Impression_Minor cuts & bruising | | | |
| Did not Occur | 94126 (99.8%) | 7142 (98.8%) | 101268 (99.7%) |
| Occurred | 168 (0.2%) | 86 (1.2%) | 254 (0.3%) |
| drug_Oxygen | | | |
| Did not Occur | 82346 (87.3%) | 7077 (97.9%) | 89423 (88.1%) |
| Occurred | 11948 (12.7%) | 151 (2.1%) | 12099 (11.9%) |
| respiratory_rate_subsequent | | | |
| Mean (SD) | 20.3 (5.94) | 18.3 (3.52) | 20.2 (5.84) |
| Median [Min, Max] | 18.0 [0, 99.0] | 18.0 [1.00, 81.0] | 18.0 [0, 99.0] |
| Missing | 26569 (28.2%) | 3097 (42.8%) | 29666 (29.2%) |
| Impression_Unable to cope | | | |
| Did not Occur | 93990 (99.7%) | 7137 (98.7%) | 101127 (99.6%) |
| Occurred | 304 (0.3%) | 91 (1.3%) | 395 (0.4%) |
| drug_Aspirin | | | |
| Did not Occur | 90409 (95.9%) | 7164 (99.1%) | 97573 (96.1%) |
| Occurred | 3885 (4.1%) | 64 (0.9%) | 3949 (3.9%) |
| Impression_Abdominal pain | | | |
| Did not Occur | 86900 (92.2%) | 6780 (93.8%) | 93680 (92.3%) |
| Occurred | 7394 (7.8%) | 448 (6.2%) | 7842 (7.7%) |
| bp_systolic_primary | | | |
| Mean (SD) | 143 (28.3) | 143 (24.4) | 143 (28.1) |
| Median [Min, Max] | 142 [0, 265] | 140 [1.00, 288] | 142 [0, 288] |
| Missing | 2991 (3.2%) | 388 (5.4%) | 3379 (3.3%) |
| bp_diastolic_primary | | | |
| Mean (SD) | 82.9 (17.7) | 86.4 (15.6) | 83.2 (17.6) |
| Median [Min, Max] | 83.0 [0, 200] | 86.0 [4.00, 182] | 83.0 [0, 200] |
| Missing | 3114 (3.3%) | 397 (5.5%) | 3511 (3.5%) |
| PulseInt | | | |
| Mean (SD) | -2.14 (11.2) | -2.68 (9.52) | -2.18 (11.1) |
| Median [Min, Max] | -1.00 [-149, 144] | -2.00 [-127, 94.0] | -1.00 [-149, 144] |
| Missing | 27964 (29.7%) | 3192 (44.2%) | 31156 (30.7%) |
| Impression_Wound Closure | | | |
| Did not Occur | 94206 (99.9%) | 7193 (99.5%) | 101399 (99.9%) |
| Occurred | 88 (0.1%) | 35 (0.5%) | 123 (0.1%) |
| epr_news_score_0 | | | |
| Did not Occur | 65881 (69.9%) | 4103 (56.8%) | 69984 (68.9%) |
| Occurred | 20807 (22.1%) | 2194 (30.4%) | 23001 (22.7%) |
| Missing | 7606 (8.1%) | 931 (12.9%) | 8537 (8.4%) |

| | | | |
|--|-------------------|--------------------|-------------------|
| bp_systolic_subsequent | | | |
| Mean (SD) | 141 (28.0) | 140 (24.2) | 141 (27.8) |
| Median [Min, Max] | 139 [0, 282] | 138 [59.0, 242] | 139 [0, 282] |
| Missing | 29145 (30.9%) | 3262 (45.1%) | 32407 (31.9%) |
| Impression_Drug overdose | | | |
| Did not Occur | 92781 (98.4%) | 7068 (97.8%) | 99849 (98.4%) |
| Occurred | 1513 (1.6%) | 160 (2.2%) | 1673 (1.6%) |
| Impression_Stroke FAST positive | | | |
| Did not Occur | 92749 (98.4%) | 7204 (99.7%) | 99953 (98.5%) |
| Occurred | 1545 (1.6%) | 24 (0.3%) | 1569 (1.5%) |
| Impression_Eye injury/eye problem | | | |
| Did not Occur | 94200 (99.9%) | 7187 (99.4%) | 101387 (99.9%) |
| Occurred | 94 (0.1%) | 41 (0.6%) | 135 (0.1%) |
| Impression_Alcohol related | | | |
| Did not Occur | 93893 (99.6%) | 7114 (98.4%) | 101007 (99.5%) |
| Occurred | 401 (0.4%) | 114 (1.6%) | 515 (0.5%) |
| Impression_Fracture/possible fracture | | | |
| Did not Occur | 91981 (97.5%) | 7174 (99.3%) | 99155 (97.7%) |
| Occurred | 2313 (2.5%) | 54 (0.7%) | 2367 (2.3%) |
| Impression_Haemorrhage/lacerations | | | |
| Did not Occur | 93896 (99.6%) | 7135 (98.7%) | 101031 (99.5%) |
| Occurred | 398 (0.4%) | 93 (1.3%) | 491 (0.5%) |
| SysBPInt | | | |
| Mean (SD) | -2.36 (17.7) | -2.84 (14.6) | -2.39 (17.5) |
| Median [Min, Max] | -1.00 [-182, 239] | -2.00 [-85.0, 147] | -1.00 [-182, 239] |
| Missing | 29901 (31.7%) | 3300 (45.7%) | 33201 (32.7%) |
| Impression_Collapse-reason unknown | | | |
| Did not Occur | 92200 (97.8%) | 7168 (99.2%) | 99368 (97.9%) |
| Occurred | 2094 (2.2%) | 60 (0.8%) | 2154 (2.1%) |
| O2Int | | | |
| Mean (SD) | 1.42 (5.12) | 0.146 (3.19) | 1.34 (5.04) |
| Median [Min, Max] | 0 [-84.0, 86.0] | 0 [-81.0, 80.0] | 0 [-84.0, 86.0] |
| Missing | 27728 (29.4%) | 3176 (43.9%) | 30904 (30.4%) |
| Impression_Cardiac Arrhythmia | | | |
| Did not Occur | 92568 (98.2%) | 7194 (99.5%) | 99762 (98.3%) |
| Occurred | 1726 (1.8%) | 34 (0.5%) | 1760 (1.7%) |
| drug_GTN | | | |
| Did not Occur | 90832 (96.3%) | 7178 (99.3%) | 98010 (96.5%) |
| Occurred | 3462 (3.7%) | 50 (0.7%) | 3512 (3.5%) |
| Impression_Minor injuries - other | | | |
| Did not Occur | 93776 (99.5%) | 7130 (98.6%) | 100906 (99.4%) |

| | | | |
|--|-------------------|---------------------|-------------------|
| Occurred | 518 (0.5%) | 98 (1.4%) | 616 (0.6%) |
| DiaBPInt | | | |
| Mean (SD) | -1.46 (13.5) | -1.83 (11.4) | -1.48 (13.4) |
| Median [Min, Max] | -1.00 [-127, 135] | -1.00 [-73.0, 84.0] | -1.00 [-127, 135] |
| Missing | 30059 (31.9%) | 3307 (45.8%) | 33366 (32.9%) |
| psyc_AVPU_Confusion | | | |
| Did not Occur | 90490 (96.0%) | 7133 (98.7%) | 97623 (96.2%) |
| Occurred | 3201 (3.4%) | 63 (0.9%) | 3264 (3.2%) |
| Missing | 603 (0.6%) | 32 (0.4%) | 635 (0.6%) |
| oxygen_saturation_subsequent | | | |
| Mean (SD) | 96.3 (3.30) | 97.2 (2.92) | 96.4 (3.28) |
| Median [Min, Max] | 97.0 [14.0, 100] | 98.0 [17.0, 100] | 97.0 [14.0, 100] |
| Missing | 27171 (28.8%) | 3146 (43.5%) | 30317 (29.9%) |
| Location_Domestic Address | | | |
| Did not Occur | 15303 (16.2%) | 1106 (15.3%) | 16409 (16.2%) |
| Occurred | 68004 (72.1%) | 5281 (73.1%) | 73285 (72.2%) |
| Missing | 10987 (11.7%) | 841 (11.6%) | 11828 (11.7%) |
| drug_Morphine.Sulphate | | | |
| Did not Occur | 89586 (95.0%) | 7129 (98.6%) | 96715 (95.3%) |
| Occurred | 4708 (5.0%) | 99 (1.4%) | 4807 (4.7%) |
| Impression_Burns | | | |
| Did not Occur | 94243 (99.9%) | 7204 (99.7%) | 101447 (99.9%) |
| Occurred | 51 (0.1%) | 24 (0.3%) | 75 (0.1%) |
| Impression_No injury or illness | | | |
| Did not Occur | 94126 (99.8%) | 7183 (99.4%) | 101309 (99.8%) |
| Occurred | 168 (0.2%) | 45 (0.6%) | 213 (0.2%) |
| drug_Entonox | | | |
| Did not Occur | 89453 (94.9%) | 6890 (95.3%) | 96343 (94.9%) |
| Occurred | 4841 (5.1%) | 338 (4.7%) | 5179 (5.1%) |
| Impression_Panic attack | | | |
| Did not Occur | 94078 (99.8%) | 7151 (98.9%) | 101229 (99.7%) |
| Occurred | 216 (0.2%) | 77 (1.1%) | 293 (0.3%) |
| drug_Chlorphenamine | | | |
| Did not Occur | 94123 (99.8%) | 7159 (99.0%) | 101282 (99.8%) |
| Occurred | 171 (0.2%) | 69 (1.0%) | 240 (0.2%) |
| Impression_Headache | | | |
| Did not Occur | 93387 (99.0%) | 7050 (97.5%) | 100437 (98.9%) |
| Occurred | 907 (1.0%) | 178 (2.5%) | 1085 (1.1%) |
| Impression_Seizures (non-EP) | | | |
| Did not Occur | 93531 (99.2%) | 7198 (99.6%) | 100729 (99.2%) |
| Occurred | 763 (0.8%) | 30 (0.4%) | 793 (0.8%) |

| | | | |
|--|-----------------|-----------------|-----------------|
| epr_news_score_1 | | | |
| Did not Occur | 69887 (74.1%) | 4518 (62.5%) | 74405 (73.3%) |
| Occurred | 16801 (17.8%) | 1779 (24.6%) | 18580 (18.3%) |
| Missing | 7606 (8.1%) | 931 (12.9%) | 8537 (8.4%) |
| Impression_Vomiting | | | |
| Did not Occur | 93084 (98.7%) | 7167 (99.2%) | 100251 (98.7%) |
| Occurred | 1210 (1.3%) | 61 (0.8%) | 1271 (1.3%) |
| Impression_Catheter problems | | | |
| Did not Occur | 93956 (99.6%) | 7181 (99.3%) | 101137 (99.6%) |
| Occurred | 338 (0.4%) | 47 (0.7%) | 385 (0.4%) |
| Impression_Pain - other | | | |
| Did not Occur | 84576 (89.7%) | 6293 (87.1%) | 90869 (89.5%) |
| Occurred | 9718 (10.3%) | 935 (12.9%) | 10653 (10.5%) |
| avpu_score_subsequent_1 | | | |
| Did not Occur | 68458 (72.6%) | 4199 (58.1%) | 72657 (71.6%) |
| Occurred | 450 (0.5%) | 2 (0.0%) | 452 (0.4%) |
| Missing | 25386 (26.9%) | 3027 (41.9%) | 28413 (28.0%) |
| RRInt | | | |
| Mean (SD) | -0.898 (3.94) | -0.744 (3.50) | -0.889 (3.91) |
| Median [Min, Max] | 0 [-83.0, 84.0] | 0 [-76.0, 57.0] | 0 [-83.0, 84.0] |
| Missing | 26941 (28.6%) | 3114 (43.1%) | 30055 (29.6%) |
| drug_Adrenaline.1:1000 | | | |
| Did not Occur | 94158 (99.9%) | 7223 (99.9%) | 101381 (99.9%) |
| Occurred | 136 (0.1%) | 5 (0.1%) | 141 (0.1%) |
| Impression_Haematemesis | | | |
| Did not Occur | 93671 (99.3%) | 7204 (99.7%) | 100875 (99.4%) |
| Occurred | 623 (0.7%) | 24 (0.3%) | 647 (0.6%) |
| Location_Care Home | | | |
| Did not Occur | 75693 (80.3%) | 6015 (83.2%) | 81708 (80.5%) |
| Occurred | 7614 (8.1%) | 372 (5.1%) | 7986 (7.9%) |
| Missing | 10987 (11.7%) | 841 (11.6%) | 11828 (11.7%) |
| avpu_score_primary_4 | | | |
| Did not Occur | 89568 (95.0%) | 7033 (97.3%) | 96601 (95.2%) |
| Occurred | 3434 (3.6%) | 66 (0.9%) | 3500 (3.4%) |
| Missing | 1292 (1.4%) | 129 (1.8%) | 1421 (1.4%) |
| obs_supplimental_oxygen_primary | | | |
| Did not Occur | 88112 (93.4%) | 6998 (96.8%) | 95110 (93.7%) |
| Occurred | 4560 (4.8%) | 78 (1.1%) | 4638 (4.6%) |
| Missing | 1622 (1.7%) | 152 (2.1%) | 1774 (1.7%) |
| Location_Other | | | |
| Did not Occur | 78858 (83.6%) | 6067 (83.9%) | 84925 (83.7%) |

| | | | |
|--|----------------|--------------|-----------------|
| Occurred | 4449 (4.7%) | 320 (4.4%) | 4769 (4.7%) |
| Missing | 10987 (11.7%) | 841 (11.6%) | 11828 (11.7%) |
| Impression_Falls | | | |
| Did not Occur | 87481 (92.8%) | 6876 (95.1%) | 94357 (92.9%) |
| Occurred | 6813 (7.2%) | 352 (4.9%) | 7165 (7.1%) |
| Impression_Other medical condition | | | |
| Did not Occur | 89214 (94.6%) | 6742 (93.3%) | 95956 (94.5%) |
| Occurred | 5080 (5.4%) | 486 (6.7%) | 5566 (5.5%) |
| Impression_Choking | | | |
| Did not Occur | 94216 (99.9%) | 7211 (99.8%) | 101427 (99.9%) |
| Occurred | 78 (0.1%) | 17 (0.2%) | 95 (0.1%) |
| epr_nok_named | | | |
| Did not Occur | 7556 (8.0%) | 706 (9.8%) | 8262 (8.1%) |
| Occurred | 86738 (92.0%) | 6522 (90.2%) | 93260 (91.9%) |
| avpu_score_subsequent_3 | | | |
| Did not Occur | 68517 (72.7%) | 4193 (58.0%) | 72710 (71.6%) |
| Occurred | 391 (0.4%) | 8 (0.1%) | 399 (0.4%) |
| Missing | 25386 (26.9%) | 3027 (41.9%) | 28413 (28.0%) |
| Impression_Dental | | | |
| Did not Occur | 94261 (100.0%) | 7213 (99.8%) | 101474 (100.0%) |
| Occurred | 33 (0.0%) | 15 (0.2%) | 48 (0.0%) |
| Impression_Bleeding PR | | | |
| Did not Occur | 93551 (99.2%) | 7192 (99.5%) | 100743 (99.2%) |
| Occurred | 743 (0.8%) | 36 (0.5%) | 779 (0.8%) |
| epr_news_score_2 | | | |
| Did not Occur | 76161 (80.8%) | 5369 (74.3%) | 81530 (80.3%) |
| Occurred | 10527 (11.2%) | 928 (12.8%) | 11455 (11.3%) |
| Missing | 7606 (8.1%) | 931 (12.9%) | 8537 (8.4%) |
| total_component_score_subsequent_14 | | | |
| Did not Occur | 59914 (63.5%) | 3791 (52.4%) | 63705 (62.8%) |
| Occurred | 5499 (5.8%) | 214 (3.0%) | 5713 (5.6%) |
| Missing | 28881 (30.6%) | 3223 (44.6%) | 32104 (31.6%) |
| Impression_Diarrhoea/Constipation | | | |
| Did not Occur | 93762 (99.4%) | 7211 (99.8%) | 100973 (99.5%) |
| Occurred | 532 (0.6%) | 17 (0.2%) | 549 (0.5%) |
| Location_Public Place | | | |
| Did not Occur | 80597 (85.5%) | 6052 (83.7%) | 86649 (85.4%) |
| Occurred | 2710 (2.9%) | 335 (4.6%) | 3045 (3.0%) |
| Missing | 10987 (11.7%) | 841 (11.6%) | 11828 (11.7%) |
| drug_Activated.Charcoal | | | |
| Did not Occur | 94145 (99.8%) | 7219 (99.9%) | 101364 (99.8%) |

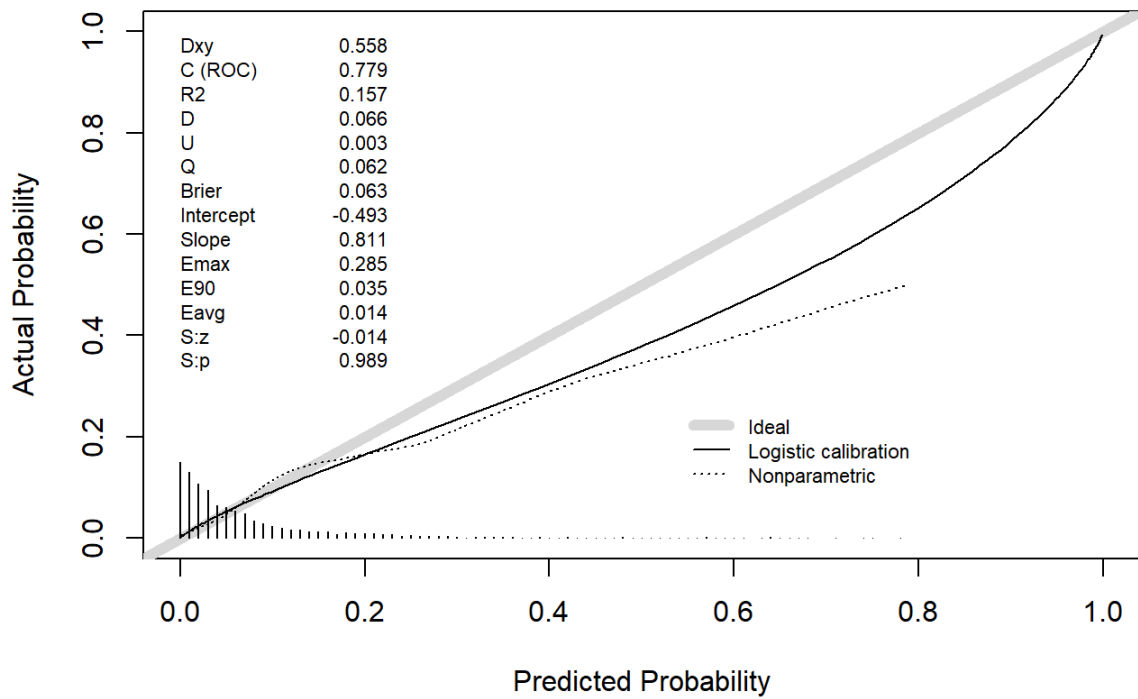
| | | | |
|---|----------------|---------------|-----------------|
| Occurred | 149 (0.2%) | 9 (0.1%) | 158 (0.2%) |
| immob_Other | | | |
| Did not Occur | 93931 (99.6%) | 7225 (100.0%) | 101156 (99.6%) |
| Occurred | 363 (0.4%) | 3 (0.0%) | 366 (0.4%) |
| total_component_score_primary_12 | | | |
| Did not Occur | 90588 (96.1%) | 6907 (95.6%) | 97495 (96.0%) |
| Occurred | 805 (0.9%) | 50 (0.7%) | 855 (0.8%) |
| Missing | 2901 (3.1%) | 271 (3.7%) | 3172 (3.1%) |
| Impression_Asthma | | | |
| Did not Occur | 93760 (99.4%) | 7165 (99.1%) | 100925 (99.4%) |
| Occurred | 534 (0.6%) | 63 (0.9%) | 597 (0.6%) |
| drug_Ondansetron | | | |
| Did not Occur | 91736 (97.3%) | 7181 (99.3%) | 98917 (97.4%) |
| Occurred | 2558 (2.7%) | 47 (0.7%) | 2605 (2.6%) |
| epr_news_score_5 | | | |
| Did not Occur | 81516 (86.4%) | 6111 (84.5%) | 87627 (86.3%) |
| Occurred | 5172 (5.5%) | 186 (2.6%) | 5358 (5.3%) |
| Missing | 7606 (8.1%) | 931 (12.9%) | 8537 (8.4%) |
| drug_Salbutamol | | | |
| Did not Occur | 89012 (94.4%) | 7103 (98.3%) | 96115 (94.7%) |
| Occurred | 5282 (5.6%) | 125 (1.7%) | 5407 (5.3%) |
| Impression_Cold & flu | | | |
| Did not Occur | 94186 (99.9%) | 7207 (99.7%) | 101393 (99.9%) |
| Occurred | 108 (0.1%) | 21 (0.3%) | 129 (0.1%) |
| psyc_CatastrophicHaemorrhage_No | | | |
| Did not Occur | 37 (0.0%) | 4 (0.1%) | 41 (0.0%) |
| Occurred | 71580 (75.9%) | 5529 (76.5%) | 77109 (76.0%) |
| Missing | 22677 (24.0%) | 1695 (23.5%) | 24372 (24.0%) |
| Impression_Hypertension | | | |
| Did not Occur | 93941 (99.6%) | 7187 (99.4%) | 101128 (99.6%) |
| Occurred | 353 (0.4%) | 41 (0.6%) | 394 (0.4%) |
| Impression_Bite/sting | | | |
| Did not Occur | 94263 (100.0%) | 7217 (99.8%) | 101480 (100.0%) |
| Occurred | 31 (0.0%) | 11 (0.2%) | 42 (0.0%) |
| Impression_Poisoning | | | |
| Did not Occur | 94040 (99.7%) | 7197 (99.6%) | 101237 (99.7%) |
| Occurred | 254 (0.3%) | 31 (0.4%) | 285 (0.3%) |

Appendix I: Hyperparameter values per cluster

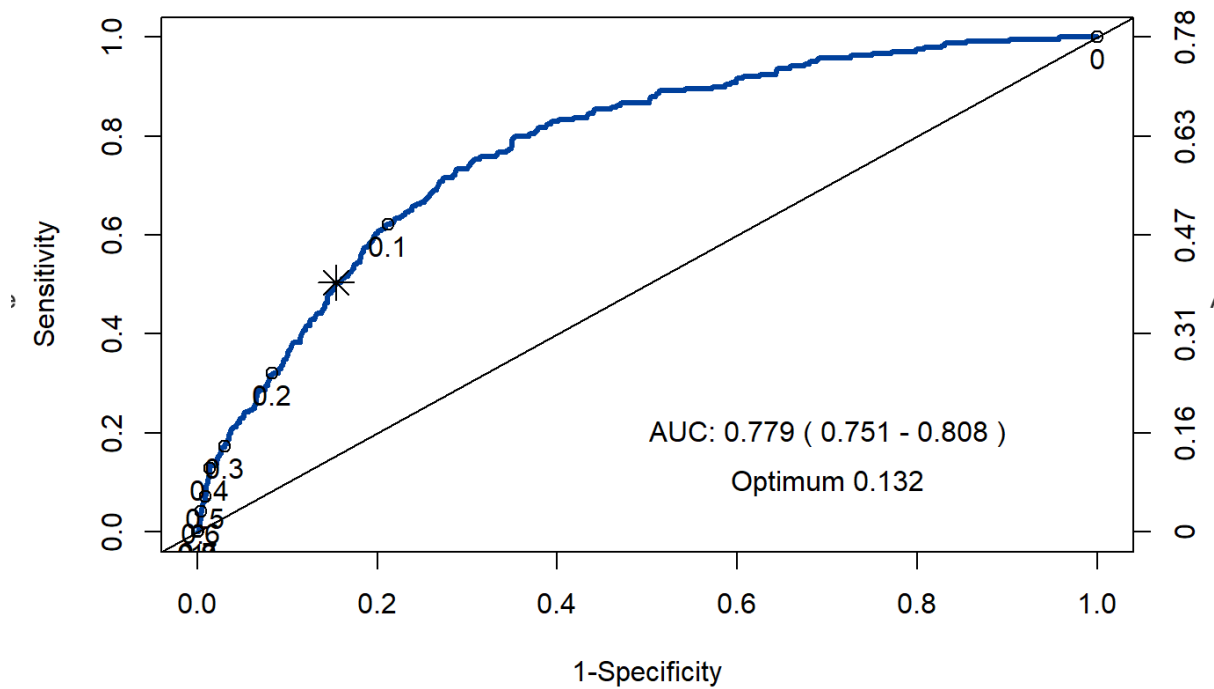
| Model | eta | max_depth | min_child | Colsample | | | scale_pos | | n_rounds |
|----------------|------|-----------|-----------|-----------|---------|-------|-----------|---------|----------|
| | | | _weight | subsample | _bytree | gamma | alpha | _weight | |
| Airedale | 0.06 | 4 | 4 | 0.9 | 0.6 | 0.5 | 0.6 | 1 | 408 |
| Barnsley | 0.06 | 3 | 4 | 1 | 0.6 | 1 | 0.7 | 0.67 | 630 |
| Bradford | 0.06 | 3 | 2 | 0.7 | 0.9 | 0.5 | 0.6 | 2.1 | 395 |
| Calderdale | 0.06 | 3 | 2 | 0.9 | 0.6 | 0.5 | 0.8 | 0.68 | 477 |
| Dewsbury | 0.06 | 3 | 4 | 0.9 | 0.6 | 1 | 0.6 | 1.49 | 463 |
| Doncaster | 0.06 | 3 | 2 | 0.9 | 0.6 | 1 | 0.7 | 1.01 | 453 |
| Harrogate | 0.08 | 3 | 4 | 0.9 | 0.6 | 0 | 0.7 | 0.85 | 467 |
| Huddersfield | 0.06 | 3 | 2 | 0.9 | 0.9 | 0 | 0.7 | 0.83 | 516 |
| Hull | 0.06 | 3 | 4 | 0.9 | 0.6 | 1 | 0.7 | 0.9 | 472 |
| Middlesborough | 0.08 | 4 | 2 | 0.9 | 0.9 | 0.5 | 0.8 | 1.37 | 261 |
| Leeds 1 | 0.06 | 3 | 4 | 0.9 | 0.6 | 0 | 0.7 | 0.68 | 485 |
| Sheffield | 0.06 | 3 | 4 | 0.9 | 0.6 | 0 | 0.6 | 1.13 | 411 |
| Wakefield | 0.06 | 4 | 2 | 0.7 | 0.6 | 1 | 0.7 | 1.06 | 330 |
| Rotherham | 0.06 | 3 | 4 | 0.9 | 0.6 | 0.5 | 0.8 | 0.81 | 524 |
| Scarborough | 0.06 | 3 | 4 | 0.9 | 0.9 | 0.5 | 0.6 | 0.41 | 577 |
| Leeds 2 | 0.06 | 4 | 4 | 0.9 | 0.9 | 0 | 0.7 | 1.26 | 348 |
| York | 0.06 | 4 | 4 | 0.9 | 0.9 | 1 | 0.8 | 0.86 | 367 |

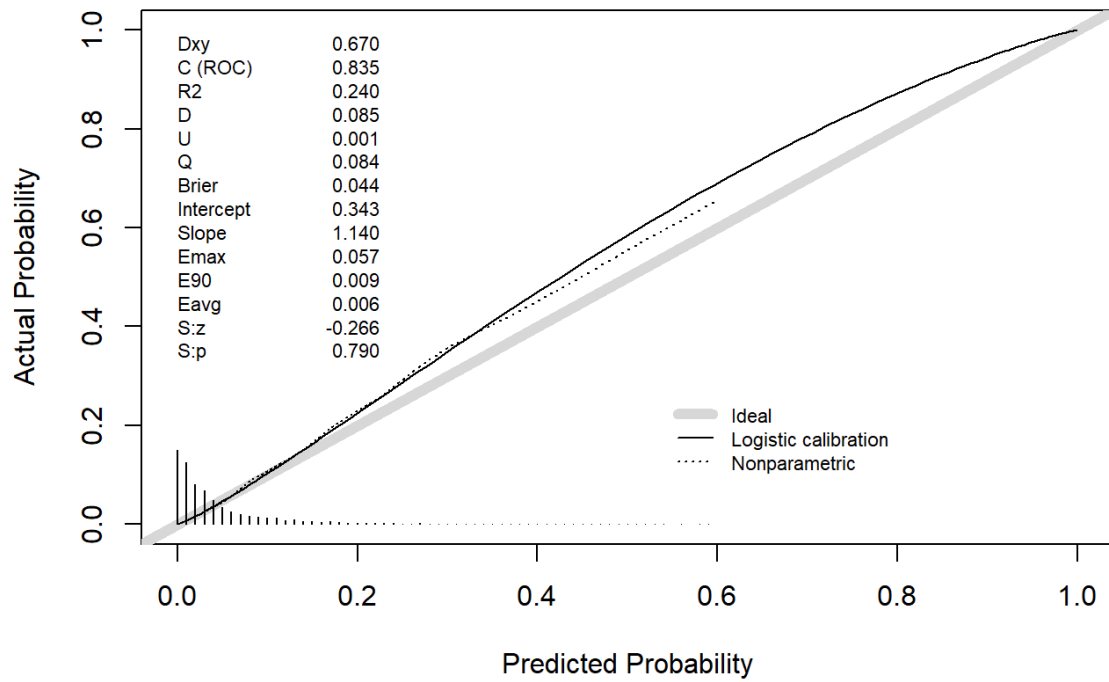
Appendix J: ROC and calibration curves for the IECVmodels

Airedale

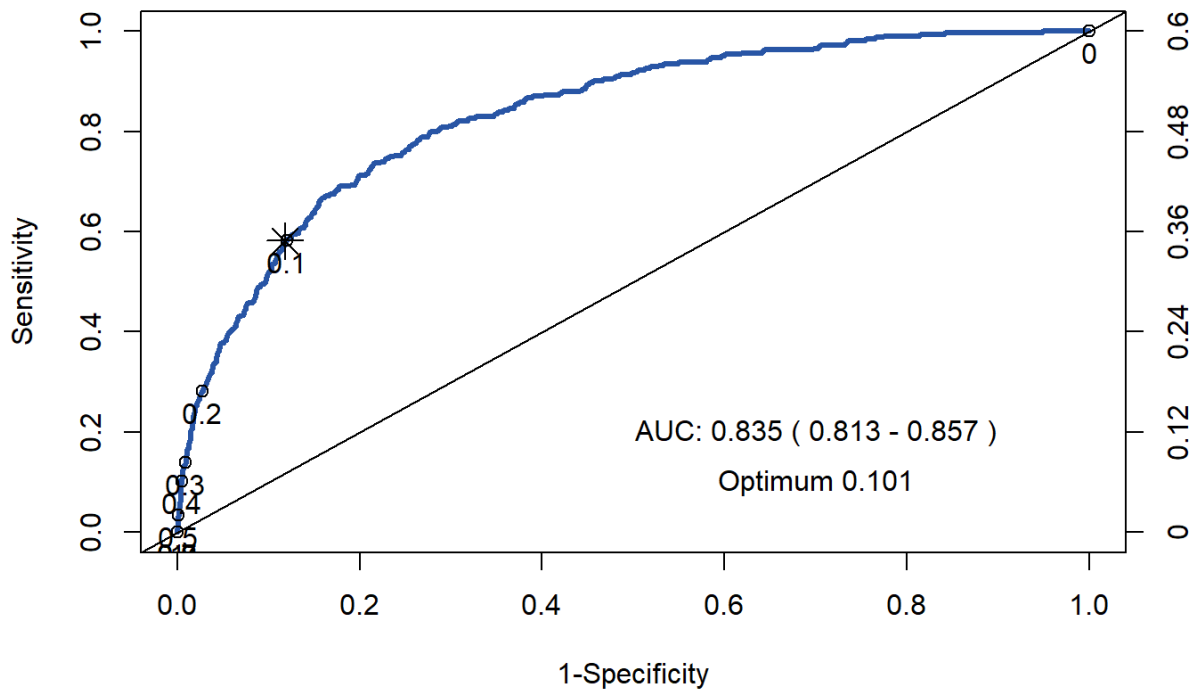


ROC curve of the Airedale model

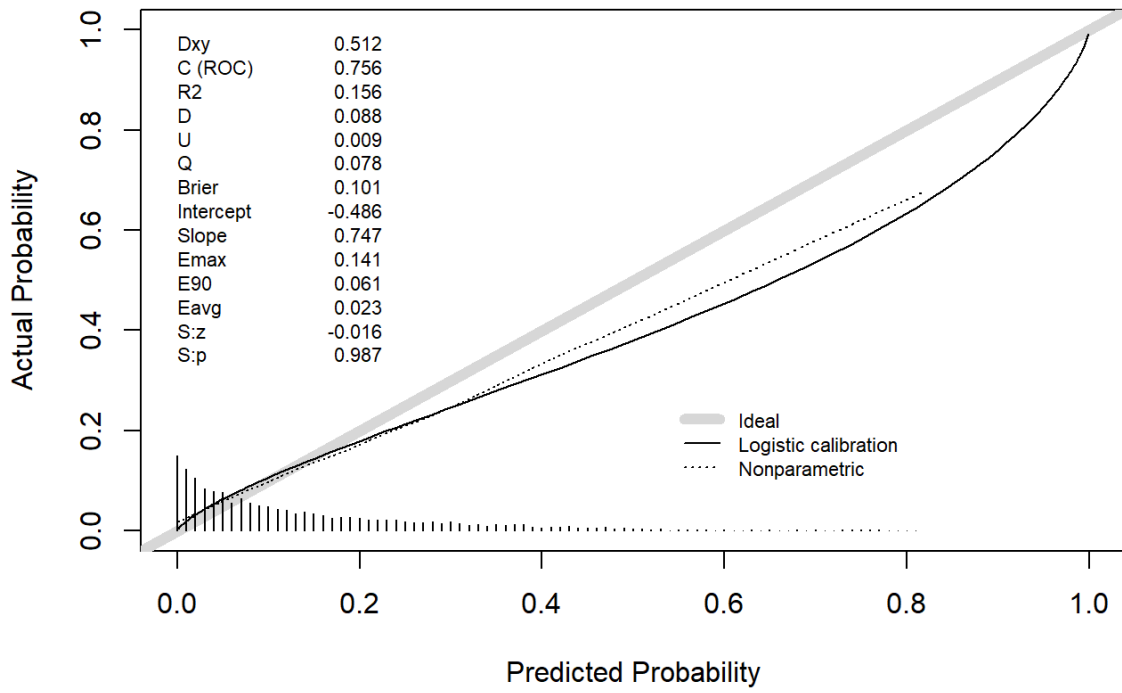




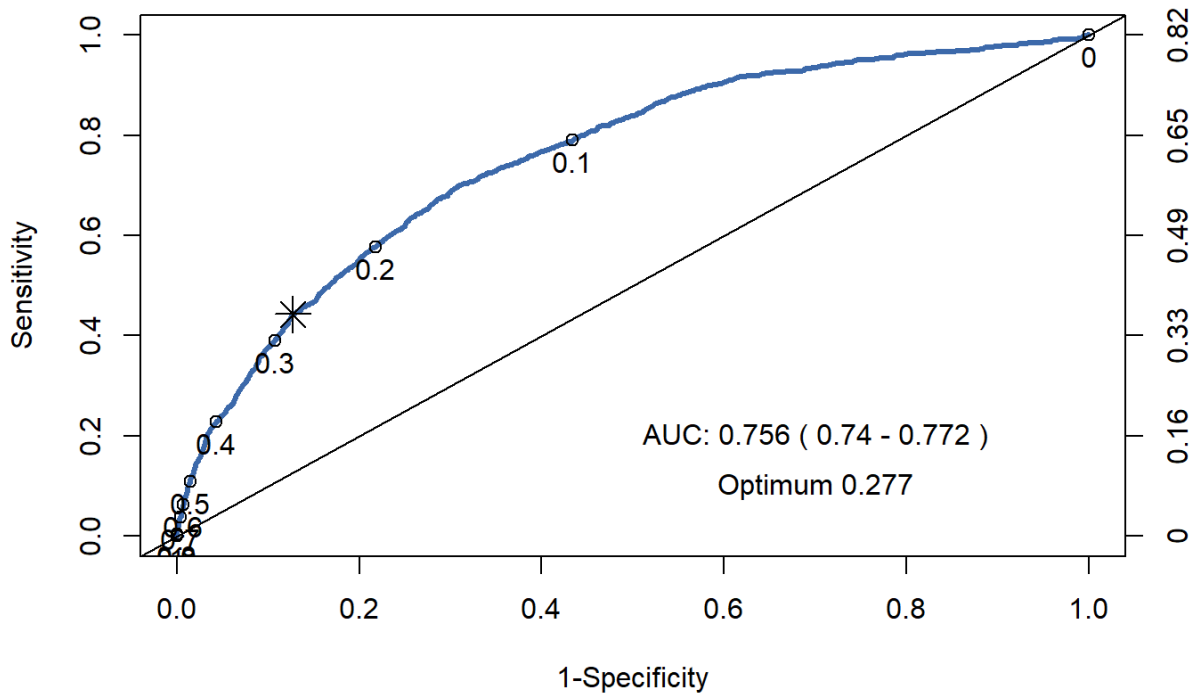
ROC curve of the Barnsley model



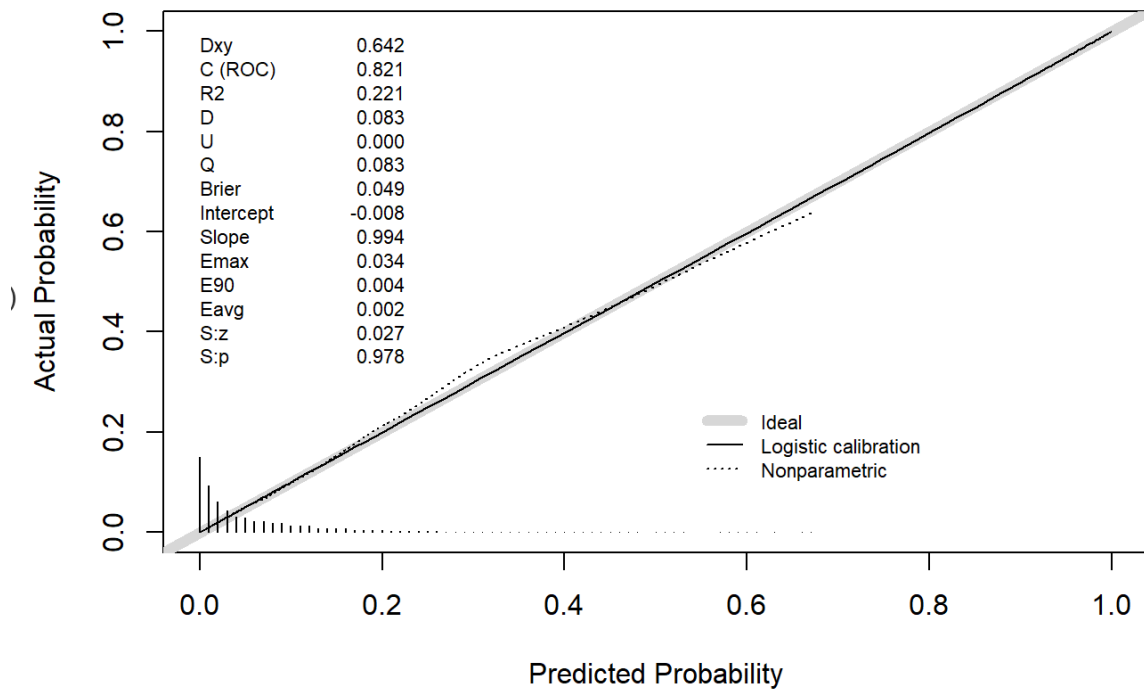
Bradford



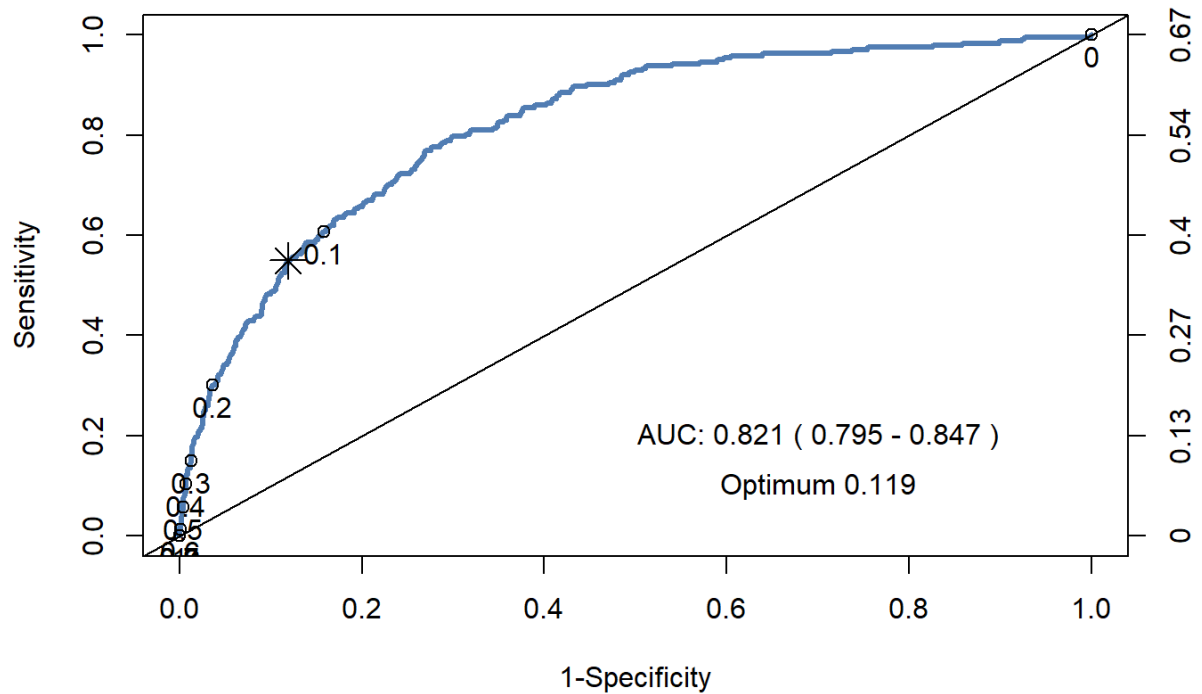
ROC curve of the Bradford model



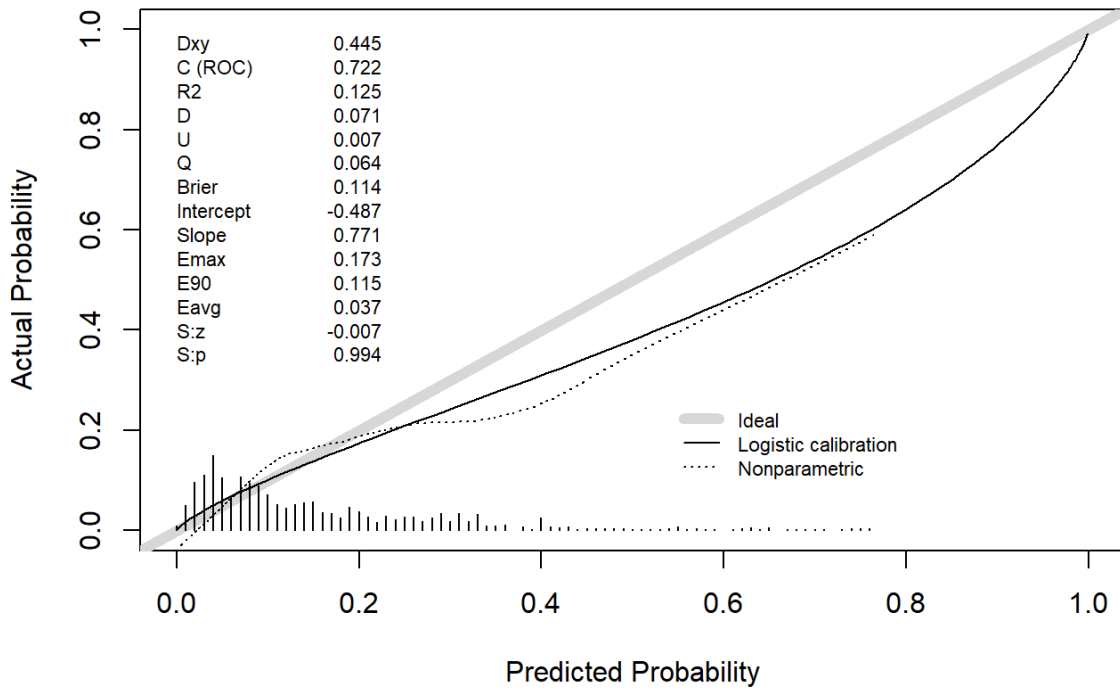
Calderdale



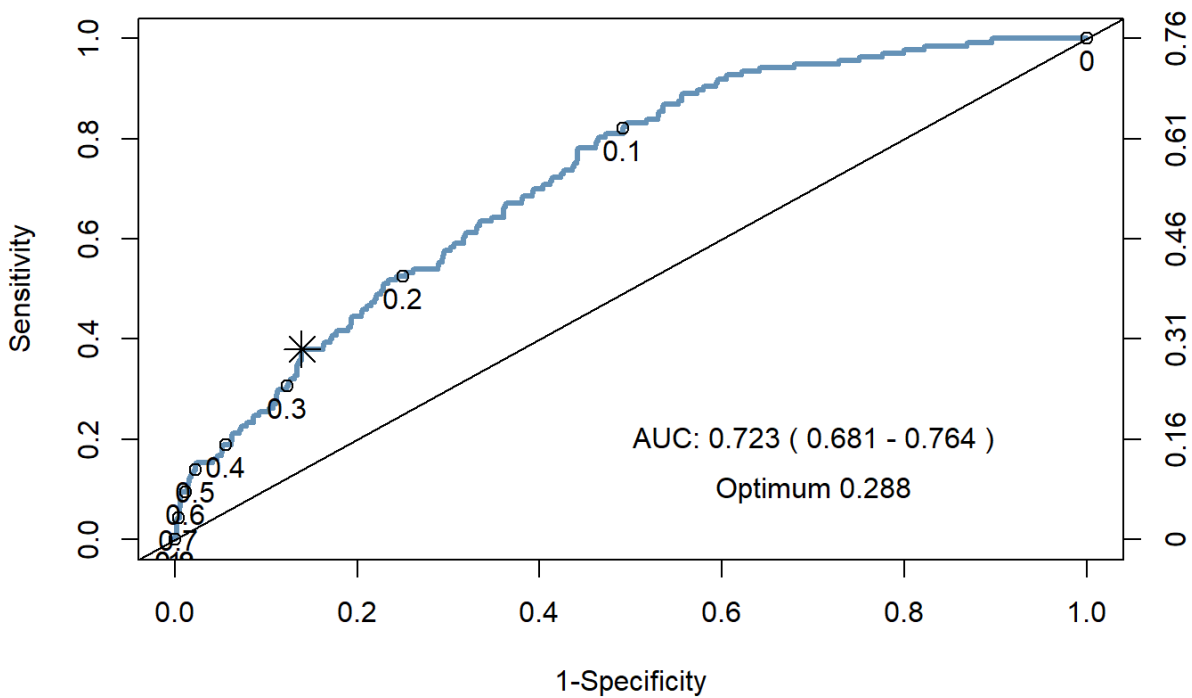
ROC curve of the Calderdale model



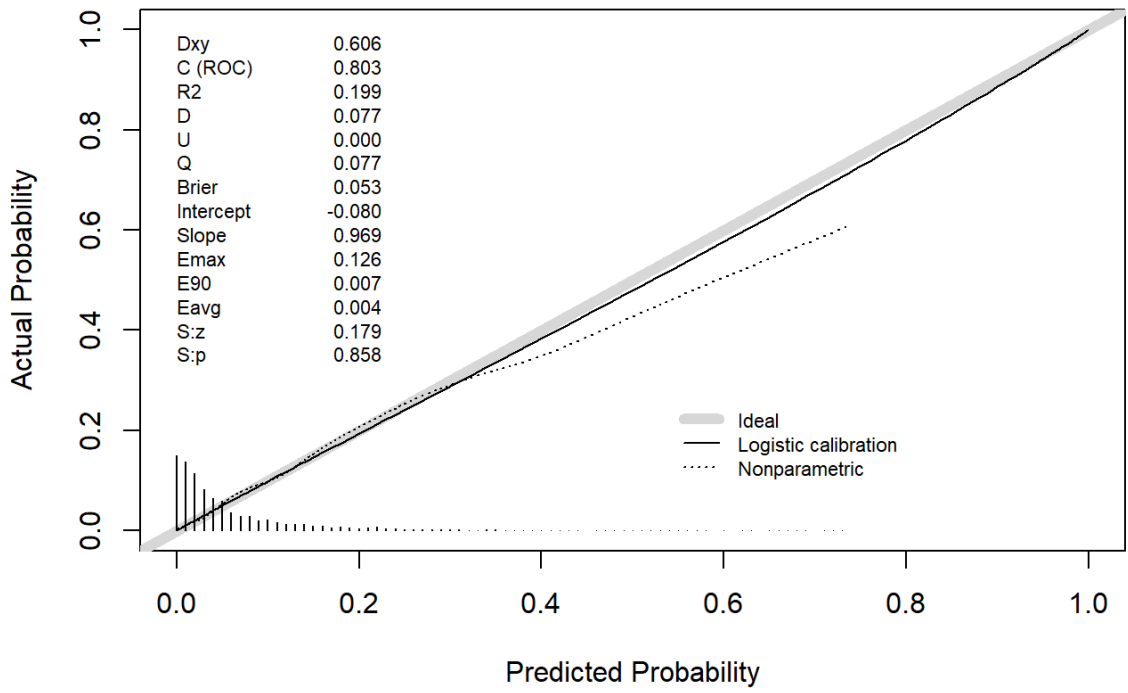
Dewsbury



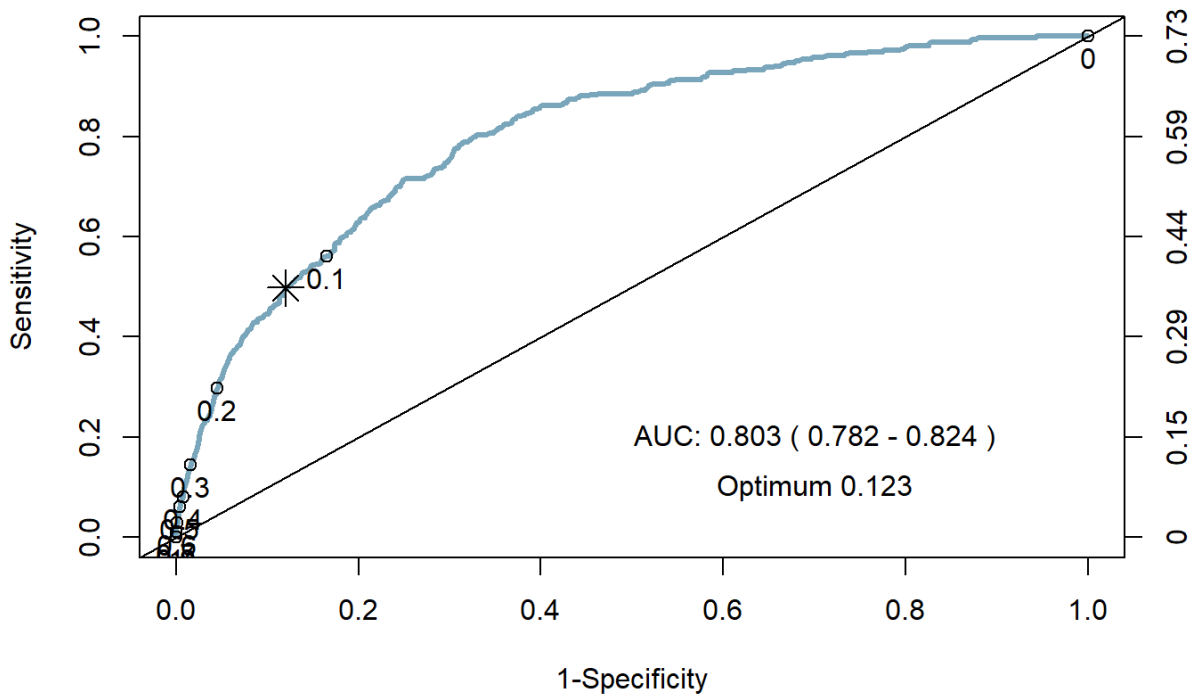
ROC curve of the Dewsbury model



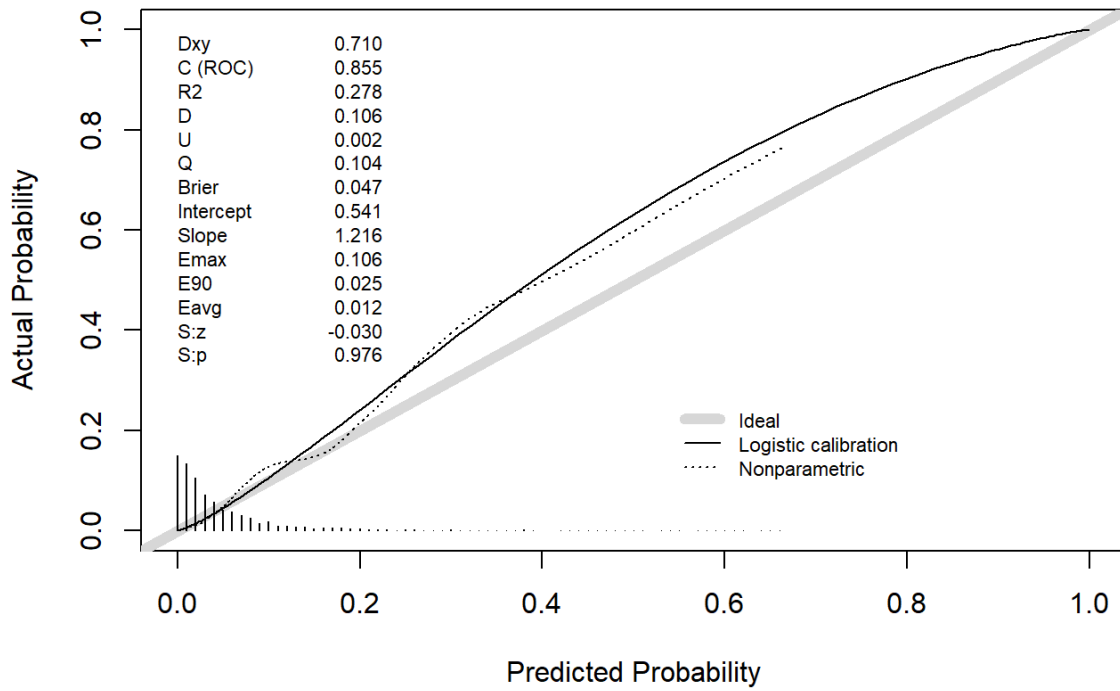
Doncaster



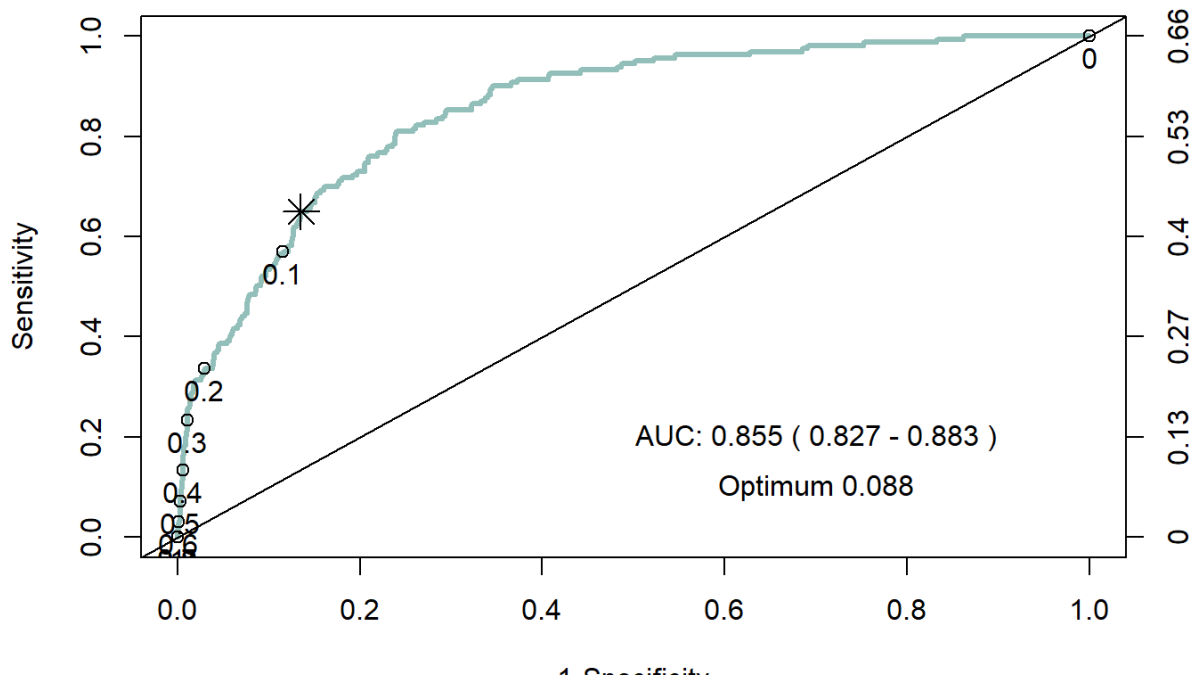
ROC curve of the Doncaster model



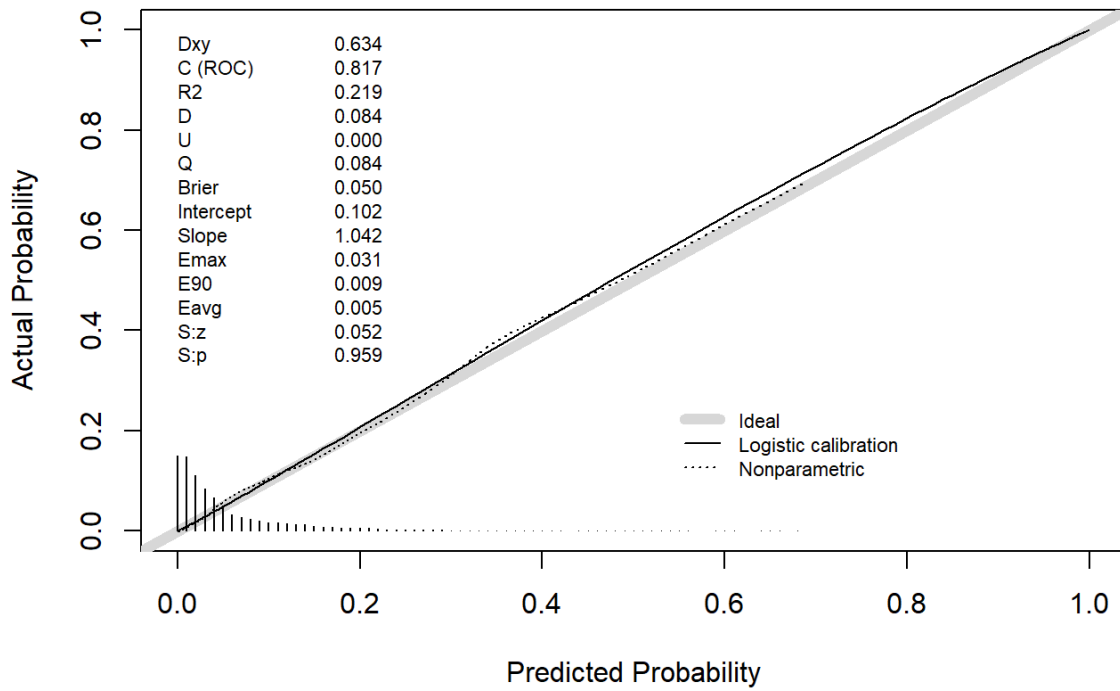
Harrogate



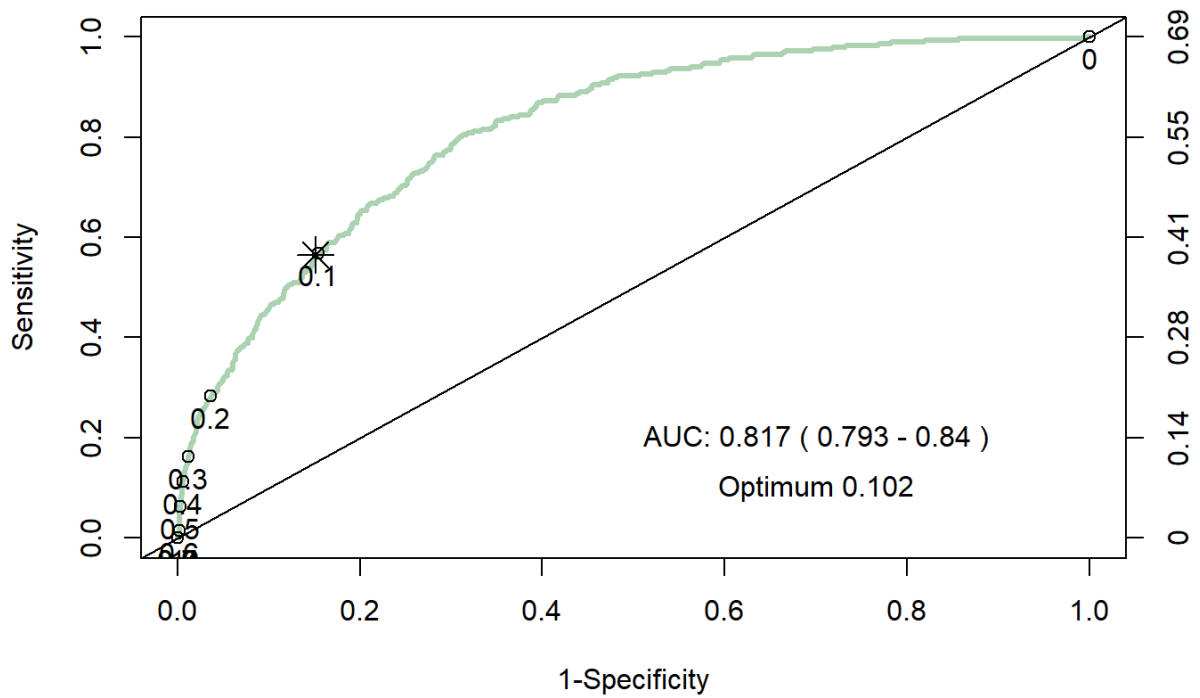
ROC curve of the Harrogate model



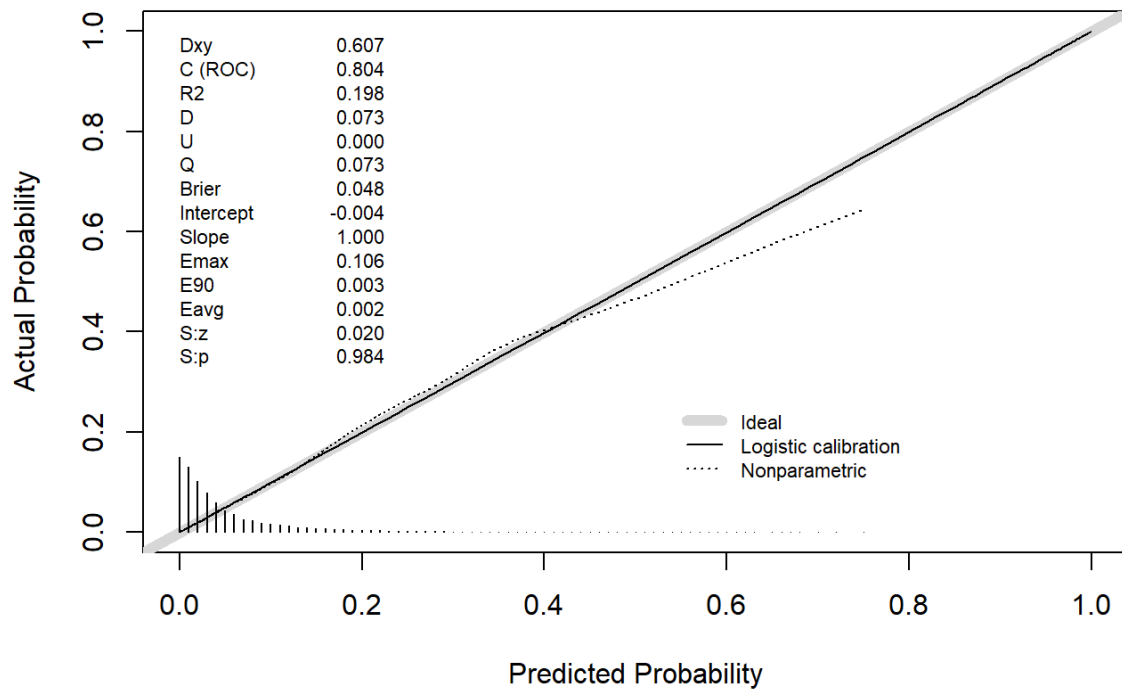
Huddersfield



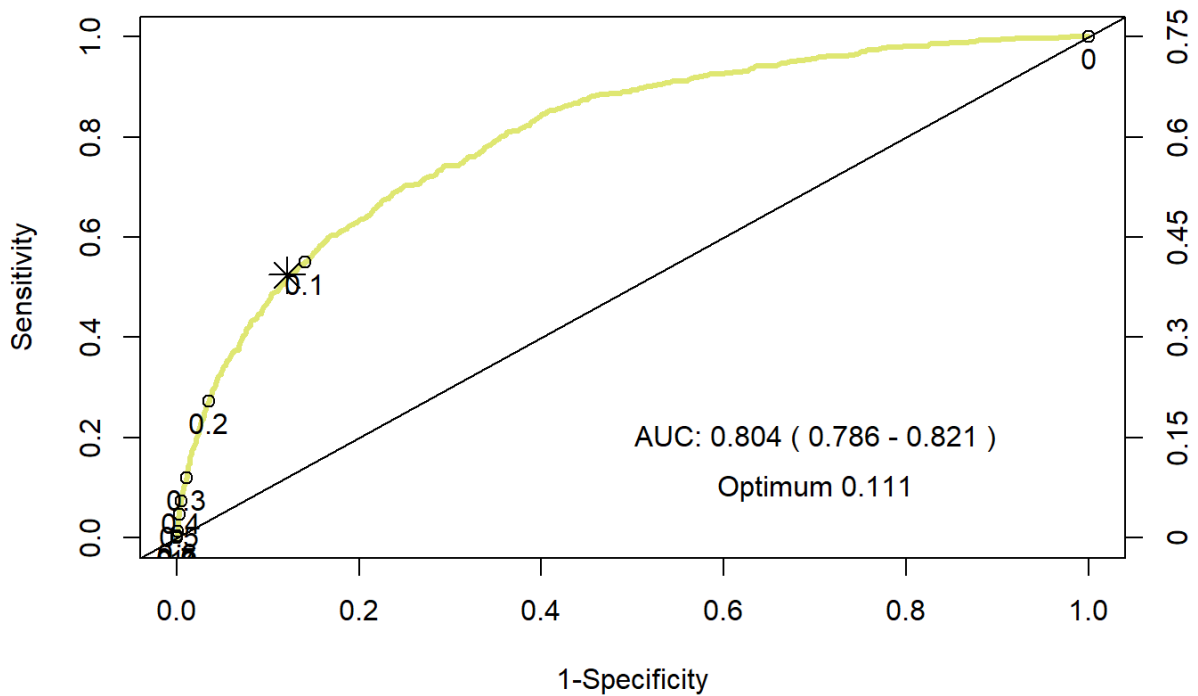
ROC curve of the Huddersfield model



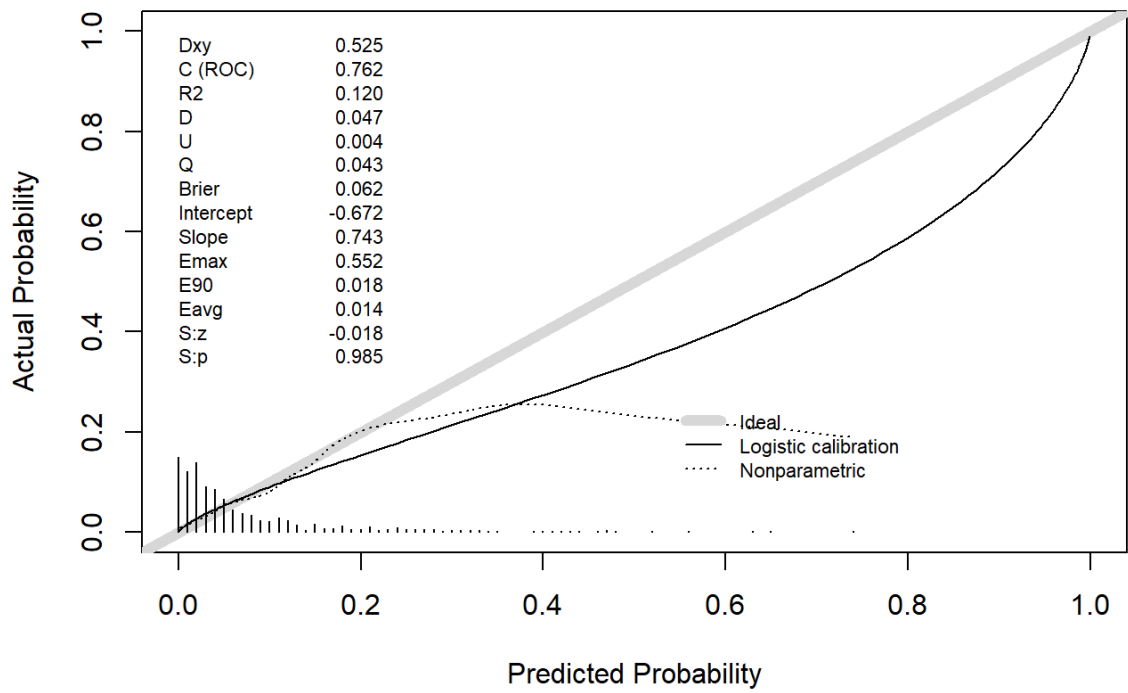
Hull



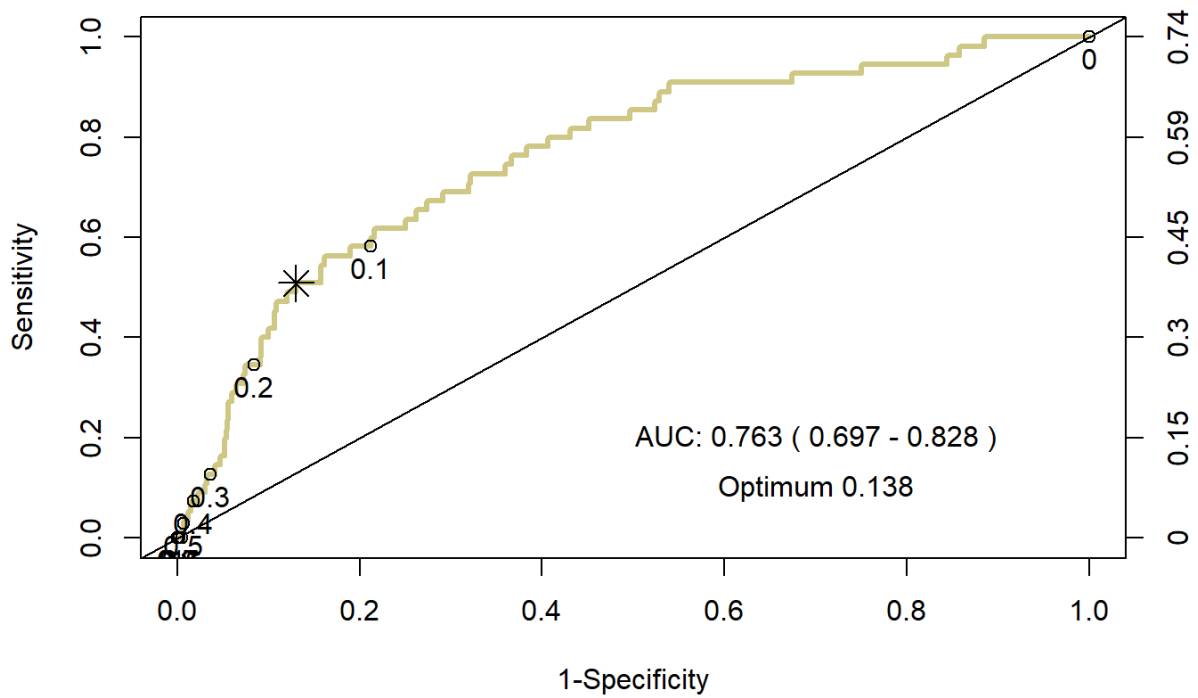
ROC curve of the Hull model



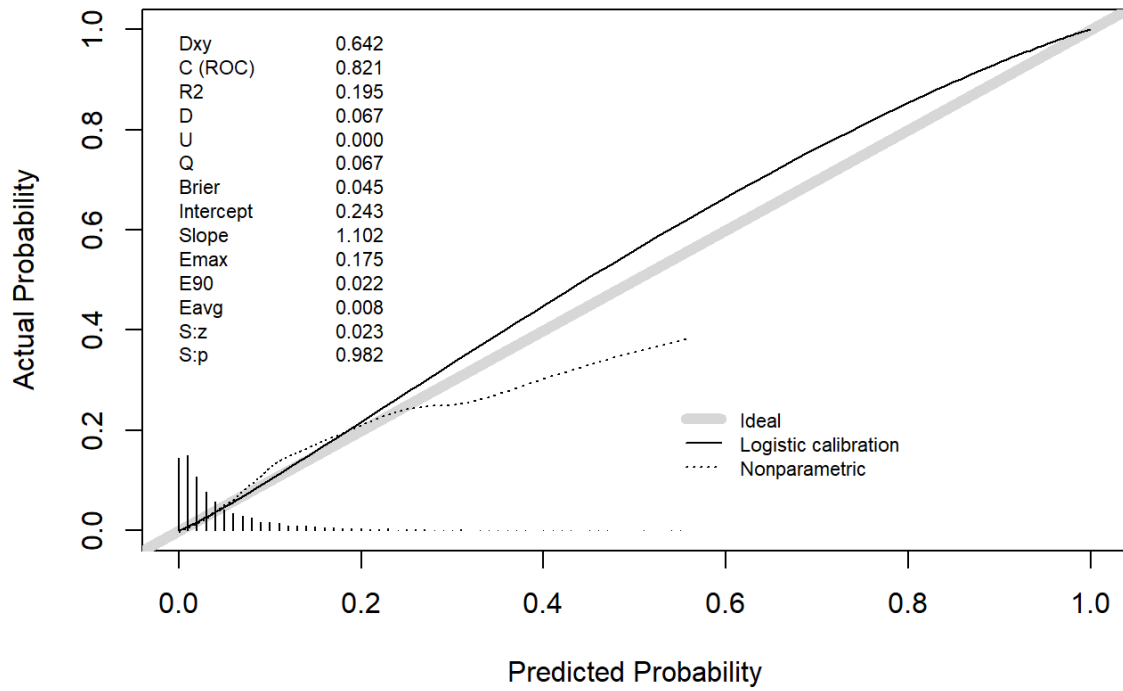
James Cook University Hospital Middlesbrough



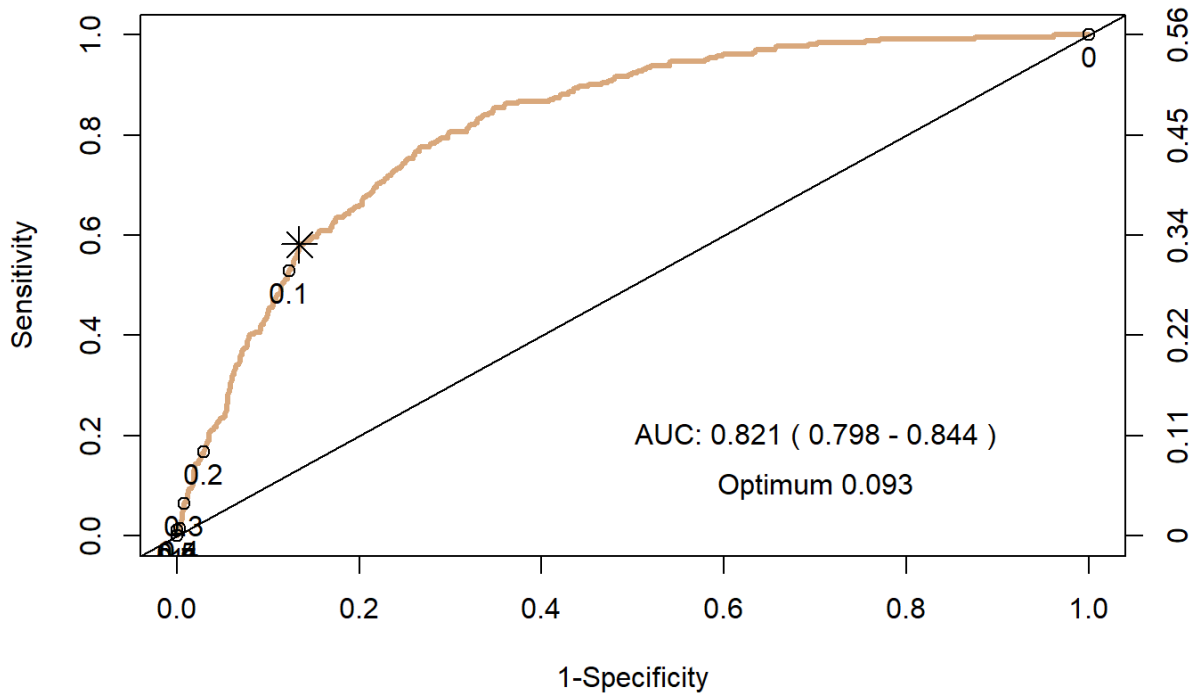
ROC curve of the Middlesbrough model



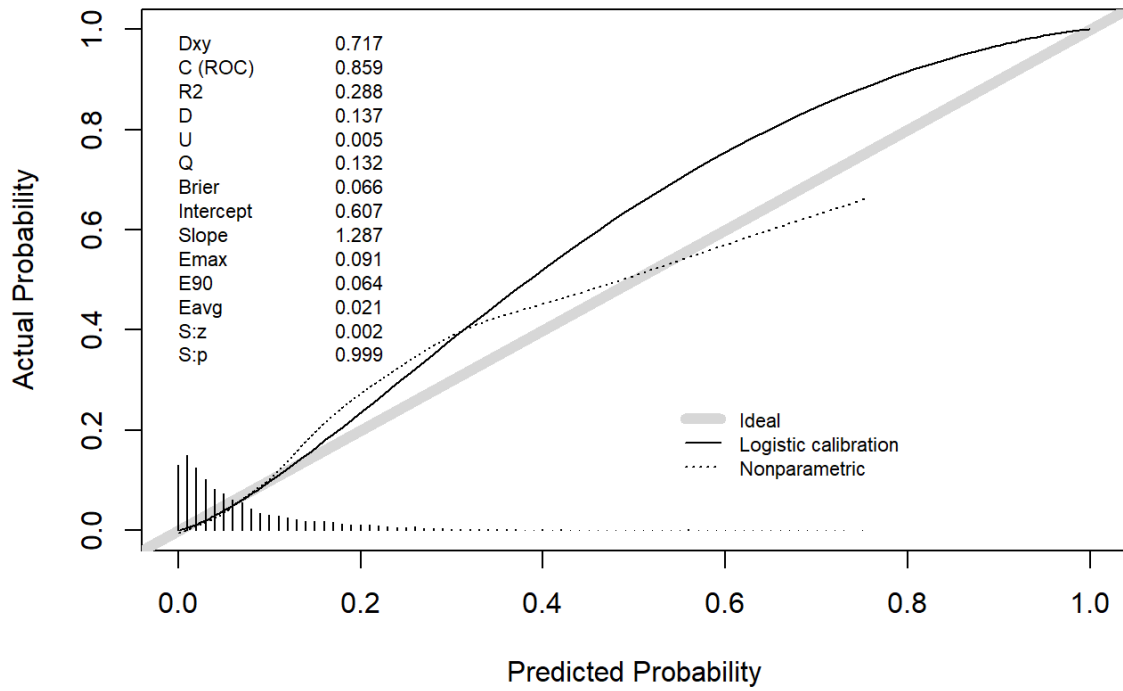
Leeds General Infirmary (LGI)



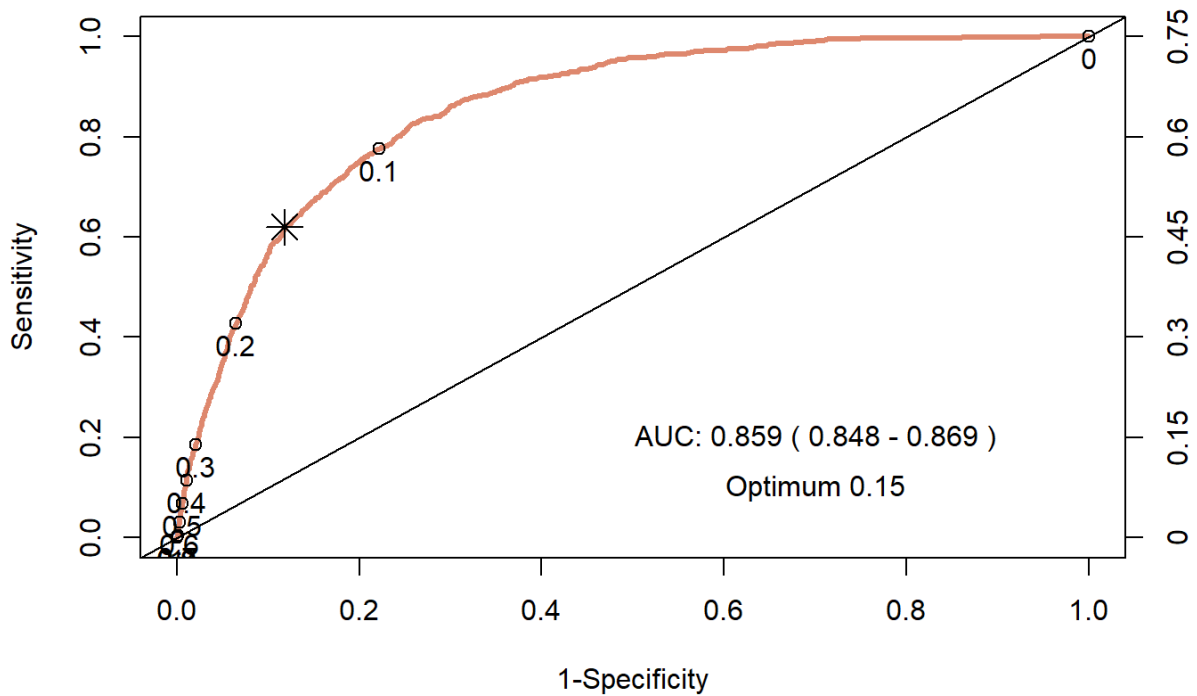
ROC curve of the Leeds 1 model



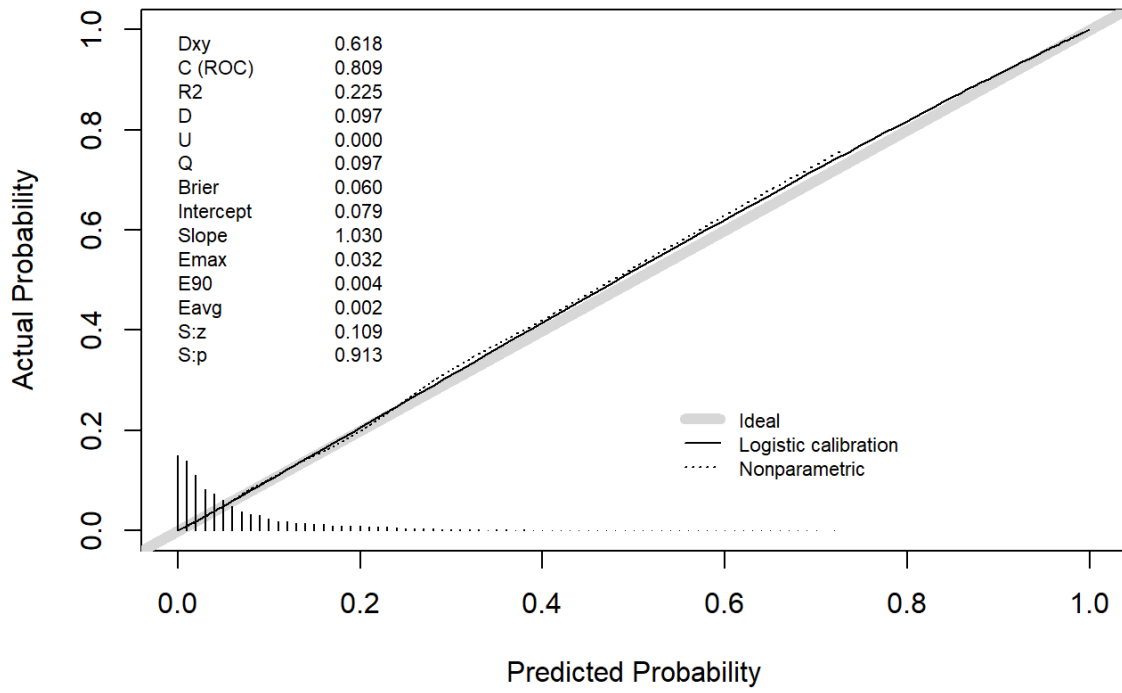
Northern General Hospital Sheffield



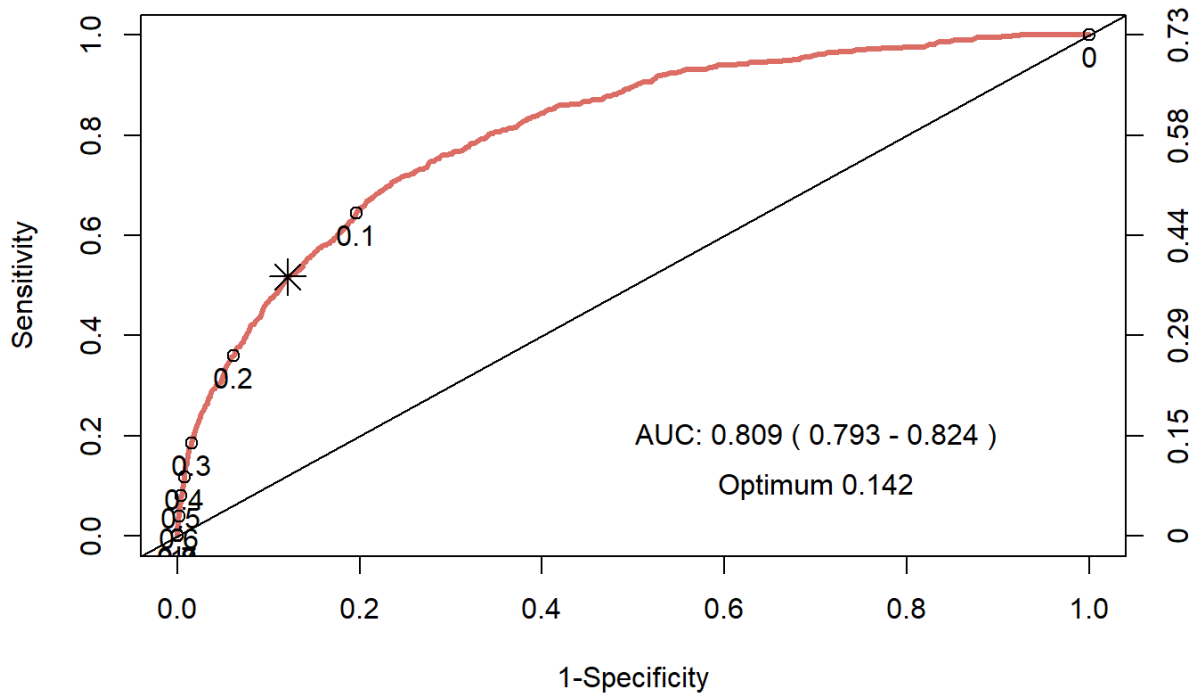
ROC curve of the Sheffield model



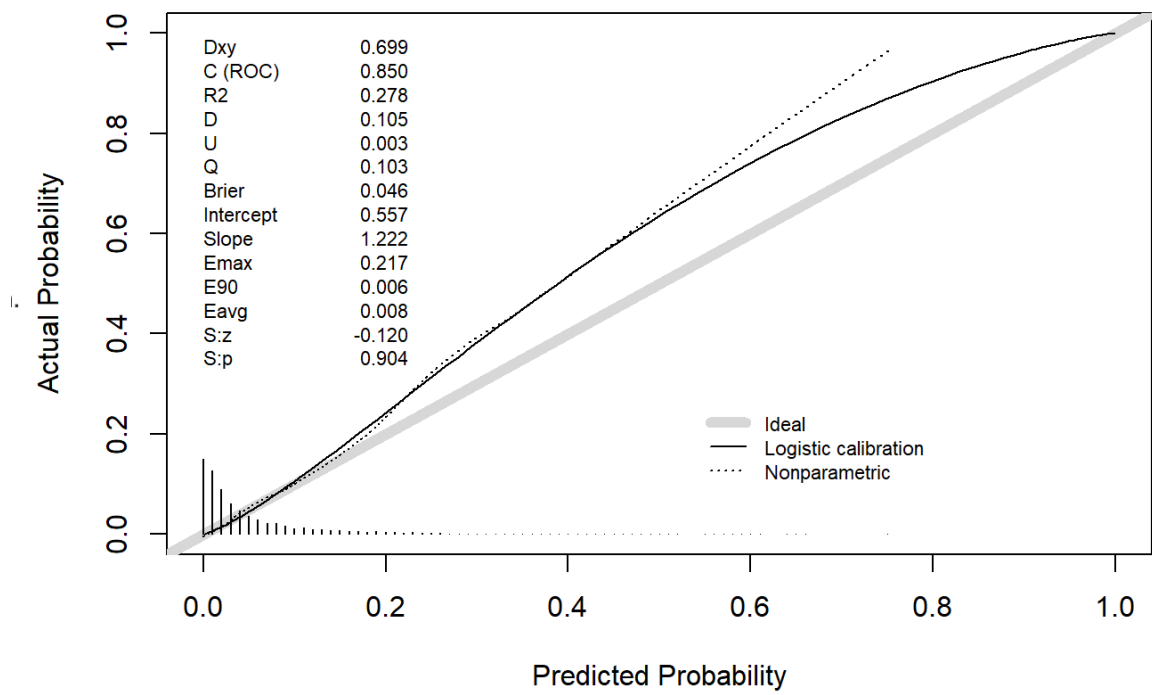
Pinderfields Hospital Wakefield



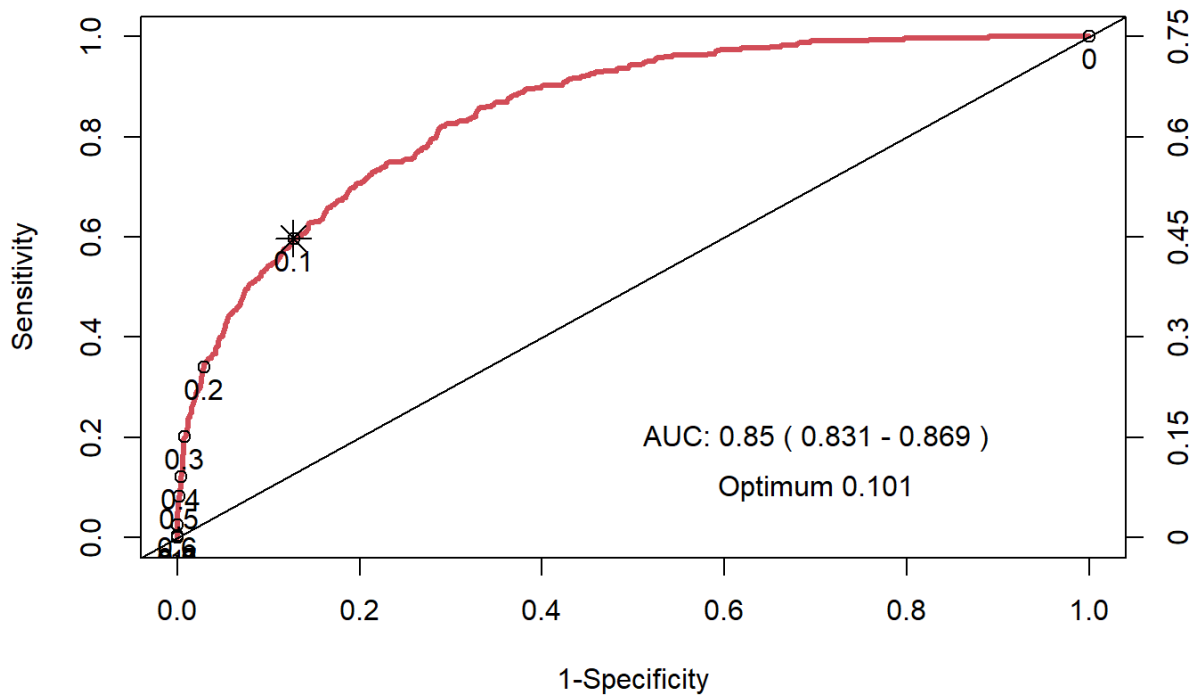
ROC curve of the Wakefield model



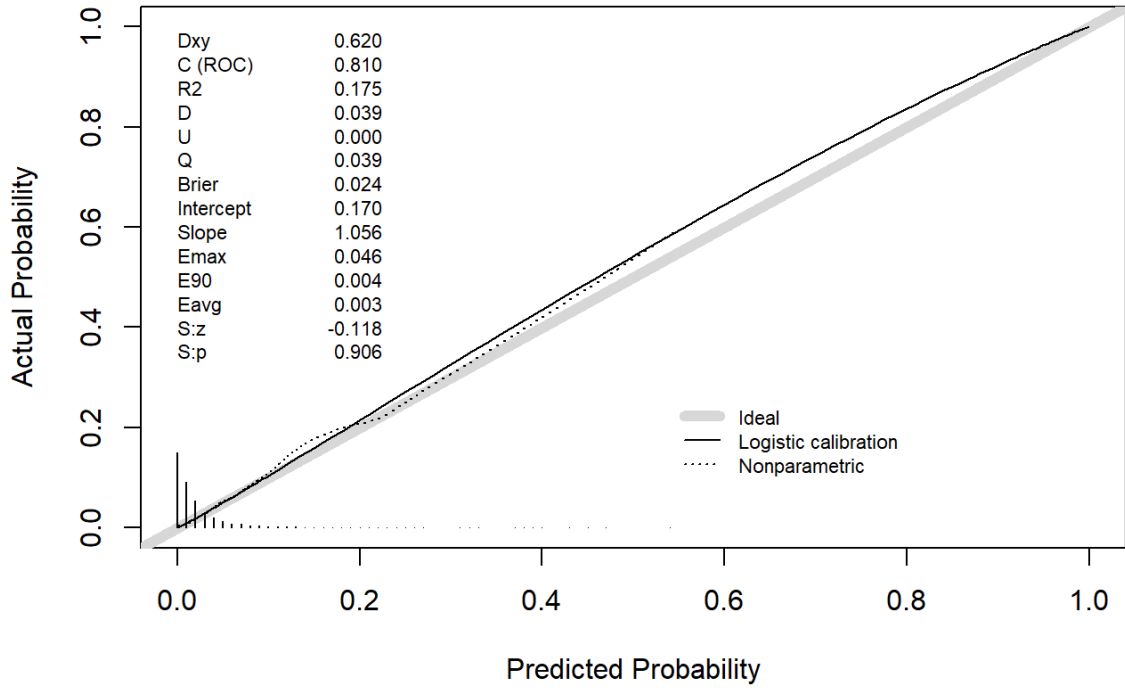
Rotherham



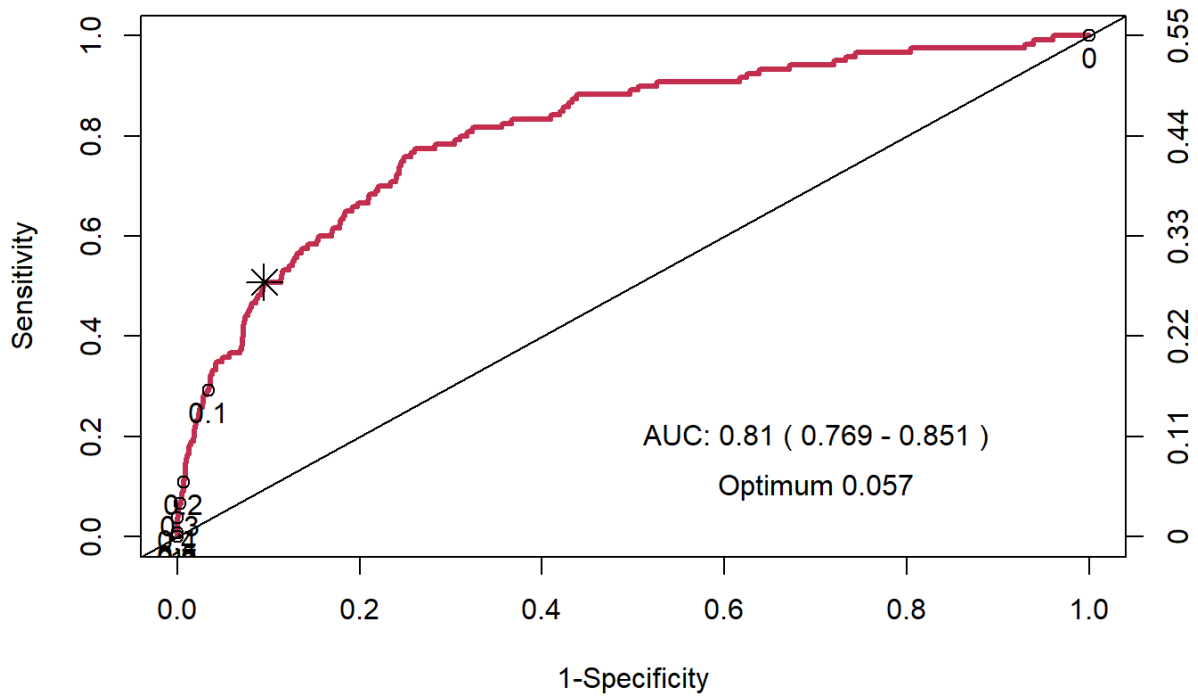
ROC curve of the Rotherham model



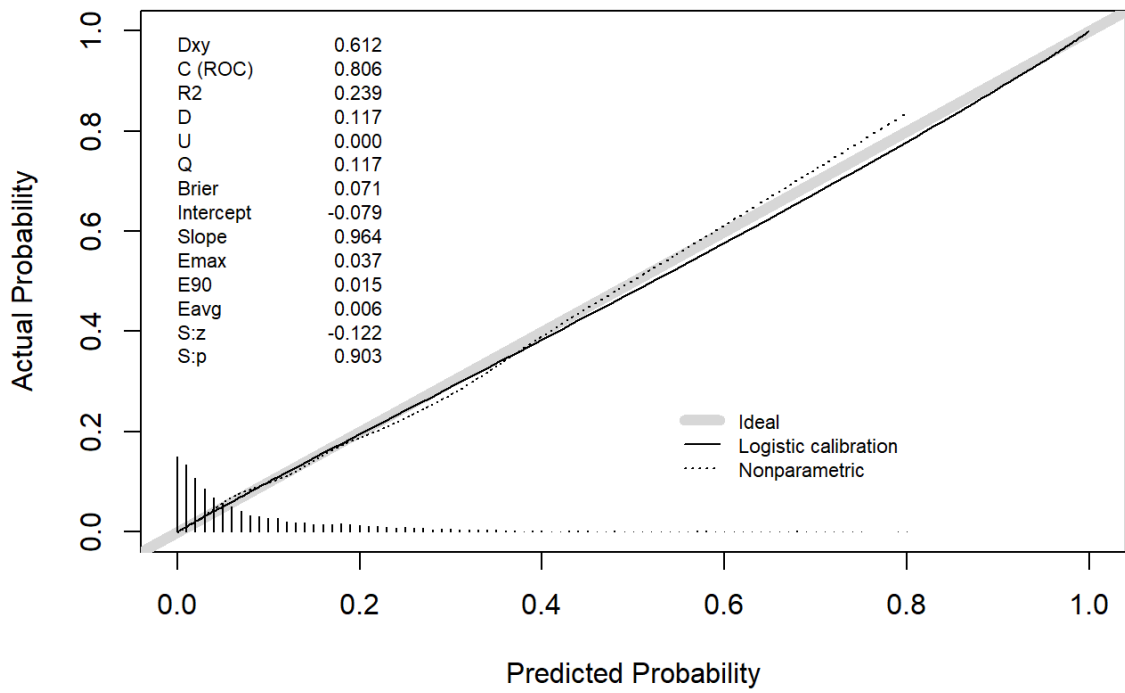
Scarborough



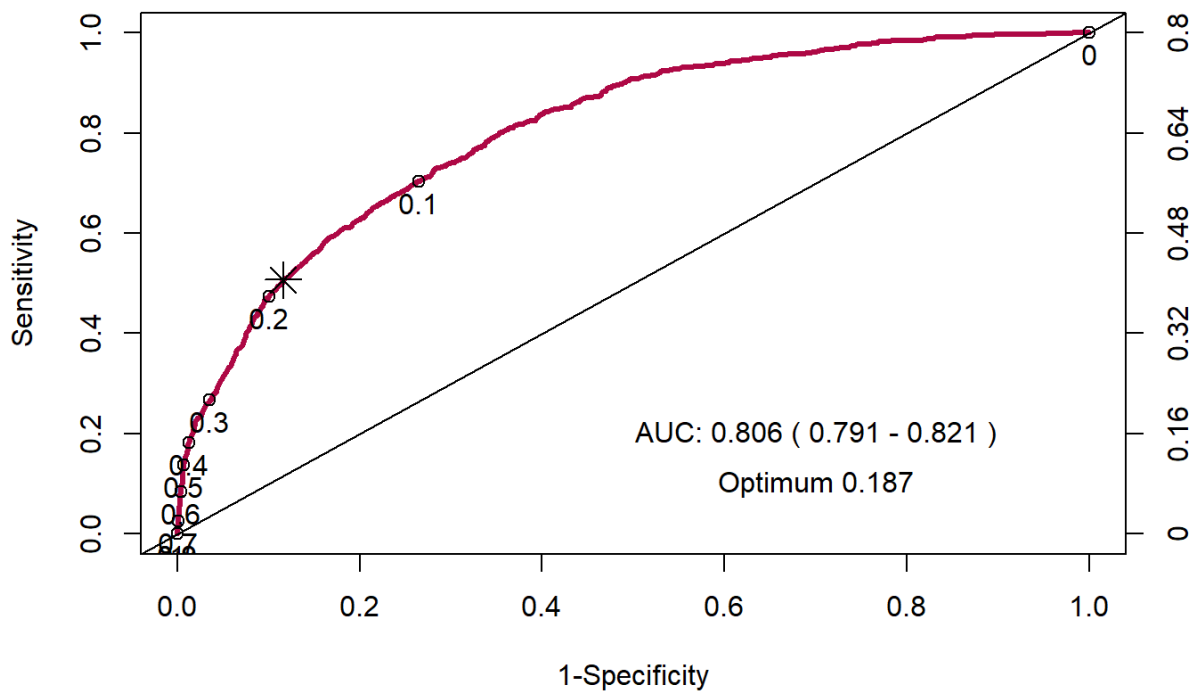
ROC curve of the Scarborough model



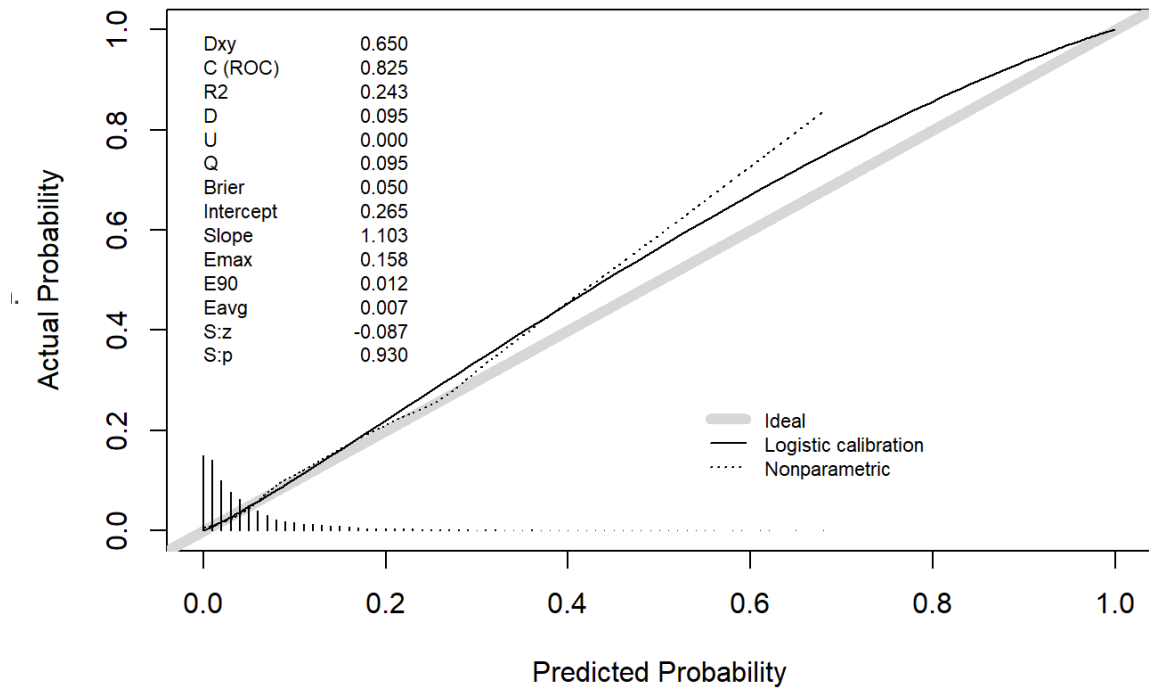
St. James's Hospital Leeds



ROC curve of the Leeds 2 model



York



ROC curve of the York model

