
Artificial Intelligence-Driven Clinical Decision Support for Antibiotic Optimisation

William James Bolton

*Degree of Doctor of Philosophy
Department of Computing
Imperial College London*

13th December, 2024

*Submitted in part fulfilment of the requirements for the degree of Doctor of Philosophy
in Computing of Imperial College London.*

*I declare that this thesis and the work described within was composed by and originated
from myself. Information from the published and unpublished work of others has been
appropriately referenced in the text.*

Supervisors: Professor Pantelis Georgiou, Professor Alison Holmes, Dr Timothy Rawson

Examiners: Professor Mauricio Barahona, Dr Charis Marwick

*The copyright of this thesis rests with the author. Unless otherwise indicated, its
contents are licensed under a Creative Commons Attribution-NonCommercial 4.0
International Licence (CC BY-NC). Under this licence, you may copy and redistribute
the material in any medium or format. You may also create and distribute modified
versions of the work. This is on the condition that: you credit the author and do not
use it, or any derivative works, for a commercial purpose. When reusing or sharing
this work, ensure you make the licence terms clear to others by naming the licence
and linking to the licence text. Where a work has been adapted, you should indicate
that the work has been changed and describe those changes. Please seek permission
from the copyright holder for uses of this work that are not included in this licence or
permitted under UK Copyright Law.*

Copyright © Dr William James Bolton 2024

To my family with all my love

*"Live as if you were to die tomorrow.
Learn as if you were to live forever."*

Gandhi

Acknowledgements

I have somewhat dreaded writing my PhD acknowledgements. Not because I want to take all the credit but because over my entire life, I have developed my scientific writing skills and hopefully I am somewhat good at it. Whereas opening up and sharing my true feelings is something I often struggle to do. As I sit here flying back from my final PhD conference in Cape Town, South Africa, staring over the beautiful African deserts, I am going to try and express my endless gratitude for the support and opportunities I have been given during this PhD and beyond.

First, to my supervisors Pantelis, Alison and Tim, thank you for believing in me, giving me the opportunity to undertake this PhD, and allowing me the freedom to pursue my own research ideas. Pantelis your drive for me to pursue excellence and aim higher contributed towards the strong outputs linked to this thesis. Alison your expertise and network have been invaluable for building meaningful relationships and collaborations. Tim, I was not formally your PhD student, but you took me under your wing and this research would not have been possible without your dedication and time towards supporting and reviewing my work, as well as your heartfelt advice on navigating the sometimes bumpy PhD journey.

I would like to thank everyone I have worked with along the way. Including but not limited to: Dr Mark Gilchrest, Mr Richard Wilson, Dr Bernard Hernandez, Dr David Antcliffe, Dr Cosmin Badea, Dr Ash Myall, Dr Paul Vasicasin and Dr Damien Ming. Indeed, the wider group from the Imperial Health Protection Research Unit (HPRU) in Antimicrobial Resistance (AMR) and Healthcare Associated Infections (HCAI) and the Centre for Antimicrobial Optimization (CAMO) have greatly supported my research. I would like to say thank you to all those involved whether it was an interesting conversation or a entertaining social event. To those in the Centre for Bio Inspired Technology, thanks for making every day enjoyable. To my friends Lewis, Ryan, Tasha, Katerina, Costanza, and Elliot, hanging out whether on away days or whilst watching rugby has been a lot of fun.

Some of the unsung heroes of research are those who ensure everything is organized and runs smoothly. To Britta, Safa, Iza, Rakhee, and Kerri, thank you for all your help along the way in facilitating research resources and opportunities, and putting up with my mad ideas. In particular, I would like to thank Britta, your light-hearted approach and endless work on the Centre for Doctoral Training (CDT) in AI for healthcare has led to an environment where PhD students can thrive. The Imperial Clinical Analytics Research and Evaluation (iCARE) team have been invaluable in providing access to data, without which this work would have never occurred, thank you. Furthermore, Dr Lea Goetz was instrumental in helping me to organise an internship with GSK.ai during my PhD. Lea thank you for the trust, opportunity and support. Researching with you, Rafael, Ed and Gabriela was thoroughly enjoyable and an amazing learning experience. I would like to thank the AI for Health CDT and the HPRU in AMR and HCAI for the PhD funding. I have treasured, travelling to and disseminating my work at international conferences, allowing me to network beyond borders and travel the globe.

I have some incredible friends who I cherish, and without which would not be able to do what I do. Doga, Simon and Josh our adventures and chats are so much fun, I hope they never stop. Special mention to Doga, 'working' in that CDT room at the start of the PhD, helped me through a tough time in my life, and created a life-long friendship. Tom, Charlie, Jack, Jenny, and Lucy, I love hanging out with you all and look forward to continuing to do so as we muddle through life. Emma, meeting you has been the highlight of the last four years of my life. You are endlessly caring, thoughtful, and have helped me more than you know. This included correcting many grammatical errors in this thesis and by teaching me how to (sometimes) slow down and chill. I can't wait for our next adventure! To those I have not mentioned by name, thank you for your friendship and all the good times.

Finally, to my family I am endlessly grateful for your love, support in me pursuing my dreams, and the experiences you have given me in life. Dad, you have always inspired me with your achievements and more importantly mentality and approach towards life. Mum, you instilled in me a work ethic I am proud of and you have been selfless in support of my education throughout my life. I will never forget how all this started with us going through my GCSE and A-level notes in the kitchen. Oli you are the smartest person I know, and I love it when we nerd out together. Tom, we find joy in many of the same things and hanging out is so much fun. It's been amazing to see you grow (not physically your already tall enough) into adulthood and show what you are really made of these last few years. To my wider family, both past and present, I cherish your love, the relationships we have and the support we provide each other. I hope to provide the same to future generations.

Funnily enough, I first visited South Africa with my family in 2015 after finishing A-levels at school. So it feels very nostalgic to have travelled there now at the end of my higher education. Perhaps I will have to make this a tradition after big life achievements. Now back to staring at this beautiful planet of ours, which I am also thankful to call home.

Abstract

Artificial intelligence (AI) is poised to revolutionise healthcare in the coming decades. AI-based clinical decision support systems (CDSS) that aim to help clinicians make complex decisions and deliver precision medicine have been an area of significant research. Such technology has been developed for infectious diseases to tackle the growing global challenge of antimicrobial resistance (AMR). To curtail the development of AMR we must optimize antimicrobial use and prolong the therapeutic life of current drugs, which is called antimicrobial stewardship. In this thesis, AI-driven CDSSs that use routinely collected electronic health record (EHR) data to optimise antibiotic use are researched, developed, and evaluated.

This thesis employed multiple research methodologies to investigate key aspects of decision-making that clinicians currently find challenging when prescribing antimicrobials. Specifically, work focused on intravenous-to-oral switching (IVOS) and stopping treatment, aspects of antimicrobial stewardship that have previously been neglected in the literature. Numerous machine learning and deep learning methods were applied with a web-based app, semi-structured interviews, and surveys, used by clinicians to evaluate the IVOS-focused CDSS and determine its acceptability and impact on decision-making. AI models were able to obtain good performance in stewardship-focused prediction tasks.

IVOS models obtained an AUROC of up to 0.80 while not being biased to individuals' protected characteristics such as sex. Algorithms that investigated stopping treatment accomplished an AUROC of 0.77 for mortality prediction. 42 individuals completed the clinician evaluation study. Often the CDSS did not influence decision-making, with participants able to identify and ignore incorrect advice. Some statistical differences were observed with the AI CDSS influencing clinicians towards not switching the route of antibiotic administration, which aligns with the cautious behaviour reported by 40% of participants.

AI-driven CDSSs hold promise to improve and optimise the complex decisions made by clinicians when prescribing antimicrobials, to benefit healthcare systems, patients, and tackle AMR.

Contents

Acknowledgements	v
Abstract	xi
1 Introduction	1
1.1 Motivation	1
1.2 Research objectives	2
1.3 Thesis organisation	2
1.4 Datasets utilised	6
References	9
2 Background	11
2.1 Chapter objectives	12
2.2 Clinical decision support can help tackle the global threat of antimicrobial resistance . .	12
2.2.1 AMR is caused by natural selection and accelerated through human behaviour . .	12
2.2.2 Antibiotic stewardship is key to mitigate AMR	14
2.2.3 AI clinical decision support can provide innovative solutions	15
2.3 Co-morbidities are a complex, growing problem	19
2.3.1 Antibiotic management in multi-morbid patients is difficult	19
2.3.2 AI co-morbid research occurs at the disease and individual level	20
2.3.3 Challenges remain in co-morbidity research	21
2.4 Moral AI for antimicrobial resistance	22
2.4.1 Inequalities exist within infectious diseases	22

2.4.2	Fairness and biases are a problem for AI in healthcare	23
2.4.3	Antimicrobial resistance raises significant ethical questions	24
2.4.4	Maximising total-average utility in healthcare	25
2.4.5	Bentham’s felicific calculus puts society first	26
2.4.6	Alternative ethical frameworks arrive at different conclusions	29
2.4.7	Antimicrobial resistance is a negative externality	30
2.4.8	Our moral obligation towards the health of future humans	31
2.4.9	Developing intelligent moral decision systems is complex	32
2.5	Conclusion	35
	References	36

3 Co-morbidity Representation in AI 57

3.1	Chapter objectives	59
3.2	Introduction	59
3.2.1	The challenges with co-morbidity representation	59
3.2.2	Current research on co-morbidity representation	60
3.2.3	Leveraging medically grounded knowledge	61
3.3	Methods and tasks	62
3.3.1	Dataset	62
3.3.2	Generating SNOMED disease embeddings	63
3.3.3	Supervised learning task	64
3.3.4	Similar patient retrieval and case-based reasoning tasks	66
3.4	Results	69
3.4.1	SNOMED disease embeddings	69
3.4.2	Supervised learning task	69
3.4.3	Similar patient retrieval task	71

3.4.4	Case-based reasoning antibiotic prescription task	75
3.5	Discussion	76
3.6	Conclusion	80
	References	82
4	Antibiotic Cessation Decision Support	87
4.1	Chapter objectives	89
4.2	Introduction	89
4.3	Methods	90
4.3.1	Dataset	90
4.3.2	Model architecture	91
4.3.3	Model development	94
4.3.4	Model evaluation	95
4.4	Results	99
4.4.1	Data	99
4.4.2	Length of stay and mortality estimation	99
4.4.2.1	Autoencoder	99
4.4.2.2	Synthetic outcome estimation	100
4.4.2.3	Pneumonia and UTIs	102
4.4.3	White blood cell count estimation	103
4.5	Discussion	105
4.6	Conclusion	109
	References	111
5	Intravenous to Oral Switch Decision Support	117
5.1	Chapter objectives	119
5.2	Introduction	119

5.3	Methods	121
5.3.1	Datasets	121
5.3.2	Feature selection	122
5.3.3	Model development	124
5.3.4	Model evaluation	125
5.3.5	Imperial College Healthcare NHS Trust	128
5.3.5.1	Point prevalence survey	129
5.4	Results	130
5.4.1	MIMIC and eICU	130
5.4.1.1	Data	130
5.4.1.2	Feature and model optimisation	132
5.4.1.3	Model evaluation	136
5.4.1.4	Interpretability	141
5.4.1.5	Fairness	141
5.4.2	Imperial College Healthcare NHS Trust	144
5.4.2.1	Data	144
5.4.2.2	Model evaluation	146
5.4.2.3	Prospective point prevalence survey evaluation	148
5.5	Discussion	152
5.6	Conclusion	160
	References	161
6	Clinician Evaluation	169
6.1	Chapter objectives	171
6.2	Introduction	171
6.3	Methods	172

6.3.1	Study and participants	172
6.3.2	Semi-structured interviews	173
6.3.3	Clinical vignette experiment	174
6.3.3.1	Clinical decision support systems	179
6.3.4	Questionnaires	180
6.4	Results	181
6.4.1	Study and participants	181
6.4.2	Semi-structured interviews	182
6.4.3	Clinical vignette experiment	186
6.4.4	Questionnaires	190
6.5	Discussion	194
6.6	Conclusion	200
	References	201
7	Prospective Outlook and Conclusion	209
7.1	Contributions and takeaways	209
7.2	Future work and progressing AI CDSSs in healthcare	211
7.3	Conclusion	218
	References	219
A	Appendix	223
A.1	Evaluation metric formulas	223
A.2	Chapter 3	224
A.3	Chapter 4	226
A.4	Chapter 5	228
A.5	Chapter 6	233

B. Publications	245
B.1 Journal articles	245
B.2 Other articles	246
B.3 Conference abstracts and posters	246
B.4 Conference oral presentations	247
B.5 Additional selected activities	248
C. Code and data	249
D. Glossary	250

List of Tables

1.1	Overview of datasets	7
2.1	Overview of research articles published since 2020 applying AI to antimicrobial optimisation. Those in bold are articles highlighted for their quality. This list is not comprehensive.	18
2.2	Overview of Bentham's felicific calculus variables and example application to starting antimicrobial treatment.	27
3.1	Mean unseen test set AUROC results for classification tasks in different populations. . . .	71
3.2	Accuracy results for the case-based reasoning antibiotic prescription task.	75
4.1	Hyperparameters used for each model and their training	94
4.2	Datasets statistics.	99
4.3	Outcome estimation results for patients in the unseen test set. * indicates statistical significance with alpha set at 0.05.	101
4.4	Outcome estimation results for patients with Pneumonia and UTIs. Results are shown for both the whole population and analysis of what if those who received long treatment durations stopped after day 7. * indicates statistical significance with alpha set at 0.05. .	103
4.5	WBC estimation results for patients in the unseen test set. * indicates statistical significance with alpha set at 0.05.	104
5.1	Dataset demographics and statistics.	131
5.2	Feature sets identified by the genetic algorithm. The first 5 were used within the short model while all the features were used in the long model. The short feature set contained the 1st, 2nd, 4th, 5th and 6th SHAP-ranked features, while the long feature set contained those same 5 alongside an additional 32. Details of the Catch22 transformation key can be found in the appendix (Table A.3.)	134
5.3	Models' hyperparameters and optimisation options.	135
5.4	Evaluation results.	137

5.5	Fairness results.	143
5.6	ICHT demographics and statistics.	145
5.7	ICHT evaluation results	146
5.8	Characteristics of patients from the point prevalence survey	149
5.9	Point prevalence survey model evaluation results	150
6.1	Details of the experimental setup and case vignettes.	174
6.2	Participants' demographic information	181
6.3	Results from coding semi-structured interviews.	185
6.4	All participants' switch choices for each patient (case vignette) by the intervention they experienced (CDSS or SOC). Chi-square statistical analysis results are also shown.	187
6.5	Technology Acceptance Model (TAM) questionnaire results. Note that the questions with a * were added after 7 clinicians had completed the study.	193
A.1	Co-morbidities for the patient in question and for those patients identified as most similar by each method for each question in the multi-morbid patient similarity survey.	224
A.2	Antibiotics selected for incomplete oral absorption analysis. * Selected due to nonlinear absorption.	228
A.3	Catch22 time-series transformation key. Note that the mean and standard deviation were also included. More information on these transformations can be found at https://feature-based-time-series-analys.gitbook.io/catch22-features/	229
A.4	Patient mean mortality by switch event temporal difference.	230
A.5	Threshold optimisation fairness results for the short model.	231
A.6	Threshold optimisation fairness results for the long model.	232
A.7	Chi-squared differences in interview coding results across demographics with a p-value below 0.20.	233
A.8	Statistically significant differences in decision-making observed for patients when stratifying by participant demographics.	234
A.9	Statistically significant differences in decision-making observed through the Cochran-Mantel-Haenszel (CMH) test when controlling for participant demographics.	234

A.10	Statistically significant differences in decision-making observed for patients when stratifying by TAM subgroups.	235
A.11	Shannon entropy values for the switch decisions made by participants for each patient. Note due to two options (i.e., switch or don't switch) the minimum and maximum possible Shannon entropy values are 0 and 1 respectively.	236
A.12	Frequency analysis of clinicians' free text switch choice explanations. Please see the associated code for all the terms used when searching for those with a *. Searches were not case-sensitive.	236
A.13	Hover duration for different components of the vignettes across all cases. Note that the low hover time for CDSS recommendations is because they were buttons only clicked on to access explanations.	237
A.14	System Usability Scale (SUS) questionnaire results.	237
A.15	Statistically significant differences observed in the Technology Acceptance Model (TAM) questionnaire by patient demographics.	238

List of Figures

2.1	Aspects of AI for antibiotic optimisation research aligned with steps in the patient pathway. The amount of research within each stage is indicated by the level of the grey triangle. The later stages of antibiotic stewardship have previously been neglected. Those areas of research investigated during this PhD are highlighted in bold. In particular, Chapter 4 discusses cessation and Chapter 5 describes AI for IV-to-oral switching.	17
3.1	Outline of the co-morbidity representation pipeline developed and tasks used for evaluation.	58
3.2	[A] The frequency distribution for historical diseases in the EHR data. [B] Histogram displaying the number of patient co-morbidities. The exponential decay on both graphs indicates the complexity of using co-morbidities in AI. Note to preserve privacy for both plots only those with counts greater than 5 were included.	69
3.3	t-SNE visualisations of the SNOMED disease embeddings with [A] the nearest neighbour SNOMED distance displayed and [B] high-level disease groups highlighted. Most diseases are embedded next to their neighbours in SNOMED CT, with related disease groups clustered together.	70
3.4	Distribution of SNOMED similarity scores from the similar patient retrieval task. [A] compares our co-morbid patient embeddings method to one hot encodings while [B] compares it to Rocheteau's score. Co-morbid patient embeddings returns patients with smaller scores.	72
3.5	Similar patient retrieval task evaluation by SNOMED similarity score [A, B] and Charlson-Jaccard index [C, D] for patients with different numbers of co-morbidities [A, C] and those with co-morbidities of varying rarity [B, D] (co-morbidities were ordered based on their frequency in the EHR, and then starting from the most common, diseases and patients with those conditions were excluded based on percentages, as such, patients with all co-morbidities are included on the left-hand side while only those with rare diseases are represented on the right-hand side).	72
3.6	Stacked bar plot showing the proportion of votes for patients identified by each method, for each question, in the survey. Co-morbid patient embeddings obtained the most votes for 6 questions with one hot encodings prevailing in the others.	73

3.7	Multiple correspondence analysis visualisations of the survey results with k-mode clusters highlighted. Notice how those in cluster 0 are more closely located implying agreement in the responses. In particular, 8 respondents are highly clustered (note that 2 respondents were identical and hence overlap on this plot). Cluster 1 on the other hand is more heterogeneous.	74
3.8	Three survey examples of the patient in question and the similar patients retrieved by each method. The method that obtained the most votes for each question is shown in bold, with the similarity of co-morbidities indicated through a traffic light colouring scheme. Co-morbid patient embeddings identifies clinically appropriate examples with analogous diseases (e.g., alcoholism and alcohol dependence in question 10).	78
4.1	Overview of the steps taken in order to develop models for antimicrobial cessation synthetic outcome and white blood cell count estimation.	88
4.2	Model illustration. [A] The encoder is trained using both a supervised loss (L_s) and reconstruction loss (L_r) [3]. [B] To estimate outcomes and WBC during testing an embedding is created for every day of each patient's stay, embedding spaces are partitioned temporally and based on whether the patient stopped or continued antibiotics. The closest k neighbours are selected as donors from each embedding space and L2 regression returns weights that minimise the square difference between the patient and the donors. A stop and continue synthetic control outcome and WBC is estimated as a weighted average of the donors' outcomes and WBC respectively.	93
4.3	Overview of an LSTM cell within the autoencoder models for any given time-step. Dimension sizes are shown in 4.1. Note a final sigmoid activation function (not shown) is applied to the output for mortality prediction.	95
4.4	Demonstration of the impact and control evaluation process for stop and continue scenarios. An antibiotic day is defined as each day the patient receives treatment as well as the day they stop. After starting antibiotics, each day the patient receives treatment acts as a stop impact and continue control. This continues until antibiotic cessation or ICU discharge. If the patient stops antibiotics during their ICU stay that initial day when no antibiotics are administered acts as a stop control and a continue impact.	96
4.5	ROC and PRC results for the RNN autoencoder on mortality classification.	100
4.6	WBC estimation RMSE results for all scenarios broken down by the day of antibiotic treatment.	104

4.7	LOS and mortality synthetic outcome estimation results for particular patients. These cases were selected as illustrative examples of four distinct patient scenarios: [A] the patient has a long course of antibiotics, [B] the patient has a short course of antibiotics, [C] the patient dies, [D] the patient survives. In A/B control estimation results show minimal deviation from the true LOS label while the stop impact estimations have a reduced LOS. Results in C/D indicate mortality estimations are temporally dynamic but with little difference between stop versus continue.	106
5.1	Overview of the steps taken to develop machine learning models for antimicrobial switch decision-making.	118
5.2	An overview of the structure of the model used to make IVOS predictions for patients in ICHT. Neural networks used the ReLU activation function and dropout. The set transformer used induced set attention followed by pooling by multihead attention (Lee et al. 2019), each with 4 heads and a hidden dimension of size 160, finally a linear decoder layer was applied.	129
5.3	MIMIC antibiotic spectrum index results for the day before switching (IV) and the day after switching (oral). Error bars indicate the standard deviation.	132
5.4	Preprocessing threshold optimisation graphs. ROC curve, Youden's index and Precision, Recall and F1-Score plots for the short model [A] (1st and 2nd threshold of 0.54 and 0.74 respectively) and the long model [B] (1st and 2nd threshold of 0.52 and 0.79 respectively).	135
5.5	Labels and predictions by IV treatment duration. Plots for the short model [A] and the long model [B].	138
5.6	Mean MIMIC patient LOS outcomes by days between the real and predicted switch event. Plots for the short model [A] and the long model [B]. A negative number on the x axis indicates the predicted switch event was before the real switch event. The opposite is true for positive numbers, while 0 means they occurred on the same day.	138
5.7	Labels, predictions, and LOS broken down by IV treatment duration. [A] demonstrates those patients on oral therapies on average have a shorter remaining LOS than those on IV. [B] shows switch labels for both the MIMIC and eICU datasets. They are generally consistent although eICU sees fewer patients switch on day 3. [C] and [D] are predictions for the transfer learning short and long models respectively on the eICU dataset. Models predict many more patients should not switch when compared to MIMIC results (Figure 5.5). [E] and [F] show how baseline predictions compare to the labels for both MIMIC and eICU.	139

5.8	Example visual representation for a particular patient. ‘Traffic light’ suggestions are initially displayed in a temporal manner as the patient progresses. If required clinicians can obtain more information for any given day of the patient’s stay. Simple textual descriptions are provided alongside detailed tabular graphics to maximise clarity. The text quickly gets across key points, while tables show similar example patients and their features, both with their relative importance and contribution respectively. Furthermore, switch labels and predictions across both thresholds are displayed. Note that the patient’s label would obviously not be available during clinical use, but is shown here to be comprehensive. Finally, the limitations and use cases for the system are clearly labelled.	142
5.9	Sensitive attribute threshold optimisation graphs. [A] indicates that while fairness is now achieved by the short model for the 30-year-old age group, 20-year-olds suffer (Supplementary Table A.5). [B] demonstrates fair performance by the short model across insurance groups. [C] shows equalised odds cannot be achieved by the short model for the native population. [D] exhibits fair performance by the long model across age groups.	144
5.10	Labels, predictions, and LOS broken down by IV treatment duration. [A] and [B] show that predictions are generally temporally balanced with the labels, but that the model starts by predicting fewer patients should switch and over time progresses to predicting more should switch. [C] and [D] shows the distribution of switch labels for the ICHT dataset. [E] and [F] demonstrate those patients on oral therapies on average have a shorter remaining LOS than those on IV.	147
5.11	Mean ICHT patient LOS outcomes by days between the real and predicted switch event. A negative number on the x axis indicates the predicted switch event was before the real switch event. The opposite is true for positive numbers, while 0 means they occurred on the same day.	148
5.12	Time difference in days for model predictions and real-world actions when compared to pharmacists’ recommendations. [A] shows all patients while [B] shows only those where at some point in time both the model and pharmacists perceived switching was appropriate. Note that in the cases where no switch prediction occurred during the patient’s stay, 1 day was added to the calculation to account for the earliest they could possibly switch.	151
5.13	Box plot of IV treatment durations by model predictions and pharmacists’ recommendations for switch and don’t switch.	152
6.1	Overview of the experimental protocol conducted with clinicians in this study [A] and the subsequent analysis undertaken [B].	170

6.2	Screenshots of the web app used for the case vignette experiment. A is a simulated patient from ICHT while B is from MIMIC. Both examples are from the video demonstration used to inform participants on how to complete the study.	176
6.3	Distribution of participants' familiarity with AI.	182
6.4	Distribution of codes for key themes explored in semi-structured interviews.	184
6.5	Participants' switch decisions for patients 6, 7 and 10 by intervention.	188
6.6	AI and guideline CDSS explainability click-through rate for each case vignette.	189
6.7	Result distribution for key questions asked in the TAM questionnaire.	192
A.1	Screenshot of a section from the multi-morbid patient similarity survey. Note that the context, instructions, and patient similarity questions given to participants were asked in a plain text format so as not to introduce any biases.	225
A.2	Results for the elbow method on K-modes clustering of the survey data. Note that the infection point occurs at two clusters.	225
A.3	Principal component analysis visualisation of all the embeddings for WBC estimation within the unseen test set.	226
A.4	Distribution of continuing and stopping antibiotic treatment by day within the test dataset. Excluding the particular patient i whose outcomes are being investigated, those who continue on a given day are within the continue donor pool, while those who stop on that day are within the stop donor pool.	226
A.5	Overall antibiotic treatment duration and length of stay for patients. The normalised count represents the number of distinct occurrences for that specific day, while the raw count represents the number of days of data available.	227
A.6	SHAP and feature value distribution for the 20 most important features. Notice how 'respiratory rate11_current_stay' and 'blood pressure mean11_current_stay' (features of importance 3 and 4 respectively) are very similar implying redundancy.	228
A.7	eICU test results broken down by hospitals.	232
A.8	Participants' stewardship responsibilities by their medical speciality. Note that those with a medical speciality of 'other' are not involved in guideline creation.	233
A.9	Distribution of TAM results by patient subgroups for statistically significant differences.	239

Chapter 1

Introduction

1.1 Motivation

*"AI is the most powerful
technology force of our time."*

Jensen Huang
CEO NVIDIA, 2020

*"The world is facing an
antibiotic apocalypse."*

Dame Sally Davies
Former Chief Medical Officer for
England, 2017

These two quotes encapsulate why I pursued this PhD. Artificial intelligence (AI) has seen rapid development in recent years and is poised to disrupt numerous industries and aspects of our lives, including healthcare (Makridakis 2017). Meanwhile, antimicrobial resistance (AMR) is one of the key challenges of our generation, resulting in millions of deaths annually with daunting projections over the next few decades (O'Neill 2014; Murray et al. 2022). Fighting natural selection is not trivial. To reduce the rate of development of AMR we must optimise the use of antimicrobials both on a societal and individual patient level. This is often called antimicrobial stewardship which is a form of precision medicine where actions are tailored to specific groups or individuals, which has been hailed as the future of healthcare (Johnson et al. 2021). However, taking a precision approach to antibiotic prescribing requires a significant amount of information, analysis, and specialist knowledge. This is particularly challenging in multi-morbid patients, where competing priorities and complex clinical presentations make optimal antibiotic management difficult. Today, 30% of the world's data is generated by the

healthcare industry (Markets n.d.), using technology such as electronic medical records. AI is therefore an essential component of revolutionising healthcare and addressing AMR by acting as a bridge between data, medical professionals, and patients.

1.2 Research objectives

Research and development into AI for antimicrobial optimisation remains in an early stage (Rawson et al. 2019). In particular, studies focused on antibiotic stewardship and the later stages of decision-making are lacking (see Section 2.2.3). This thesis aims to fill these gaps in the literature and tackle challenges such as co-morbidities through understanding whether AI clinical decision support systems (CDSSs) can be used to support antibiotic prescribing. CDSSs augment clinicians' complex decision-making processes and have been adopted alongside electronic health records, but applications involving AI and those that specifically recommend an action remain limited (Sutton et al. 2020). Given this bottleneck of AI adoption in infectious disease, this research focuses on proving the potential benefit of AI applications to real clinical problems, and understanding if such systems can actually influence clinical decision-making. Therefore, the research hypothesis of this PhD is:

*Artificial intelligence-based clinical decision support systems
utilising routinely collected electronic health record data
can support antibiotic stewardship*

1.3 Thesis organisation

This thesis is organised into seven chapters. A brief overview of the chapters and the research questions they tackle is presented below.

Chapter 2: Background

Overview: This chapter provides an introduction to the topics covered in this thesis, including infectious disease, antimicrobial resistance, AI clinical decision support systems, co-morbidities and moral AI.

Research Questions: How can we mitigate AMR? What decision support systems currently exist? Why are co-morbidities complex? What moral issues are raised by AMR and AI? How can ethical frameworks help us work towards intelligent moral decision systems?

Contributions: An overview of the current literature on AI-driven clinical decision support systems for antimicrobial optimisation. Application of an ethical framework to ensure a moral approach towards AI decision support for AMR.

Chapter 3: Co-morbidity Representation in AI

Overview: This chapter presents a novel approach to representing co-morbidity data in AI systems which is important for AI applications in healthcare including infectious diseases.

Research Questions: Why is comorbidity representation challenging? How can we use expert medical knowledge to represent patients and their co-morbidities? What clinically relevant tasks can such representations support?

Contributions: A pipeline for creating disease and patient embeddings using the fundamental structure of a common clinical coding system (SNOMED CT). Novel tasks and metrics for evaluating models that use co-morbidities in machine learning.

Chapter 4: Antibiotic Cessation Decision Support

Overview: In this chapter, an AI model is researched to investigate the impact of stopping antibiotic treatment on patient outcomes and treatment response.

Research Questions: Why is the optimal antibiotic treatment duration poorly understood? How can AI be used to estimate the impact of stopping or continuing antibiotic treatment?

Contributions: Bi-directional long-short-term memory (LSTM) autoencoder models that create embedding representations of patients for each day and predict clinically relevant variables (e.g., outcomes and treatment response). A machine learning and synthetic control-based method for estimations under alternative antibiotic treatment scenarios.

Chapter 5: Intravenous to Oral Switch Decision Support

Overview: In this chapter, an AI approach towards predicting when a patient can switch from intravenous (IV) to oral antibiotic treatment is investigated.

Research Questions: Why is early oral switching not frequently undertaken? What are the potential benefits of early oral switching? What performance can an AI model achieve when predicting a patient's route of administration? Can such an approach be generalisable to different healthcare settings? How can we improve models' fairness and provide interpretability? How do AI predictions compare temporally to antimicrobial pharmacists' recommendations?

Contributions: A pipeline for developing simple, generalisable, interpretable and unbiased machine learning models using routinely collected temporal clinical data. Rigorous evaluation of models that predict when a patient may be suitable for switching from IV to oral antibiotic treatment across numerous datasets.

Chapter 6: Clinician evaluation

Overview: This chapter uses clinicians to evaluate an AI CDSS through a research study using a web-based application, case vignettes, interviews and surveys.

Research Questions: What are the nuances and challenges of evaluating AI CDSSs for antibiotic stewardship? How can an AI CDSS influence antimicrobial decision-making? What factors impact the adoption of such technology? What are healthcare professionals' perspectives on AI?

Contributions: A demo web app interoperating the AI CDSSs for IV to oral switching developed in Chapter 5. A research study with 42 clinicians recruited indicating how the AI CDSS can influence antibiotic prescribing decisions as well as barriers to implementation and adoption.

Chapter 7: Prospective Outlook and Conclusion

Overview: This chapter further discusses the research presented in this thesis and provides a prospective outlook on the field of AI decision support systems in healthcare more broadly including future research directions and how to potentially realise clinical impact.

Contributions: A summary of the contributions and key takeaways from this thesis. A prospective outlook on the future of AI CDSSs in healthcare.

Appendix

The appendix contains details of the evaluation metrics used and additional tables and figures resulting from the research.

Publications

Lists the journal and conference publications achieved with this research.

Code and data

Notes how to access the code produced by this research and the datasets used.

Glossary

Contains a list of acronyms and their meanings used within this thesis.

1.4 Datasets utilised

The research presented in this thesis utilised multiple datasets, the details of which are presented in Table 1.1. MIMIC-IV (4th version of the Medical Information Mart for Intensive Care database) is a large de-identified real-world clinical dataset that is publicly available for clinical research (Goldberger et al. 2000; Johnson et al. 2022). It contains data for patients admitted to the Beth Israel Deaconess Medical Center (BIDMC) in Boston, Massachusetts, between 2008 and 2019. This dataset was utilised to research AI models to support antimicrobial switching and cessation as discussed in Chapters 4 and 5. MIMIC-IV was often the initial dataset driving the research in this thesis given it is freely available, large with respect to both the number of patients and their associated variables, and frequently utilised in the AI for healthcare

Table 1.1: Overview of datasets

	MIMIC-IV	eICU	iCARE ICHT COVID-19 Database	iCARE ICHT COVID-19 Database 2023
Release date	2020	2018	2021	June 2023
Time span	2008 – 2019	2014 – 2015	2020 – 2021	2020 – June 2023
Patient count	180,733 (50,920 in ICU)	200,859	126,521 (5,455 in ICU)	170,376 (9,851 in ICU)
Stay count	431,231 (73,181 in ICU)	139,367	297,391 (6,054 in ICU)	522,509 (11,089 in ICU)
Hospitals	Beth Israel Deaconess Medical Center	335 ICUs across 208 hospitals	Imperial College Healthcare NHS Trust	
Geographical spread	Single hospital in Boston, USA	Nationwide, USA	Single hospital in London, UK	
Focus	General and ICU patients	ICU	General and ICU patients	
Data	Admission information, demographics, diagnoses, vital signs, laboratory measurements, prescriptions, microbiology (note that this is not inclusive and differences in the data available and how it is recorded exist between datasets)			
Accessibility	Publicly available	Publicly available	Limited access	

community. The eICU Collaborative Research Database contains data for patients admitted to numerous intensive care units (ICUs) across the United States from 2014 to 2015 (Goldberger et al. 2000; Pollard et al. 2018; Pollard et al. 2019). This dataset was used to assess the external validity of our methodological approach towards models that support antimicrobial switching developed in Chapter 5. eICU was selected in this case due to its focus on the ICU, and similarity to MIMIC with regards to being from the USA and the variables available. However, it has broader generalizability given the diverse hospital settings included in the dataset. Finally, the Imperial College Healthcare NHS Trusts (ICHT) Clinical Analytics, Research and Evaluation (iCARE) environment contains numerous datasets that encompass de-identified real-world clinical data from those with an inpatient encounter at ICHT. The original COVID-19 Database accessed contained data for patients with an inpatient encounter from 1st January 2020 onwards. An updated version of this dataset containing admissions from a broader time frame was released in 2023. The first dataset was used in multiple research projects. In Chapter 3 patients and their co-morbidities were extracted and our methodology was assessed through

tasks. In Chapter 5 models were developed to support antimicrobial switching in the ICHT population, and in Chapter 6 simulated case vignettes were created. The updated ICHT dataset released in 2023 was used for prospective evaluation of the ICHT intravenous to oral switch model in Chapter 5. The iCARE datasets were used for these purposes given the presence of SNOMED diagnosis codes and to extrapolate the research to a more relevant and local setting. All work complied with the data use and ethical regulations required to access each respective dataset. Additional data resources used are discussed in their respective chapters' methods section. The diversity of datasets used in this thesis demonstrates the depth, breadth, and generalisability of this research.

References

1. Makridakis, Spyros (June 2017). "The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms". In: *Futures* 90, pp. 46–60. ISSN: 0016-3287. doi: 10.1016/j.futures.2017.03.006. URL: <https://www.sciencedirect.com/science/article/pii/S0016328717300046>.
2. O'Neill, Jim (2014). "Antimicrobial Resistance: Tackling a Crisis for the Health and Wealth of Nations". In: *Review on Antimicrobial Resistance*, pp. 1–16.
3. Murray, Christopher JL et al. (Feb. 2022). "Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis". In: *The Lancet* 399.10325, pp. 629–655. ISSN: 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(21)02724-0. URL: [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(21\)02724-0/fulltext](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)02724-0/fulltext).
4. Johnson, Kevin B. et al. (Jan. 2021). "Precision Medicine, AI, and the Future of Personalized Health Care". In: *Clinical and Translational Science* 14.1, pp. 86–93. ISSN: 1752-8054. doi: 10.1111/cts.12884. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7877825/>.
5. Markets, RBC Capital (n.d.). *RBC Capital Markets | Navigating the Changing Face of Healthcare Episode*. URL: https://www.rbccm.com/en/gib/healthcare/episode/the_healthcare_data_explosion.
6. Rawson, Timothy M. et al. (June 2019). "Artificial intelligence can improve decision-making in infection management". In: *Nature Human Behaviour* 3.6. ISSN: 2397-3374. doi: 10.1038/s41562-019-0583-9. URL: <https://www.nature.com/articles/s41562-019-0583-9>.
7. Sutton, Reed T. et al. (Feb. 2020). "An overview of clinical decision support systems: benefits, risks, and strategies for success". In: *npj Digital Medicine* 3.1. ISSN: 2398-6352. doi: 10.1038/s41746-020-0221-y. URL: <https://www.nature.com/articles/s41746-020-0221-y>.
8. Goldberger, A. L. et al. (June 2000). "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals". In: *Circulation* 101.23, E215–220. ISSN: 1524-4539. doi: 10.1161/01.cir.101.23.e215.
9. Johnson, Alistair et al. (2022). *MIMIC-IV*. doi: 10.13026/s6n6-xd98. URL: <https://physionet.org/content/mimiciv/2.0/>.
10. Pollard, Tom J. et al. (Sept. 2018). "The eICU Collaborative Research Database, a freely available multi-center database for critical care research". In: *Scientific Data* 5.1, p. 180178. ISSN: 2052-4463. doi: 10.1038/sdata.2018.178. URL: <https://www.nature.com/articles/sdata2018178>.
11. Pollard, Tom Joseph et al. (2019). *The eICU Collaborative Research Database (version 2.0)*. PhysioNet. doi: 10.13026/C2WM1R. URL: <https://doi.org/10.13026/C2WM1R>.

Chapter 2

Background

Abstract

This chapter provides an overview of the existing literature for the topics covered in this thesis. First, an introduction to infectious disease and the problem of antimicrobial resistance (AMR) is presented, followed by a summary of artificial intelligence-based clinical decision support systems (AI CDSSs) as potential innovative solutions, with a review of the existing literature. Next, multi-morbidity as another key problem in healthcare, particularly for infectious diseases, is discussed. Finally, issues surrounding inequalities and biases within infectious disease and AI systems are voiced with a particular focus on the ethical questions raised by AMR.

This chapter's discussion on ethical AI for AMR was published in Nature Machine Intelligence as "Developing moral AI to support decision-making about antimicrobial use" (Bolton et al. 2022).

2.1 Chapter objectives

- *Provide an introduction to infectious disease and antimicrobial resistance*
- *Review the latest literature on AI clinical decision support systems for antibiotic optimisation*
- *Describe the challenges associated with co-morbidities*
- *Discuss the ethical implications of AI and AMR as well as how we can ensure moral decision-making*

2.2 Clinical decision support can help tackle the global threat of antimicrobial resistance

2.2.1 AMR is caused by natural selection and accelerated through human behaviour

Infectious diseases are caused by pathogenic microorganisms that can spread from one individual to another. The discovery of antimicrobials in the early 20th century that could kill or inhibit the growth of pathogenic microorganisms caused a revolution, making previously fatal diseases easily treatable. However, what was once thought to be a 'silver bullet' solution to infectious diseases has been brought into question in recent decades with significant outbreaks, including the emergence of highly resistant pathogens, such as Methicillin-resistant *Staphylococcus aureus* (MRSA, Bloom et al. 2019). In healthcare facilities, guidelines and procedures are regularly implemented for infection prevention and control, but due to the frequent

use of invasive devices such as ventilators and urinary catheters, healthcare-associated infections frequently occur. Studies have shown that around 10% of hospital patients acquire a healthcare-associated infection, with that figure increasing for intensive care unit (ICU) patients given invasive procedures and their often immunocompromised status (Haque et al. 2018).

The consumption of antimicrobials has increased substantially over the past few decades with a 65% increase in global defined daily doses between 2000 and 2015, leading to 34.8 billion antimicrobials being prescribed daily (Klein et al. 2018). On a population level, the frequency with which antimicrobials are used is directly related to the rate of development of resistance due to natural selection. When a selective pressure such as an antimicrobial is applied to a population of pathogens, those individuals with traits that provide a survival advantage (i.e., resistance) are more likely to reproduce, and hence over time, these resistant traits become more common in the population. Given the short generation time of most pathogens, resistance can emerge and flourish rapidly. Furthermore, genetic material that confers resistance can be passed between bacterial species through horizontal gene transfer. Resistance to one specific antibiotic often leads to resistance to different antibiotics that belong to the same class of drugs (WHO 2015). Unfortunately, due to limited commercial interest and investment, recent research on the development of new antibiotics has been inadequate, with 43 currently in clinical development, of which only 27 are innovative agents (WHO 2021). As such, AMR has been highlighted by the World Health Organization and the United Nations General Assembly as a global threat, akin to that posed by climate change (WHO 2015; UN 2016). When patients have a drug-resistant infection, their length of hospital stay, mortality rate, and the cost of their treatment all increase. It is estimated that AMR costs the UK's National Health Service (NHS) £180 million each year (House of Commons 2018). In 2019 and 2021 bacterial AMR directly caused an estimated 1.27 and 1.14 million deaths, respectively (Murray et al. 2022; Naghavi et al. 2024). This mortality rate is expected to grow to 1.91 million attributable deaths and 8.22 million associated deaths by 2050 (Naghavi et al. 2024), which will cost the global economy trillions (O'Neill 2016).

A major driver of AMR is the inappropriate use of antibiotics in humans, either through prolonged or unnecessary treatment (Holmes et al. 2016). The reason behind this is complex and

multifactorial. Firstly, many individuals who prescribe antibiotics do not have formal training in antimicrobials, stewardship or AMR (Rawson et al. 2016a). Practitioners focus on their own discipline to improve the health of a patient and often overuse antibiotics, particularly for preventive medicine (Charani et al. 2013; Charani et al. 2019). In addition, at the early stages of an infection, treatment decisions are often based on educated guesses because of limited information. Frequently, this results in the use of broad-spectrum antibiotics, which target a wide range of bacteria but conduce more AMR. This frequent use of broad-spectrum agents and congregation of vulnerable individuals makes secondary care hospitals particularly problematic with regards to driving the development of resistance (Cocker et al. 2024).

2.2.2 Antibiotic stewardship is key to mitigate AMR

A number of solutions have been proposed to prevent the future mortality and increased health-care costs associated with AMR. Emphasis has been on rapid and point-of-care diagnostics, which can identify the causative organism and its susceptibility profile faster than traditional culture-based microbiology diagnostic methods, which rely on growing bacteria and therefore have a minimum incubation time (Wang et al. 2022). However, the implementation of such technology has been limited beyond emergency use in outbreaks due to their significant cost (Hays et al. 2019). Other approaches involve using enhanced surveillance to contribute to the evidence base, detect the emergence of novel resistance, and inform decision-making. Crucial for this thesis and as outlined in the UK's five-year national action plan to tackle AMR (GOV.UK 2024), approaches towards optimising antimicrobial use and prolonging their useful therapeutic life are essential. This process forms part of antimicrobial stewardship which can be defined as 'A coherent set of actions which promote using antimicrobials responsibly' (Dyar et al. 2017). In particular, the Department of Health and Social Care report highlighted that stewardship programmes should be strengthened, relevant data should be appropriately recorded, and evidence-based guidance and technology-driven innovations should be promoted (GOV.UK 2024). All of these link to CDSSs and in particular those utilising AI. With robust data and evidence, they can be part of a stewardship programme to ensure that antimicrobials

are only used when appropriate. To support the development of novel technologies, sustainable economic models must be developed to incentivise investment in this area (Health et al. 2019; European Commission 2024) and awareness among clinicians must be increased through education and training to promote adoption and behavioural change (Davey et al. 2017).

2.2.3 AI clinical decision support can provide innovative solutions

CDSSs are software designed to provide information to healthcare professionals, patients, or other individuals to make informed clinical decisions. CDSSs that are currently in use are generally very simple for example providing warnings, reminding clinicians of local guidelines, and monitoring vital signs (McGregor et al. 2006a; Demonchy et al. 2014). With the advent of AI though and the ever-increasing prevalence of electronic health records (EHRs), numerous CDSSs have now been developed with models achieving accuracy on par and even above that of human experts (Esteva et al. 2019; Topol 2019). However, currently these AI CDSSs have seen limited uptake within actual clinical practice. This is due to numerous reasons including limited integration with current healthcare systems, CDSSs often being complex and difficult to use, questions about the liability of such systems, issues such as explainability and interoperability, as well as the immature nature of the evaluation and regulation of such technologies (Khairat et al. 2018; Levy-Fix et al. 2019; Magrabi et al. 2019; Sutton et al. 2020). Research has also shown that clinician opinions and the attitude of organisations toward innovation are key drivers behind the adoption of CDSSs (Liberati et al. 2017). There are successful case studies though, particularly in diagnosis using imaging modalities, such as in radiology or stroke detection. (Mello-Thoms et al. 2023; NICE 2024). AI CDSSs are therefore poised to impact clinical practice, but challenges remain.

Machine learning and deep learning techniques have been applied to EHR data to optimise antimicrobial use and tackle the problem of AMR. The majority of work focuses on the prediction of resistance (Jeon et al. 2023; Kherabi et al. 2024), the diagnoses of bacterial infections (Hernandez et al. 2017; Rawson et al. 2021), and empiric antimicrobial therapy selection (Kanjilal et al. 2020; Stracy et al. 2022). Some studies in the literature also look to predict

treatment response, patient outcomes (Zoabi et al. 2021; Cai et al. 2023), the gram stain of the infecting organism and the source of infection (Hernandez Perez 2019). Finally, a few look to optimise antimicrobial dosing (Imai et al. 2020; Huang et al. 2021) or derive novel prescribing policies (Beaudoin et al. 2016; Boominathan et al. 2020). As far as we are aware, there is only one other study (published in early 2024) that has looked at AI to support the de-escalation of antibiotics through switching or stopping treatment (Tran-The et al. 2024). Historical research has also approached these problems through the use of more traditional rule-based systems (Hulgan et al. 2004; McGregor et al. 2006b). More broadly other work has investigated the detection of prescribing errors (Lee et al. 2022) or potential side effects (Zitnik et al. 2018), but these studies did not focus specifically on antimicrobials or resistance.

An overview of some of the literature published since 2020 is shown in Table 2.1 (ordered by the type of decision support). For an overview of the research before this, please see appropriate reviews (Rawson et al. 2017a; Luz et al. 2020; Peiffer-Smadja et al. 2020; Anahtar et al. 2021; He et al. 2021). Prior research focuses primarily on secondary care, specifically the ICU and emergency department (ED), this is in part due to the greater availability of data in these settings. Most studies focus on one specific type of infection, such as urinary tract infections. Interestingly studies often include a large number of clinical variables. While this, in theory, should lead to more nuanced and potentially effective models, this high dimensionality means models require a lot of data during training, and in real-life clinical practice where time is often critical, and electronic infrastructure not interoperable, it may be difficult to accumulate such diverse information. Finally, as discussed above, the evaluation and validation of such systems is often insufficient leading to limited clinical uptake. To date we are not aware of any publications describing the prospective clinical validation of such models, outside of those focusing on resistance prediction or sepsis diagnosis (Elligsen et al. 2021; Adams et al. 2022). Although, some other commercialised technologies exist such as AntibioKos from nosotech, TREAT-Essential from TREAT systems, and LUMED from Biomerieux. Table 2.1 shows that most studies use a proprietary dataset, often from one hospital, limiting their generalisation, particularly given the dynamic temporal and geographic nature of infecting organisms and re-

sistance patterns. This lack of accessible and standardised data can also hinder comparisons between different models and methodological approaches. The type of data processing and AI utilised is usually restricted to more traditional machine learning algorithms and methods such as logistic regression and decision trees rather than more modern and powerful deep learning approaches. Figure 2.1 shows some areas of research on AI for antibiotic optimisation aligned with steps in the infection patient pathway, and illustrates the amount of research in the literature for each stage. Given the lack of studies focusing on the de-escalation of antibiotic treatment and stewardship-focused decisions, this thesis aims to advance research into AI CDSSs within these fields, overcome the shortcomings of current systems, and develop effective solutions to sustainability tackle AMR.

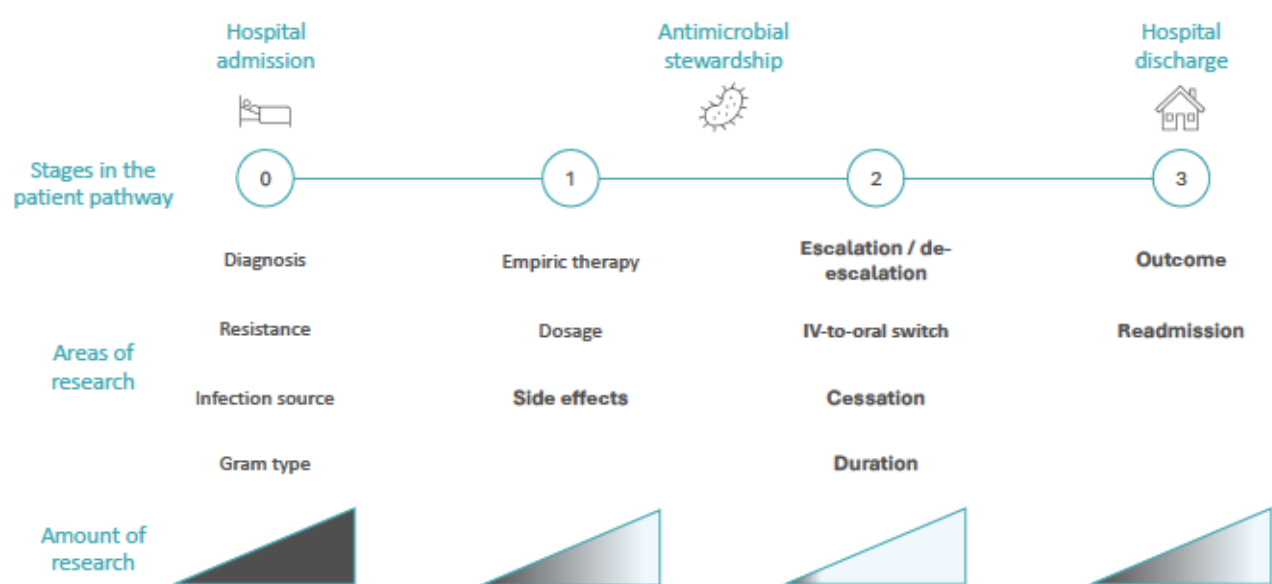


Figure 2.1: Aspects of AI for antibiotic optimisation research aligned with steps in the patient pathway. The amount of research within each stage is indicated by the level of the grey triangle. The later stages of antibiotic stewardship have previously been neglected. Those areas of research investigated during this PhD are highlighted in bold. In particular, Chapter 4 discusses cessation and Chapter 5 describes AI for IV-to-oral switching.

Table 2.1: Overview of research articles published since 2020 applying AI to antimicrobial optimisation. Those in **bold** are articles highlighted for their quality. This list is not comprehensive.

Reference	Decision support	AI	Data
Eickelberg et al. 2020	Infection risk / diagnosis prediction	Random forest	MIMIC-III (n=19,633)
Feng et al. 2023	Infection risk / diagnosis prediction	Boosted decision trees	3 datasets combined (n=293,109)
Wong et al. 2020	Infection diagnosis	Decision trees, LASSO, logistic regression	Proprietary dataset (n=715)
Rawson et al. 2021	Infection diagnosis	Gaussian naive bayes, support vector machine, neural network	Proprietary dataset (n=15,599)
Seheult et al. 2023	Urine culture positivity (diagnosis)	Decision tree	Proprietary dataset (n=38,361)
Moran et al. 2020	Resistance prediction	XGBoost	Proprietary dataset (n=15,695)
Feretzakis et al. 2021	Resistance prediction	AutoML	Proprietary dataset (n=11,496)
Corbin et al. 2022	Resistance prediction	Logistic regression, random forest, gradient-boosted decision trees	Proprietary dataset (n=8,342) and MIMIC-III (n=15,806)
Rich et al. 2022	Resistance prediction	Decision trees, boosted logistic regression, random forest	Proprietary dataset (n=9,990)
Boerman et al. 2022	Resistance prediction	Logistic regression, gradient-boosted decision trees	Proprietary dataset (n=4,885)
Tharmakulasingam et al. 2023	Resistance prediction	Transformer	AMR-UTI (n=116,902)
Kim et al. 2023	Resistance prediction	Logistic regression, extreme gradient boosting, random forest, ensemble methods	Proprietary dataset (n=10,474)
Lee et al. 2023	Resistance prediction	Gradient-boosted decision tree	Proprietary dataset (n=550)
Cavallaro et al. 2023	Resistance prediction	Gradient-boosted decision tree	Proprietary dataset (n=5,190)
Stracy et al. 2022	Post-treatment resistance prediction	Logistic regression	Proprietary dataset (n=147,714)
Kanjilal et al. 2020	Resistance prediction and empiric treatment recommendation	Logistic regression	MIMIC-III (n=15,806)
Qin et al. 2024	Combination treatment recommendation	Neural network	Proprietary dataset (n=5,173)
Imai et al. 2020	Optimal vancomycin dosing	Decision tree	Proprietary dataset (n=654)
Huang et al. 2021	Optimal vancomycin dosing	Extreme gradient boosting	Proprietary dataset (n=722)
Smith et al. 2020	Optimal meropenem and polymyxin B dosing regimens	Genetic algorithm	Proprietary dataset (n=51)
Batisti Biffignandi et al. 2024	Minimum inhibitory concentration prediction	Elastic net, random forest	3 datasets combined (n=4,367)
Pascucci et al. 2021	Antimicrobial susceptibility detection	Convolutional neural network	Proprietary dataset (n=18,000)
Bollapragada et al. 2024	Antimicrobial susceptibility detection	Convolutional neural network	Manually created dataset (n=500)
Tran-The et al. 2024	IV to oral, discontinuation, de-escalation	Extreme gradient boosting and light gradient boosting machine	Proprietary dataset (n=306,193)
Cai et al. 2023	Treatment efficacy prediction	Neural network	Proprietary dataset (n=1,043)
Zoabi et al. 2021	Outcome prediction	Light gradient boosting machine	Proprietary dataset (n=7,889)

2.3 Co-morbidities are a complex, growing problem

2.3.1 Antibiotic management in multi-morbid patients is difficult

Co-morbidities defined as chronic long-term medical conditions are a major challenge in health-care (Gijzen et al. 2001; Riaño et al. 2019). Co-morbid conditions put individuals at a high risk of developing an infection with a greater chance of poor outcomes (Huttunen et al. 2013; Charani et al. 2015). They are associated with an increased duration of treatment, length of hospital stay, chance of readmission, and risk of death. Furthermore, the prevalence of multi-morbidity (the presence of concurrent medical conditions in an individual) is continuing to increase globally given the ageing population (Medical Sciences 2018; Pearson-Stuttard et al. 2019). The National Institute of Health and Care Excellence (NICE) and the National Institute for Health Research (NIHR) in the UK have highlighted it as a highly complex priority research area (NICE 2019; NIHR 2020). Inappropriate antibiotic regimes and dosing in multi-morbid patients is a major underlying factor driving AMR, and for the healthcare sector, infections in multi-morbid populations have significant financial implications through increased bed days and greater rehabilitation following recovery (McDaid et al. 2015; Ticinesi et al. 2015; Weir et al. 2015). In addition, numerous medications (polypharmacy) often have significant interactions with antibiotics. This increased complexity means that antibiotic decisions for these patients often go beyond the scope of routine clinical guidelines. Precision prescribing in co-morbid patients is further compounded by a lack of evidence to support decision-making in these patient populations, meaning ultimately many multi-morbid patients fail to be treated appropriately, leading to inequitable care (Grace 2012; Polso et al. 2014; Boyd et al. 2016; Rawson et al. 2017b). It is therefore clear that enhanced evidence and personalised approaches to antibiotic management would help address the complexity associated with infections in multi-morbid patients (Vincent et al. 2016; Rawson et al. 2017a).

2.3.2 AI co-morbid research occurs at the disease and individual level

A range of methodologies have been employed to carry out research on co-morbidities. The Charlson co-morbidity index (Charlson et al. 1987; Deyo et al. 1992) is a widely adopted tool that quantifies the burden of chronic conditions on patients and predicts mortality risk through weighted scores. It assigns some specific co-morbidities to 17 different categories and was developed through retrospective analysis and expert consensus. A large part of the literature focuses on clustering (Violán et al. 2019; Haug et al. 2020; Beaney et al. 2024a; Beaney et al. 2024b) or finite mixture models (Zador et al. 2019) to elucidate groups of patients and understand relationships between diseases through unsupervised, pairwise, probabilistic and factorization methods (Hassaine et al. 2020; Violán et al. 2020; Majnarić et al. 2021). Subsequent analysis then aims to improve disease understanding and management (Majnarić et al. 2021). Within infectious diseases, there has been limited co-morbidity-focused research to date (Cheng et al. 2020; Papin et al. 2021), with a particular lack of understanding in how to optimise antimicrobial prescribing in specific co-morbid populations such as those who are obese (Grace 2012; Charani et al. 2015; Rawson et al. 2017b).

On an individual level, AI-based clinical decision support systems can use co-morbidity information alongside other data to make clinically useful patient-specific predictions. Often disease data is fed into such models using one hot encodings, where categorical data is represented as binary vectors (0=no, 1=yes), allowing algorithms to process such data numerically (Nielsen et al. 2019). While modern AI methods are sophisticated and can have a substantial number of parameters, they still suffer from the combinatorial explosion of many diseases and a lack of data on those with rare diseases or a high degree of multi-morbidity. Modelling time is important when considering co-morbidities since disease occurrence and risk are highly variable over the course of an individual's life. Several papers have tackled this temporal aspect through different approaches, such as including a time dimension within tensors (Zhou et al. 2014; Afshar et al. 2020). Deep learning and in particular attention mechanisms hold great promise in this area as large datasets of patients and the long-range temporal relationships between their co-morbidities can be interrogated. Using this approach Li et al. showed im-

pressive performance in predicting a patient's likelihood of disease in the future from EHR data (Li et al. 2020). Kraljevic et al. took this approach further with their pipeline called Foresight which uses a generative pre-trained transformer to model patient timelines and predict future medical events such as disorders and procedures (Kraljevic et al. 2023). Representations developed during training also allow operations such as similarity or distance to be calculated (Beam et al. 2019), which is useful for case-based reasoning. This is an alternative approach to clinical decision support where historical examples are collated to help solve current problems. It has seen great use in healthcare (Noll et al. 2022), but ensuring patients recalled are similar to the current patient in question with regards to their chronic conditions is non-trivial.

Representing an EHR or a patient as a graph has also gained traction in recent years with the advent of graph neural networks (Choi et al. 2020). Once developed graphs can be used for prediction tasks such as a patient's co-morbidity risk or likely outcome (Zhang et al. 2017a; Rocheteau et al. 2021). While the majority of AI research into multi-morbidity focuses on disease predictions, there is also an issue with polypharmacy where numerous medications taken by patients to treat multiple diseases can result in adverse drug reactions. This is a significant problem and some research has looked into modelling these side effects and determining optimal prescribing (Zhang et al. 2017b; Zitnik et al. 2018).

2.3.3 Challenges remain in co-morbidity research

Current healthcare structures and pathways are designed to treat one specific disease. These processes will need to be adapted to cope with the high prevalence of multi-morbidity among individuals if we are to reach truly personalised medicine. The literature highlights that multi-morbidity is not random and in fact creates distinct clusters given regular disease co-occurrences, risk factors, and causes (Onder et al. 2020; Whitty et al. 2020). Co-morbidity information is therefore vital for AI-CDSSs but challenges exist using co-morbidity data in AI systems. These include sparsity, missingness, combinatorial complexity due to the large number of unique diseases, a lack of data particularly for those with rare diseases or com-

plex co-morbidity combinations, and heterogeneity in how chronic conditions are recorded. More research is therefore needed to fully comprehend the complexities of concurrent medical conditions and develop effective CDSSs (Fraccaro et al. 2015).

2.4 Moral AI for antimicrobial resistance

2.4.1 Inequalities exist within infectious diseases

Research has shown a strong association between poor socioeconomic status, increased risk of infection and AMR (ECDC 2013; Williams et al. 2016; Oestergaard et al. 2017; Alividza et al. 2018; Donnelly et al. 2018; Colzani 2019). People of different ethnic backgrounds have different infection-related risks and outcomes. Inequalities within society lead to socioeconomically deprived populations having a greater number of hospitalisations and worse outcomes (Fiscella et al. 2004). Socioeconomic status affects the healthcare an individual can access and receive, as well as their health outcomes (Caballo et al. 2021). Socioeconomic status is affected by social constructs such as race, ethnicity, indigeneity, gender, as well as education. Significant health inequalities between different population groups drive the poor health outcomes that are reported in people from ethnic minority backgrounds (GOV.UK 2020; Prats-Urbe et al. 2020). Higher rates of co-morbidities are observed in those from ethnic minority groups (Daw 2017; Whitty et al. 2020), further causing these populations to be disproportionately affected. For example, a UK report has highlighted that individuals from ethnic minority backgrounds are more likely to have diabetes at a younger age (Harrison et al. 2020). The impact of ethnicity and co-morbidities was emphasised during the COVID-19 pandemic, where patients from ethnic minority groups or individuals with co-morbidities saw significantly worse treatment outcomes than those from a white background or with no underlying health conditions (Pan et al. 2020). In particular black and south Asian populations, obesity, cardiac disease, chronic obstructive pulmonary disease and severe asthma were all associated with increased hospitalisation or mortality (Academy 2021; Docherty et al. 2020; Iaccarino et al. 2020; Niedzwiedz et al. 2020; Patel et al. 2020; Prats-Urbe et al. 2020; Raharja et al. 2020; Williamson et al. 2020).

2.4.2 Fairness and biases are a problem for AI in healthcare

Fairness aims to ensure that AI systems operate impartially and equitably, avoiding biases that can lead to discrimination or unequal treatment of individuals. Biases often exist in AI systems, as there is a lack of data for minority groups, datasets are ingrained with societal biases, and models learn spurious correlations (Goetz et al. 2024). For example, during the COVID-19 pandemic, a large number of AI models were developed to predict the diagnosis and prognosis of COVID-19 patients (Assaf et al. 2020; Dashti et al. 2020; Wang et al. 2020; Yadaw et al. 2020). However, these models suffered from bias and lack of generalisation due to the limited nature of training datasets (Wynants et al. 2020). Generalisation is a key challenge for AI systems, particularly those operating in healthcare (Goetz et al. 2024), and for clinicians to have trust in an AI system, they must understand on which patients the technology can be used, similar to how drugs have a specific indication of use. The selective deployment of AI models has been proposed as a solution to the generalisation problem (Vandersluis et al. 2024). But allowing AI models to only be used for particular patient groups raises further ethical questions on fairness and may exacerbate existing inequalities.

To develop fair AI systems, diverse and representative datasets are needed that contain information on wider social and cultural determinants of health, given that often these overlooked variables have a significant impact on patients and their outcomes (Chen et al. 2020). The traditional saying of 'a model is only as good as the data used to train it' holds firmly true in healthcare, but by carefully considering the data used we can prevent harm and inequality for minority and disadvantaged groups. Several methods have been developed to promote the fairness of AI systems. These techniques can be applied at pre-, in-, and post-processing stages depending on the particular problem being tackled. Methods include reweighing, resampling, fairness constraints, alternative thresholds, and fair adversarial representation learning (Kamiran et al. 2012; Hardt et al. 2016; Beutel et al. 2017; Zafar et al. 2017a; Zafar et al. 2017b; Madras et al. 2018). Complementary fairness metrics then enable the potential biases within a model to be tested; for example, demographic parity, equality of opportunity, equalised odds, and predictive parity (Gajane et al. 2018; Verma et al. 2018). Research focusing on infections

in minority populations or those from deprived backgrounds is of paramount importance. AI solutions must not be biased and should consider important multifactorial variables of disease, including sex, race, occupation and living situation (WHO 2018; Khalatbari-Soltani et al. 2020; Leslie et al. 2021). By obtaining a greater understanding of how different populations experience and receive healthcare for infections, more inclusive and effective decision-making is possible, which should facilitate better patient outcomes and a fair approach to addressing AMR. Responsible practices and ensuring fairness are therefore essential when working on AI in healthcare to ensure that the potential benefits of the technology are balanced against potential biases and ethical values.

2.4.3 Antimicrobial resistance raises significant ethical questions

Antimicrobials are a staple of modern medicine, and hence the development of AMR raises significant ethical and societal implications (Littmann et al. 2015; Parsonage et al. 2017). Most notably, how can a moral balance be obtained between the needs of an individual patient and those of wider and future society? On the one hand, restricting the use of antimicrobials may enable their continued effectiveness across society, while their current use can grant a patient an immediate benefit. When making decisions about whether or not to prescribe antimicrobials to a patient, clinicians should evaluate the potential risks and rewards (Butler et al. 2001; Brink et al. 2020). Treating the patient may reduce suffering and disease progression, but risks driving the evolution of AMR both at the individual and population level, as well as possibly causing side effects, while not prescribing antimicrobials has the opposite risk and reward profile. National and local guidelines are important in decision-making by helping clinicians follow evidence-based practice and understand when certain drugs should be administered (NICE 2015). However, this decision-making process is highly individualised and unfortunately frequently undertaken with limited information, meaning it is unclear which antimicrobial the infectious pathogen is susceptible to, or even if the patient could recover without treatment. The evolution of resistance is also random and hence prescribing an antimicrobial may help the patient recover with no negative consequences. This results in complex decisions that are

not clear-cut, the consequences of which may or may not help the patient and harm future populations. As such, the morally right decision is often unclear, with those who do not have expertise in infectious disease or resistance often just focusing on the individual they are treating (Rawson et al. 2016b).

2.4.4 Maximising total-average utility in healthcare

The question of how to build moral AI decision-making systems is very important and many approaches can be found in the literature (Badea 2021). Building a decision system based on an underlying moral paradigm, called the top-down approach in AI, is one of the most common ways of doing this (Badea 2021). The main moral paradigms employed in AI ethics are utilitarianism (a type of consequentialism), deontology and virtue ethics.

Utilitarian ethicists believe that an action is only 'good' if it creates utility (frequently measured as happiness) (Sinnott-Armstrong 2003). This can be linked to healthcare as by maximising health, doctors will in turn increase happiness through improving quality of life and extending survival. Multiple evaluation techniques including total, average, and minimum utility exist. Total utility evaluates an action by adding up the utility brought to all people, average utility considers the average utility across a population, and minimum utility acknowledges those with the lowest utility value. When considering health, minimum utility is almost impossible to define given the immeasurable value of life. Total utilitarianism suffers from the 'addition paradox' whereby just increasing the population size would maximise utility even if each individual has a very low utility value. Average utility gets around this but still suffers from problems such as utility monsters, who gain significantly greater sums of utility from actions (Sinnott-Armstrong 2003). Sidgwick proposed a hybrid total-average view which aims to maximise the number of people who are currently alive and their average happiness (Sinnott-Armstrong 2003). This aligns with the UK's General Medical Council (GMC) duties of a doctor where the objective is to maximise that patient's health and care, while also considering wider society through the duty to 'Protect and promote the health of patients and the public' (GMC 2019). In my opinion, maximising total-average utility is of greatest importance in the context

of AMR and healthcare, given the aligned aims of extending life and providing equality, and so I will assume it is the most relevant measure of happiness in this context.

2.4.5 Bentham's felicific calculus puts society first

Bentham's felicific calculus provides a framework upon which the utility of an action can be measured (Bentham 1789; Post et al. 2022). 7 variables are included within the calculation:

1. Intensity – how strong is the pleasure?
2. Duration – how long will the pleasure last?
3. Certainty or uncertainty – how likely or unlikely is it that the pleasure will occur?
4. Propinquity – how soon will the pleasure occur?
5. Fecundity – the likelihood of further sensations of the same kind
6. Purity – the likelihood of not being followed by opposite sensations
7. Extent – how many people will be affected?

Table 2.2: Overview of Bentham's felicific calculus variables and example application to starting antimicrobial treatment.

Variables	Description	Exemplar of starting antimicrobial treatment	Corresponding ad-hoc utility value
Intensity	How strong is the pleasure?	Treating a relevant infection with antimicrobials has the potential to save that person's life	Highly positive utility
Duration	How long will the pleasure last?	Any extension of life is immeasurable while it is reasonable AMR will continue in the near-term future	Positive utility
Certainty or uncertainty	How likely or unlikely is it that the pleasure will occur?	Limited information often means treatment may or may not be helpful and there is always an inherent risk of developing AMR	Neutral utility, without more information
Propinquity	How soon will the pleasure occur?	Treatment can be effective immediately however the same is true for the evolution of AMR	Neutral utility, without more information
Fecundity	The likelihood of further sensations of the same kind	-	Unable to assign
Purity	The likelihood of not being followed by opposite sensations	-	Unable to assign
Extent	How many people will be affected?	Prescribing antimicrobials affects the patient and those close to them, while the development of AMR is a certainty and may affect everyone, causing significant suffering and mortality	Immense negative utility

Table 2.2 provides an illustration of this calculus and its application to the decision of starting antimicrobial treatment. Using this algorithm, we can attempt to understand whether prescribing antimicrobials is morally justified. Assigning a numerical value to the results of the felicific calculus is notoriously difficult (Rzepka et al. 2013), but we can try to assign the criteria in a more ad-hoc manner to hypothesise what the framework would prescribe. The variables intensity, duration, certainty, propinquity and extent are of greatest interest (Bentham 1789). Treating a relevant infection with antimicrobials has the potential to save that person's life, and as such the intensity and duration of utility can be highly positive. As noted above limited information often means treatment may or may not be helpful and there is always an inherent risk of developing AMR, therefore it is difficult to assign a certainty utility value.

Treatment can be effective immediately; however, the same is true for the evolution of AMR, with existing evidence that it is already having a significant impact on patients (Dadgostar 2019). As such we are again unable to assign a utility value to propinquity. The development of AMR may affect everyone, causing significant death and suffering (O'Neill 2016), hence an immense negative utility value must be assigned to extent. When considered together, on the one hand, prescribing antimicrobials has a very high potential intensity and duration if the patient is cured, and zero if they are not cured. But by using antimicrobials we are risking the evolution of AMR, which would have an incredibly high extent in its negative utility due to everyone who will suffer from its consequences in the future, and the potential for high negative intensity and duration if patients die or have reduced quality of life because of AMR. Thus, this framework suggests that to justify prescribing antimicrobials, the intensity and duration utility gained from one individual must outweigh the negative impact on everyone else. The only scenario in which this could exist is with a utility monster, which seems improbable in our reality and to maximise total-average happiness we must look to maintain all life. Therefore, because of the extent component, overall, the felicific calculus indicates the moral action is to reduce antimicrobial prescribing and tackle AMR to maximise society's utility.

By reducing antimicrobial prescribing and consequently AMR, a great positive utility benefit is gained in the intensity, duration, and extent variables for those who would otherwise be affected. Importantly, though, certainty and propinquity cannot be defined without more information. As such, a probabilistic analysis should be performed on the chance of AMR increasing for any given prescription versus the chance that the patient recovers or dies. AI models can support this and therefore contribute towards utility evaluations through estimating the impact of the selected agent on the development of AMR versus the likelihood of clinical efficacy. Indeed, expert opinion and guidelines concur with this and note antimicrobials should only be used when there is likely to be a clear benefit so that we can sustainably manage AMR moving forward (Holmes et al. 2016; Singer 2017; PHE 2020; NICE 2021). In conclusion, Bentham's framework suggests that decisions should put greater emphasis on wider society as opposed to the individual, in order to reduce the downstream suffering of others, but that more contextual information is also required. By combining frameworks such as Bentham's felicific

calculus with AI-based CDSSs, we can ensure that the utility of an antimicrobial prescription can be quantified to understand potential individual and societal implications.

2.4.6 Alternative ethical frameworks arrive at different conclusions

In deontology or duty-based ethics, an action is morally good or bad based on if it acts in accordance with our duty (Alexander et al. 2020). Duty-based ethics is very amenable to rule-based decision-making, as such one could attempt to figure out what the perfect and imperfect duties are in this situation, based on, for instance, Kant's categorical imperative. The UK's GMC 'duties of a doctor' or the Hippocratic oath can be considered a more deontology-focused approach given care of the patient is of primary concern (GMC 2019). However, these principles were designed to be generalist and are not tailored to the significant ethical dilemma posed by antimicrobial prescribing.

Virtue ethics on the other hand focuses on the moral character of the person carrying out actions and assesses them based on comparison with a virtuous person who possesses and lives the virtues (Hursthouse et al. 2016). In this context virtues would need to be defined and the question of 'What would a virtuous person do?' considered. One could argue that virtuous clinicians can act as moral exemplars for complex decision-making (Hindocha et al. 2021), but common moral dilemmas such as those seen in the trolley problem are likely to arise in the context of AMR. For example, weighing up the potential number of lives lost and taking action to reduce the number of deaths versus having personal responsibility for an individual's outcome, and whether acting under a high-pressure situation is the same as not acting.

Applied ethical theories can also be explored, for example, the four principles of medical ethics (autonomy, beneficence, non-maleficence and justice) which are commonly used as a platform upon which moral agents in healthcare should act (Beauchamp et al. 1979). When applying these principles at the societal level, one can consider that they should translate across time (Meyer 2020). In this case 'justice' defined as the obligation of fairness in the distribution of benefits and risk, means that there is a responsibility to provide equal and fair

care to everyone no matter if they are alive yet or not. Furthermore, since modern medicine only began in the 19th century with breakthrough discoveries from Louis Pasteur and Florence Nightingale (Sakai et al. 2022), one can argue that the vast majority of individuals who will need care will not yet have lived or be alive. Therefore, for antimicrobials which can be considered a finite and limited resource given the development of AMR, should we not limit their use so that their associated risks and benefits are fairly distributed? This notion is in agreement with the utilitarian findings previously discussed of putting society first and fairly distributing resources and happiness. However, it should be noted that there is significant discussion in the literature on whether or not the four principles of medical ethics are appropriate (Gillon 1994; Holm 1995; Lawrence 2007; Huxtable 2013). By comparing different ethical theories, we can insinuate they may have contrasting viewpoints on what is considered morally right with regards to prescribing antimicrobials.

2.4.7 Antimicrobial resistance is a negative externality

In economics, a negative externality occurs when producing or consuming a good causes a negative impact on a third party not directly involved in the transaction. This concept frequently occurs in healthcare: for example, the private monetary costs to the individual suffering from obesity are less than the costs of obesity-related illness paid by society of which many individuals are non-obese (Cawley 2015). The literature notes that the development of AMR will have a significant financial impact on the global economy (O'Neill 2016). Cheap commonly used antimicrobials which we rely on may become obsolete, simple medical procedures could become dangerous, and patients would ultimately require longer stays in hospital (Parsonage et al. 2017). AMR is therefore a negative externality, whereby a prescription of antimicrobials between a doctor and their patient may negatively impact everyone else in society. Furthermore, the low cost of the antimicrobials currently used does not account for the future financial implications of AMR. To overcome this externality, we must reduce antimicrobial usage to curb the development of AMR. This aligns with the economics of healthcare resource allocation to distribute finances in a fair and objective way that reflects local healthcare needs and reduces

inequalities (NHS n.d.; Barr et al. 2014).

AMR has also been cited as an example of the 'tragedy of the commons', whereby the development of resistant organisms through antimicrobial consumption reduces the drug's efficacy, and hence the long-term value of the antimicrobial commons shared across society (Hollis et al. 2015). This theory states that if individuals act according to their own self-interest, then the future value of the resource is reduced. To overcome this, a number of solutions can be employed, including governmental legalisation that penalises overuse, taxes unnecessary use, and financially rewards innovation, as well as local guidelines and stewardship programmes (Tarrant et al. 2019). The problem with applying AMR to the 'tragedy of the commons' model is that while there is a clear threat of resource depletion and few innovative alternative drugs, it is extremely difficult to define boundaries with regards to responsible antimicrobial use. We must try to find solutions though, as if we do nothing, there will be considerable future costs, and the absence of effective antimicrobials will result in greater health inequalities. As such, it is the morally right thing to do to address the increasing negative externality and the 'tragedy of the commons' of AMR.

2.4.8 Our moral obligation towards the health of future humans

The main counterargument to the above reasoning is that future generations should not be considered indefinitely, given that it is not a certainty that they will ever exist. As such, antimicrobials should be prescribed when there is the potential for a patient to benefit to maximise current utility. Furthermore, another viewpoint is that given the rate of technological change over the last century, and the development of AI, future generations will themselves be able to innovate and overcome the problem of AMR. It is true to some extent that we cannot realistically consider human populations far in the distant future, given the uncertainty surrounding their existence and potential for technological advancement. However, we know that humans in the near future will face similar problems and have a fundamental interest in health and survival. For instance, based on our current understanding, the development of AMR is a certainty in the near future (Littmann et al. 2015). Through Darwinian evolution, microorganisms

have shown endless versatility in adapting to antimicrobials. With no breakthrough solutions on the horizon, it seems AMR is a case of not if but when. Even if we treat all infections appropriately, AMR will continue to develop and more people will eventually be at risk of worse outcomes. If we cannot solve this emerging crisis now, then we must do everything in our power to reduce the rate of development of AMR, so that it is feasible for future generations to similarly consider the health of their descendants. Our moral obligation to future generations is an ethical topic itself that is the subject of much debate (Meyer 2020). A lot of the literature surrounding this topic is focused on climate change and environmental damage, with the idea that we have a responsibility to protect our livable environment (United Nations University n.d.; Godofsky et al. 2010). I strongly agree with this and believe if an argument can be made for physical requirements such as oxygen and water, then this idea can be extended to health. Indeed, it is recent advancements in technology and the availability of data that has allowed us humans to understand the impact of our decisions on society, in a way that was not previously quantifiable. Thus, because of this awareness, we must act and take the survival of our species into our own hands.

2.4.9 Developing intelligent moral decision systems is complex

Developing moral AI-driven CDSSs that incorporate ethical frameworks may support the wider adoption of such systems. However, further technological and infection-specific clinical factors need to be considered (Lysaght et al. 2019; Grote et al. 2020). Technical issues with AI such as accountability (Price et al. 2019), transparency (explainability and interpretability) (Barredo Arrieta et al. 2020; Hindocha et al. 2021) and bias (Wynants et al. 2020) have been extensively discussed in the literature, but what is morally acceptable from the perspective of an AI-assisted antimicrobial prescribing decision is yet to be fully defined. Model fairness is particularly important within infectious diseases where people of different ethnic backgrounds have different infection-related risks and outcomes, as discussed in Section 2.4.1. Another critical consideration is how information should be combined to reach a morally good decision. Data required for decision-making are unlikely to be processed through one model. For exam-

ple, one system may output the anticipated antimicrobial risk and reward profile to the patient while another produces a prediction for the likelihood of AMR development. As such, some sort of aggregation model (Badea 2021) or an experienced clinician will be required, which becomes particularly complex when patients' preferences are taken into account, as is the case with shared decision-making (Butler et al. 2001). To make a morally good choice, machine ethics must also be considered (Anderson et al. 2007). This topic covers a large area; however, if we are to assume some sort of moral framework or rules will be used (at least initially when training), to define what is ethical or not, then the question of which ethical framework is most appropriate arises (Badea 2021). Following a utilitarian perspective would lean towards reducing the prescribing of known and suspected drivers of AMR for the benefit of society. Meanwhile, deontology in alignment with the UK GMC, may say that it is our duty to 'make the care of your patient your first concern' (GMC 2019). This leads on to who is accountable for AI-based decisions, what happens if a clinician ignores or disagrees with these, or if recommendations lead to unexpected negative consequences. Clinicians are currently responsible for their actions and in case of dispute must provide rationale and justification. All of these issues require further analysis and consideration as we proceed to develop intelligent and moral decision systems.

Within infectious diseases, AI CDSS are increasingly being developed to optimise antimicrobial use. However, what is considered an appropriate antimicrobial prescription is difficult to define. Expert opinion often deviates depending on the specific scenario, while relying on a single methodology, such as an antibiogram, can be unreliable or not correlate with the expected response. This uncertainty hinders antimicrobial optimisation and the development of AI-based CDSSs but highlights why further research and a holistic, well-considered decision is required. In addition, the evolutionary process of pathogenic microorganisms and thus the development of AMR occurs on a human life timescale. Hence, when tackling AMR, AI should be temporally dynamic so that it is sensitive to microbiological evolution. Furthermore, systems should be geographically revised given infectious diseases, resistance rates, and antimicrobial availability vary dramatically by region (Holmes et al. 2016). Moreover, heterogeneity is needed with regards to antimicrobial prescribing, because uniform treatment drives

AMR (Holmes et al. 2016). These factors increase the importance of algorithmic transparency and accessibility of live local medical data. Creating moral AI to support optimal antimicrobial prescribing is therefore a complex but crucial endeavour to try to counteract AMR.

Moral frameworks have been designed to help humans make ethical decisions. As our species advances into a new age with AI we must consider how to ensure such 'intelligent' decision systems promote moral decision-making. I suggest that a utilitarian approach is most appropriate for antimicrobial decision-making given the number of individuals potentially affected, alignment with current best practice, Bentham's felicific calculus analysis, and the principles of medical ethics indicating that antimicrobial resources should be fairly distributed. Using the felicific calculus, we have shown that from a utilitarian perspective, we must reduce the downstream suffering of others to maximise total-average utility across society. In addition, considering the temporal nature of medical ethics, we can reach a similar finding that the morally right decision is to fairly distribute resources and address the future negative externality of AMR. It should be pointed out that I am not advocating that humans completely cease to use antimicrobials today and instead save them for future generations. That would imply that they have a greater value than patients today, which is certainly not the case if we are aiming to achieve moral equality. Instead, this highlights that from a moral perspective, we must fairly distribute the finite resource of antimicrobials, only use them when there is a high likelihood of benefit, and try to tackle AMR to fulfil our responsibility towards the health of future generations. AI CDSSs can potentially help realise this vision by providing prescribers with a high degree of confidence in their optimised prescribing so they fulfil their 'duties of a doctor' while driving a behaviour change so that only those patients who will benefit receive antibiotics to curb the evolution of AMR and protect wider society.

2.5 Conclusion

Antimicrobial stewardship is a key part of tackling the growing global threat of AMR. AI CDSSs hold potential to support optimal antimicrobial use, but further research is required to realise true benefit, with complexities such as co-morbidities posing unique challenges. Research and development of AI-based CDSS should focus on areas where it can have the greatest impact on supporting optimised and moral antimicrobial prescribing. Within utilitarianism Table 2.2 indicates that the variables of interest are certainty and propinquity. From an antimicrobial perspective, this means AI providing insight into the likelihood of drug effectiveness and resistance developing to support moral decisions regarding antimicrobial optimisation. For AMR spatial and temporal considerations are also critical and all aspects of the complex decision-making path from initial consideration to cessation should be supported. Finally, technological aspects of AI such as accountability, interpretability, and biases, require further exploration, and what is deemed appropriate in infectious disease must be determined based on appropriate ethical and scientific factors. Decision-making in antimicrobial prescribing is frequent and both morally and technically complex. By applying ethical theories to specific scenarios and taking a proactive approach towards inequalities we can ensure AI-based CDSS tackle global problems, such as the emerging AMR crisis, in a fair, moral way.

References

1. Bolton, William J. et al. (Nov. 2022). "Developing moral AI to support decision-making about antimicrobial use". In: *Nature Machine Intelligence* 4.11, pp. 912–915. ISSN: 2522-5839. doi: 10.1038/s42256-022-00558-5. URL: <https://www.nature.com/articles/s42256-022-00558-5>.
2. Bloom, David E. and Cadarette, Daniel (Mar. 2019). "Infectious Disease Threats in the Twenty First Century: Strengthening the Global Response". In: *Frontiers in Immunology* 10, p. 549. ISSN: 1664-3224. doi: 10.3389/fimmu.2019.00549. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6447676/>.
3. Haque, Mainul et al. (Nov. 2018). "Health care-associated infections – an overview". In: *Infection and Drug Resistance* 11, pp. 2321–2333. ISSN: 1178-6973. doi: 10.2147/IDR.S177247. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6245375/>.
4. Klein, Eili Y. et al. (Apr. 2018). "Global increase and geographic convergence in antibiotic consumption between 2000 and 2015". In: *Proceedings of the National Academy of Sciences* 115.15, E3463–E3470.
5. WHO (May 2015). *Global action plan on antimicrobial resistance*. Tech. rep. World Health Organization. URL: <https://ahpsr.who.int/publications/i/item/global-action-plan-on-antimicrobial-resistance>.
6. WHO (Apr. 2021). *Antibacterial agents in clinical and preclinical development: an overview and analysis*. Tech. rep. World Health Organization. URL: <https://www.who.int/publications-detail-redirect/9789240021303>.
7. UN (Sept. 2016). *Political Declaration of the High-Level Meeting of the General Assembly on Antimicrobial Resistance*. Tech. rep. United Nations. URL: <http://digitallibrary.un.org/record/842813>.
8. House of Commons (Sept. 2018). *Health and Social Care Committee – Oral evidence – Antimicrobial resistance*. URL: <https://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/health-and-social-care-committee/antimicrobial-resistance/oral/88745.html>.
9. Murray, Christopher JL et al. (Feb. 2022). "Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis". In: *The Lancet* 399.10325, pp. 629–655. ISSN: 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(21)02724-0. URL: [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(21\)02724-0/fulltext](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)02724-0/fulltext).
10. Naghavi, Mohsen et al. (Sept. 2024). "Global burden of bacterial antimicrobial resistance 1990–2021: a systematic analysis with forecasts to 2050". In: *The Lancet* 404.10459, pp. 1199–1226. ISSN: 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(24)01867-1. URL: [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(24\)01867-1/fulltext](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(24)01867-1/fulltext).

11. O'Neill, Jim (May 2016). *Tackling Drug-Resistant Infections Globally: final report and recommendations*. Tech. rep. The Review on Antimicrobial Resistance. URL: https://amr-review.org/sites/default/files/160525%5C_Final%5C%20paper%5C_with%5C%20cover.pdf.
12. Holmes, Alison H et al. (Jan. 2016). "Understanding the mechanisms and drivers of antimicrobial resistance". In: *The Lancet* 387.10014, pp. 176–187. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(15)00473-0. URL: <http://www.sciencedirect.com/science/article/pii/S0140673615004730>.
13. Rawson, Timothy M. et al. (Nov. 2016a). "Exploring the coverage of antimicrobial stewardship across UK clinical postgraduate training curricula". In: *Journal of Antimicrobial Chemotherapy* 71.11, pp. 3284–3292. ISSN: 0305-7453. DOI: 10.1093/jac/dkw280. URL: <https://doi.org/10.1093/jac/dkw280>.
14. Charani, E. et al. (July 2013). "Understanding the Determinants of Antimicrobial Prescribing Within Hospitals: The Role of "Prescribing Etiquette"". In: *Clinical Infectious Diseases* 57.2, pp. 188–196. ISSN: 1058-4838. DOI: 10.1093/cid/cit212. URL: <https://doi.org/10.1093/cid/cit212>.
15. Charani, E. et al. (Sept. 2019). "Antibiotic prescribing in general medical and surgical specialties: a prospective cohort study". In: *Antimicrobial Resistance & Infection Control* 8.1, p. 151. ISSN: 2047-2994. DOI: 10.1186/s13756-019-0603-6. URL: <https://doi.org/10.1186/s13756-019-0603-6>.
16. Cocker, Derek et al. (Oct. 2024). "Healthcare as a driver, reservoir and amplifier of antimicrobial resistance: opportunities for interventions". In: *Nature Reviews Microbiology* 22.10, pp. 636–649. ISSN: 1740-1534. DOI: 10.1038/s41579-024-01076-4. URL: <https://www.nature.com/articles/s41579-024-01076-4>.
17. Wang, Hui, Zhang, Wenhong, and Tang, Yi-Wei (Dec. 2022). "Clinical microbiology in detection and identification of emerging microbial pathogens: past, present and future". In: *Emerging Microbes & Infections* 11.1, pp. 2579–2589. ISSN: null. DOI: 10.1080/22221751.2022.2125345. URL: <https://doi.org/10.1080/22221751.2022.2125345>.
18. Hays, John P. et al. (2019). "The successful uptake and sustainability of rapid infectious disease and antimicrobial resistance point-of-care testing requires a complex 'mix-and-match' implementation package". In: *European Journal of Clinical Microbiology & Infectious Diseases* 38.6, pp. 1015–1022. ISSN: 0934-9723. DOI: 10.1007/s10096-019-03492-4. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6520316/>.
19. GOV.UK (May 2024). *Confronting antimicrobial resistance 2024 to 2029*. URL: <https://www.gov.uk/government/publications/uk-5-year-action-plan-for-antimicrobial-resistance-2024-to-2029/confronting-antimicrobial-resistance-2024-to-2029>.
20. Dyar, O. J. et al. (Nov. 2017). "What is antimicrobial stewardship?" In: *Clinical Microbiology and Infection* 23.11, pp. 793–798. ISSN: 1198-743X. DOI: 10.1016/j.cmi.2017.08.026. URL: [https://www.clinicalmicrobiologyandinfection.com/article/S1198-743X\(17\)30489-5/fulltext](https://www.clinicalmicrobiologyandinfection.com/article/S1198-743X(17)30489-5/fulltext).

21. Health, Department of and Care, Social (2019). *UK 5-year action plan for antimicrobial resistance 2019 to 2024*. URL: <https://www.gov.uk/government/publications/uk-5-year-action-plan-for-antimicrobial-resistance-2019-to-2024>.
22. European Commission (May 2024). *EU Action on Antimicrobial Resistance – European Commission*. URL: https://health.ec.europa.eu/antimicrobial-resistance/eu-action-antimicrobial-resistance_en.
23. Davey, Peter et al. (Feb. 2017). "Interventions to improve antibiotic prescribing practices for hospital inpatients". In: *The Cochrane Database of Systematic Reviews* 2, p. CD003543. ISSN: 1469-493X. DOI: 10.1002/14651858.CD003543.pub4.
24. McGregor, Jessina C. et al. (2006a). "Impact of a Computerized Clinical Decision Support System on Reducing Inappropriate Antimicrobial Use: A Randomized Controlled Trial". In: *Journal of the American Medical Informatics Association : JAMIA* 13.4, pp. 378–384. ISSN: 1067-5027. DOI: 10.1197/jamia.M2049. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1513678/>.
25. Demonchy, Elisa et al. (Oct. 2014). "Impact of a computerized decision support system on compliance with guidelines on antibiotics prescribed for urinary tract infections in emergency departments: a multicentre prospective before-and-after controlled interventional study". In: *The Journal of Antimicrobial Chemotherapy* 69.10, pp. 2857–2863. ISSN: 1460-2091. DOI: 10.1093/jac/dku191.
26. Esteva, Andre et al. (Jan. 2019). "A guide to deep learning in healthcare". In: *Nature Medicine* 25.1, pp. 24–29. ISSN: 1546-170X. DOI: 10.1038/s41591-018-0316-z. URL: <https://www.nature.com/articles/s41591-018-0316-z>.
27. Topol, Eric J. (Jan. 2019). "High-performance medicine: the convergence of human and artificial intelligence". In: *Nature Medicine* 25.1, pp. 44–56. ISSN: 1546-170X. DOI: 10.1038/s41591-018-0300-7. URL: <https://www.nature.com/articles/s41591-018-0300-7>.
28. Khairat, Saif et al. (Apr. 2018). "Reasons For Physicians Not Adopting Clinical Decision Support Systems: Critical Analysis". In: *JMIR Medical Informatics* 6.2, e24. ISSN: 2291-9694. DOI: 10.2196/medinform.8912. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5932331/>.
29. Levy-Fix, Gal, Kuperman, Gilad J., and Elhadad, Noémie (June 2019). "Machine Learning and Visualization in Clinical Decision Support: Current State and Future Directions". In: *arXiv:1906.02664 [cs, stat]*. URL: <http://arxiv.org/abs/1906.02664>.
30. Magrabi, Farah et al. (Aug. 2019). "Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications". In: *Yearbook of Medical Informatics* 28.1, pp. 128–134. ISSN: 0943-4747, 2364-0502. DOI: 10.1055/s-0039-1677903. URL: <http://www.thieme-connect.de/DOI/DOI?10.1055/s-0039-1677903>.

31. Sutton, Reed T. et al. (Feb. 2020). "An overview of clinical decision support systems: benefits, risks, and strategies for success". In: *npj Digital Medicine* 3.1. ISSN: 2398-6352. doi: 10.1038/s41746-020-0221-y. URL: <https://www.nature.com/articles/s41746-020-0221-y>.
32. Liberati, Elisa G. et al. (Sept. 2017). "What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation". In: *Implementation science: IS* 12.1, p. 113. ISSN: 1748-5908. doi: 10.1186/s13012-017-0644-2.
33. Mello-Thoms, Claudia and Mello, Carlos A B (Oct. 2023). "Clinical applications of artificial intelligence in radiology". In: *British Journal of Radiology* 96.1150, p. 20221031. ISSN: 0007-1285. doi: 10.1259/bjr.20221031. URL: <https://doi.org/10.1259/bjr.20221031>.
34. NICE (Jan. 2024). *Overview | Artificial intelligence (AI)-derived software to help clinical decision making in stroke | Guidance*. URL: <https://www.nice.org.uk/guidance/dg57>.
35. Jeon, D. et al. (Apr. 2023). "Clinical prediction tools for identifying antimicrobial-resistant organism (ARO) carriage on hospital admissions: a systematic review". In: *Journal of Hospital Infection* 134, pp. 11–26. ISSN: 0195-6701. doi: 10.1016/j.jhin.2023.01.003. URL: <https://www.sciencedirect.com/science/article/pii/S0195670123000129>.
36. Kherabi, Yousra et al. (Apr. 2024). "Machine learning to predict antimicrobial resistance: future applications in clinical practice?" In: *Infectious Diseases Now* 54.3, p. 104864. ISSN: 2666-9919. doi: 10.1016/j.idnow.2024.104864. URL: <https://www.sciencedirect.com/science/article/pii/S2666991924000198>.
37. Hernandez, Bernard et al. (Dec. 2017). "Supervised learning for infection risk inference using pathology data". In: *BMC Medical Informatics and Decision Making* 17.1, p. 168. ISSN: 1472-6947. doi: 10.1186/s12911-017-0550-1. URL: <https://doi.org/10.1186/s12911-017-0550-1>.
38. Rawson, Timothy M et al. (Mar. 2021). "Supervised machine learning to support the diagnosis of bacterial infection in the context of COVID-19". In: *JAC-Antimicrobial Resistance* 3.1, dlab002. ISSN: 2632-1823. doi: 10.1093/jacamr/dlab002. URL: <https://doi.org/10.1093/jacamr/dlab002>.
39. Kanjilal, Sanjat et al. (Nov. 2020). "A decision algorithm to promote outpatient antimicrobial stewardship for uncomplicated urinary tract infection". In: *Science Translational Medicine* 12.568, eaay5067. doi: 10.1126/scitranslmed.aay5067. URL: <https://www.science.org/doi/10.1126/scitranslmed.aay5067>.
40. Stracy, Mathew et al. (Feb. 2022). "Minimizing treatment-induced emergence of antibiotic resistance in bacterial infections". In: *Science* 375.6583, pp. 889–894. doi: 10.1126/science.abg9868. URL: <https://www.science.org/doi/10.1126/science.abg9868>.
41. Zoabi, Yazeed et al. (Oct. 2021). "Predicting bloodstream infection outcome using machine learning". In: *Scientific Reports* 11.1, p. 20101. ISSN: 2045-2322. doi: 10.1038/s41598-021-99105-2. URL: <https://www.nature.com/articles/s41598-021-99105-2>.

42. Cai, Tommaso et al. (Feb. 2023). "Artificial Intelligence Can Guide Antibiotic Choice in Recurrent UTIs and Become an Important Aid to Improve Antimicrobial Stewardship". In: *Antibiotics* 12.2, p. 375. ISSN: 2079-6382. doi: 10.3390/antibiotics12020375. URL: <https://www.mdpi.com/2079-6382/12/2/375>.
43. Hernandez Perez, Bernard (2019). PhD thesis. Imperial College London. doi: 10.25560/73000. URL: <https://spiral.imperial.ac.uk/handle/10044/1/73000>.
44. Imai, Shungo et al. (2020). "A New Algorithm Optimized for Initial Dose Settings of Vancomycin Using Machine Learning". In: *Biological and Pharmaceutical Bulletin* 43.1, pp. 188–193. doi: 10.1248/bpb.b19-00729.
45. Huang, Xiaohui et al. (June 2021). "Prediction of vancomycin dose on high-dimensional data using machine learning techniques". In: *Expert Review of Clinical Pharmacology* 14.6, pp. 761–771. ISSN: 1751-2433. doi: 10.1080/17512433.2021.1911642. URL: <https://doi.org/10.1080/17512433.2021.1911642>.
46. Beaudoin, Mathieu et al. (Mar. 2016). "Evaluation of a machine learning capability for a clinical decision support system to enhance antimicrobial stewardship programs". In: *Artificial Intelligence in Medicine* 68, pp. 29–36. ISSN: 0933-3657. doi: 10.1016/j.artmed.2016.02.001. URL: <https://www.sciencedirect.com/science/article/pii/S0933365716300574>.
47. Boominathan, Soorajath et al. (Aug. 2020). "Treatment Policy Learning in Multiobjective Settings with Fully Observed Outcomes". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '20. New York, NY, USA: Association for Computing Machinery, pp. 1937–1947. ISBN: 978-1-4503-7998-4. doi: 10.1145/3394486.3403245. URL: <https://doi.org/10.1145/3394486.3403245>.
48. Tran-The, Tam et al. (Jan. 2024). "Development of machine learning algorithms for scaling-up antibiotic stewardship". In: *International Journal of Medical Informatics* 181, p. 105300. ISSN: 1386-5056. doi: 10.1016/j.ijmedinf.2023.105300. URL: <https://www.sciencedirect.com/science/article/pii/S1386505623003180>.
49. Hulgán, T et al. (2004). "Oral quinolones in hospitalized patients: an evaluation of a computerized decision support intervention". In: *Journal of internal medicine* 256.4, pp. 349–357.
50. McGregor, Jessina C et al. (2006b). "Impact of a computerized clinical decision support system on reducing inappropriate antimicrobial use: a randomized controlled trial". In: *Journal of the American Medical Informatics Association* 13.4, pp. 378–384.
51. Lee, Seunghee et al. (Jan. 2022). "Hybrid Method Incorporating a Rule-Based Approach and Deep Learning for Prescription Error Prediction". In: *Drug Safety* 45.1, pp. 27–35. ISSN: 1179-1942. doi: 10.1007/s40264-021-01123-6. URL: <https://doi.org/10.1007/s40264-021-01123-6>.

52. Zitnik, Marinka, Agrawal, Monica, and Leskovec, Jure (July 2018). "Modeling polypharmacy side effects with graph convolutional networks". In: *Bioinformatics* 34.13, pp. i457–i466. ISSN: 1367-4803. doi: 10.1093/bioinformatics/bty294. URL: <https://doi.org/10.1093/bioinformatics/bty294>.
53. Rawson, T. M. et al. (Aug. 2017a). "A systematic review of clinical decision support systems for antimicrobial management: are we failing to investigate these interventions appropriately?" In: *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 23.8, pp. 524–532. ISSN: 1469-0691. doi: 10.1016/j.cmi.2017.02.028.
54. Luz, C. F. et al. (Oct. 2020). "Machine learning in infection management using routine electronic health records: tools, techniques, and reporting of future technologies". In: *Clinical Microbiology and Infection* 26.10, pp. 1291–1299. ISSN: 1198-743X. doi: 10.1016/j.cmi.2020.02.003. URL: <http://www.sciencedirect.com/science/article/pii/S1198743X20300823>.
55. Peiffer-Smadja, N. et al. (May 2020). "Machine learning for clinical decision support in infectious diseases: a narrative review of current applications". In: *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 26.5, pp. 584–595. ISSN: 1469-0691. doi: 10.1016/j.cmi.2019.09.009.
56. Anahtar, Melis N., Yang, Jason H., and Kanjilal, Sanjat (June 2021). "Applications of Machine Learning to the Problem of Antimicrobial Resistance: an Emerging Model for Translational Research". In: *Journal of Clinical Microbiology* 59.7, 10.1128/jcm.01260–20. doi: 10.1128/jcm.01260–20. URL: <https://journals.asm.org/doi/full/10.1128/jcm.01260-20>.
57. He, Sheng, Leanse, Leon G., and Feng, Yanfang (Nov. 2021). "Artificial intelligence and machine learning assisted drug delivery for effective treatment of infectious diseases". In: *Advanced Drug Delivery Reviews* 178, p. 113922. ISSN: 0169-409X. doi: 10.1016/j.addr.2021.113922. URL: <https://www.sciencedirect.com/science/article/pii/S0169409X2100315X>.
58. Elligsen, Marion et al. (July 2021). "Improving Decision Making in Empiric Antibiotic Selection (IDEAS) for Gram-negative Bacteremia: A Prospective Clinical Implementation Study". In: *Clinical Infectious Diseases* 73.2, e417–e425. ISSN: 1058-4838. doi: 10.1093/cid/ciaa921. URL: <https://doi.org/10.1093/cid/ciaa921>.
59. Adams, Roy et al. (July 2022). "Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis". In: *Nature Medicine* 28.7, pp. 1455–1460. ISSN: 1546-170X. doi: 10.1038/s41591-022-01894-0. URL: <https://www.nature.com/articles/s41591-022-01894-0>.

60. Eickelberg, Garrett, Sanchez-Pinto, L. Nelson, and Luo, Yuan (Sept. 2020). "Predictive modeling of bacterial infections and antibiotic therapy needs in critically ill adults". In: *Journal of Biomedical Informatics* 109, p. 103540. ISSN: 1532-0464. doi: 10.1016/j.jbi.2020.103540. URL: <https://www.sciencedirect.com/science/article/pii/S1532046420301684>.
61. Feng, T. et al. (2023). "Machine learning-based clinical decision support for infection risk prediction". In: *Frontiers in Medicine* 10. ISSN: 2296-858X. doi: 10.3389/fmed.2023.1213411.
62. Wong, Joshua Guoxian et al. (Nov. 2020). "Risk prediction models to guide antibiotic prescribing: a study on adult patients with uncomplicated upper respiratory tract infections in an emergency department". In: *Antimicrobial Resistance & Infection Control* 9.1, p. 171. ISSN: 2047-2994. doi: 10.1186/s13756-020-00825-3. URL: <https://doi.org/10.1186/s13756-020-00825-3>.
63. Seheult, Jansen N. et al. (May 2023). "Development, Evaluation, and Multisite Deployment of a Machine Learning Decision Tree Algorithm To Optimize Urinalysis Parameters for Predicting Urine Culture Positivity". In: *Journal of Clinical Microbiology* 61.6, e00291-23. doi: 10.1128/jcm.00291-23. URL: <https://journals.asm.org/doi/10.1128/jcm.00291-23>.
64. Moran, Ed et al. (Sept. 2020). "Towards personalized guidelines: using machine-learning algorithms to guide antimicrobial selection". In: *Journal of Antimicrobial Chemotherapy* 75.9, pp. 2677-2680. ISSN: 0305-7453. doi: 10.1093/jac/dkaa222. URL: <https://doi.org/10.1093/jac/dkaa222>.
65. Feretzakis, Georgios et al. (July 2021). "Machine Learning for Antibiotic Resistance Prediction: A Prototype Using Off-the-Shelf Techniques and Entry-Level Data to Guide Empiric Antimicrobial Therapy". In: *Healthcare Informatics Research* 27.3, pp. 214-221. doi: 10.4258/hir.2021.27.3.214. URL: <https://synapse.koreamed.org/articles/1147754>.
66. Corbin, Conor K. et al. (Apr. 2022). "Personalized antibiograms for machine learning driven antibiotic selection". In: *Communications Medicine* 2.1, pp. 1-14. ISSN: 2730-664X. doi: 10.1038/s43856-022-00094-8. URL: <https://www.nature.com/articles/s43856-022-00094-8>.
67. Rich, Shannan N. et al. (Oct. 2022). "Development of a Prediction Model for Antibiotic-Resistant Urinary Tract Infections Using Integrated Electronic Health Records from Multiple Clinics in North-Central Florida". In: *Infectious Diseases and Therapy* 11.5, pp. 1869-1882. ISSN: 2193-6382. doi: 10.1007/s40121-022-00677-x. URL: <https://doi.org/10.1007/s40121-022-00677-x>.
68. Boerman, Anneroo W. et al. (Jan. 2022). "Using machine learning to predict blood culture outcomes in the emergency department: a single-centre, retrospective, observational study". In: *BMJ Open* 12.1, e053332. ISSN: 2044-6055, 2044-6055. doi: 10.1136/bmjopen-2021-053332. URL: <https://bmjopen.bmj.com/content/12/1/e053332>.

69. Tharmakulasingam, Mukunthan et al. (2023). "TransAMR: An Interpretable Transformer Model for Accurate Prediction of Antimicrobial Resistance Using Antibiotic Administration Data". In: *IEEE Access* 11, pp. 75337–75350. ISSN: 2169-3536. doi: 10.1109/ACCESS.2023.3296221.
70. Kim, C. et al. (2023). "Translation of Machine Learning-Based Prediction Algorithms to Personalised Empiric Antibiotic Selection: A Population-Based Cohort Study". In: *International Journal of Antimicrobial Agents* 62.5. ISSN: 0924-8579. doi: 10.1016/j.ijantimicag.2023.106966.
71. Lee, Hyun-Gyu et al. (Feb. 2023). "Machine learning model for predicting ciprofloxacin resistance and presence of ESBL in patients with UTI in the ED". In: *Scientific Reports* 13.1, p. 3282. ISSN: 2045-2322. doi: 10.1038/s41598-023-30290-y. URL: <https://www.nature.com/articles/s41598-023-30290-y>.
72. Cavallaro, Massimo et al. (Jan. 2023). "Informing antimicrobial stewardship with explainable AI". In: *PLOS Digital Health* 2.1, e0000162. ISSN: 2767-3170. doi: 10.1371/journal.pdig.0000162. URL: <https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000162>.
73. Qin, Jia'an et al. (May 2024). "Antibiotic combinations prediction based on machine learning to multi-centre clinical data and drug interaction correlation". In: *International Journal of Antimicrobial Agents* 63.5, p. 107122. ISSN: 0924-8579. doi: 10.1016/j.ijantimicag.2024.107122. URL: <https://www.sciencedirect.com/science/article/pii/S0924857924000402>.
74. Smith, N. M. et al. (Sept. 2020). "Using machine learning to optimize antibiotic combinations: dosing strategies for meropenem and polymyxin B against carbapenem-resistant *Acinetobacter baumannii*". In: *Clinical Microbiology and Infection* 26.9, pp. 1207–1213. ISSN: 1198-743X. doi: 10.1016/j.cmi.2020.02.004. URL: <https://www.sciencedirect.com/science/article/pii/S1198743X20300835>.
75. Batisti Biffignandi, Gherard et al. (2024). "Optimising machine learning prediction of minimum inhibitory concentrations in *Klebsiella pneumoniae*". In: *Microbial Genomics* 10.3, p. 001222. ISSN: 2057-5858. doi: 10.1099/mgen.0.001222. URL: <https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.001222>.
76. Pascucci, Marco et al. (Feb. 2021). "AI-based mobile application to fight antibiotic resistance". In: *Nature Communications* 12.1, p. 1173. ISSN: 2041-1723. doi: 10.1038/s41467-021-21187-3. URL: <https://www.nature.com/articles/s41467-021-21187-3>.
77. Bollapragada, Keerthi Priya, Dodda, Akhila Reddy, and Dodda, Venkata Rama Koti Reddy (2024). "Development of Deep Learning based Antimicrobial Susceptibility Test Smart System". In: *IETE Journal of Research* 0.0, pp. 1–19. ISSN: 0377-2063. doi: 10.1080/03772063.2024.2353353. URL: <https://doi.org/10.1080/03772063.2024.2353353>.
78. Gijzen, Ronald et al. (July 2001). "Causes and consequences of comorbidity: A review". In: *Journal of Clinical Epidemiology* 54.7, pp. 661–674. ISSN: 0895-4356. doi: 10.1016/S0895-4356(00)00363-2.

79. Riaño, David, Peleg, Mor, and Teije, Annette ten (Sept. 2019). "Ten years of knowledge representation for health care (2009–2018): Topics, trends, and challenges". In: *Artificial Intelligence in Medicine* 100, p. 101713. ISSN: 0933-3657. doi: 10.1016/j.artmed.2019.101713.
80. Huttunen, R. and Syrjänen, J. (Mar. 2013). "Obesity and the risk and outcome of infection". In: *International Journal of Obesity* 37.3, pp. 333–340. ISSN: 1476-5497. doi: 10.1038/ijo.2012.62. URL: <https://www.nature.com/articles/ijo201262/>.
81. Charani, Esmita et al. (Oct. 2015). "Antimicrobial therapy in obesity: a multicentre cross-sectional study". In: *Journal of Antimicrobial Chemotherapy* 70.10, pp. 2906–2912. ISSN: 0305-7453. doi: 10.1093/jac/dkv189. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4566962/>.
82. Medical Sciences, The Academy of (Apr. 2018). *Multimorbidity: a priority for global health research* | The Academy of Medical Sciences. Tech. rep. URL: <https://acmedsci.ac.uk/policy/policy-projects/multimorbidity>.
83. Pearson-Stuttard, Jonathan, Ezzati, Majid, and Gregg, Edward W. (Dec. 2019). "Multimorbidity—a defining challenge for health systems". In: *The Lancet Public Health* 4.12, e599–e600. ISSN: 2468-2667. doi: 10.1016/S2468-2667(19)30222-1. URL: [https://www.thelancet.com/journals/lanpub/article/PIIS2468-2667\(19\)30222-1/abstract](https://www.thelancet.com/journals/lanpub/article/PIIS2468-2667(19)30222-1/abstract).
84. NICE (Sept. 2019). *Multimorbidity and polypharmacy* | Advice. URL: <https://www.nice.org.uk/advice/ktt18/chapter/Key-points>.
85. NIHR (Apr. 2020). *NIHR Strategic Framework for Multiple Long-Term Conditions (Multimorbidity) MLTC-M Research*. URL: <https://www.nihr.ac.uk/documents/nihr-strategic-framework-for-multiple-long-term-conditions-multimorbidity-mltc-m-research/24639>.
86. McDaid, David and Park, A-La (2015). "Counting All the Costs: The Economic Impact of Comorbidity". In: *Comorbidity of Mental and Physical Disorders* 179, pp. 23–32. doi: 10.1159/000365941. URL: <https://www.karger.com/Article/FullText/365941>.
87. Ticinesi, Andrea et al. (Oct. 2015). "Multimorbidity in elderly hospitalised patients and risk of Clostridium difficile infection: a retrospective study with the Cumulative Illness Rating Scale (CIRS)". In: *BMJ Open* 5.10, e009316. ISSN: 2044-6055, 2044-6055. doi: 10.1136/bmjopen-2015-009316. URL: <https://bmjopen.bmj.com/content/5/10/e009316>.
88. Weir, D. L. et al. (Mar. 2015). "The impact of multimorbidity on short-term events in patients with community-acquired pneumonia: prospective cohort study". In: *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 21.3, 264.e7–264.e13. ISSN: 1469-0691. doi: 10.1016/j.cmi.2014.11.002.

89. Grace, Edward (June 2012). "Altered vancomycin pharmacokinetics in obese and morbidly obese patients: what we have learned over the past 30 years". In: *Journal of Antimicrobial Chemotherapy* 67.6, pp. 1305–1310. ISSN: 0305-7453. doi: 10.1093/jac/dks066. URL: <https://academic.oup.com/jac/article/67/6/1305/691495>.
90. Polso, A. K., Lassiter, J. L., and Nagel, J. L. (Dec. 2014). "Impact of hospital guideline for weight-based antimicrobial dosing in morbidly obese adults and comprehensive literature review". In: *Journal of Clinical Pharmacy and Therapeutics* 39.6, pp. 584–608. ISSN: 1365-2710. doi: 10.1111/jcpt.12200.
91. Boyd, Sara Elizabeth et al. (Dec. 2016). "Information provision for antibacterial dosing in the obese patient: a sizeable absence?" In: *Journal of Antimicrobial Chemotherapy* 71.12, pp. 3588–3592. ISSN: 0305-7453. doi: 10.1093/jac/dkw324. URL: <https://academic.oup.com/jac/article/71/12/3588/2631283>.
92. Rawson, Timothy M. et al. (Mar. 2017b). "Vancomycin therapy in secondary care; investigating factors that impact therapeutic target attainment". In: *Journal of Infection* 74.3, pp. 320–324. ISSN: 0163-4453. doi: 10.1016/j.jinf.2016.12.011. URL: <https://www.sciencedirect.com/science/article/pii/S0163445316303425>.
93. Vincent, Jean-Louis et al. (May 2016). "Advances in antibiotic therapy in the critically ill". In: *Critical Care* 20.1, p. 133. ISSN: 1364-8535. doi: 10.1186/s13054-016-1285-6. URL: <https://doi.org/10.1186/s13054-016-1285-6>.
94. Charlson, M. E. et al. (1987). "A new method of classifying prognostic comorbidity in longitudinal studies: development and validation". In: *Journal of Chronic Diseases* 40.5, pp. 373–383. ISSN: 0021-9681. doi: 10.1016/0021-9681(87)90171-8.
95. Deyo, Richard A., Cherkin, Daniel C., and Ciol, Marcia A. (June 1992). "Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases". In: *Journal of Clinical Epidemiology* 45.6, pp. 613–619. ISSN: 0895-4356. doi: 10.1016/0895-4356(92)90133-8.
96. Violán, Concepción et al. (Aug. 2019). "Soft clustering using real-world data for the identification of multi-morbidity patterns in an elderly population: cross-sectional study in a Mediterranean population". In: *BMJ open* 9.8, e029594. ISSN: 2044-6055. doi: 10.1136/bmjopen-2019-029594.
97. Haug, Nina et al. (Mar. 2020). "High-risk multimorbidity patterns on the road to cardiovascular mortality". In: *BMC medicine* 18.1, p. 44. ISSN: 1741-7015. doi: 10.1186/s12916-020-1508-1.
98. Beaney, Thomas et al. (Jan. 2024a). "Assigning disease clusters to people: A cohort study of the implications for understanding health outcomes in people with multiple long-term conditions". In: *Journal of Multimorbidity and Comorbidity* 14, p. 26335565241247430. ISSN: 2633-5565. doi: 10.1177/26335565241247430. URL: <https://doi.org/10.1177/26335565241247430>.

99. Beaney, Thomas et al. (May 2024b). "Identifying multi-resolution clusters of diseases in ten million patients with multimorbidity in primary care in England". In: *Communications Medicine* 4.1, pp. 1–15. issn: 2730-664X. doi: 10.1038/s43856-024-00529-4. URL: <https://www.nature.com/articles/s43856-024-00529-4>.
100. Zador, Zsolt et al. (July 2019). "Multimorbidity states associated with higher mortality rates in organ dysfunction and sepsis: a data-driven analysis in critical care". In: *Critical Care* 23.1, p. 247. issn: 1364-8535. doi: 10.1186/s13054-019-2486-6.
101. Hassaine, Abdelaali et al. (Sept. 2020). "Untangling the complexity of multimorbidity with machine learning". In: *Mechanisms of Ageing and Development* 190, p. 111325. issn: 0047-6374. doi: 10.1016/j.mad.2020.111325. URL: <https://www.sciencedirect.com/science/article/pii/S0047637420301214>.
102. Violán, Concepción et al. (Oct. 2020). "Five-year trajectories of multimorbidity patterns in an elderly Mediterranean population using Hidden Markov Models". In: *Scientific Reports* 10, p. 16879. issn: 2045-2322. doi: 10.1038/s41598-020-73231-9. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7547668/>.
103. Majnarić, Ljiljana Trtica et al. (Feb. 2021). "AI and Big Data in Healthcare: Towards a More Comprehensive Research Framework for Multimorbidity". In: *Journal of Clinical Medicine* 10.4, p. 766. issn: 2077-0383. doi: 10.3390/jcm10040766. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7918668/>.
104. Cheng, Lily Shui-kuen et al. (Jan. 2020). "Bacterial co-infections and antibiotic prescribing practice in adults with COVID-19: experience from a single hospital cluster". In: *Therapeutic Advances in Infectious Disease* 7, p. 2049936120978095. issn: 2049-9361. doi: 10.1177/2049936120978095. URL: <https://doi.org/10.1177/2049936120978095>.
105. Papin, Grégory et al. (Aug. 2021). "Clinical and biological clusters of sepsis patients using hierarchical clustering". In: *PLOS ONE* 16.8, e0252793. issn: 1932-6203. doi: 10.1371/journal.pone.0252793. URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0252793>.
106. Nielsen, Annelaura B. et al. (June 2019). "Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish National Patient Registry and electronic patient records". In: *The Lancet. Digital Health* 1.2, e78–e89. issn: 2589-7500. doi: 10.1016/S2589-7500(19)30024-X.
107. Zhou, Jiayu et al. (Aug. 2014). "From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records". In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '14. New York, NY, USA: Association for Computing Machinery, pp. 135–144. doi: 10.1145/2623330.2623711. URL: <https://doi.org/10.1145/2623330.2623711>.

108. Afshar, Ardavan et al. (Apr. 2020). "TASTE: temporal and static tensor factorization for phenotyping electronic health records". In: *Proceedings of the ACM Conference on Health, Inference, and Learning*. CHIL '20. New York, NY, USA: Association for Computing Machinery, pp. 193–203. doi: 10.1145/3368555.3384464. URL: <https://doi.org/10.1145/3368555.3384464>.
109. Li, Yikuan et al. (Apr. 2020). "BEHRT: Transformer for Electronic Health Records". In: *Scientific Reports* 10.1, p. 7155. ISSN: 2045-2322. doi: 10.1038/s41598-020-62922-y. URL: <https://www.nature.com/articles/s41598-020-62922-y>.
110. Kraljevic, Zeljko et al. (Jan. 2023). *Foresight – Generative Pretrained Transformer (GPT) for Modelling of Patient Timelines using EHRs*. doi: 10.48550/arXiv.2212.08072. URL: <http://arxiv.org/abs/2212.08072>.
111. Beam, Andrew L. et al. (Nov. 2019). "Clinical Concept Embeddings Learned from Massive Sources of Multi-modal Medical Data". In: *Biocomputing 2020*. World Scientific, pp. 295–306. doi: 10.1142/9789811215636_0027. URL: https://www.worldscientific.com/doi/abs/10.1142/9789811215636%5C_0027.
112. Noll, Richard, Schaaf, Jannik, and Storf, Holger (2022). "The Use of Computer-Assisted Case-Based Reasoning to Support Clinical Decision-Making – A Scoping Review". In: *Case-Based Reasoning Research and Development*. Ed. by Mark T. Keane and Nirmalie Wiratunga. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 395–409. ISBN: 978-3-031-14923-8. doi: 10.1007/978-3-031-14923-8_26.
113. Choi, Edward et al. (Jan. 2020). "Learning the Graphical Structure of Electronic Health Records with Graph Convolutional Transformer". In: *arXiv:1906.04716 [cs, stat]*. URL: <http://arxiv.org/abs/1906.04716>.
114. Zhang, Jinghe, Gong, Jiaqi, and Barnes, Laura (July 2017a). "HCNN: Heterogeneous Convolutional Neural Networks for Comorbid Risk Prediction with Electronic Health Records". In: *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pp. 214–221. doi: 10.1109/CHASE.2017.80.
115. Rocheteau, Emma et al. (Jan. 2021). "Predicting Patient Outcomes with Graph Representation Learning". In: *arXiv:2101.03940 [cs]*. URL: <http://arxiv.org/abs/2101.03940>.
116. Zhang, Yutao et al. (Aug. 2017b). "LEAP: Learning to Prescribe Effective and Safe Treatment Combinations for Multimorbidity". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '17. New York, NY, USA: Association for Computing Machinery, pp. 1315–1324. doi: 10.1145/3097983.3098109. URL: <https://doi.org/10.1145/3097983.3098109>.
117. Onder, Graziano et al. (Sept. 2020). "Facing multimorbidity in the precision medicine era". In: *Mechanisms of Ageing and Development* 190, p. 111287. ISSN: 0047-6374. doi: 10.1016/j.mad.2020.111287. URL: <https://www.sciencedirect.com/science/article/pii/S004763742030083X>.

118. Whitty, Christopher J. M. et al. (Jan. 2020). "Rising to the challenge of multimorbidity". In: *BMJ* 368, p. l6964. ISSN: 1756-1833. doi: 10.1136/bmj.l6964. URL: <https://www.bmj.com/content/368/bmj.l6964>.
119. Fraccaro, Paolo et al. (Jan. 2015). "Adoption of Clinical Decision Support in Multimorbidity: A Systematic Review". In: *JMIR Medical Informatics* 3.1, e3503. doi: 10.2196/medinform.3503. URL: <https://medinform.jmir.org/2015/1/e4>.
120. ECDC (Oct. 2013). *Health inequalities, the financial crisis, and infectious disease in Europe*. Tech. rep. URL: <https://www.ecdc.europa.eu/en/publications-data/health-inequalities-financial-crisis-and-infectious-disease-europe>.
121. Williams, David R., Priest, Naomi, and Anderson, Norman (Apr. 2016). "Understanding Associations between Race, Socioeconomic Status and Health: Patterns and Prospects". In: *Health psychology : official journal of the Division of Health Psychology, American Psychological Association* 35.4, pp. 407–411. ISSN: 0278-6133. doi: 10.1037/hea0000242. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4817358/>.
122. Oestergaard, Louise Bruun et al. (Aug. 2017). "The associations between socioeconomic status and risk of Staphylococcus aureus bacteremia and subsequent endocarditis – a Danish nationwide cohort study". In: *BMC Infectious Diseases* 17.1, p. 589. ISSN: 1471-2334. doi: 10.1186/s12879-017-2691-3. URL: <https://doi.org/10.1186/s12879-017-2691-3>.
123. Alividza, Vivian et al. (Aug. 2018). "Investigating the impact of poverty on colonization and infection with drug-resistant organisms in humans: a systematic review". In: *Infectious Diseases of Poverty* 7.1, p. 76. ISSN: 2049-9957. doi: 10.1186/s40249-018-0459-7. URL: <https://doi.org/10.1186/s40249-018-0459-7>.
124. Donnelly, John P. et al. (June 2018). "Association of Neighborhood Socioeconomic Status With Risk of Infection and Sepsis". In: *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 66.12, pp. 1940–1947. ISSN: 1537-6591. doi: 10.1093/cid/cix1109.
125. Colzani, Edoardo (Aug. 2019). "Beyond morbidity and mortality: the burden of infectious diseases on health-care services". In: *Epidemiology and Infection* 147, e251. ISSN: 0950-2688. doi: 10.1017/S0950268819001298. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6805732/>.
126. Fiscella, Kevin and Williams, David R (2004). "Health disparities based on socioeconomic inequities: implications for urban health care". In: *Academic Medicine* 79.12, pp. 1139–1147.
127. Caballo, Brian et al. (2021). "The effects of socioeconomic status on the quality and accessibility of health-care services". In: *Across The Spectrum of Socioeconomics: Issue IV* 236, pp. 1–16.
128. GOV.UK (Sept. 2020). *Socioeconomic status*. URL: <https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/demographics/socioeconomic-status/latest%5C#ethnic-groups-by-socio-economic-status>.

129. Prats-Urbe, Albert, Paredes, Roger, and Prieto-Alhambra, Daniel (June 2020). "Ethnicity, comorbidity, socioeconomic status, and their associations with COVID-19 infection in England: a cohort analysis of UK Biobank data". In: *medRxiv*, p. 2020.05.06.20092676. doi: 10.1101/2020.05.06.20092676. URL: <https://www.medrxiv.org/content/10.1101/2020.05.06.20092676v3>.
130. Daw, Jonathan (Jan. 2017). "Contribution of Four Comorbid Conditions to Racial/Ethnic Disparities in Mortality Risk". In: *American Journal of Preventive Medicine* 52.1S1, S95–S102. ISSN: 1873–2607. doi: 10.1016/j.amepre.2016.07.036.
131. Harrison, Ewen et al. (Apr. 2020). *Investigating associations between ethnicity and outcome from COVID-19 – report to SAGE, 25 April 2020*. Tech. rep. URL: <https://www.gov.uk/government/publications/covid-19-investigating-associations-between-ethnicity-and-outcome-from-covid-19-report-to-sage-25-april-2020>.
132. Pan, Daniel et al. (2020). "The impact of ethnicity on clinical outcomes in COVID-19: a systematic review". In: *EClinicalMedicine* 23.
133. Academy, British (2021). *The COVID Decade: understanding the long-term societal impacts of COVID-19*. Tech. rep. URL: <https://www.thebritishacademy.ac.uk/publications/covid-decade-understanding-the-long-term-societal-impacts-of-covid-19/>.
134. Docherty, Annemarie B. et al. (May 2020). "Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study". In: *BMJ* 369, p. m1985. ISSN: 1756–1833. doi: 10.1136/bmj.m1985. URL: <https://www.bmj.com/content/369/bmj.m1985>.
135. Iaccarino, Guido et al. (Aug. 2020). "Age and Multimorbidity Predict Death Among COVID-19 Patients: Results of the SARS-RAS Study of the Italian Society of Hypertension". In: *Hypertension (Dallas, Tex.: 1979)* 76.2, pp. 366–372. ISSN: 1524–4563. doi: 10.1161/HYPERTENSIONAHA.120.15324.
136. Niedzwiedz, Claire L. et al. (May 2020). "Ethnic and socioeconomic differences in SARS-CoV-2 infection: prospective cohort study using UK Biobank". In: *BMC Medicine* 18.1, p. 160. ISSN: 1741–7015. doi: 10.1186/s12916-020-01640-8. URL: <https://doi.org/10.1186/s12916-020-01640-8>.
137. Patel, Aatish et al. (Oct. 2020). "Investigating the association between ethnicity and health outcomes in SARS-CoV-2 in a London secondary care population". In: *PLOS ONE* 15.10, e0240960. ISSN: 1932–6203. doi: 10.1371/journal.pone.0240960. URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0240960>.

138. Raharja, Antony, Tamara, Alice, and Kok, Li Teng (Nov. 2020). "Association Between Ethnicity and Severe COVID-19 Disease: a Systematic Review and Meta-analysis". In: *Journal of Racial and Ethnic Health Disparities*, pp. 1–10. ISSN: 2197-3792. doi: 10.1007/s40615-020-00921-5. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7659894/>.
139. Williamson, Elizabeth J. et al. (Aug. 2020). "Factors associated with COVID-19-related death using OpenSAFELY". In: *Nature* 584.7821, pp. 430–436. ISSN: 1476-4687. doi: 10.1038/s41586-020-2521-4.
140. Goetz, Lea et al. (May 2024). "Generalization—a key challenge for responsible AI in patient-facing clinical applications". In: *npj Digital Medicine* 7.1, pp. 1–4. ISSN: 2398-6352. doi: 10.1038/s41746-024-01127-3. URL: <https://www.nature.com/articles/s41746-024-01127-3>.
141. Assaf, Dan et al. (Aug. 2020). "Utilization of machine-learning models to accurately predict the risk for critical COVID-19". In: *Internal and Emergency Medicine*, pp. 1–9. ISSN: 1828-0447. doi: 10.1007/s11739-020-02475-0. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7433773/>.
142. Dashti, Hesam et al. (Nov. 2020). "Abstract 17400: Using Electronic Health Records for Predicting Hospitalization of COVID-19 Patients in Massachusetts". In: *Circulation* 142.Suppl_3, A17400–A17400. doi: 10.1161/circ.142.suppl_3.17400. URL: https://www.ahajournals.org/doi/10.1161/circ.142.suppl_3.17400.
143. Wang, Linda, Lin, Zhong Qiu, and Wong, Alexander (Nov. 2020). "COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images". In: *Scientific Reports* 10.1, p. 19549. ISSN: 2045-2322. doi: 10.1038/s41598-020-76550-z. URL: <https://www.nature.com/articles/s41598-020-76550-z>.
144. Yadaw, Arjun S et al. (Oct. 2020). "Clinical features of COVID-19 mortality: development and validation of a clinical prediction model". In: *The Lancet. Digital Health* 2.10, e516–e525. ISSN: 2589-7500. doi: 10.1016/S2589-7500(20)30217-X. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7508513/>.
145. Wynants, Laure et al. (Apr. 2020). "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal". In: *BMJ* 369, p. m1328. ISSN: 1756-1833. doi: 10.1136/bmj.m1328. URL: <https://www.bmj.com/content/369/bmj.m1328>.
146. Vandersluis, Robert and Savulescu, Julian (2024). "The selective deployment of AI in healthcare". In: *Bioethics* 38.5, pp. 391–400. ISSN: 1467-8519. doi: 10.1111/bioe.13281. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/bioe.13281>.
147. Chen, Min, Tan, Xuan, and Padman, Rema (2020). "Social determinants of health in electronic health records and their impact on analysis and risk prediction: a systematic review". In: *Journal of the American Medical Informatics Association* 27.11, pp. 1764–1773.

148. Kamiran, Faisal and Calders, Toon (Oct. 2012). "Data preprocessing techniques for classification without discrimination". In: *Knowledge and Information Systems* 33.1, pp. 1–33. ISSN: 0219-3116. doi: 10.1007/s10115-011-0463-8. URL: <https://doi.org/10.1007/s10115-011-0463-8>.
149. Hardt, Moritz, Price, Eric, and Srebro, Nathan (Oct. 2016). "Equality of Opportunity in Supervised Learning". In: *arXiv:1610.02413 [cs]*. URL: <http://arxiv.org/abs/1610.02413>.
150. Beutel, Alex et al. (July 2017). "Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations". In: *arXiv:1707.00075 [cs]*. URL: <http://arxiv.org/abs/1707.00075>.
151. Zafar, Muhammad Bilal et al. (Mar. 2017a). "Fairness Constraints: Mechanisms for Fair Classification". In: *arXiv:1507.05259 [cs, stat]*. URL: <http://arxiv.org/abs/1507.05259>.
152. Zafar, Muhammad Bilal et al. (Apr. 2017b). "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment". In: *Proceedings of the 26th International Conference on World Wide Web*, pp. 1171–1180. doi: 10.1145/3038912.3052660. URL: <http://arxiv.org/abs/1610.08452>.
153. Madras, David et al. (Oct. 2018). "Learning Adversarially Fair and Transferable Representations". In: *arXiv:1802.06309 [cs, stat]*. URL: <http://arxiv.org/abs/1802.06309>.
154. Gajane, Pratik and Pechenizkiy, Mykola (May 2018). "On Formalizing Fairness in Prediction with Machine Learning". In: *arXiv:1710.03184 [cs, stat]*. URL: <http://arxiv.org/abs/1710.03184>.
155. Verma, Sahil and Rubin, J. (2018). "Fairness Definitions Explained". In: *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. doi: 10.1145/3194770.3194776.
156. WHO (2018). "Tackling antimicrobial resistance (AMR) together: working paper 5.0: enhancing the focus on gender and equity". In: URL: <https://apps.who.int/iris/handle/10665/336977>.
157. Khalatbari-Soltani, Saman et al. (Aug. 2020). "Importance of collecting data on socioeconomic determinants from the early stage of the COVID-19 outbreak onwards". In: *J Epidemiol Community Health* 74.8, pp. 620–623. ISSN: 0143-005X, 1470-2738. doi: 10.1136/jech-2020-214297. URL: <https://jech.bmj.com/content/74/8/620>.
158. Leslie, David et al. (Mar. 2021). "Does 'AI' stand for augmenting inequality in the era of covid-19 healthcare?" In: *BMJ* 372, n304. ISSN: 1756-1833. doi: 10.1136/bmj.n304. URL: <https://www.bmj.com/content/372/bmj.n304>.
159. Littmann, Jasper and Viens, A. M. (Nov. 2015). "The Ethical Significance of Antimicrobial Resistance". In: *Public Health Ethics* 8.3, pp. 209–224. ISSN: 1754-9973. doi: 10.1093/phe/phv025. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4638062/>.

160. Parsonage, Ben et al. (Nov. 2017). "Control of Antimicrobial Resistance Requires an Ethical Approach". In: *Frontiers in Microbiology* 8. ISSN: 1664-302X. doi: 10.3389/fmicb.2017.02124. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5673829/>.
161. Butler, Christopher C. et al. (Sept. 2001). "Antibiotics and shared decision-making in primary care". In: *Journal of Antimicrobial Chemotherapy* 48.3, pp. 435–440. ISSN: 0305-7453. doi: 10.1093/jac/48.3.435. URL: <https://academic.oup.com/jac/article/48/3/435/736084>.
162. Brink, Adrian John and Richards, Guy (Oct. 2020). "Best practice: antibiotic decision-making in ICUs". In: *Current Opinion in Critical Care* 26.5, pp. 478–488. ISSN: 1070-5295. doi: 10.1097/MCC.0000000000000752. URL: https://journals.lww.com/co-criticalcare/Abstract/2020/10000/Best%5C_practice%5C_%5C_antibiotic%5C_decision%5C_making%5C_in%5C_ICUs.8.aspx.
163. NICE (Aug. 2015). *Antimicrobial stewardship: systems and processes for effective antimicrobial medicine use | Guidance*. URL: <https://www.nice.org.uk/guidance/ng15>.
164. Rawson, Timothy Miles et al. (Dec. 2016b). "Mapping the decision pathways of acute infection management in secondary care among UK medical physicians: a qualitative study". In: *BMC Medicine* 14.1, p. 208. ISSN: 1741-7015. doi: 10.1186/s12916-016-0751-y. URL: <https://doi.org/10.1186/s12916-016-0751-y>.
165. Badea, Cosmin (Sept. 2021). "Have a break from making decisions, have a MARS: The Multi-valued Action Reasoning System". In: *arXiv:2109.03283 [cs]*. URL: <http://arxiv.org/abs/2109.03283>.
166. Sinnott-Armstrong, Walter (May 2003). "Consequentialism". In: *Stanford Encyclopedia of Philosophy Archive*. URL: <https://plato.stanford.edu/archives/sum2019/entries/consequentialism/>.
167. GMC (Apr. 2019). *The duties of a doctor registered with the General Medical Council*. URL: <https://www.gmc-uk.org/ethical-guidance/ethical-guidance-for-doctors/good-medical-practice/duties-of-a-doctor>.
168. Bentham, Jeremy (1789). *An Introduction to the Principles of Morals and Legislation*. Dover Philosophical Classics.
169. Post, Benjamin et al. (June 2022). *Breaking Bad News in the Era of Artificial Intelligence and Algorithmic Medicine: An Exploration of Disclosure and its Ethical Justification using the Hedonic Calculus*. doi: 10.48550/arXiv.2207.01431. URL: <http://arxiv.org/abs/2207.01431>.
170. Rzepka, Rafal and Araki, Kenji (Dec. 2013). "Possible Usage of Sentiment Analysis for Calculating Vectors of Felific Calculus". In: *2013 IEEE 13th International Conference on Data Mining Workshops*, pp. 967–970. doi: 10.1109/ICDMW.2013.70.
171. Dadgostar, Porooshat (Dec. 2019). "Antimicrobial Resistance: Implications and Costs". In: *Infection and Drug Resistance* 12, pp. 3903–3910. ISSN: 1178-6973. doi: 10.2147/IDR.S234610. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6929930/>.

172. Singer, Mervyn (May 2017). "Antibiotics for Sepsis: Does Each Hour Really Count, or Is It Incestuous Amplification?" In: *American Journal of Respiratory and Critical Care Medicine* 196.7, pp. 800–802. ISSN: 1073-449X. doi: 10.1164/rccm.201703-0621ED. URL: <https://www.atsjournals.org/doi/full/10.1164/rccm.201703-0621ED>.
173. PHE (Feb. 2020). *Summary of antimicrobial prescribing guidance – managing common infections*. Tech. rep. Public Health England, p. 118. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/906668/Common_Infect_PHE_context_references_and_rationale_August_2020.pdf.
174. NICE (2021). *Antimicrobial stewardship | Topic*. URL: <https://www.nice.org.uk/guidance/health-protection/communicable-diseases/antimicrobial-stewardship>.
175. Alexander, Larry and Moore, Michael (Oct. 2020). "Deontological Ethics". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. URL: <https://plato.stanford.edu/archives/sum2021/entries/ethics-deontological/>.
176. Hursthouse, Rosalind and Pettigrove, Glen (Dec. 2016). "Virtue Ethics". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. URL: <https://plato.stanford.edu/archives/win2018/entries/ethics-virtue/>.
177. Hindocha, Sumeet and Badea, Cosmin (Sept. 2021). "Moral exemplars for the virtuous machine: the clinician's role in ethical artificial intelligence for healthcare". In: *AI and Ethics*. ISSN: 2730-5961. doi: 10.1007/s43681-021-00089-6. URL: <https://doi.org/10.1007/s43681-021-00089-6>.
178. Beauchamp, Tom L and Childress, James F (1979). *Principles of biomedical ethics*. New York: Oxford University Press. ISBN: 978-0-19-502487-6 978-0-19-502488-3.
179. Meyer, Lukas (2020). "Intergenerational Justice". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2020. Metaphysics Research Lab, Stanford University. URL: <https://plato.stanford.edu/archives/sum2020/entries/justice-intergenerational/>.
180. Sakai, Tatsuo and Morimoto, Yuh (Oct. 2022). "The History of Infectious Diseases and Medicine". In: *Pathogens* 11.10, p. 1147. ISSN: 2076-0817. doi: 10.3390/pathogens11101147. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9609744/>.
181. Gillon, R. (July 1994). "Medical ethics: four principles plus attention to scope". In: *BMJ* 309.6948, p. 184. ISSN: 0959-8138, 1468-5833. doi: 10.1136/bmj.309.6948.184. URL: <https://www.bmj.com/content/309/6948/184>.
182. Holm, S. (Dec. 1995). "Not just autonomy—the principles of American biomedical ethics." In: *Journal of Medical Ethics* 21.6, pp. 332–338. ISSN: 0306-6800, 1473-4257. doi: 10.1136/jme.21.6.332. URL: <https://jme.bmj.com/content/21/6/332>.

183. Lawrence, Dana J. (Jan. 2007). "The Four Principles of Biomedical Ethics: A Foundation for Current Bioethical Debate". In: *Journal of Chiropractic Humanities* 14, pp. 34–40. ISSN: 1556-3499. doi: 10.1016/S1556-3499(13)60161-8. URL: <https://www.sciencedirect.com/science/article/pii/S1556349913601618>.
184. Huxtable, Richard (June 2013). "For and against the four principles of biomedical ethics". In: *Clinical Ethics* 8.2-3, pp. 39–43. ISSN: 1477-7509. doi: 10.1177/1477750913486245. URL: <https://doi.org/10.1177/1477750913486245>.
185. Cawley, John (Sept. 2015). "An economy of scales: A selective review of obesity's economic causes, consequences, and solutions". In: *Journal of Health Economics* 43, pp. 244–268. ISSN: 0167-6296. doi: 10.1016/j.jhealeco.2015.03.001. URL: <https://www.sciencedirect.com/science/article/pii/S0167629615000296>.
186. NHS (n.d.). *NHS England Allocations*. URL: <https://www.england.nhs.uk/allocations/>.
187. Barr, Ben, Bambra, Clare, and Whitehead, Margaret (May 2014). "The impact of NHS resource allocation policy on health inequalities in England 2001–11: longitudinal ecological study". In: *BMJ* 348, g3231. ISSN: 1756-1833. doi: 10.1136/bmj.g3231. URL: <https://www.bmj.com/content/348/bmj.g3231>.
188. Hollis, Aidan and Maybarduk, Peter (2015). "Antibiotic Resistance Is a Tragedy of the Commons That Necessitates Global Cooperation". In: *The Journal of Law, Medicine & Ethics: A Journal of the American Society of Law, Medicine & Ethics* 43 Suppl 3, pp. 33–37. ISSN: 1748-720X. doi: 10.1111/jlme.12272.
189. Tarrant, Carolyn et al. (2019). "Optimizing antibiotic prescribing: collective approaches to managing a common-pool resource". In: *Clinical Microbiology and Infection* 25.11, pp. 1356–1363.
190. United Nations University (n.d.). *Duty to future generations*. URL: <https://archive.unu.edu/unupress/unupbooks/uu25ee/uu25ee0m.htm>.
191. Godofsky, Jessica and Winston, M. (2010). "Future generations and the right to survival: A deontological analysis of the moral obligations of present to future people". In: *TCNJ Journal of student scholarship* 12, pp. 1–9.
192. Lysaght, Tamra et al. (Sept. 2019). "AI-Assisted Decision-making in Healthcare". In: *Asian Bioethics Review* 11.3, pp. 299–314. ISSN: 1793-9453. doi: 10.1007/s41649-019-00096-0. URL: <https://doi.org/10.1007/s41649-019-00096-0>.
193. Grote, Thomas and Berens, Philipp (Mar. 2020). "On the ethics of algorithmic decision-making in healthcare". In: *Journal of Medical Ethics* 46.3, pp. 205–211. ISSN: 0306-6800, 1473-4257. doi: 10.1136/medethics-2019-105586. URL: <https://jme.bmj.com/content/46/3/205>.
194. Price II, W. Nicholson, Gerke, Sara, and Cohen, I. Glenn (Nov. 2019). "Potential Liability for Physicians Using Artificial Intelligence". In: *JAMA* 322.18, pp. 1765–1766. ISSN: 0098-7484. doi: 10.1001/jama.2019.15064. URL: <https://doi.org/10.1001/jama.2019.15064>.

195. Barredo Arrieta, Alejandro et al. (June 2020). "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58, pp. 82–115. issn: 1566-2535. doi: 10.1016/j.inffus.2019.12.012. URL: <https://www.sciencedirect.com/science/article/pii/S1566253519308103>.
196. Anderson, Michael and Anderson, Susan Leigh (Dec. 2007). "Machine Ethics: Creating an Ethical Intelligent Agent". In: *AI Magazine* 28.4, pp. 15–15. issn: 2371-9621. doi: 10.1609/aimag.v28i4.2065. URL: <https://ojs.aaai.org/index.php/aimagazine/article/view/2065>.

Chapter 3

Co-morbidity Representation in AI

Abstract

The appropriate representation of data is critical for AI, especially in healthcare, where numerous key challenges exist including data quality, heterogeneity and biases. Representing patient co-morbidities or long-term medical conditions is non-trivial due to their heterogeneity, sparsity and combinatorial complexity. Co-morbidities are an essential piece of clinical information that can drastically alter a patient's treatment, management and outcome. For infectious diseases, they increase an individual's susceptibility, make antibiotic prescriptions more challenging, and increase the risk of complications and poor outcomes.

A novel pipeline was developed to address such pitfalls by utilising the structure of Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), the most extensive clinical vocabulary in the world, to create disease and co-morbid patient embeddings which can be used in downstream AI applications. Through human and automatic evaluation methods for numerous tasks including classification, similar patient retrieval, and antibiotic selection, we show that our method outperforms existing approaches. This research highlights that appropriate extraction and representation of clinical knowledge can advance AI systems that use routinely collected electronic health record (EHR) data such as co-morbidities towards precision antibiotic prescribing.

The work discussed in this chapter was accepted and presented in the Association for the Advancement of Artificial Intelligence 2024 Health Intelligence Workshop (AAAI W3PHIAI-24) and jointly accepted for publication by Springer Nature in a collected work titled 'AI for Health Equity and Fairness: Leveraging AI to Address Social Determinants of Health' (Bolton et al. 2024). It was also presented at the European Society of Clinical Microbiology and Infectious

Diseases (ESCMID) Global 2024 conference. Figure 3.1 shows a graphical abstract of this chapter’s research.

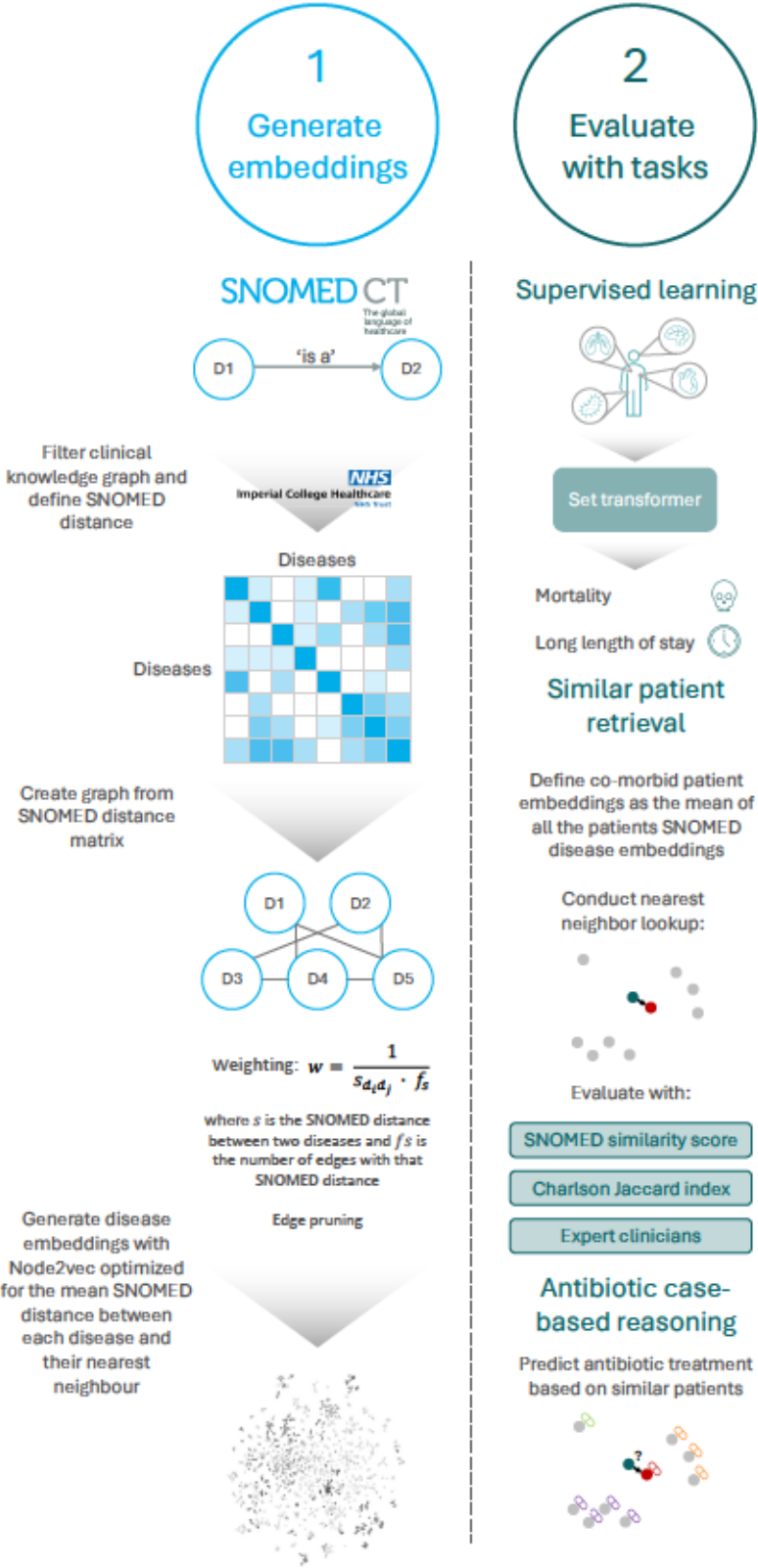


Figure 3.1: Outline of the co-morbidity representation pipeline developed and tasks used for evaluation.

3.1 Chapter objectives

- *Provide an overview of the challenges with using co-morbidity data and existing research on co-morbidity representation*
- *Develop a pipeline to generate disease and co-morbid patient embeddings from the previously untapped fundamental structure of a common clinical coding system*
- *Create a novel patient pair similarity metric*
- *Demonstrate the utility of the embeddings to AI in healthcare through supervised learning, similar patient retrieval, and antibiotic case-based reasoning tasks*

3.2 Introduction

3.2.1 The challenges with co-morbidity representation

Co-morbidities are defined here as chronic long-term medical conditions are a major challenge in healthcare (Gijssen et al. 2001; Riaño et al. 2019). They can drastically alter a patient's treatment plan, are associated with an increased infection and mortality rate, place a huge financial burden on healthcare services, and their prevalence is increasing due to the ageing population (Salive 2013; Divo et al. 2014; Skou et al. 2022). Co-morbidities are a fundamental part of illustrating a patient in both a clinical and an AI setting. However, challenges exist with representing and using co-morbidity data in AI systems. The large number of unique diseases, their sparsity, missingness, and frequent confounding means learning useful co-morbidity relationships even through powerful machine or deep learning methods is difficult. This combined with limited dataset sizes, particularly for those with rare diseases or complex co-morbidity

combinations, can hinder models' generalizability. Furthermore, heterogeneity in how medical information is recorded across different healthcare systems and institutions introduces additional complexities when training and deploying AI models. Developing methods and pipelines to mitigate these problems and appropriately represent co-morbidities is therefore of great importance to maximise the utility of healthcare-focused AI systems.

3.2.2 Current research on co-morbidity representation

Knowledge representation in healthcare is a broad topic that covers the formalisation and organisation of medical information for AI systems (Riaño et al. 2019). For ontologies such as SNOMED CT, AI techniques, including natural language processing, have previously been used to automatically identify diseases from free text (Chang et al. 2021) and predict future patient trajectories (Alaa et al. 2022; Kraljevic et al. 2023; Beaney et al. 2024). Prior work from Choi et al used an attention mechanism to represent medical concepts as a combination of their ancestors in a medical ontology (Choi et al. 2017). We progress this literature by focusing on representing patients and their long-term health conditions through a generalisable pipeline that leverages the architecture of SNOMED CT. Graphs are a way to structure data using nodes and edges (Scarselli et al. 2009). Several methods have been developed to embed graphical information in a vector latent space. Node2vec (Grover et al. 2016), subgraph2vec (Narayanan et al. 2016), and graph2vec (Narayanan et al. 2017) all convert high-dimensional graph-structured data into low-dimensional continuous vector spaces while trying to maintain graphical structure. Node2vec uses random walks across the nodes in a graph, while subgraph2vec and graph2vec use a feature extraction process before embeddings are created. Such methods have previously been applied to medical claim and co-occurrence data (Choi et al. 2016). Knowledge graphs are a type of graph that contain entities as nodes and their relationships as edges. In healthcare, there are many knowledge graphs such as SNOMED CT, which are created by combining a wide spectrum of data including expert consensus, electronic health records (EHRs), guidelines, and published literature (Cui et al. 2023). For example, PrimeKG (Chandak et al. 2023), is a large multimodal knowledge graph that contains connections from diseases to a wide range of

medically relevant factors, including drugs, phenotypes, and genotypes. AI research on such data is diverse, ranging from drug design to clinical decision support (Cui et al. 2023). Pertinent to this chapter, recent research looked into creating embeddings from knowledge graphs with the aim of understanding drug-drug interactions (Yu et al. 2021; Su et al. 2022).

Identifying similar co-morbid patients is useful for case-based reasoning where past patient cases are used to help solve current problems. Simple approaches have defined patient similarity based on their overlapping diagnosis codes (Diaz Ochoa et al. 2022) and the Jaccard index (Hu et al. 2022) when encoded using the one hot methodology where categorical data are represented through binary vectors (0=no, 1=yes). Rocheteau et al. on the other hand, approached the challenges associated with patients' diagnoses data by utilising clinical reasoning to define a patient similarity score and subsequently generating a patient graph for downstream AI applications (Rocheteau et al. 2021a). All of these patient similarity approaches require exact disease matches between patients and hence will suffer from coding quality issues and miss similar patients with slightly different diagnoses. As such, they are not generalisable and are sensitive to data heterogeneity.

3.2.3 Leveraging medically grounded knowledge

Existing AI research on co-morbid patients does not address the problems of combinatorial complexity, sparsity, missingness, heterogeneity and a lack of data, and hence lacks appropriate representation (Rocheteau et al. 2021a; Bolton et al. 2024). In this chapter, we investigate a novel approach that utilises the underlying knowledge and structure contained within a healthcare ontology to generate disease and patient embeddings and demonstrate their utility in downstream AI applications. Specifically, Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) the world's most comprehensive clinical healthcare terminology (Millar 2016) is used. Originally developed by the College of American Pathologists and the National Health Service (NHS), SNOMED CT is widely used around the world in EHRs to represent clinically relevant information. In England, all NHS healthcare providers are required to use SNOMED CT (NHS 2023) for standardised record keeping and seamless information exchange

to improve patient care. SNOMED CT contains a wealth of known clinical information about how diseases are related and interact. By utilising this publicly available, expert-curated knowledge graph, we can create embeddings that contain implicit information on the relationship between diseases. Incorporating external knowledge and creating meaningful embeddings from medically grounded data helps overcome the problems of sparsity and heterogeneity, allowing powerful deep learning methods to be utilised and precision healthcare to become a reality.

3.3 Methods and tasks

3.3.1 Dataset

EHR data from the Imperial College Healthcare NHS Trusts (ICHT) Clinical Analytics, Research and Evaluation (iCARE) environment was used in this research. In particular, the 'problem' table in iCARE was used as it contains data on patients' chronic diseases (i.e., their co-morbidities) and their date of diagnosis. Each time a new co-morbidity was added to a patient's 'problem' list, it was considered a new patient instance, this was done to maximise the dataset size. Those patients with no co-morbidities listed in the 'problem' table were not included in this chapter's work. SNOMED CT International Edition (version released on 31/07/2022) was downloaded from the NHS Digital Technology Reference Update Distribution. The structure of SNOMED CT is highly complex (SNOMED International 2023). Provided below is a high-level overview, so the methodologies employed in this research can be understood. 'Concepts' in SNOMED CT represent a specific piece of clinical information. Concepts are categorised into groups, including 'body structures', 'disorders', 'procedures' and 'pharmaceutical products'. All concepts are organised into a hierarchy where they are connected by 'relationships'. Relationships are also classified into groups by their 'type'. Relationships of type 'is a' are used to define the main hierarchy, for example, 'Diabetes mellitus is a disorder of glucose metabolism'. 'Attribute' relationships, on the other hand, allow for more nuanced connections to be made as needed, such as between a disease and a side effect. In this research, we focus on using the main

hierarchical structure of SNOMED CT ('is a' relationships) for those concepts that could be used to document patients' co-morbidities ('disorders').

3.3.2 Generating SNOMED disease embeddings

SNOMED CT was filtered for concepts classified as 'disorders' and relationships of type 'is a', to capture all diseases, their relationships to each other, and to remove all other information. The 'SNOMED distance' was defined and calculated for each unique disease pair as the shortest path length between the two concepts in the SNOMED CT hierarchy. Patients and their historic co-morbidity diagnoses were then extracted from the ICHT EHR and the SNOMED data was filtered to only those co-morbidities present in the EHR. This filtering was applied as it ensures the disease embeddings are relevant to those patients in the EHR, and from experimentation returned significantly better embeddings than simply applying node2vec directly to the whole hierarchical structure of SNOMED CT. Since diagnoses within the EHR were only used as a filter, our approach is adaptable to any number of diagnoses and different healthcare institutions. Based on the filtered SNOMED data, a connected undirected graph was created where each co-morbidity in the EHR dataset with > 10 edges with a SNOMED distance of ≤ 2 was a node and a weighted edge existed between each node if their SNOMED distance was < 6 . These preprocessing steps were hyperparameters tuned to optimise the metric discussed below and to prevent diseases with few strongly weighted edges connected to them from skewing the resulting embeddings. Edges were weighted (w) based on the following formula:

$$w = \frac{1}{s_{d_i, d_j} \cdot f_s}$$

Where s_{d_i, d_j} is the SNOMED distance between two diseases (d_i and d_j) and f_s is the number of edges with that SNOMED distance. This ensured that random walks were more likely to traverse edges with smaller SNOMED distances compared to the more common longer-range connections. Node2vec (Grover et al. 2016) was then applied to this graph to generate embeddings for each disease, the graphical structure meaning the embeddings were learnt based on the fundamental clinical knowledge and architecture of SNOMED CT. The mean

SNOMED distance between each node and their nearest neighbour was utilised as a metric for hyperparameter optimisation. Minimising this metric enabled us to optimise node2vec as we aimed to create embeddings whereby those diseases nearby in SNOMED CT are close in the embedding space. This preserves the graphical structure of SNOMED and creates meaningful disease embeddings. Optuna was used for hyperparameter optimisation (Akiba et al. 2019). The final hyperparameters for node2vec were an embedding dimension of size 128, $p=0.1$, $q=6.0$, window size=10, 10 walks per node, and a walk length of 1000.

3.3.3 Supervised learning task

SNOMED disease embeddings can be utilized as feature inputs for machine learning and deep learning models, which when trained in a supervised way can make clinically useful predictions. Given a particular patient may have any number of co-morbidities, and that their relationship is unordered, a set transformer (Lee et al. 2019) was chosen as an appropriate permutation invariant attention-based neural network model. We undertook two classification prediction tasks, firstly for in-hospital mortality within a year (assuming no loss to follow-up) and secondly for a long in-patient length of stay (LOS), defined as > 7 days. A Long LOS was defined as this to provide a reasonable class balance and because the average LOS in the NHS is around 7 days (The Health Foundation 2023). For both tasks only patients' historical co-morbidity information was used as features, no information from a patient's stay or any acute conditions were included to ensure we directly evaluated the utility of our generated embeddings. Such predictive models could be useful for planning and preventative medicine before a patient enters a hospital. Area Under the Receiver Operating Characteristic curve (AUROC, formula in the Appendix Section A.1) results were generated for the whole population and those without the 75% most common co-morbidities using 10-fold stratified cross-validation. Stratification preserves the percentage of samples with each label across dataset folds, while cross-validation ensures a robust performance estimation. We compared our method to logistic regression with patients' co-morbidities encoded using the traditional one-hot methodology, logistic regression based on patients' Charlson co-morbidity categories, and a set transformer

with random SNOMED disease embeddings. Logistic regression was implemented with Scikit-learn (a machine learning library for the Python programming language, Pedregosa et al. 2011), and the 'class_weight' parameter set to balanced. This model was chosen as a baseline because it was the best-performing Scikit-learn model on the one-hot encoded data. Note that passing this data format into a set transformer was computationally intractable in iCARE due to the matrix size. The presence of 17 Charlson co-morbidity categories (Charlson et al. 1987; Deyo et al. 1992) was determined for each patient using the Unified Medical Language System (UMLS) metathesaurus International Classification of Diseases (ICD) to SNOMED map (Unified Medical Language System 2023). Random SNOMED disease embeddings were created using NumPy's (Harris et al. 2020) random number generator, and constrained to within the maximum and minimum values of the original embeddings. While the information within the embedding was randomized, each disease was still associated with a particular embedding. This removes the disease location and relationship information and enables the capability of our generated embeddings to be evaluated. Our set transformer architecture included a single 'Induced Set Attention Block' followed by a 'Pooling by Multihead Attention' operation (Lee et al. 2019) and finally a simple linear decoder layer. It was created using Pytorch (a machine learning library for the Python programming language, Paszke et al. 2019) and had 4 heads with a hidden dimension of size 160. Hyperparameter optimisation of the set transformer is tractable but was not conducted due to computational limits within the iCARE environment and because the objective of these tasks is to show our embeddings can be informative features rather than to create the best prediction model. For training a batch size of 512, the Adam optimiser (Kingma et al. 2014) and binary cross entropy with logits loss were utilized. Statistical analysis used an alpha of 0.05 with a paired t-test or the Wilcoxon signed-rank test to determine if groups came from the same population, depending on the normality of the data distribution, which was determined using the Shapiro-Wilk test.

3.3.4 Similar patient retrieval and case-based reasoning tasks

Co-morbid patient embeddings were intuitively defined as the mean of all the SNOMED disease embeddings for that particular individual. This then allows for similar patients to be retrieved through a simple nearest neighbour lookup (Maneewongvatana et al. 1999) based on the Euclidean distance. We compared our method to nearest neighbour lookup with traditional one hot encodings, and the patient similarity score from Rocheteau et al. (Rocheteau et al. 2021a). In this research, they create a diagnosis matrix $D \in \mathbb{R}^{N \times m}$ where N is the number of patients and m is the amount of unique diagnoses. The similarity score M_{kl} between patients k and l is then defined as:

$$M_{kl} = a \sum_{\mu=1}^m (D_{k\mu} D_{l\mu} (d_{\mu}^{-1} + c)) - \sum_{\mu=1}^m (D_{k\mu} + D_{l\mu})$$

Where d_{μ} is the frequency of diagnosis μ , and a and c are hyperparameters. Through grid search, $c = 0.5$, and $a = 5$ returned the best results in our data set. This method rewards shared diseases, in particular those that are rare, and penalises higher numbers of co-morbidities. The most closely matched patient is identified through having the greatest score.

A similar patient retrieval task was defined as finding the most comparable patient within the EHR for any given unique co-morbid patient, where no identical match was possible. This tests the methods in difficult circumstances where similarity is a more meaningful and important clinical question, as opposed to simple cases where similarity is more trivial. This task was performed on 1000 randomly selected patients due to time and computational limitations within the iCARE environment. Determining whether patients retrieved using similarity methods are comparable is challenging (Sharafoddini et al. 2017). To evaluate this task, we propose a novel metric called ‘SNOMED similarity’. To determine SNOMED similarity for a particular patient pair $(p1, p2)$, their co-morbidities and respective SNOMED distances are used to create a distance matrix $(S_{p1,p2})$. Specifically, within $S_{p1,p2}$ the i^{th} and j^{th} entry is $s_{d_i^{p1}, d_j^{p2}}$, i.e., the SNOMED distance between $p1$ ’s i^{th} disease and $p2$ ’s j^{th} disease. We then match each disease of $p1$ with the diseases of $p2$ so that the matching minimised this equation,

where $A \in \mathbb{R}^{n \times m}$:

$$f(A) = \sum_{i=1}^n \min_{j \in \{1, \dots, m\}} (1 - \frac{1}{A_{ij} + 1})$$

The same operation was applied to p_2 , and the resulting numbers were summed to give the final SNOMED similarity score:

$$\text{sim}_{p_1, p_2} = f(S_{p_1, p_2}) + f(S_{p_2, p_1})$$

This methodology ensures our evaluation metric returns a single score where identical diseases result in 0 while more dissimilar co-morbidities add to the score proportionally. Furthermore, this approach is permutation invariant and by applying the operation to each individual, a mismatch in the number of co-morbidities will serve as a penalty. As such, patients with similar diseases will return low SNOMED similarity scores, while those that are dissimilar will return higher scores. Given the potential bias of using the SNOMED distance in our pipeline and assessment, we defined the Charlson-Jaccard index as another evaluation metric. As before Charlson co-morbidity categories (Charlson et al. 1987; Deyo et al. 1992) were determined for each patient through the UMLS metathesaurus ICD to SNOMED map (Unified Medical Language System 2023). The Jaccard index between the categories for the particular patient in question and the similar patient identified by each method was then calculated, where a higher score indicates a more similar match. By grouping multiple diagnostic codes into categories, this approach overcomes the exact disease match limitation of a straightforward Jaccard index on the one hot encodings. However, a drawback is that the Charlson co-morbidity index (Charlson et al. 1987; Deyo et al. 1992) does not cover all co-morbidities and hence some information is discarded. Additional analysis was conducted to understand how robust the methods are to variation in the degree of multi-morbidity and the rarity of patients' conditions.

To complement the automatic nature of the SNOMED similarity score and Charlson-Jaccard index, we employed expert human evaluation for a random subset of these unique co-morbid patient examples. More specifically, a 10-question multiple choice survey was created (Supplementary Table A.1). They were requested to select the patient (A, B, or C) they believed was the most similar to the patient in the question with regards to their co-morbidities.

Each answer represented a patient identified as most similar by the three methodologies, allowing us to evaluate the method that clinicians perceived as the most appropriate in those scenarios. The questions were presented in plain text format, no information on the methods was included, and their positioning was changed for each question to avoid introducing bias (example shown in the Supplementary Figure A.1). In two cases (questions 4 and 10) one hot encodings and Rocheteau's similarity score returned the same patient, and so these questions were presented as a binary choice to avoid duplication. Clinicians were recruited through email and additional optional free text questions were presented after each patient similarity question and at the end of the survey, to understand how they made their decisions, their definition of patient similarity, and their medical speciality. Beyond the analysis of the votes and free text responses, the Chi-Square test of independence ($\alpha = 0.05$) was used to determine whether the distribution of votes was significantly different between medical specialities. Furthermore, additional investigations were conducted to understand the variability in choices through the Gini index, clustering of K modes (Huang 1998), and multiple correspondence analysis (Greenacre et al. 2006) as a dimensionality reduction technique for plotting.

Finally, matching similar patients was extended to a case-based reasoning task, where new problems are solved using the solutions of similar problems in the past. In this task, for each patient (problem), the empiric antimicrobial treatment prescribed to their most closely matched patient (solution) was defined as their predicted antibiotic and class of antibiotic. These predictions were then compared with the real antimicrobial prescription for that patient with exact match accuracy used for evaluation. This task was run for the most common antibiotics (co-amoxiclav and ceftriaxone) and their classes (penicillins and cephalosporins) for those patients with diabetes and pneumonia to determine an appropriate empiric antimicrobial treatment. These antibiotics and co-morbidities were chosen to restrict the scope of the task and demonstrate that case-based reasoning is possible with this methodological approach. Including all possibilities would yield random results given the primary drivers behind the choice of initial antimicrobial agent is the patients' suspected diagnosis rather than their co-morbidities.

3.4 Results

3.4.1 SNOMED disease embeddings

2,133 diseases from 95,157 co-morbid patient instances were extracted from the EHR. The most prevalent disease was hypertension followed by asthma, varicella (historical diagnosis) and diabetes mellitus type 2. Figure 3.2 shows the frequency distribution for diseases and a histogram of the number of patient co-morbidities. Both show an exponentially decreasing trend, highlighting that combinatorial complexity makes using co-morbidity information in AI difficult. When generating SNOMED disease embeddings, node2vec optimisation returned a mean SNOMED distance between each node and their nearest neighbour of 1.23. Analysis of t-SNE visualisation (Figure 3.3) with disease groups highlighted demonstrates that related conditions are located close in the embedding space. For example, viral and bacterial infectious diseases are clustered together, and heart diseases are adjacent to hypertensive disorders.

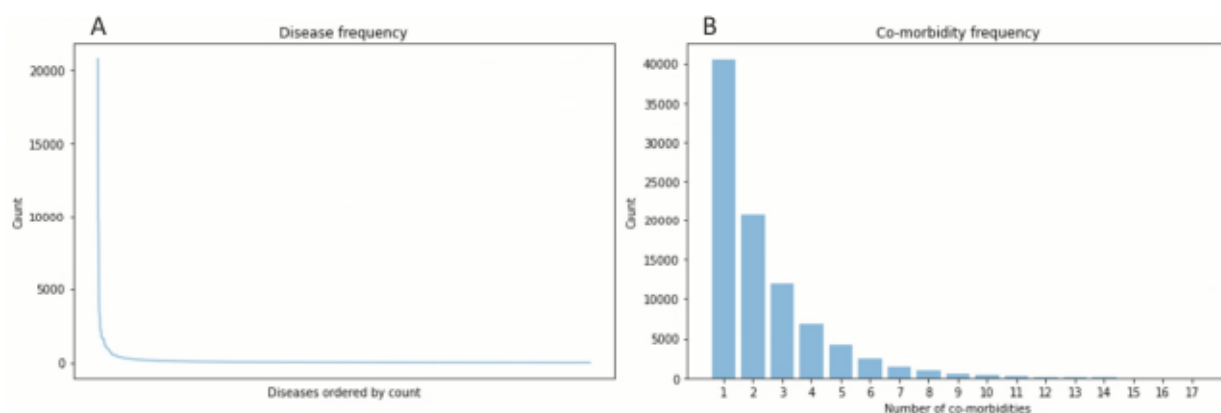


Figure 3.2: [A] The frequency distribution for historical diseases in the EHR data. [B] Histogram displaying the number of patient co-morbidities. The exponential decay on both graphs indicates the complexity of using co-morbidities in AI. Note to preserve privacy for both plots only those with counts greater than 5 were included.

3.4.2 Supervised learning task

The results of the supervised learning tasks are shown in Table 3.1. Our method of utilising SNOMED disease embeddings obtained the highest mean AUROC in all scenarios. Charlson

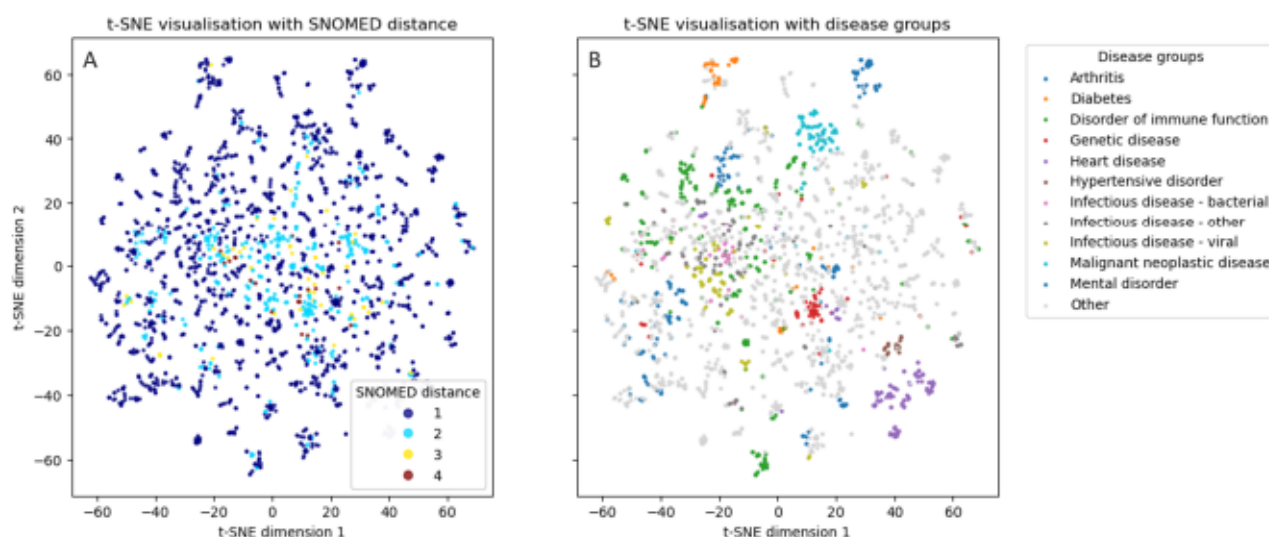


Figure 3.3: t-SNE visualisations of the SNOMED disease embeddings with [A] the nearest neighbour SNOMED distance displayed and [B] high-level disease groups highlighted. Most diseases are embedded next to their neighbours in SNOMED CT, with related disease groups clustered together.

co-morbidity categories contain less nuanced disease information and so as expected achieved the lowest AUROC results, with random performance on those with rare co-morbidities that do not fall into one of the 17 Charlson co-morbidity categories (Charlson et al. 1987; Deyo et al. 1992). Statistically significant differences (all p-values <0.03) were observed between all methods on the long LOS prediction task for the overall population using the Wilcoxon Signed-rank test. For mortality classification, a paired t-test only indicated statistical differences between the Charlson co-morbidity categories and all other methods (all p-values <0.01), and the SNOMED disease embeddings and one hot encodings (p-value <0.01). The relationship between co-morbidities and death is firmly established, and hence the superior performance on the mortality prediction task throughout is expected. Length of stay prediction is a much more nuanced task (Rocheteau et al. 2021b) and would probably require additional data regarding the patient's stay to see further improvements. Both supervised tasks were challenging with a large label imbalance. The year mortality rate was 2.53% and only 5.80% of the patients had a long LOS. Overall, these results show that our SNOMED disease embedding-set transformer method is generalisable and has utility for clinically relevant predictions.

Table 3.1: Mean unseen test set AUROC results for classification tasks in different populations.

Features	Year mortality		Long length of stay	
	Overall	Rarest co-morbidities	Overall	Rarest co-morbidities
Charlson co-morbidity categories	0.65 (SD 0.01)	0.50 (SD <0.01)	0.60 (SD 0.01)	0.50 (SD 0.03)
One hot encodings	0.79 (SD 0.02)	0.80 (SD 0.23)	0.72 (SD 0.01)	0.55 (SD 0.11)
Random SNOMED disease embeddings	0.80 (SD 0.03)	0.56 (SD 0.33)	0.74 (SD 0.02)	0.52 (SD 0.23)
SNOMED disease embeddings	0.82 (SD 0.02)	0.85 (SD 0.14)	0.75 (SD 0.01)	0.61 (SD 0.20)

3.4.3 Similar patient retrieval task

Using co-morbid patient embeddings to retrieve similar examples resulted in a mean SNOMED similarity score of 1.78 (SD 1.90) and a Charlson-Jaccard index of 0.84 (SD 0.34). One hot encodings obtained a mean score and index of 4.40 (SD 2.32) and 0.88 (SD 0.30), respectively, while Rocheteau's method achieved 3.52 (SD 3.26) and 0.69 (SD 0.20), respectively. Statistically significant results (all p-values <0.01) from the Wilcoxon Signed-rank test (alpha = 0.05) indicated that for both metrics, all the methods come from different populations. In addition, our method was over 10^4 times faster (averaged 0.01 s) in finding the patient most closely matched versus using Rocheteau's score or one hot encodings. The distribution of SNOMED similarity scores is shown in Figure 3.4. Our method returned more similar patients and showed a narrower distribution. The results based on patients with different numbers of co-morbidities and the rarity of their co-morbidities are shown in Figure 3.5. Co-morbid patient embeddings demonstrate consistent performance across patients with different degrees of rarity and multi-morbidity. In particular, our method is robust as the patient's number of co-morbidities increases, while one hot encodings see a linear increase in SNOMED similarity score. For the mean Charlson-Jaccard index metric, the co-morbid patient embeddings and one hot encodings show superior coincident behaviour compared to Rocheteau's method. Note that this metric is not suitable for those with only very rare co-morbidities (right side of Figure 3.5 [D]) as the lack of diseases that fit the Charlson categories causes the scores to trend towards 1.

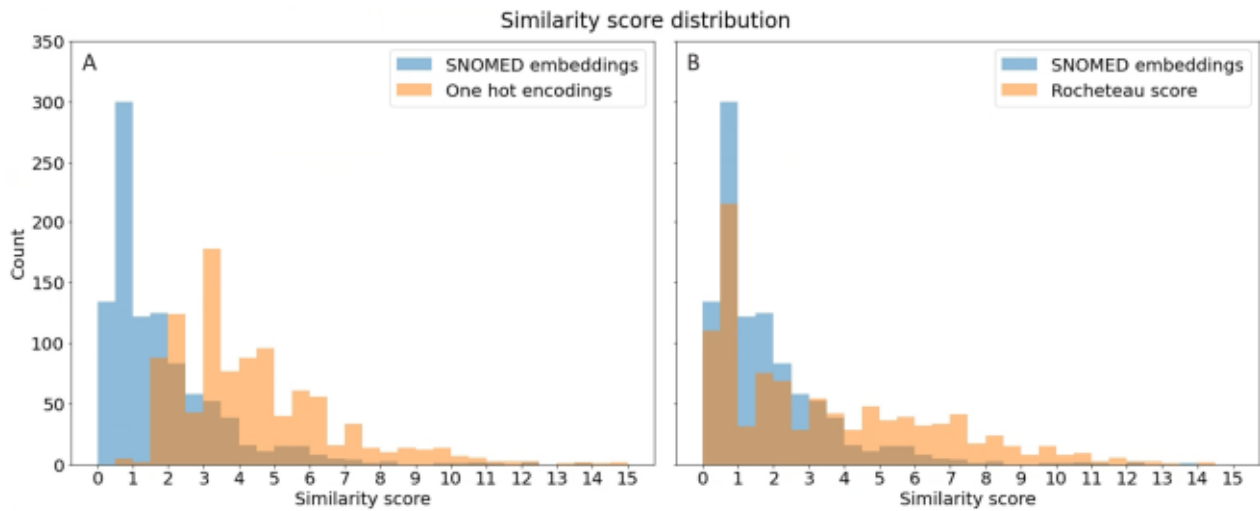


Figure 3.4: Distribution of SNOMED similarity scores from the similar patient retrieval task. [A] compares our co-morbid patient embeddings method to one hot encodings while [B] compares it to Rocheteau’s score. Co-morbid patient embeddings returns patients with smaller scores.

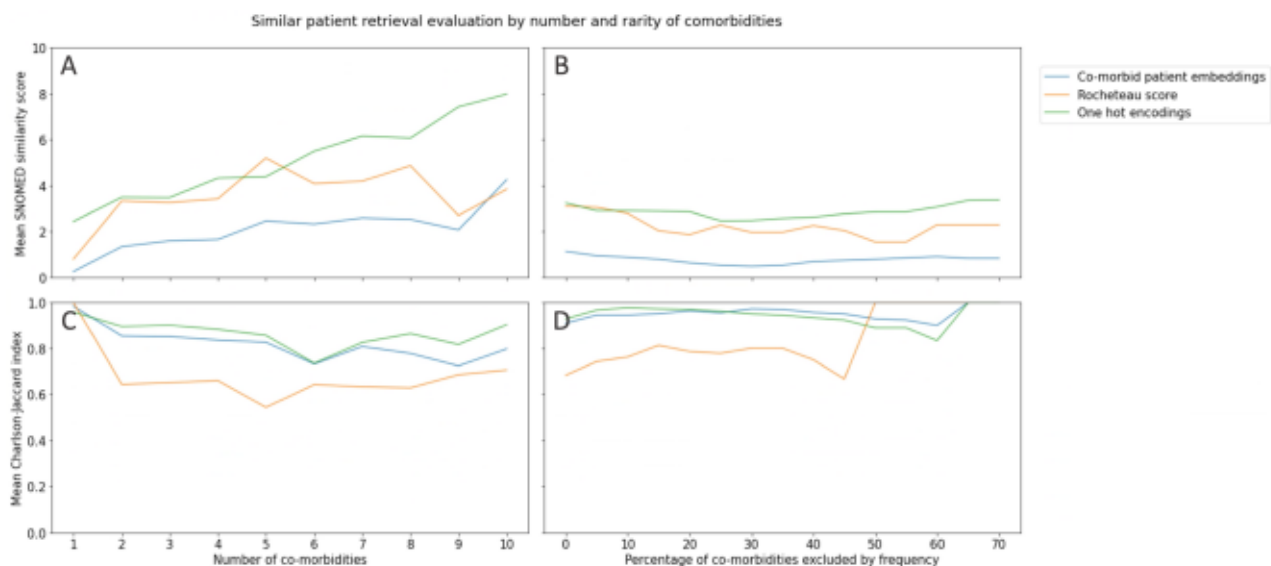


Figure 3.5: Similar patient retrieval task evaluation by SNOMED similarity score [A, B] and Charlson-Jaccard index [C, D] for patients with different numbers of co-morbidities [A, C] and those with co-morbidities of varying rarity [B, D] (co-morbidities were ordered based on their frequency in the EHR, and then starting from the most common, diseases and patients with those conditions were excluded based on percentages, as such, patients with all co-morbidities are included on the left-hand side while only those with rare diseases are represented on the right-hand side).

For human expert evaluation, 20 clinicians were recruited. The distribution of responses for each question in the multi-morbid patient similarity survey is shown in Figure 3.6. Patients identified by our method were selected as the most similar in 60% of the questions and received the majority of the total votes (53%). In the remaining 4 questions, the patients identified by one hot encodings were preferred. For the 6 questions the co-morbid patient embedding method won, it received on average 68% (SD 16.93) of the votes and had a mean winning margin of 40% (SD 32.09). On the other hand, one hot encodings was victorious by a smaller margin of 20% (SD 10.80) with fewer votes (mean 55%, SD 9.13). This method saw a highly variable number of votes across questions. For 4 questions, it received $\leq 10\%$ of the votes. Our method utilising co-morbid patient embeddings though, never got $\leq 15\%$. This consistency correlates well with the distribution of the SNOMED similarity scores shown in Figure 3.4. In the survey, 8 medical specialities were represented, from infectious diseases to general practitioners and internal medicine, but no statistically significant differences were observed between these groups, which is likely due to a limited sample size. The results from the automatic SNOMED similarity score and expert human evaluators are aligned and demonstrate the ability of our method to identify analogous co-morbid patients.

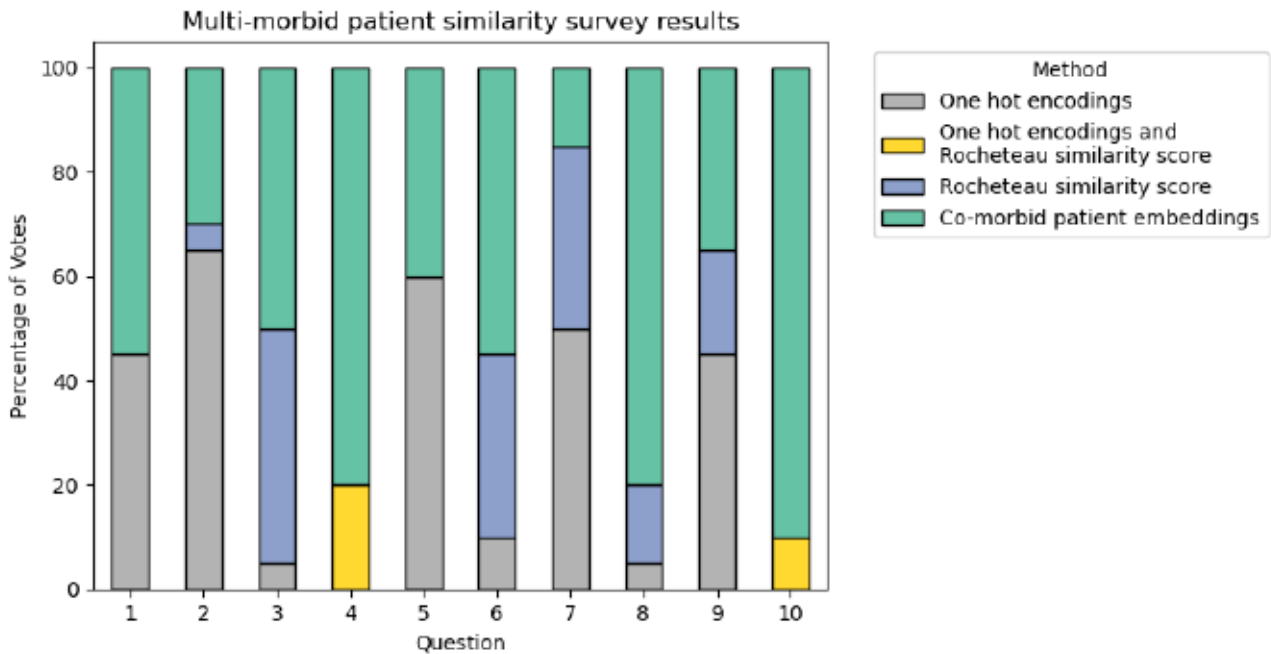


Figure 3.6: Stacked bar plot showing the proportion of votes for patients identified by each method, for each question, in the survey. Co-morbid patient embeddings obtained the most votes for 6 questions with one hot encodings prevailing in the others.

Variation in clinical responses and decision-making is expected (Bate et al. 2012). In this survey, reasonable heterogeneity was observed with a mean Gini index across questions of 0.46 (SD 0.14). The elbow method (Supplementary Figure A.2) was applied to K-modes clustering (Huang 1998) to find the most meaningful number of clusters within the data. 2 clusters were identified as the optimum. A multiple correspondence analysis (Greenacre et al. 2006) plot of these clusters is shown in Figure 3.7.

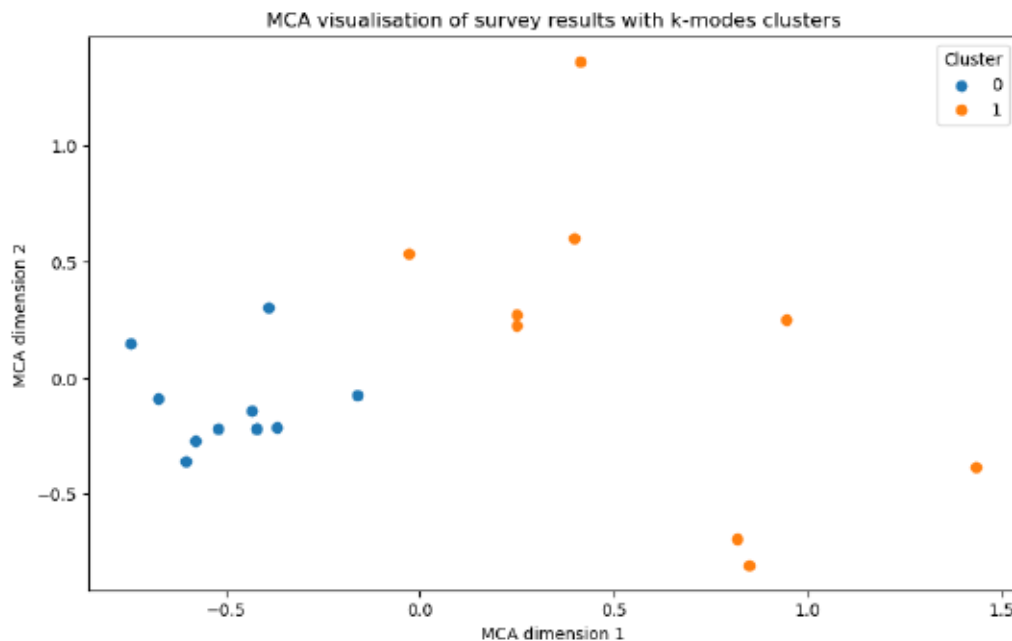


Figure 3.7: Multiple correspondence analysis visualisations of the survey results with k-mode clusters highlighted. Notice how those in cluster 0 are more closely located implying agreement in the responses. In particular, 8 respondents are highly clustered (note that 2 respondents were identical and hence overlap on this plot). Cluster 1 on the other hand is more heterogeneous.

Free text responses from clinicians were examined. When asked to describe how and why they made their decisions a few key themes emerged, beyond the obvious matching of similar conditions. In particular, 7 respondents noted that certain conditions, such as diabetes, cardiovascular and those that are more chronic, are of high priority, while others are not required to be present in both patients. One respondent stated 'some diagnoses are core, and influence many aspects of care (diabetes) so must be included', while another said they 'deprioritized 'benign' co-morbidities such as gastroesophageal reflux disease'. Furthermore, 6 participants mentioned that it is crucial to include the diseases that are most life-threatening. Responding to "How would you define patient similarity?", 10 participants mentioned that

similar co-morbidities are a requirement as investigated in this research. However, other important factors were also mentioned including demographics (age, sex, and ethnicity), as well as circumstance and contextual considerations.

3.4.4 Case-based reasoning antibiotic prescription task

Assigning antibiotic prescriptions through similar patients retrieved via co-morbid patient embeddings achieved an accuracy of 0.78 and 0.79 for specific antibiotics (co-amoxiclav and ceftriaxone) and antibiotic classes (penicillins and cephalosporins) respectively (Table 3.2). Rocheteau’s method obtained 0.80 and 0.78 respectively, and one hot encodings got 0.75 in both scenarios.

Table 3.2: Accuracy results for the case-based reasoning antibiotic prescription task.

Method	Antibiotic	Antibiotic class
One hot encodings	0.75	0.75
Rocheteau’s similarity score	0.80	0.78
Co-morbid patient embeddings	0.78	0.79

3.5 Discussion

Co-morbidity data can be hard to use in AI systems. By supplementing data with external medically grounded knowledge and creating meaningful embeddings, our approach overcomes some of the existing challenges. For example, the embeddings are not influenced by dataset size, the number of diseases or their rareness, and they are adaptable to variation in clinical documentation when different SNOMED disorder codes are used to record the same chronic condition (e.g., Diabetes versus Diabetes Type 1 or 2), as these diseases will be closely located in the embedding space. Furthermore, by exploiting a well-adopted common clinical coding language, our approach is inherently generalisable, and we mitigate bias concerns (Gerke et al. 2020) and integration issues which are common in healthcare due to systems lacking interoperability. EHR data is not required beyond filtering for diseases of interest, overcoming limitations frequently seen with such data including a lack of examples of rare diseases and highly multi-morbid patients. Furthermore, generating SNOMED disease embeddings using our method is computationally efficient and only needs to be run once. Unlike the graph-based attention method (Choi et al. 2017), our embeddings are task-agnostic and can be applied across various settings without requiring re-training or labelled data. By addressing the problem of co-morbidity representation in AI systems, we envisage that our approach could be useful to AI applications that use case-based reasoning or decision support to provide precision care.

The tasks were designed to cover common challenges seen in the AI healthcare domain, including supervised predictions of patient outcomes, patient matching and case-based reasoning. In the supervised learning task the difference in performance between the SNOMED disease embeddings and the random embeddings on the rarest co-morbidities demonstrates that the positional information of our embeddings is useful and enables the model to effectively approximate those co-morbidity instances with a low frequency in the training data. The random embeddings performed reasonably well when evaluated on the overall dataset, indicating that the set transformer can learn useful relationships between diseases and a label even if the information within each item in the set is uninformative. The results of the one

hot encodings were good considering the simplicity of the approach. However, its behaviour would likely not scale with an increasing number of diseases (SNOMED CT contains > 70,000 'disorders' (Lyudovik et al. 2019)). For the similar patient retrieval task, one hot encodings performed well on questions 2 and 5. In question 2 the other methods identified patients that were overly complex. Question 5 was a scenario involving a patient with many co-morbid conditions (10), where two participants who voted for one hot encodings said that none of the patients shown were very similar. Patients from Rocheteau's similarity score were selected the least overall but performed well in cases where the patient had many long-term conditions. Co-morbid patient embeddings was selected by the majority of participants in cases ranging in complexity from 2 co-morbidities (question 10) to highly multi-morbid patients with 8 diseases (question 6). In the two binary choice questions where one hot encodings and Rocheteau's similarity score returned the same patient (questions 4 and 10) co-morbid patient embeddings performed particularly well obtaining 80% and 90% of the votes respectively. Upon clustering, results showed that the 13 participants in cluster 0 were more concurrent, while those in group 1 showed a greater variation in their responses. In particular, 8 clinicians in cluster 0 showed considerable agreement in their votes, which aligned with the overall results discussed above. This investigation shows that although there is heterogeneity in opinions for this type of clinical problem, appropriate research methodologies can be used to obtain a consensus and understand respondents' differences. The results for the case-based reasoning antibiotic prescription task were similar between our method and Rocheteau's score, with both outperforming the more simple and strict matching of one hot encodings. Demonstrating that novel methods that consider subtle differences between patients can enhance case-based reasoning. Improvements are likely needed for a real-world decision support system, but the task still serves as a proof of concept for how such technology can identify similar patients, enable case-based reasoning, and support antibiotic prescribing recommendations. Overall, the results from all tasks highlight the potential benefit of our approach, pipeline and embeddings for downstream AI applications that utilise patient co-morbidity data.

Co-morbidities									
Question 8 patient	Gestational diabetes mellitus	Hypertensive disorder	Pre-eclampsia	Varicella					
Co-morbid patient embeddings	Gestational diabetes mellitus	Pregnancy-induced hypertension	Pre-eclampsia	Varicella					
Rocheteau score	Gestational diabetes mellitus	Hypertensive disorder	-	Varicella					
One hot encodings	Gestational diabetes mellitus	-	Pre-eclampsia	Varicella					
Question 9 patient	Asthma	Hypertensive disorder	Osteoarthritis	Type 2 diabetes	Hypercholesterolemia	Anemia	Gastroesophageal reflux disease	Hypothyroidism	
Co-morbid patient embeddings	Asthma	Hypertensive disorder	Osteoarthritis	Diabetes	Hyperlipidemia	Anemia	Gastroesophageal reflux disease	Obstructive sleep apnea	
Rocheteau score	Asthma	Hypertensive disorder	Rheumatoid arthritis	Diabetes	Hypercholesterolemia	Anemia	Gastroesophageal reflux disease	Hypothyroidism	Coronary arteriosclerosis Pulmonary embolism Chronic kidney disease
One hot encodings	Asthma	Hypertensive disorder	Osteoarthritis	Type 2 diabetes	Hypercholesterolemia				
Question 10 patient	Osteoarthritis	Alcoholism							
Co-morbid patient embeddings	Osteoarthritis	Alcohol dependence							
Rocheteau score	Osteoarthritis	Alcoholism	Peripheral nerve entrapment						
One hot encodings	Osteoarthritis	Alcoholism	Peripheral nerve entrapment						

Figure 3.8: Three survey examples of the patient in question and the similar patients retrieved by each method. The method that obtained the most votes for each question is shown in bold, with the similarity of co-morbidities indicated through a traffic light colouring scheme. Co-morbid patient embeddings identifies clinically appropriate examples with analogous diseases (e.g., alcoholism and alcohol dependence in question 10).

Three patient similarity cases from the survey are shown in Figure 3.8. Question 8 serves as an example of the utility of our method as it is able to discern that ‘Pregnancy-induced hypertension’ is comparable to ‘Hypertensive disorder’ especially given the wider context of the patient in question, who had a number of pregnancy-related complications. This enables a highly comparable patient to be returned, while the other methods exact matching approach means nuances are missed, and hence compromises between co-morbidities must be made. For question 9 our method retrieves a patient whose main clinical difference is a lack of hypothyroidism. The most similar patient obtained using Rocheteau’s method also

has many overlapping conditions, but these are alongside other diseases. For example, they have rheumatoid arthritis which is associated with a different treatment to osteoarthritis and their other additional co-morbidities places them at a higher risk of other conditions, such as cardiovascular disease. The more simplistic matching of the one hot encoding method obtained the most votes for this question, highlighting the variability in clinicians' opinions and expectations. Finally, in Question 10 our method is able to find an almost identically matched patient, despite the different coding terms used. However, other methods fail to account for the similarity between alcoholism and alcohol dependence and hence identify a more chronically unwell patient. These examples illustrate the advantage of co-morbid patient embeddings in accounting for nuanced differences between patients' diseases. Alongside results from the antibiotic prescription task, this indicates that our similar patient retrieval methodology could be useful to healthcare case-based reasoning technologies, by enabling more nuanced and relevant patients to be identified, particularly for cases with significant multi-morbidity or rarity.

The main limitation of this study is that we did not include additional patient information in our tasks. This was done deliberately to highlight how informative co-morbidity information can be through our methodological approach, rather than to create the best prediction model. However, as mentioned by clinicians, many other important factors must be taken into account when making clinical decisions and therefore developing and evaluating AI systems in healthcare. Given the generalisable nature of our methodology, future work should seek to address this by embedding additional clinically relevant information, such as demographics, medications, and clinical variables, while ensuring fairness. Furthermore, the SNOMED embeddings themselves could potentially be improved by considering that edges should have different magnitudes at different depths of the SNOMED CT hierarchy (Wu et al. 1994). Another limitation of this work is that both our embedding method and the evaluation metric for the similar patient retrieval task use the SNOMED distance. This introduces a potential bias towards our approach, as although the methodologies are completely independent, both use the underlying distances between disorders in SNOMED CT. This metric was developed and used as traditional approaches towards evaluating if patients are comparable such as the Jaccard

similarity, require exact disease matches which, as discussed, is not appropriate. To account for this potential bias, we conducted an additional evaluation with the Charlson-Jaccard index and expert human evaluators on a subset. An additional limitation of the similar patient retrieval task and survey is the sample size. For the supervised learning task, one drawback of our approach is that set transformers using our SNOMED disease embeddings are much greater in complexity than logistic regression, meaning they require greater computational resources and are potentially less interpretable. Future work includes representing patients with more information than just the mean of their diseases. For example, by weighting conditions according to the Charlson co-morbidity index (Charlson et al. 1987; Deyo et al. 1992), to ensure that those co-morbidities considered a high-priority are suitably accounted for, or by using more sophisticated methods such as a set transformer variational autoencoder to prevent information such as the patients' number of co-morbidities from being lost. Moreover, due to no data being present in iCARE on if or when patients recovered from their chronic conditions, this research does not consider the temporal evolution of a patient's co-morbid conditions, but this is an avenue for future work. From free text survey responses, it is clear that clinicians believe patient similarity can be useful to help inform treatment and predict future risks. However, given the diversity of factors reported as important by participants, it is clear this is a complex and multi-factorial task with more work needed for a gold standard definition. In particular, future work should focus on matching those patients with similar high-priority and life-threatening conditions. These approaches can then be extended to support further tasks exploring case-based reasoning for antibiotic optimisation. Methodologies and research on clinically informed and data-driven strategies, as employed in this study, progresses work on the representation of healthcare data, with the aim of supporting AI applications that will ultimately benefit patients.

3.6 Conclusion

This chapter describes a novel pipeline to extract and utilise the information contained within SNOMED CT to represent co-morbidities, multi-morbid patients, and support healthcare-focused AI systems. We demonstrated the utility of our approach in classification, similar

patient retrieval, and case-based reasoning tasks with automatic and human evaluation techniques. Our approach leverages medical knowledge that was previously untapped within the AI research community and is able to overcome many of the problems associated with using disease data in AI, such as rarity and heterogeneity, through creating standardised, informative and clinically grounded embeddings that allow for generalisation across both patients and their diseases. Our method advances medical knowledge representation and can support AI applications that involve co-morbidity data, such as those combating antimicrobial resistance, with the ultimate aim of advancing such systems towards clinical benefit. This research highlights that by appropriately utilising routinely collected patient co-morbidity data, AI-driven systems can support optimal antibiotic prescribing in the diverse range of patients seen in real clinical practice.

References

1. Bolton, William J. et al. (2024). "Co-morbidity Representation in Artificial Intelligence: Tapping into Unused Clinical Knowledge". In: *AI for Health Equity and Fairness: Leveraging AI to Address Social Determinants of Health*. Ed. by Arash Shaban-Nejad, Martin Michalowski, and Simone Bianco. Cham: Springer Nature Switzerland, pp. 173–196. ISBN: 978-3-031-63592-2. doi: 10.1007/978-3-031-63592-2_14. URL: https://doi.org/10.1007/978-3-031-63592-2_14.
2. Gijzen, Ronald et al. (July 2001). "Causes and consequences of comorbidity: A review". In: *Journal of Clinical Epidemiology* 54.7, pp. 661–674. ISSN: 0895-4356. doi: 10.1016/S0895-4356(00)00363-2.
3. Riaño, David, Peleg, Mor, and Teije, Annette ten (Sept. 2019). "Ten years of knowledge representation for health care (2009–2018): Topics, trends, and challenges". In: *Artificial Intelligence in Medicine* 100, p. 101713. ISSN: 0933-3657. doi: 10.1016/j.artmed.2019.101713.
4. Salive, Marcel E. (2013). "Multimorbidity in older adults". In: *Epidemiologic Reviews* 35, pp. 75–83. ISSN: 1478-6729. doi: 10.1093/epirev/mxs009.
5. Divo, Miguel J., Martinez, Carlos H., and Mannino, David M. (Oct. 2014). "Ageing and the epidemiology of multimorbidity". In: *European Respiratory Journal* 44.4, pp. 1055–1068. ISSN: 0903-1936, 1399-3003. doi: 10.1183/09031936.00059814.
6. Skou, Søren T. et al. (July 2022). "Multimorbidity". In: *Nature Reviews Disease Primers* 8.1, pp. 1–22. ISSN: 2056-676X. doi: 10.1038/s41572-022-00376-4.
7. Chang, Eunsuk and Mostafa, Javed (Sept. 2021). "The use of SNOMED CT, 2013–2020: a literature review". In: *Journal of the American Medical Informatics Association* 28.9, pp. 2017–2026. ISSN: 1527-974X. doi: 10.1093/jamia/ocab084.
8. Alaa, Asem, Mayer, Erik, and Barahona, Mauricio (Dec. 2022). "ICE-NODE: Integration of Clinical Embeddings with Neural Ordinary Differential Equations". In: *Proceedings of the 7th Machine Learning for Healthcare Conference*. PMLR, pp. 537–564. URL: <https://proceedings.mlr.press/v182/alaa22a.html>.
9. Kraljevic, Zeljko et al. (Jan. 2023). *Foresight – Generative Pretrained Transformer (GPT) for Modelling of Patient Timelines using EHRs*. doi: 10.48550/arXiv.2212.08072. URL: <http://arxiv.org/abs/2212.08072>.
10. Beaney, Thomas et al. (July 2024). "Comparing natural language processing representations of coded disease sequences for prediction in electronic health records". In: *Journal of the American Medical Informatics Association* 31.7, pp. 1451–1462. ISSN: 1527-974X. doi: 10.1093/jamia/ocae091. URL: <https://doi.org/10.1093/jamia/ocae091>.

11. Choi, Edward et al. (Aug. 2017). "GRAM: Graph-based Attention Model for Healthcare Representation Learning". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '17. New York, NY, USA: Association for Computing Machinery, pp. 787–795. ISBN: 978-1-4503-4887-4. DOI: 10.1145/3097983.3098126. URL: <https://dl.acm.org/doi/10.1145/3097983.3098126>.
12. Scarselli, Franco et al. (Jan. 2009). "The Graph Neural Network Model". In: *IEEE Transactions on Neural Networks* 20.1, pp. 61–80. ISSN: 1941-0093. DOI: 10.1109/TNN.2008.2005605.
13. Grover, Aditya and Leskovec, Jure (Aug. 2016). "node2vec: Scalable Feature Learning for Networks". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. New York, NY, USA: Association for Computing Machinery, pp. 855–864. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939754.
14. Narayanan, Annamalai et al. (June 2016). *subgraph2vec: Learning Distributed Representations of Rooted Sub-graphs from Large Graphs*. DOI: 10.48550/arXiv.1606.08928.
15. Narayanan, Annamalai et al. (July 2017). *graph2vec: Learning Distributed Representations of Graphs*. DOI: 10.48550/arXiv.1707.05005.
16. Choi, Youngduck, Chiu, Chill Yi-I, and Sontag, David (July 2016). "Learning Low-Dimensional Representations of Medical Concepts". In: *AMIA Summits on Translational Science Proceedings 2016*, pp. 41–50. ISSN: 2153-4063.
17. Cui, Hejie et al. (Aug. 2023). *A Survey on Knowledge Graphs for Healthcare: Resources, Applications, and Promises*. DOI: 10.48550/arXiv.2306.04802.
18. Chandak, Payal, Huang, Kexin, and Zitnik, Marinka (Feb. 2023). "Building a knowledge graph to enable precision medicine". In: *Scientific Data* 10.1, p. 67. ISSN: 2052-4463. DOI: 10.1038/s41597-023-01960-3.
19. Yu, Yue et al. (Sept. 2021). "SumGNN: multi-typed drug interaction prediction via efficient knowledge graph summarization". In: *Bioinformatics* 37.18, pp. 2988–2995. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab207.
20. Su, Xiaorui et al. (May 2022). "Attention-based Knowledge Graph Representation Learning for Predicting Drug-drug Interactions". In: *Briefings in Bioinformatics* 23.3, bbac140. ISSN: 1477-4054. DOI: 10.1093/bib/bbac140.
21. Diaz Ochoa, Juan G. and Mustafa, Faizan E. (Sept. 2022). "Graph neural network modelling as a potentially effective method for predicting and analyzing procedures based on patients' diagnoses". In: *Artificial Intelligence in Medicine* 131, p. 102359. ISSN: 0933-3657. DOI: 10.1016/j.artmed.2022.102359.

22. Hu, Zhixu et al. (Mar. 2022). "Network analytics and machine learning for predicting length of stay in elderly patients with chronic diseases at point of admission". In: *BMC Medical Informatics and Decision Making* 22.1, p. 62. ISSN: 1472-6947. doi: 10.1186/s12911-022-01802-z.
23. Rocheteau, Emma et al. (Jan. 2021a). "Predicting Patient Outcomes with Graph Representation Learning". In: *arXiv:2101.03940 [cs]*. URL: <http://arxiv.org/abs/2101.03940>.
24. Millar, Jane (2016). "The Need for a Global Language - SNOMED CT Introduction". In: *Studies in Health Technology and Informatics* 225, pp. 683–685. ISSN: 1879-8365.
25. NHS (May 2023). *SNOMED CT*.
26. SNOMED International (2023). *SNOMED CT Starter Guide*.
27. Akiba, Takuya et al. (2019). "Optuna: A Next-generation Hyperparameter Optimization Framework". In: *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
28. Lee, Juho et al. (May 2019). *Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks*. doi: 10.48550/arXiv.1810.00825.
29. The Health Foundation (Dec. 2023). *What's driving increasing length of stay in hospitals since 2019?* URL: <https://www.health.org.uk/publications/long-reads/what-s-driving-increasing-length-of-stay-in-hospitals-since-2019>.
30. Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.
31. Charlson, M. E. et al. (1987). "A new method of classifying prognostic comorbidity in longitudinal studies: development and validation". In: *Journal of Chronic Diseases* 40.5, pp. 373–383. ISSN: 0021-9681. doi: 10.1016/0021-9681(87)90171-8.
32. Deyo, Richard A., Cherkin, Daniel C., and Ciol, Marcia A. (June 1992). "Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases". In: *Journal of Clinical Epidemiology* 45.6, pp. 613–619. ISSN: 0895-4356. doi: 10.1016/0895-4356(92)90133-8.
33. Unified Medical Language System (2023). *Metathesaurus*.
34. Harris, Charles R. et al. (Sept. 2020). "Array programming with NumPy". In: *Nature* 585.7825, pp. 357–362. doi: 10.1038/s41586-020-2649-2.
35. Paszke, Adam et al. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems* 32. Curran Associates, Inc., pp. 8024–8035.
36. Kingma, Diederik P. and Ba, Jimmy (Dec. 2014). *Adam: A Method for Stochastic Optimization*. URL: <https://arxiv.org/abs/1412.6980v9>.

37. Maneewongvatana, Songrit and Mount, David M. (Jan. 1999). *Analysis of approximate nearest neighbor searching with clustered point sets*. doi: 10.48550/arXiv.cs/9901013.
38. Sharafoddini, Anis, Dubin, Joel A., and Lee, Joon (Mar. 2017). "Patient Similarity in Prediction Models Based on Health Data: A Scoping Review". In: *JMIR Medical Informatics* 5.1, e6730. doi: 10.2196/medinform.6730.
39. Huang, Zhexue (Sept. 1998). "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values". In: *Data Mining and Knowledge Discovery* 2.3, pp. 283–304. issn: 1573-756X. doi: 10.1023/A:1009769707641.
40. Greenacre, Michael and Blasius, Jorg (June 2006). *Multiple Correspondence Analysis and Related Methods*. CRC Press. isbn: 978-1-4200-1131-9.
41. Rocheteau, Emma, Liò, Pietro, and Hyland, Stephanie (2021b). "Temporal Pointwise Convolutional Networks for Length of Stay Prediction in the Intensive Care Unit". In: *Proceedings of the Conference on Health, Inference, and Learning*. CHIL '21. Virtual Event, USA: Association for Computing Machinery, pp. 58–68. isbn: 9781450383592. doi: 10.1145/3450439.3451860.
42. Bate, Louise et al. (Oct. 2012). "How clinical decisions are made". In: *British Journal of Clinical Pharmacology* 74.4, pp. 614–620. issn: 0306-5251. doi: 10.1111/j.1365-2125.2012.04366.x.
43. Gerke, Sara, Minssen, Timo, and Cohen, Glenn (Jan. 2020). "Chapter 12 – Ethical and legal challenges of artificial intelligence-driven healthcare". In: *Artificial Intelligence in Healthcare*. Ed. by Adam Bohr and Kaveh Memarzadeh. Academic Press, pp. 295–336. isbn: 978-0-12-818438-7. doi: 10.1016/B978-0-12-818438-7.00012-5.
44. Lyudovyyk, Olga and Weng, Chunhua (Aug. 2019). "SNOMEDtxt: Natural Language Generation from SNOMED Ontology". In: *Studies in health technology and informatics* 264, pp. 1263–1267. issn: 0926-9630. doi: 10.3233/SHTI190429.
45. Wu, Zhibiao and Palmer, Martha (June 1994). *Verb Semantics and Lexical Selection*. doi: 10.48550/arXiv.cmp-1g/9406033.

Chapter 4

Antibiotic Cessation Decision Support

Abstract

It is challenging on an individual patient level to determine when stopping antimicrobial treatment is most appropriate. Unnecessary exposure causes side effects and drives antimicrobial resistance (AMR), while ceasing too early can result in an ineffective prescription. A large proportion of the extended antimicrobial durations seen in clinical practice currently are due to prescribers' fears of patient deterioration, poor outcomes, and potential litigation. We therefore hypothesised that an artificial intelligence-driven model could estimate the impact of changing antimicrobial treatment on a patient's response or outcome. By understanding the potential impact of their actions, clinicians may be able to identify for an individual patient when it is appropriate to stop antimicrobial treatment, which on a population level could help reduce the excessive use of antibiotics and address the problem of AMR.

We investigated this by developing a model that utilises a recurrent neural network autoencoder and a synthetic control-based approach to estimate patients' intensive care unit (ICU) length of stay (LOS), mortality outcome and total white blood cell count (WBC) as a proxy for treatment response, for any given day, under the alternative scenarios of if they were to stop versus continue antibiotic treatment. We extracted data from the routinely collected electronic health records (EHRs) in the MIMIC-IV database for patients who received intravenous antibiotic treatment during an ICU admission. Control days where our model should reproduce labels demonstrated minimal difference for both stopping and continuing scenarios, indicating estimations are reliable (LOS results of 0.24 and 0.42 days mean delta, 1.93 and 3.76 root mean squared error (RMSE) respectively). Meanwhile, impact days where we assess the potential effect of the unobserved scenario showed that stopping antibiotic therapy earlier had

a statistically significant shorter LOS and a reduction in WBC (mean reduction 2.71 days and $0.71 \times 10^3/\mu\text{L}$, p-value <0.01). No impact on mortality was observed.

A significant proportion of the work discussed in this chapter was published in *Frontiers in Digital Health* under the title 'Machine learning and synthetic outcome estimation for individualised antimicrobial cessation' (Bolton et al. 2022). In addition, it was accepted for an oral presentation at the Digital Solutions for Infections Disease session of the European Congress of Clinical Microbiology and Infectious Diseases (ECCMID) 2023. Figure 4.1 shows a graphical abstract of the approach and research methodologies employed in this chapter.

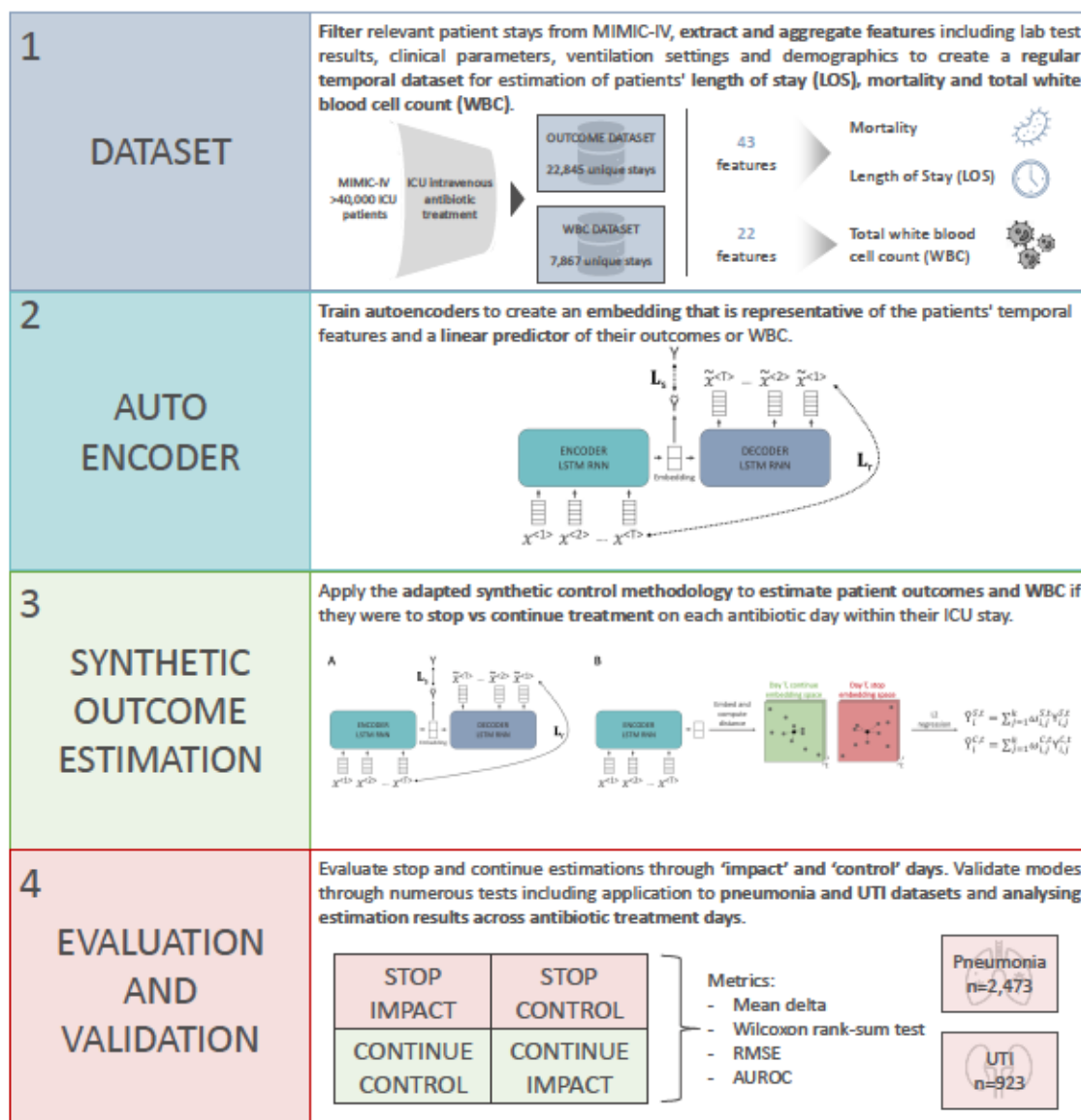


Figure 4.1: Overview of the steps taken in order to develop models for antimicrobial cessation synthetic outcome and white blood cell count estimation.

4.1 Chapter objectives

- *Develop a machine learning and synthetic control-based method for estimations under alternative antibiotic treatment scenarios*
- *Train bi-directional long-short-term memory (LSTM) autoencoder models to predict patient outcomes and WBC*
- *Test and analyse outcome and WBC estimations for stop and continue scenarios*

4.2 Introduction

One challenge in treating a patient with a bacterial infection is determining when it is appropriate to stop antibiotic treatment (Tamma et al. 2019). The decision to cease antibiotics too early can result in the patient's condition worsening, while unnecessary exposure increases the risk of toxicity (Langford et al. 2017) and drives the evolution of AMR (Holmes et al. 2016; Mo et al. 2023). Even extending treatment by a short amount of time can have a significant impact on the population level and enhance the development of resistance (Spellberg 2016). Furthermore, excessive treatment is responsible for most avoidable antibiotic adverse events, including gastrointestinal distress and allergic reactions (Vaughn et al. 2019; Curran et al. 2022). Numerous studies have shown that on a population level shorter treatment durations are often non-inferior to longer ones (Hanretty et al. 2018; Royer et al. 2018; Spellberg et al. 2019; Wald-Dickler et al. 2019; Yahav et al. 2019). The challenge is that the resulting recommendations do not take into account the characteristics of the individual patient or specific scenarios. It is difficult for clinicians to have confidence in individualised treatment decisions for their patients, when there is a poor understanding of the factors that facilitate or inhibit an individual from receiving a shorter duration of antibiotic therapy. Therefore, durations are

often unnecessarily extended (Janssen et al. 2022) and decided by habit or arbitrarily based on population evidence (Sax 2010). The cessation of antibiotics should be a collective and data-driven decision, as choices are made in a more favourable environment once time has passed from presentation and significant amounts of information have been gathered. Despite this, systems to help support individualised antibiotic duration and cessation decision-making are often neglected and under-researched, with little innovation in this area (Peiffer-Smadja et al. 2020; Charani et al. 2021).

Given that standard of care uses clinical factors to determine if a patient should stop antibiotics or not, we hypothesised that an AI-based CDSS using routinely collected EHR data may be able to assist in decision-making regarding stopping antibiotics. This would support individualised antibiotic cessation decisions and overcome prescriber concerns of poor patient outcomes and patient deterioration that is likely a major driver of overtreatment (Rawson et al. 2016; Pandolfo et al. 2022). We approached this problem by estimating clinical outcomes and treatment response under alternative scenarios with the aim of showing non-inferiority or a direct benefit of antibiotic cessation. More specifically, a machine learning and synthetic control-based approach was developed to estimate the LOS, mortality, and total WBC of patients for any given day, if they were to stop versus continue antibiotic treatment.

4.3 Methods

4.3.1 Dataset

MIMIC-IV a publicly available large de-identified real-world clinical dataset was used in this research (Goldberger et al. 2000; Johnson et al. 2021). Input features including lab test results, vitals and assessment scales were extracted, analysed, and selected based on prevalence, correlation, as well as the advice of infectious disease doctors and critical care consultants. Some features were calculated on the basis of other variables. Specifically, cumulative overall antibiotic treatment length was determined for each day of each ICU stay that considered

consecutive treatment days regardless of the antibiotic administered. In addition, whether the patient received antibiotic re-treatment and their age at ICU admission were also computed. Preprocessing was applied to features including outliers being removed and values normalised, as well as missing values forward filled or highlighted. The features were aggregated by day for each unique stay using the mean to create a regular temporal dataset. In general, there was a high degree of missingness, and so patients with greater than 50% of values missing each day were removed. The resulting dataset for mortality and ICU LOS prediction contained 43 input features while WBC estimation used 22 features. Labels were extracted for each patient. WBC was temporally dynamic, while mortality (binary) and LOS (continuous value) were static. For the prediction of mortality and LOS, the patient population was filtered to those who received intravenous antibiotic treatment for <21 days during an ICU stay, while WBC used a filter of <8 days. These filters were applied due to the exponential decrease in patients as the length of treatment increased.

4.3.2 Model architecture

The objective of our model was to estimate the patients' LOS, mortality, and WBC for any given day, if they were to stop versus continue antibiotic treatment. The model used a bi-directional long-short-term memory (LSTM) autoencoder which takes in a sequence of patient input features ($x^{<1>}, x^{<2>} \dots x^{<T>}$), creates an embedding representation and outputs a sequence of reconstructed features ($\tilde{x}^{<T>} \dots \tilde{x}^{<2>}, \tilde{x}^{<1>}$). This autoencoder was trained through two loss functions (Qian et al. 2021), which were summed together to create a combined loss for backpropagation. First, the reconstruction loss L_r was calculated by the RMSE between the outputs that are trying to reproduce the inputs and the real input data. Secondly, a supervised learning loss L_s is calculated by performing a linear transformation of the embedding representation ($\tilde{Y}$) to try to predict the real label (Y) and taking either the RMSE loss for the LOS and WBC labels or the binary cross entropy (BCE) loss for mortality classification. L_s ensures that the embedding space created by the autoencoder is a good linear predictor of the label of interest, which is important for the subsequent adapted synthetic control method.

Overall, an embedding representation was created that considers the history of a patient and is representative of their state on that day.

Once the autoencoder was trained and an embedding representation for each antibiotic day in all patient stays had been created, the adapted synthetic control approach (Abadie et al. 2003) was utilised, where the act of stopping or continuing treatment on a particular day is considered an intervention and each patient acts as a singular unit. This method is useful when evaluating an intervention using randomised controlled trials is challenging, as is the case with antibiotic cessation, and hence retrospective observational data is assessed. Synthetic controls have frequently been applied to understand public health interventions (Kreif et al. 2016; Bottell et al. 2018) but their use within digital health research is limited. Here we wanted to know what the predicted outcomes or treatment response would be if a particular patient was to stop versus continue antibiotics on a given day within their ICU stay. To this extent two synthetic controls are created, one can be labelled the 'stop synthetic control' which is based on subjects who stopped antibiotics on that particular day, and the second labelled the 'continue synthetic control' which is created from subjects who continue antibiotic treatment on that particular day. To achieve this for each day (t), two separate donor pools are created based on subjects' associated embedding representation and antibiotic treatment status. In other words, those who continue antibiotics on day t are partitioned into the 'continue' embedding space while those who stop antibiotics are placed in the 'stop' embedding space. In this way the estimated outcomes and WBC for stopping and continuing on day t are driven by representative donors who experienced analogous treatment. To create the stop and continue synthetic controls for a particular patient i the k most closely related embedding representations from each relevant donor pool are selected based on a distance metric (in this research $k=10$ and Euclidean distance was used for both stop and continue estimations). Given embeddings are representative of the patients' state those donors selected will be similar, giving a considered insight into potential alternative outcomes or treatment response under antibiotic temporality. Ridge regression functions where d are the embedding dimensions, j are the donors and z represents the particular patient i 's embedding for a given dimension and time is then applied

to the subject and their respective stop and continue donor embeddings ($x_{j,d}^{S,t}$ and $x_{j,d}^{C,t}$):

$$Loss_i^{S,t} = \sum_{d=1}^D \left[z_{i,d}^t - \sum_{j=1}^k x_{j,d}^{S,t} w_{i,j}^{S,t} \right]^2 + \sum_{j=1}^k w_{i,j}^{S,t^2} \text{ for stop estimations}$$

$$Loss_i^{C,t} = \sum_{d=1}^D \left[z_{i,d}^t - \sum_{j=1}^k x_{j,d}^{C,t} w_{i,j}^{C,t} \right]^2 + \sum_{j=1}^k w_{i,j}^{C,t^2} \text{ for continue estimations}$$

Ridge regression returns two sets of weights ($w_{i,j}^{S,t}$ for 'stop' and $w_{i,j}^{C,t}$ for 'continue') that minimise the square difference between the subject of interest and the selected units in the donor pools ($Y_{i,j}^{S,t}$ for 'stop' and $Y_{i,j}^{C,t}$ for 'continue'). The objective of this L2 regularization is to fairly distribute weights across the donors for stop and continue estimations. Finally the stop and continue synthetic control estimations ($\tilde{Y}_i^{S,t}$ and $\tilde{Y}_i^{C,t}$ respectively) for the particular patient i are computed from the weighted average of donor labels. To this extent during estimation for a given patient i , we assume we know the label for all other patients within the dataset. Estimations are made for each patient on each relevant antibiotic day of their stay if they were to stop versus continue antibiotic treatment. An overview of the model's architecture and this process for stop and continue outcome and WBC estimation is shown in Figure 4.2.

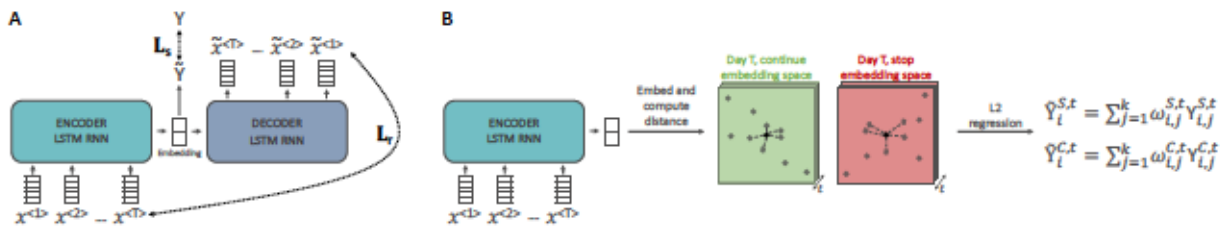


Figure 4.2: Model illustration. [A] The encoder is trained using both a supervised loss (L_s) and reconstruction loss (L_r) [3]. [B] To estimate outcomes and WBC during testing an embedding is created for every day of each patient's stay, embedding spaces are partitioned temporally and based on whether the patient stopped or continued antibiotics. The closest k neighbours are selected as donors from each embedding space and L2 regression returns weights that minimise the square difference between the patient and the donors. A stop and continue synthetic control outcome and WBC is estimated as a weighted average of the donors' outcomes and WBC respectively.

4.3.3 Model development

The model was applied to the MIMIC-IV EHR dataset, which was randomly split based on the patient's 'stay_id' into training, validation, and testing sets (70%, 15%, 15%, respectively). PyTorch (a machine learning library for the Python programming language, Paszke et al. 2019) was used to create a bi-directional long short-term memory (LSTM) recurrent neural network (RNN) with a custom dataset class to extract labels and features. In order to address the mortality class imbalance (Table 4.2) over-sampling was used during training. To be specific, those cases with positive mortality were replicated 3 times within the custom dataset class to achieve a more balanced mortality rate of 51.90% within the training dataset. The Adam optimiser (Kingma et al. 2014) was used with binary cross entropy loss for classification, and mean squared error loss for regression. Training used 50 epochs, during which the model with the best performance on the validation dataset (RMSE for LOS and WBC or Area Under the Receiver Operating Characteristic curve for mortality prediction) was selected as the final model. Three separate LSTM autoencoder models were trained on their respective train datasets to create embedding representations relevant to patients' LOS, mortality and WBC. Models were evaluated using functions and metrics from the TorchMetrics, Scikit-learn, and SciPy Python libraries. Further details of the models and their hyperparameters are shown in Table 4.1 and Figure 4.3.

Table 4.1: Hyperparameters used for each model and their training

Hyperparameter	LOS model	Mortality model	WBC model
Learning rate	0.0001	0.00001	0.001
Batch size	512	512	256
Input dimension	43	43	22
Number of LSTM layers	5	2	1
Bidirectional encoder	True	True	True

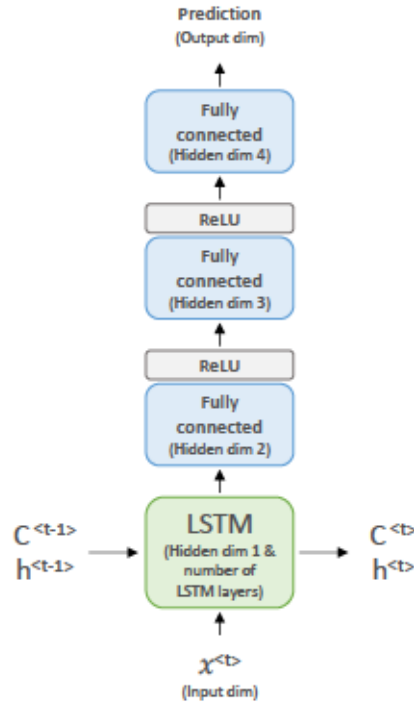


Figure 4.3: Overview of an LSTM cell within the autoencoder models for any given time-step. Dimension sizes are shown in 4.1. Note a final sigmoid activation function (not shown) is applied to the output for mortality prediction.

4.3.4 Model evaluation

Commonly with the synthetic control method the delta difference between the single unit and the counterfactual in the pre-intervention period is minimised and the treatment effect is then observed in the post-intervention period. For our research question, this is not possible due to the nature of stopping antibiotics being the final event at one point in time, after which the patient is not applicable to our research population or question. An analogue can be applied for this study, where we define 'control' and 'impact' days that are equivalent to the pre and post-intervention periods. For estimating outcomes and WBC when continuing antibiotics, all the days the patient actually continues antibiotics are 'control' days where we expect minimal difference between the true label and our estimation. On the other hand, on the single day the patient stops antibiotics, we can assess the 'impact' if they were to instead continue. When estimating outcomes and WBC upon stopping antibiotics, the reverse is true, whereby each day antibiotics were continued the 'impact' of stopping can be assessed and the final day where

the patient stops treatment acts as a 'control'. Note that it is not possible to define this for every patient, as not every individual will stop antibiotics during their stay in the ICU. The percentage of patients who stopped antibiotic treatment during their ICU stay is shown in Table 4.2. Patients' outcomes and WBC are estimated in the same way for impact and control days as discussed in Subsection 4.3.2. However, for control days, we know the real label and so can compare our estimations, while for impact days the real label is unknown. Each day therefore acts as both a 'control' and an 'impact' across the two 'stop' and 'continue' scenario estimations. An outline of this is shown in Figure 4.4 and the number of continue and stop donors for each day in the outcome (LOS and mortality) test dataset is illustrated in the Supplementary Figure A.4.

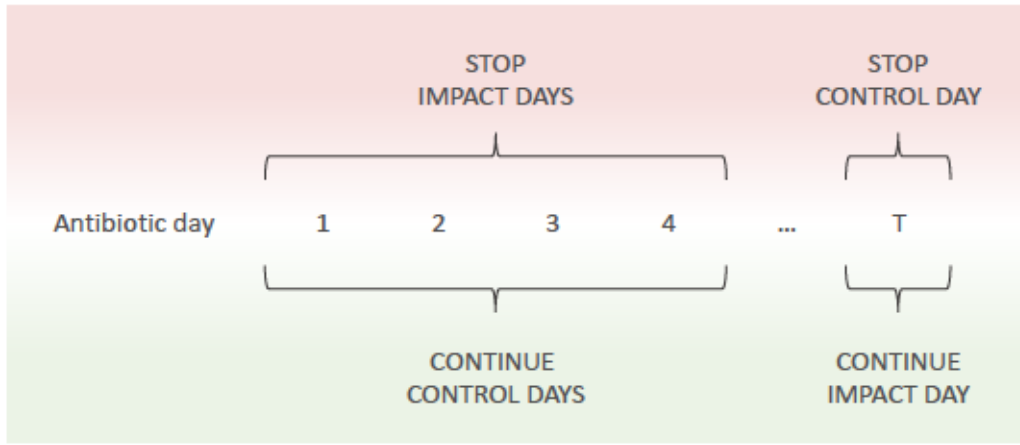


Figure 4.4: Demonstration of the impact and control evaluation process for stop and continue scenarios. An antibiotic day is defined as each day the patient receives treatment as well as the day they stop. After starting antibiotics, each day the patient receives treatment acts as a stop impact and continue control. This continues until antibiotic cessation or ICU discharge. If the patient stops antibiotics during their ICU stay that initial day when no antibiotics are administered acts as a stop control and a continue impact.

The mean delta is calculated to evaluate the difference between the real labels and the estimations, though the following formulas, where T_i is the number of days the patient receives antibiotics:

$$\mu\Delta^S = \frac{1}{n} \sum_{i=1}^n \left[\frac{1}{T_i} \sum_{t=1}^{T_i} \left[Y_i^{S,t} - \hat{Y}_i^{S,t} \right] \right] \text{ for stop estimations}$$

$$\mu\Delta^C = \frac{1}{n} \sum_{i=1}^n \left[\frac{1}{T_i} \sum_{t=1}^{T_i} \left[Y_i^{C,t} - \hat{Y}_i^{C,t} \right] \right] \text{ for continue estimations}$$

Minimal difference should be seen on the control days where our model aims to reproduce the labels, while on impact days you can assess the effect of the unobserved scenario. According to the formula, a positive result indicates that the estimation is lower than the actual value, while a negative result means the estimation is higher than the actual value. Statistical analysis can be used to determine whether the difference between the true LOS and WBC labels and the estimated outcomes is statistically significant. Given the non-normal data distribution, the non-parametric Wilcoxon rank-sum (Mann-Whitney U) test was used with the alpha set at 0.05. Furthermore, the mean absolute percentage error (MAPE) and mean absolute error (MAE) can be calculated using the following notation:

$$MAPE^S = \frac{1}{n} \sum_{i=1}^n \left[\frac{1}{T_i} \sum_{t=1}^{T_i} \frac{|Y_i^{S,t} - \tilde{Y}_i^{S,t}|}{Y_i^{S,t}} \right] \quad \text{and} \quad MAE^S = \frac{1}{n} \sum_{i=1}^n \left[\frac{1}{T_i} \sum_{t=1}^{T_i} |Y_i^{S,t} - \tilde{Y}_i^{S,t}| \right]$$

respectively for stop estimations

$$MAPE^C = \frac{1}{n} \sum_{i=1}^n \left[\frac{1}{T_i} \sum_{t=1}^{T_i} \frac{|Y_i^{C,t} - \tilde{Y}_i^{C,t}|}{Y_i^{C,t}} \right] \quad \text{and} \quad MAE^C = \frac{1}{n} \sum_{i=1}^n \left[\frac{1}{T_i} \sum_{t=1}^{T_i} |Y_i^{C,t} - \tilde{Y}_i^{C,t}| \right]$$

respectively for continue estimations

Standard machine learning metrics can also be used to evaluate model prediction performance. For LOS and WBC regression estimation, the RMSE is used. For the mortality classification task, the Area Under the Receiver Operating Characteristic curve (AUROC) is most appropriate given the class imbalance (Table 4.2), but accuracy, precision, recall, F1 score and Area Under the Precision Recall curve (AUPRC) can also be calculated. The details and formulas for these metrics are provided in the Appendix (Section A.1). Metrics were calculated as global averages, across all samples, meaning that every day of antibiotic treatment within each patient's stay is considered equally. 95% confidence intervals were calculated through 1000 bootstrapped samples on the test set with n=1000 for mortality metrics and the sum of

the squared errors method for LOS RMSE.

To further validate our outcome estimation findings, we applied our model to patients in MIMIC who were diagnosed with pneumonia or a urinary tract infection (UTI). The effects of short versus longer antibiotic treatment regimes have been extensively studied in pneumonia and UTIs. In general, research supports the notion that shorter durations of antibiotic treatments are non-inferior to longer ones in these infections, especially for non-complicated cases (Gier et al. 1995; Peterson et al. 2008; Dimopoulos et al. 2013; Pugh et al. 2015; Royer et al. 2018; Drekonja et al. 2021). Based on this evidence and antimicrobial prescribing guidelines (NICE 2018; NICE 2019) we defined a long treatment duration as any patient receiving antibiotics for more than 7 days and applied our model to estimate their outcomes if they were to instead stop treatment after 7 days. In addition, there is increasing evidence that even shorter courses of antibiotics can be used in such infections, particularly pneumonia (Royer et al. 2018; NICE 2019). Hence, we investigated the estimated outcomes of those patients who received the standard of care 7 days of treatment, for slightly shorter treatment durations (5 or 6 days).

4.4 Results

4.4.1 Data

For outcome estimation, 18,988 patients, associated with 22,845 unique ICU stays, were included across datasets. The mortality rate was 18.60% and the mean LOS 5.63 days. 7,867 unique ICU stays were identified for WBC estimation, the mean WBC was 10.88 with an interquartile range of 5.50. An overview of statistics for each dataset is shown in Table 4.2

Table 4.2: Datasets statistics.

Statistic	Outcome estimation datasets						WBC estimation datasets			
	Overall	Train	Validation	Test	Pneumonia	UTI	Overall	Train	Validation	Test
Number of stays	22,854	15,991	3,427	3,427	2,473	923	7,867	5,507	1,180	1,180
Mean length of treatment (days)	4.38	4.38	4.48	4.30	6.95	4.77	2.85	2.86	2.81	2.84
Length of treatment standard deviation	3.32	3.32	3.48	3.18	4.28	3.55	1.60	1.62	1.54	1.59
Mortality rate	18.60	18.47	18.30	19.52	24.02	18.96	-	-	-	-
Mean LOS (days)	5.63	5.62	5.74	5.55	9.05	5.50	-	-	-	-
LOS standard deviation	4.23	4.24	4.31	4.15	5.19	4.32	-	-	-	-
Spearman's correlation between LOS and treatment length	0.72	0.72	0.72	0.73	0.73	0.74	-	-	-	-
Mean WBC ($\times 10^3/\mu\text{L}$)	-	-	-	-	-	-	10.88	10.88	10.89	10.86
WBC standard deviation	-	-	-	-	-	-	4.19	4.21	4.13	4.20
Percentage of patients that stopped treatment during their ICU stay	41.56	41.64	41.17	41.55	31.95	26.54	86.41	86.69	85.59	85.93

4.4.2 Length of stay and mortality estimation

4.4.2.1 Autoencoder

Through a linear transformation of a given patient day embedding, outcome estimations could be made on the unseen test set (3,427 unique ICU stays). The LOS model achieved an RMSE of 3.88 (95% CI 3.84–3.92), while the mortality estimation model obtained an AUROC of 0.77 (95% CI 0.73–0.80) (accuracy 0.73 (95% CI 0.71–0.75), precision 0.44 (95% CI 0.36–0.46), recall 0.67

(95% CI 0.61–0.72), specificity 0.75 (95% CI 0.72–0.78), F1 0.53 (95% CI 0.46–0.56), AUPRC 0.55 (95% CI 0.42–0.56)), Figure 4.5, indicates that the model was relatively effective at balancing false positive and false negative mortality predictions.

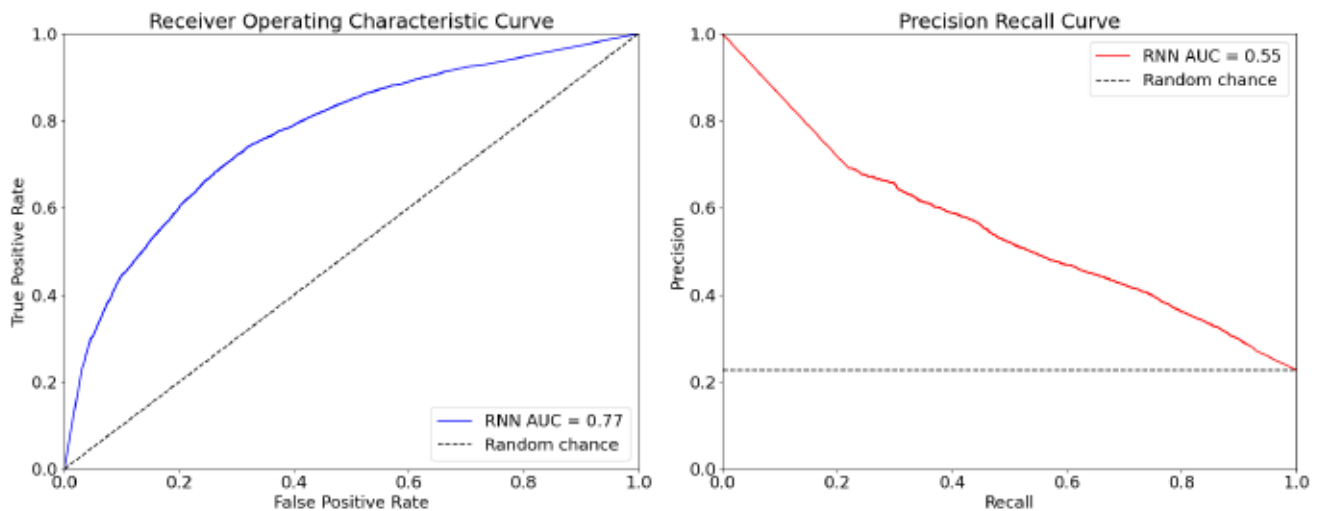


Figure 4.5: ROC and PRC results for the RNN autoencoder on mortality classification.

4.4.2.2 Synthetic outcome estimation

LOS and mortality estimation results on the unseen test set are shown in Table 4.3. For LOS estimation on control days, the mean delta under both stopping and continuing scenarios was 0.24 and 0.42 days, respectively, showing a minimal difference between predictions and the ground truth labels. Furthermore, a MAPE of 0.26, MAE of 1.32 and RMSE of 1.93 for stop control days shows that the corresponding impact estimations are more reliable. On impact days, stopping earlier had a statistically significant shorter LOS (mean difference 2.71 days, p -value <0.01). This indicates that on average, the LOS estimates for stopping antibiotics earlier are shorter in duration than those when the patient continues antibiotics. For mortality, no impact was observed by stopping or extending antibiotic treatment with the mean delta being similar for impact and control days. Estimations had modest performance with an average AUROC of 0.67 and accuracy of 0.82.

Table 4.3: Outcome estimation results for patients in the unseen test set. * indicates statistical significance with alpha set at 0.05.

SCENARIO	DAY(S)	LOS				Mortality		
		Mean delta (days, p-value)	MAPE	MAE	RMSE	Mean delta	MAE	AUROC
STOP	IMPACT	2.71*, <0.01	0.36	3.30	4.80	0.06	0.25	0.66
	CONTROL	0.24, 0.60	0.26	1.32	1.93	0.05	0.15	0.72
CONTINUE	IMPACT	-2.09*, <0.01	0.77	2.85	3.16	0.05	0.18	0.67
	CONTROL	0.42*, 0.01	0.48	2.72	3.76	0.07	0.24	0.64

Estimations were made for each day of each patient's stay within all the extracted data (i.e., train, validation, and testing sets combined), to understand if results would deviate by dataset size. For LOS reliable estimations were once again obtained (mean stop control difference of 0.33 days and mean continue control difference of 0.42 days). Continuing showed no given impact (mean difference of -0.30 days) while stopping once again showed a significant impact with a mean reduction of 1.87 days. Little difference in mortality estimations was observed between stop and continue controls and impacts (stop impact - 0.03, stop control - 0.03, continue control - 0.05, continue impact - 0.05). Mortality predictions were relatively reliable with a mean AUROC of 0.72.

To show the importance of the temporality in our predictions we created estimations for each antibiotic day of each patient's stay, without segregating the embedding space (by time or by antibiotic treatment given they are mutually dependent). The resulting estimations had a mean LOS difference of 2.60 days from the true labels, a RMSE of 5.05, and a statistically significant difference in medians (p-value of <0.01).

The performance of the model on subjects towards the edges of the distribution in terms of the correlation between LOS and overall antibiotic treatment length was investigated. Subjects in the 10th and 90th percentiles were selected leading to a smaller Spearman's

correlation of 0.35. As expected given the dataset size ($n=686$) and donor distribution, results were quite poor with a mean stop control difference of 2.92 days and a mean continue control difference of 2.13 days. The impact of stopping early though was still much greater than the control at 4.36 days mean difference.

4.4.2.3 Pneumonia and UTIs

2,473 stays were identified in which the patient was diagnosed with pneumonia, with a mean LOS of 9.05 days and a mean duration of antibiotic treatment of 6.95 days. Overall estimation results on this whole pneumonia population reflected the wider dataset and are shown in Table 4.4. When focusing in on those with long treatment durations and the question of what if they stopped after 7 days of treatment, statistically significant results show that average LOS were 2.82 days shorter when stopping earlier. No differences in estimated mortality were observed; however, estimations were consistent across groups with an average AUROC of 0.75. No significant difference in LOS or mortality was estimated for pneumonia patients that received the standard of care 7 days treatment if they had slightly shorter treatment durations of 5 or 6 days.

For UTIs, 923 patient stays were selected having a mean LOS and antibiotic treatment duration of 5.50 and 4.77 days, respectively. Once again, overall estimation results (Table 4.4) were similar to previous findings with trustworthy controls, early stopping being associated with shorter LOS and no difference in mortality, but reliable estimations (AUROC ranging from 0.63 to 0.87). Estimates for stopping after 7 days for those with long treatment durations did show a positive impact in terms of reduced LOS (mean difference 2.08 days, p -value <0.01). The stop control where we expected to see minimal difference showed a larger mean deviation of 1.04 days, but statistical analysis showed that the medians between control estimations and labels were not significantly different. Mortality estimations here were for the most part reliable, a high predictive performance on stop and continue controls was achieved with an AUROC of 0.93 and 0.78 respectively, but a lower score for the stop impact of 0.52. When analysing those patients who received the standard of care 7 days treatment, for slightly

Table 4.4: Outcome estimation results for patients with Pneumonia and UTIs. Results are shown for both the whole population and analysis of what if those who received long treatment durations stopped after day 7. * indicates statistical significance with alpha set at 0.05.

INFECTION	ANALYSIS	SCENARIO	DAY(S)	LOS		Mortality	
				Mean delta (days, p-value)	RMSE	Mean delta	AUROC
Pneumonia	Whole dataset	STOP	IMPACT	3.72*, <0.01	5.87	0.00	0.71
			CONTROL	0.26, 0.47	2.14	0.07	0.76
		CONTINUE	IMPACT	-2.79*, <0.01	3.65	0.10	0.69
			CONTROL	0.49*, <0.01	4.01	0.05	0.68
	Long treatment durations stopping after 7 days	STOP	IMPACT	2.82*, <0.01	4.65	-0.03	0.74
			CONTROL	0.43, 0.08	2.11	0.05	0.80
		CONTINUE	IMPACT	-	-	-	-
			CONTROL	0.41, 0.21	3.47	0.05	0.73
UTI	Whole dataset	STOP	IMPACT	2.36*, <0.01	4.70	0.14	0.63
			CONTROL	0.36, 0.89	2.04	0.07	0.87
		CONTINUE	IMPACT	-1.91*, <0.01	3.26	0.03	0.79
			CONTROL	0.38, 0.05	3.82	0.04	0.71
	Long treatment durations stopping after 7 days	STOP	IMPACT	2.08*, <0.01	4.35	0.30	0.52
			CONTROL	1.04, 0.23	2.42	0.17	0.93
		CONTINUE	IMPACT	-	-	-	-
			CONTROL	0.26, 0.05	3.48	0.05	0.76

shorter treatment durations (5 or 6 days), a statistically significant result was observed where the estimated LOS outcomes were on average longer by 1.45 days if the patients stopped antibiotics slightly earlier (p-value <0.01, RMSE 2.72).

4.4.3 White blood cell count estimation

The standard autoencoder achieved a RMSE of 3.33 for WBC prediction on the unseen test set. For the test dataset synthetic WBC estimations an embedding was created for every day of each patient's stay (Supplementary Figure A.3). Results on control days show estimations are reliable with a RMSE of 3.24 and 3.44 for stopping and continuing, respectively. As expected,

the mean delta on control days was minimal at <0.01 for stopping and 0.02 for continuing. However, on impact days, statistically significant differences were observed with 0.71 and -0.31 for stopping and continuing, respectively (Table 4.5). The WBC estimation results were also consistent over time, as shown in Figure 4.6.

Table 4.5: WBC estimation results for patients in the unseen test set. * indicates statistical significance with alpha set at 0.05.

SCENARIO	DAY(S)	WBC			
		Mean delta (days, p-value)	MAPE	MAE	RMSE
STOP	IMPACT	0.71*, <0.01	0.34	2.99	3.89
	CONTROL	0.00, 0.06	0.32	2.50	3.24
CONTINUE	IMPACT	-0.31*, <0.01	0.33	2.53	3.18
	CONTROL	0.02*, 0.01	0.32	2.70	3.44

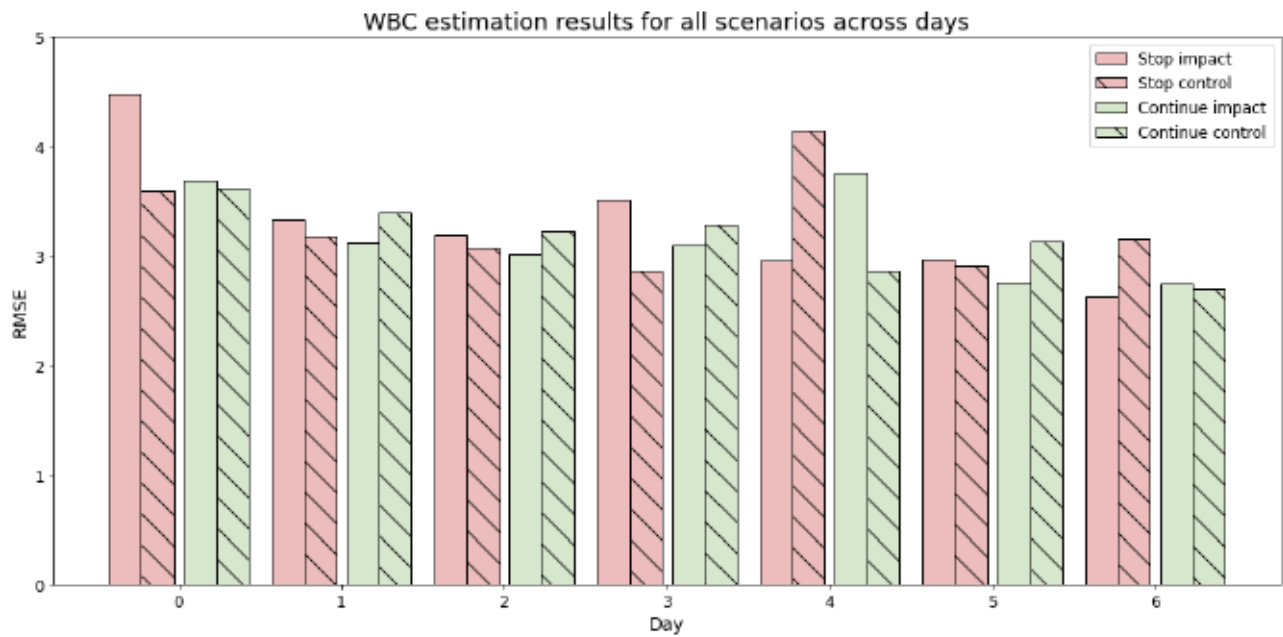


Figure 4.6: WBC estimation RMSE results for all scenarios broken down by the day of antibiotic treatment.

4.5 Discussion

We demonstrate that our RNN autoencoder and synthetic control-based approach trained on a large ICU EHR dataset can estimate patient outcomes and WBC under the alternative scenarios of stopping versus continuing antibiotic treatment. The results across experiments were consistent, with stop control days often showing the greatest performance, indicating that our stop impact estimations, which occur on days where the true label upon stopping is unknown, are more reliable. The stop impact results from this retrospective study show that stopping antibiotics earlier may be associated with a statistically significant average reduction in LOS of 2.71 days and can decrease WBC. These results are likely due to those that stop treatment being more healthy and hence in the future it would be interesting to investigate and disentangle the behavioural and biological components of this. Overall, minimal impact on mortality was observed, which reassures that early cessation can be a safe stewardship option. But this result is also to be expected given death can be caused by a large number of factors beyond those included as model features. Figure 4.7 shows some specific illustrative examples of patient LOS and mortality estimations. The pneumonia dataset demonstrated particularly positive results with overall and stopping on day 7 results indicating that antibiotic cessation can have a significant impact on LOS in this population (mean difference 3.72 and 2.82 days, respectively). This reflects current clinical thinking that shorter treatments are optimal for pneumonia infections (Dimopoulos et al. 2013; Pugh et al. 2015; Royer et al. 2018; NICE 2019). However, a balance must be struck with antibiotic treatment durations. The UTI analysis indicated courses shorter than 7 days may be detrimental to the patient and that the current standard of care treatment duration is likely appropriate for those in an ICU. Given most UTIs can be treated outside of an ICU environment and this research used ICU data this result may also be driven by a lack of embeddings for those more healthy UTI patients. Care must therefore be taken to prioritize the best interests of patients and the public in addressing current infections and the threat of antimicrobial resistance (AMR). This includes accounting for multi-factorial impacts on infectious diseases, such as patients' co-morbidities (Chapter 3) and the ethical considerations of antimicrobial prescriptions (Section 2.4).

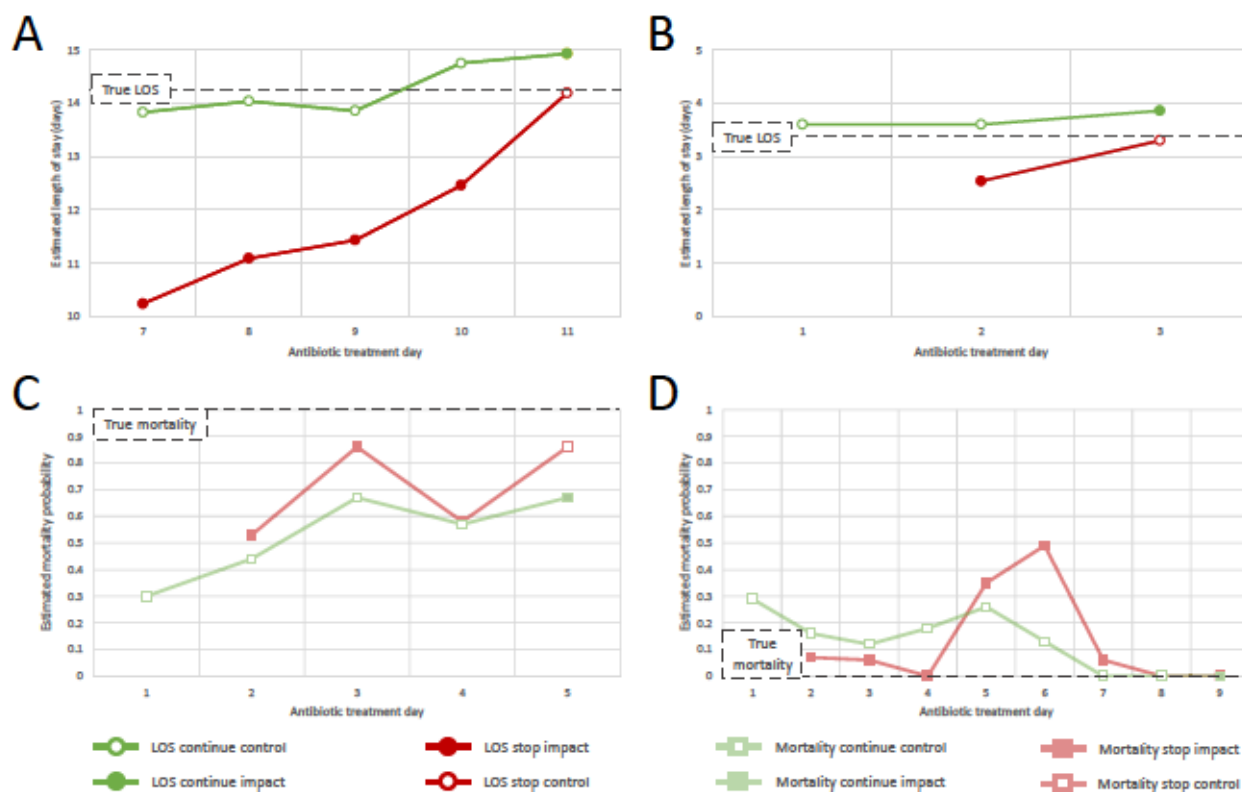


Figure 4.7: LOS and mortality synthetic outcome estimation results for particular patients. These cases were selected as illustrative examples of four distinct patient scenarios: [A] the patient has a long course of antibiotics, [B] the patient has a short course of antibiotics, [C] the patient dies, [D] the patient survives. In A/B control estimation results show minimal deviation from the true LOS label while the stop impact estimations have a reduced LOS. Results in C/D indicate mortality estimations are temporally dynamic but with little difference between stop versus continue.

The methodological approach towards the problem of antibiotic cessation discussed in this chapter is novel. This model can in principle help with individualised antibiotic cessation decisions as it takes into account numerous patient characteristics and the specific treatment scenario with regard to patient outcomes and treatment response, factors that previously could not be considered together in their entirety. We approached the problem of antibiotic cessation from the perspective of making a clinically useful tool designed to estimate direct measures that may influence clinical decision-making in alternative scenarios. We believe that it could be useful for prescribing physicians during their daily clinical rounds to compare between stop and continue estimates and to understand when it is appropriate to stop antibiotic treatment. In particular, this system should help show that shorter treatment durations can be safe and support individualised antimicrobial decision-making through hard outcome and treat-

ment response estimation. From a behaviour change perspective, this approach may provide reassurance to support early cessation of therapy, while promoting improved knowledge and understanding on the issue of antimicrobial optimisation and stewardship (Pauwels et al. 2021). It should be noted that too short a course of antibiotics can cause harm and have negative knock-on effects. As such, the objective of this research was to help optimise antimicrobial use and determine the most appropriate duration of antibiotic treatment for each individual patient. An important and outstanding question is how clinicians treating a patient would adopt the recommendations provided by such a system and if it would influence antimicrobial clinical decision-making. Holistically, we believe that antibiotic cessation is a collective, data-driven decision, meaning a CDSS in this area can have a greater influence and acceptance by end users. However, the degree to which this tool would be accepted and work alongside clinical decision-making behaviour requires investigation.

We have shown that our methodological approach and models can reliably predict different labels across classification and regression tasks. As well as estimate alternative patient outcomes and WBC depending on their antibiotic treatment status. Based on our results, the size and consistency of the dataset used, and hence the number of available donors, are strongly related to the reliability of outputs. Experiments utilising small datasets often led to poor results, as there were not enough suitable patients within a given embedding space to create an appropriate synthetic estimation. On the other hand, there does seem to be a ceiling above which more instances are not necessary. For example, similar results were obtained across the pneumonia, test and whole datasets even though they had a range of sizes of 2,476, 3,427 and 22,845 patient stays, respectively. As such, we can infer that this method is likely to produce suitable estimations if several thousand patient examples are available. Although this should be reasonable for most clinical scenarios, it does act as a dataset constraint when evaluating less common infections, where potentially more interesting nuanced findings could be made.

The quality of the initial autoencoder model is another significant implication that determines the performance. The standard autoencoder model without the synthetic control methods applied achieved higher performance on the LOS prediction task than estimations generated

without segregating the embedding space (RMSE of 3.88 and 5.05 respectively). This first confirms that the model has been trained to adequately represent the patient in the embedding space with respect to their outcome. Secondly, the temporal aspect of the embeddings assists with synthetic outcome estimations, and finally the subsequent synthetic outcome estimation methodology applied ensures that outputs can be clinically applicable with regard to antibiotic treatment. As such, the autoencoder is critical for appropriate temporal representations and subsequent estimations.

It is important to note that for the LOS estimation task there was a high degree of correlation between LOS and overall treatment length in the datasets (Table 4.2, Supplementary Figure A.5). This was expected since those patients who are less sick will likely receive fewer antibiotics and leave the ICU sooner. Although the model architecture was designed to account for this, through representative and segregated embeddings, it is still likely that the model 'learnt' this association, causing some confounding. Results on outliers when there is reduced correlation still illustrate that stopping can impact LOS outcomes, even if the predictions themselves are not reliable in this situation given the skewed dataset analysed. Numerous factors influence ICU LOS, therefore, even if the model predicts stopping antibiotics could be neutral or beneficial, other factors may make this impossible. However, our results and the strong correlation observed between antibiotic treatment duration and LOS in this dataset mean that this model can act as a proxy with the ultimate objective of reducing unnecessary antibiotic use.

This research has several limitations. We focused on addressing what would happen if antibiotic cessation occurred earlier during a patient's ICU stay. The synthetic control methodology was chosen and adapted as it allows us to address this problem, while causal discovery seems intractable, and causal inference models rely on assumptions including all relevant confounders being observed which are often too stringent for real-world application. MAPE and MAE LOS estimation results are in the region of days, which could limit clinical utility, but are comparable to those of recent research (Rocheteau et al. 2021). Unlike most synthetic control applications we don't have an extensive pre-intervention period making confidence in results more challenging. Furthermore, one of our analogues stop 'control' days would not be

available on a patient-specific level during clinical use, due to the nature of cessation occurring after treatment, which could hinder the necessary post-market surveillance once implemented. Other types of interpretability such as case-based reasoning and being able to investigate selected donors to see if they are clinically meaningful could counteract this. Secondly, the use of historical EHR data to estimate synthetic outcomes and WBC, means all our predictions are biased based on past antibiotic prescribing policies. These methodological approaches were required to answer our question of interest but mean that historical approaches towards antimicrobial stewardship govern our model's outputs. The analysis of such a large dataset along with estimations being the weighted average of donors does, however, mitigate this to some extent. In conjunction with this, the analysis presented here is of a macro scale; however, to realise the potential of this approach for true antimicrobial optimisation more nuanced, relative and individualised studies will be required. Finally, given the high degree of missingness in the dataset, a number of clinically important characteristics had to be excluded. In particular, research shows Procalcitonin (PCT) and C-reactive protein (CRP) are useful biomarkers for determining when it is safe and appropriate to stop antibiotic therapy (Coelho et al. 2007; Oliveira et al. 2013; Rhee 2016; Schuetz et al. 2017). Neither of these were included as features or predicted as labels due to insufficient data. As such, this approach and the subsequent results could potentially be more powerful if applied to a more complete dataset focused on a narrow type of infection with defined variables and outcomes of interest.

4.6 Conclusion

This chapter discusses the development of an AI-driven model to estimate patient outcomes and treatment response under the contrasting scenarios of stopping or continuing antibiotic treatment. Retrospective results are in line with previous clinical studies, demonstrating that shorter antibiotic treatment durations are often non-inferior. Such an approach can act as an alternative towards traditional AI clinical decision support methods which focus on direct predictions of diagnoses, treatments or actions, by instead providing prescribers with an understanding of the potential impact of their actions. Such information is essential for optimised and

ethical antimicrobial prescribing as discussed in Section 2.4. This work therefore progresses the research field with an innovative and appropriate CDSS for infectious diseases. Further research is needed to discern the ability of such a tool to influence antimicrobial decision-making, but it holds promise to support antimicrobial optimisation and reduce the excessive use of antibiotics to tackle AMR.

References

1. Bolton, William J. et al. (2022). "Machine learning and synthetic outcome estimation for individualised antimicrobial cessation". In: *Frontiers in Digital Health* 4. ISSN: 2673-253X. URL: <https://www.frontiersin.org/articles/10.3389/fdgth.2022.997219>.
2. Tamma, Pranita D., Miller, Melissa A., and Cosgrove, Sara E. (Jan. 2019). "Rethinking How Antibiotics Are Prescribed: Incorporating the 4 Moments of Antibiotic Decision Making Into Clinical Practice". In: *JAMA* 321.2, pp. 139–140. ISSN: 0098-7484. DOI: 10.1001/jama.2018.19509. URL: <https://doi.org/10.1001/jama.2018.19509>.
3. Langford, Bradley J. and Morris, Andrew M. (Oct. 2017). "Is it time to stop counselling patients to "finish the course of antibiotics"?" In: *Canadian Pharmacists Journal : CPJ* 150.6, pp. 349–350. ISSN: 1715-1635. DOI: 10.1177/1715163517735549. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5661683/>.
4. Holmes, Alison H et al. (Jan. 2016). "Understanding the mechanisms and drivers of antimicrobial resistance". In: *The Lancet* 387.10014, pp. 176–187. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(15)00473-0. URL: <http://www.sciencedirect.com/science/article/pii/S0140673615004730>.
5. Mo, Yin et al. (June 2023). "Implications of reducing antibiotic treatment duration for antimicrobial resistance in hospital settings: A modelling study and meta-analysis". In: *PLOS Medicine* 20.6, e1004013. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1004013. URL: <https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1004013>.
6. Spellberg, Brad (Sept. 2016). "The New Antibiotic Mantra—"Shorter Is Better"". In: *JAMA Internal Medicine* 176.9, pp. 1254–1255. ISSN: 2168-6106. DOI: 10.1001/jamainternmed.2016.3646. URL: <https://doi.org/10.1001/jamainternmed.2016.3646>.
7. Vaughn, Valerie M. et al. (Aug. 2019). "Excess Antibiotic Treatment Duration and Adverse Events in Patients Hospitalized With Pneumonia". In: *Annals of Internal Medicine* 171.3, pp. 153–163. ISSN: 0003-4819. DOI: 10.7326/M18-3640. URL: <https://www.acpjournals.org/doi/full/10.7326/M18-3640>.
8. Curran, Jennifer et al. (Apr. 2022). "Estimating daily antibiotic harms: an umbrella review with individual study meta-analysis". In: *Clinical Microbiology and Infection* 28.4, pp. 479–490. ISSN: 1198-743X. DOI: 10.1016/j.cmi.2021.10.022. URL: [https://www.clinicalmicrobiologyandinfection.com/article/S1198-743X\(21\)00624-8/fulltext?dgcid=raven_jbs_etoc_email#secsectitle0155](https://www.clinicalmicrobiologyandinfection.com/article/S1198-743X(21)00624-8/fulltext?dgcid=raven_jbs_etoc_email#secsectitle0155).
9. Hanretty, Alexandra M. and Gallagher, Jason C. (2018). "Shortened Courses of Antibiotics for Bacterial Infections: A Systematic Review of Randomized Controlled Trials". In: *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* 38.6, pp. 674–687. ISSN: 1875-9114. DOI: 10.1002/phar.2118. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/phar.2118>.

10. Royer, Stephanie et al. (May 2018). "Shorter Versus Longer Courses of Antibiotics for Infection in Hospitalized Patients: A Systematic Review and Meta-Analysis". In: *Journal of hospital medicine* 13.5, pp. 336–342. ISSN: 1553-5592. doi: 10.12788/jhm.2905. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5945333/>.
11. Spellberg, Brad and Rice, Louis B. (Aug. 2019). "Duration of Antibiotic Therapy: Shorter Is Better". In: *Annals of internal medicine* 171.3, pp. 210–211. ISSN: 0003-4819. doi: 10.7326/M19-1509. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6736742/>.
12. Wald-Dickler, Noah and Spellberg, Brad (Oct. 2019). "Short-course Antibiotic Therapy—Replacing Constantine Units With "Shorter Is Better"". In: *Clinical Infectious Diseases* 69.9, pp. 1476–1479. ISSN: 1058-4838. doi: 10.1093/cid/ciy1134. URL: <https://doi.org/10.1093/cid/ciy1134>.
13. Yahav, Dafna et al. (Sept. 2019). "Seven Versus 14 Days of Antibiotic Therapy for Uncomplicated Gram-negative Bacteremia: A Noninferiority Randomized Controlled Trial". In: *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 69.7, pp. 1091–1098. ISSN: 1537-6591. doi: 10.1093/cid/ciy1054.
14. Janssen, Robin M. E. et al. (May 2022). "Why we prescribe antibiotics for too long in the hospital setting: a systematic scoping review". In: *Journal of Antimicrobial Chemotherapy*, dkac162. ISSN: 0305-7453. doi: 10.1093/jac/dkac162. URL: <https://doi.org/10.1093/jac/dkac162>.
15. Sax, Paul (Oct. 2010). *How to Figure Out the Length of Antibiotic Therapy*. URL: <https://blogs.jwatch.org/hiv-id-observations/index.php/how-to-figure-out-the-length-of-antibiotic-therapy/2010/10/22/>.
16. Peiffer-Smadja, N. et al. (May 2020). "Machine learning for clinical decision support in infectious diseases: a narrative review of current applications". In: *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 26.5, pp. 584–595. ISSN: 1469-0691. doi: 10.1016/j.cmi.2019.09.009.
17. Charani, Esmita et al. (Aug. 2021). "Optimising antimicrobial use in humans – review of current evidence and an interdisciplinary consensus on key priorities for research". In: *The Lancet Regional Health – Europe* 7, p. 100161. ISSN: 2666-7762. doi: 10.1016/j.lanepe.2021.100161. URL: <https://www.sciencedirect.com/science/article/pii/S2666776221001381>.
18. Rawson, Timothy Miles et al. (Dec. 2016). "Mapping the decision pathways of acute infection management in secondary care among UK medical physicians: a qualitative study". In: *BMC Medicine* 14.1, p. 208. ISSN: 1741-7015. doi: 10.1186/s12916-016-0751-y. URL: <https://doi.org/10.1186/s12916-016-0751-y>.

19. Pandolfo, Alyssa M. et al. (Mar. 2022). "Understanding decisions about antibiotic prescribing in ICU: an application of the Necessity Concerns Framework". In: *BMJ Quality & Safety* 31.3, pp. 199–210. ISSN: 2044-5415, 2044-5423. doi: 10.1136/bmjqs-2020-012479. URL: <https://qualitysafety.bmj.com/content/31/3/199>.
20. Goldberger, A. L. et al. (June 2000). "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals". In: *Circulation* 101.23, E215–220. ISSN: 1524-4539. doi: 10.1161/01.cir.101.23.e215.
21. Johnson, Alistair et al. (2021). *MIMIC-IV*. doi: 10.13026/s6n6-xd98. URL: <https://physionet.org/content/mimiciv/1.0/>.
22. Qian, Zhaozhi et al. (2021). "SyncTwin: Treatment Effect Estimation with Longitudinal Outcomes". In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., pp. 3178–3190.
23. Abadie, Alberto and Gardeazabal, Javier (Mar. 2003). "The Economic Costs of Conflict: A Case Study of the Basque Country". In: *American Economic Review* 93.1, pp. 113–132. doi: 10.1257/000282803321455188. URL: <https://www.aeaweb.org/articles?id=10.1257/000282803321455188>.
24. Kreif, Noémi et al. (Dec. 2016). "Examination of the Synthetic Control Method for Evaluating Health Policies with Multiple Treated Units". In: *Health Economics* 25.12, pp. 1514–1528. ISSN: 1057-9230. doi: 10.1002/hec.3258. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5111584/>.
25. Bouttell, Janet et al. (Aug. 2018). "Synthetic control methodology as a tool for evaluating population-level health interventions". In: *J Epidemiol Community Health* 72.8, pp. 673–678. ISSN: 0143-005X, 1470-2738. doi: 10.1136/jech-2017-210106. URL: <https://jech.bmj.com/content/72/8/673>.
26. Paszke, Adam et al. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems* 32. Curran Associates, Inc., pp. 8024–8035.
27. Kingma, Diederik P. and Ba, Jimmy (2014). "Adam: A Method for Stochastic Optimization". In: URL: <http://arxiv.org/abs/1412.6980>.
28. Gier, R. de et al. (Sept. 1995). "A sequential study of intravenous and oral Fleroxacin for 7 or 14 days in the treatment of complicated urinary tract infections". In: *International Journal of Antimicrobial Agents* 6.1, pp. 27–30. ISSN: 0924-8579. doi: 10.1016/0924-8579(95)00011-V. URL: <https://www.sciencedirect.com/science/article/pii/092485799500011V>.
29. Peterson, Janet et al. (Jan. 2008). "A Double-Blind, Randomized Comparison of Levofloxacin 750 mg Once-Daily for Five Days With Ciprofloxacin 400/500 mg Twice-Daily for 10 Days for the Treatment of Complicated Urinary Tract Infections and Acute Pyelonephritis". In: *Urology* 71.1, pp. 17–22. ISSN: 0090-4295. doi: 10.1016/j.urology.2007.09.002. URL: <https://www.sciencedirect.com/science/article/pii/S0090429507021139>.

30. Dimopoulos, George et al. (Dec. 2013). "Short- vs long-duration antibiotic regimens for ventilator-associated pneumonia: a systematic review and meta-analysis". In: *Chest* 144.6, pp. 1759–1767. issn: 1931-3543. doi: 10.1378/chest.13-0076.
31. Pugh, Richard et al. (Aug. 2015). "Short-course versus prolonged-course antibiotic therapy for hospital-acquired pneumonia in critically ill adults". In: *The Cochrane Database of Systematic Reviews* 8, p. CD007577. issn: 1469-493X. doi: 10.1002/14651858.CD007577.pub3. url: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7025798/>.
32. Drekonja, Dimitri M. et al. (July 2021). "Effect of 7 vs 14 Days of Antibiotic Therapy on Resolution of Symptoms Among Afebrile Men With Urinary Tract Infection: A Randomized Clinical Trial". In: *JAMA* 326.4, pp. 324–331. issn: 0098-7484. doi: 10.1001/jama.2021.9899. url: <https://doi.org/10.1001/jama.2021.9899>.
33. NICE (Oct. 2018). "Overview | Urinary tract infection (lower): antimicrobial prescribing | Guidance". In: url: <https://www.nice.org.uk/guidance/ng109>.
34. NICE (Sept. 2019). "Overview | Pneumonia (hospital-acquired): antimicrobial prescribing | Guidance". In: url: <https://www.nice.org.uk/guidance/ng139>.
35. Pauwels, Ines et al. (Sept. 2021). "Assessing the impact of the Global Point Prevalence Survey of Antimicrobial Consumption and Resistance (Global-PPS) on hospital antimicrobial stewardship programmes: results of a worldwide survey". In: *Antimicrobial Resistance & Infection Control* 10.1, p. 138. issn: 2047-2994. doi: 10.1186/s13756-021-01010-w. url: <https://doi.org/10.1186/s13756-021-01010-w>.
36. Rocheteau, Emma, Liò, Pietro, and Hyland, Stephanie (2021). "Temporal Pointwise Convolutional Networks for Length of Stay Prediction in the Intensive Care Unit". In: *Proceedings of the Conference on Health, Inference, and Learning*. CHIL '21. Virtual Event, USA: Association for Computing Machinery, pp. 58–68. isbn: 9781450383592. doi: 10.1145/3450439.3451860.
37. Coelho, Luís et al. (2007). "Usefulness of C-reactive protein in monitoring the severe community-acquired pneumonia clinical course". In: *Critical Care* 11.4, R92. issn: 1364-8535. doi: 10.1186/cc6105. url: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2206486/>.
38. Oliveira, Carolina F. et al. (Oct. 2013). "Procalcitonin versus C-reactive protein for guiding antibiotic therapy in sepsis: a randomized trial". In: *Critical Care Medicine* 41.10, pp. 2336–2343. issn: 1530-0293. doi: 10.1097/CCM.0b013e31828e969f.
39. Rhee, Chanu (Dec. 2016). "Using Procalcitonin to Guide Antibiotic Therapy". In: *Open Forum Infectious Diseases* 4.1, ofw249. issn: 2328-8957. doi: 10.1093/ofid/ofw249. url: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5414114/>.

40. Schuetz, Philipp et al. (Oct. 2017). "Procalcitonin to initiate or discontinue antibiotics in acute respiratory tract infections". In: *The Cochrane Database of Systematic Reviews* 2017.10, p. CD007498. ISSN: 1469-493X. doi: 10.1002/14651858.CD007498.pub3. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6485408/>.

Chapter 5

Intravenous to Oral Switch Decision Support

Abstract

Maximising targeted oral therapy and reducing the use of indwelling vascular devices for intravenous (IV) administration is a key part of preventing healthcare-associated infections and conducting antimicrobial stewardship. Oral therapy has frequently demonstrated equal effectiveness to IV and has additional benefits including reduced side effects, cost savings, and increased patient comfort. Despite IV-to-oral switching (IVOS) being a well-established stewardship focus the uptake of practising early oral switching remains low with limited data-driven research in this area. Part of the reason for this is a poor understanding on an individual patient level when someone is suitable for switching from IV-to-oral antibiotic treatment. We hypothesised that a simple, fair, and interpretable artificial intelligence (AI) model could support clinicians with IVOS decisions for individual patients.

To tackle this non-trivial and not standardised problem we developed machine learning models to predict when a patient could switch based on routinely collected clinical parameters. Two publicly available datasets focused on the ICU setting were used, with the research also expanded to data originating from Imperial College Healthcare NHS Trust (ICHT). Results show our data processing pipelines and models are generalisable across datasets, with the best model achieving a mean Area Under the Receiver Operating Characteristic curve (AUROC) of 0.80 (SD 0.01) on the MIMIC hold-out set while not being biased to individuals' protected characteristics. Interpretability methodologies were utilised to create clinically useful visual explanations. Models trained on ICHT were prospectively evaluated with new data and patients from a point prevalence survey.

This chapter’s research was published in Nature Communications as ‘Personalising intravenous to oral antibiotic switch decision-making through fair interpretable machine learning’ (Bolton 2023; Bolton et al. 2024). Furthermore, it was presented at the European Society of Clinical Microbiology and Infectious Diseases (ESCMID) Global 2023, 2024 and 2025. Figure 5.1 provides a graphical abstract overview of this chapter.

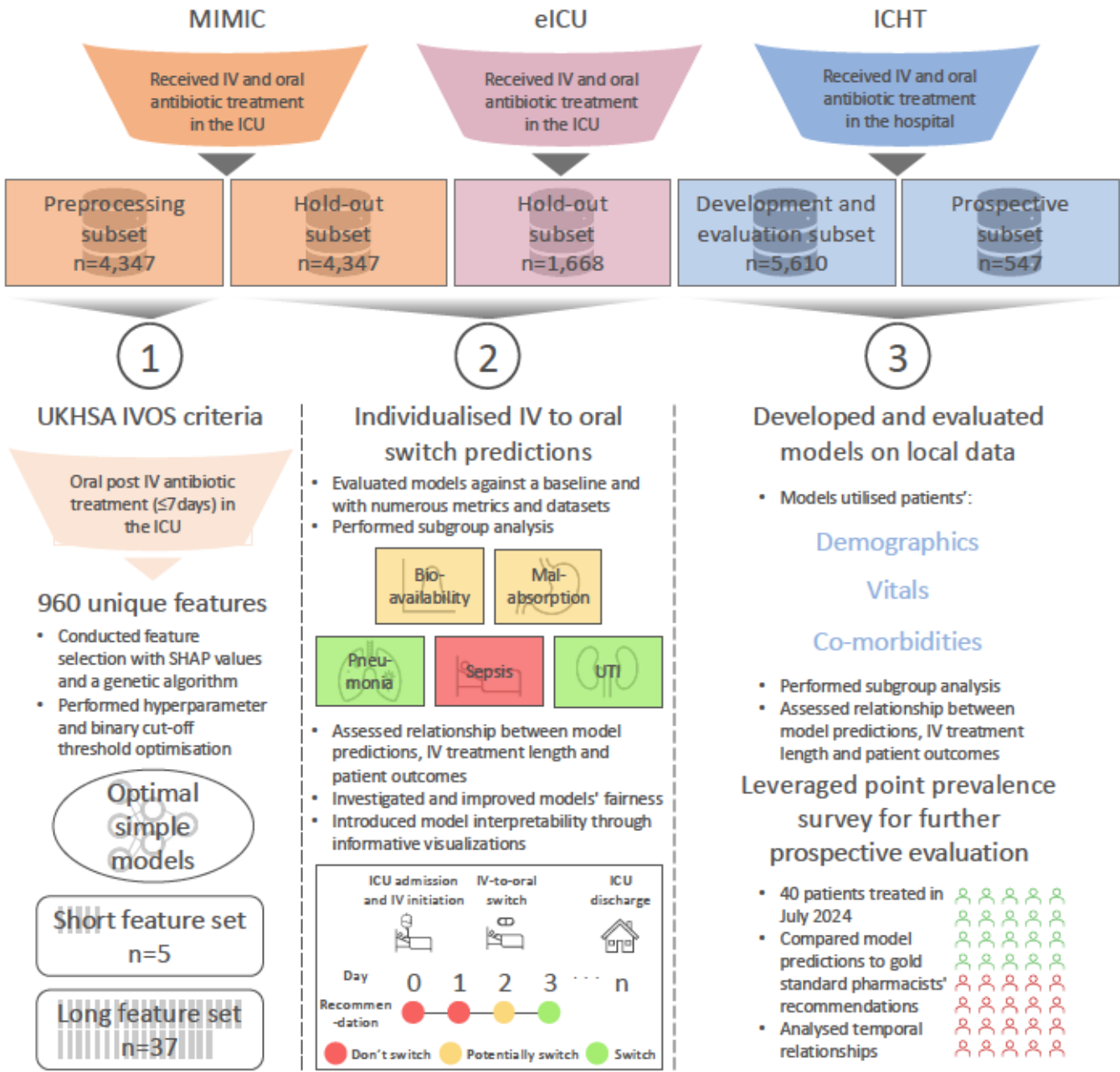


Figure 5.1: Overview of the steps taken to develop machine learning models for antimicrobial switch decision-making.

5.1 Chapter objectives

- *Develop a data pre-processing and machine learning pipeline that utilises the temporal nature of routinely collected clinical data and results in informative feature sets and simple IVOS predictive models*
- *Retrospectively analyse the potential impact of early switching on patient outcomes*
- *Assess the generalisability of our models through evaluations on subgroups and different datasets*
- *Create visualisations to help users interpret the models' predictions*
- *Investigate and mitigate biases within our models*
- *Conduct an initial prospective evaluation of a model*

5.2 Introduction

To combat antimicrobial resistance (AMR), antimicrobial stewardship aims to optimise drug use to prolong current therapeutic effectiveness (Dyar et al. 2017). One key aspect of antimicrobial stewardship is the route of administration. Hospitalised patients are commonly given empirical intravenous (IV) antibiotic therapy upon admission due to their higher bioavailability, and uncertainty surrounding a potential infection (Cyriac et al. 2014). Later in the treatment regime, once the patient has stabilised and the infection is understood, their antibiotics may be switched to an oral route of administration. There is a well-described focus to switch from IV-to-oral administration as early as possible and to use more oral drugs when appropriate, as they are often equally effective and can reduce side effects such as thrombus, or line occlusion

(Oosterheert et al. 2006; UK Health Security Agency 2015; Davar et al. 2022; Gilchrist et al. 2022). In a range of infectious diseases that were traditionally treated with IV only (e.g., bacteremia, endocarditis, and bone and joint infections), recent studies have demonstrated that oral therapy can be non-inferior to IV in efficacy (Iversen et al. 2019; Li et al. 2019; Spellberg et al. 2020; Platts et al. 2022; Wald-Dickler et al. 2022; Kaasch et al. 2023). Furthermore, reducing the unnecessary use of indwelling IV devices is a well-established patient safety and infection prevention priority to minimise the risk of healthcare-associated infections (HCAIs) (Ray-Barruel et al. 2019). Beyond the infection complications of IV catheters, oral administration is more convenient for the patient, reduces the workload of nurses, and allows for easy discharge from the hospital. Furthermore, oral therapy is cheaper and more cost-effective (McMeekin et al. 2020). The UK Health Security Agency has published national antimicrobial IV-to-oral switch (IVOS) criteria for early switch (UK Health Security Agency 2022). These requirements were developed based on expert consensus and primarily revolve around the patient's clinical and infection markers improving as well as specific points with regard to absorption, bioavailability, and infection type.

Despite significant evidence in support of IVOS, the uptake of early oral therapy remains low, primarily due to fear and a lack of awareness, education, and empowerment (Shrayteh et al. 2014; Rawson et al. 2016; Hospenthal et al. 2020; Buis et al. 2022; Pandolfo et al. 2022; Hamilton et al. 2024). Clinicians are often guided by experience and memories of early switch failures will have a greater imprint than successes (Sendi et al. 2023). These behavioural factors and the historical perception of the superiority of IVs have hindered IVOS programmes. Furthermore, beyond guidelines (UK Health Security Agency 2022) there is a lack of research into data-driven IV-to-oral decision support systems (Rawson et al. 2017). Given this, we investigated whether a machine learning-based clinical decision support system (CDSS) could assist antibiotic switch decision-making at the individual patient level. More specifically, neural network models were developed to predict, based on routinely collected clinical parameters, whether a patient could be suitable for switching from IV-to-oral antibiotics on any given day. This research utilised the co-morbidity representation pipeline previously discussed in this thesis in Chapter 3.

Many CDSSs that use machine learning have been developed to help with other aspects of antimicrobial use (Rawson et al. 2017; Peiffer-Smadja et al. 2020; Bolton et al. 2022a); however, limited clinical use and adoption has been observed (D'Hondt et al. 2022). As such, when tackling this problem, we wanted to ensure that our CDSS solution was simple, fair, interpretable, generalisable, and prospectively validated to maximise the ability for clinical translation. By simple we mean the AI approach can be understood by non-experts, to this extent a different and more traditional machine learning methodological approach was taken in this work when compared to Chapter 4, to reduce complexity and confounding and streamline research translation. Fair infers model performance is not biased to particular sensitive attributes or protected characteristics, while interpretability means that predictions can be more easily understood, explained, and trusted. Finally, a model is generalisable if it can potentially be applied to many healthcare settings with consistent results, with prospective validation requiring evaluation on new future datasets. We imagine that by providing individualised antibiotic switch estimations, such a system could support patient-centric decisions and provide assurance on whether switching could be appropriate or not in a given clinical context.

5.3 Methods

5.3.1 Datasets

For the first part of this chapter's research, two publicly available large de-identified real-world clinical datasets containing routinely collected EHR information were used. MIMIC-IV (Goldberger et al. 2000; Johnson et al. 2021), was used for feature selection, model optimisation, and hold-out testing. Meanwhile, the eICU Collaborative Research Database (Goldberger et al. 2000; Pollard et al. 2018; Pollard et al. 2019), was used for transfer learning to confirm generalisability. For both datasets, the patient population was filtered to those who received IV and oral antibiotic treatment within the intensive care unit (ICU) (IV treatment was limited to < 8 days due to the exponential decrease in the frequency of patients). Unfortunately, the datasets used in this research do not contain explicit information on if, when, or why

an IV-to-oral switch was considered. However, by utilising the available prescribing data and taking what the clinicians actually did as a label we can approximate the prescribing behaviour and train a machine learning model. We therefore focused on making a route of administration prediction for each day the patient was on antibiotics given clinical decisions regarding antimicrobial treatment are most often made on a daily basis (Agency for Healthcare Research and Quality 2019). As such negative switch labels were defined as each day a patient was on IV antibiotics, while positive labels were defined as every other day (i.e., where the patient was on oral but not IV antibiotics). The antibiotic spectrum index (ASI) (Gerber et al. 2017) was used to assess the average breadth of activity for IV and oral treatment regimes. Larger values indicate a broader spectrum, while smaller values correlate with a more targeted activity. By looking at the antibiotic spectrum index (ASI) on the day before switching and the first day of only oral administration we can understand how a change in route of administration is most often associated with the ASI.

5.3.2 Feature selection

Our aim was to make models that through utilising routinely available patient vitals could act as a starting point for the decision-making process and flag when a switch could be considered for a particular patient. The UK Health Security Agency (UKHSA) IVOS criteria (UK Health Security Agency 2022) were analysed and ten related features were extracted from the datasets. Specifically: heart rate, respiratory rate, temperature, O2 saturation pulseoxymetry, systolic blood pressure, diastolic blood pressure, mean blood pressure, GCS motor response, GCS verbal response, and GCS motor eye opening. White Cell Count and C-Reactive Protein were excluded due to missing data, the requirement of a blood test, and UKHSA guidelines stating that they should be considered but are not necessary for a switch. Other important aspects of the guidelines, such as infection type and absorption status, were also not included as input features to the model as much of this data was unavailable or collected in a way that makes it unsuitable for machine learning. Furthermore, the evidence surrounding these is constantly changing (Spellberg et al. 2020; Wald-Dickler et al. 2022; Korzilius et al. 2023). Our aim was

to create a simple and generalisable model that uses only routinely available patient data and has the potential to be used in many different healthcare settings. The Canonical Time-series Characteristics (Catch22) methodology (Lubba et al. 2019) (along with the mean and variance) was utilised through `sktime` (Löning et al. 2019; Löning et al. 2022) to transform temporal data into daily tabular values. Catch22 provides a set of diverse, interpretable and minimally redundant feature representations for temporal data, allowing for the dynamical properties of time series to be easily and concisely mapped to feature vectors for use in non-temporal models. This temporal-to-tabular transformation was done for each specific day and the entire current stay. In addition, the difference between the transformed values for a given day and the preceding day was calculated. SHapley Additive exPlanations (SHAP) values (Lundberg et al. 2017) and a genetic algorithm (Vafaie et al. 1992) were then used for feature selection. Specifically, a large neural network with 851,729 trainable parameters was initially trained, SHAP values were calculated, and those features with a value greater than or equal to 0.5 were selected for use in the genetic algorithm (given running the genetic algorithm on all 960 features would be extremely computationally expensive). A genetic algorithm mimics the process of natural selection to solve optimization problems and find the best-performing set of features. It begins by initializing a population of individuals, each representing a set of features, which are all evaluated using a fitness function. The algorithm then iteratively selects individuals based on their fitness (those with greater fitness are more likely to be selected), applies genetic operators such as crossover (combining parts of two individuals) and mutation (randomly altering parts of an individual), and iterates. Over time the population evolves and converges towards a near-optimal individual. The genetic algorithm was optimised for the AUROC and was run twice. Once for a simple set of 5 features and the second time without a limitation on the number of features. 10 iterations with 50 individuals and 25 iterations with 20 individuals were used, respectively. By using this feature engineering and selection approach with non-temporal neural network models we can create simple, interpretable models that have a low resource footprint and can be utilised by clinicians in real-world healthcare settings.

5.3.3 Model development

The MIMIC-IV EHR dataset was randomly split (50%, 50%) based on patients' ICU 'stay_id' into a preprocessing and a hold-out set in order to generalise switching prescribing behaviour and get a reliable unbiased estimate of the models' performance given the selected hyperparameters and feature set. The preprocessing set was split randomly into training, validation, and testing sets for feature selection as discussed above with Pytorch (a machine learning library for the Python programming language, Paszke et al. 2019) used to create the neural networks. After feature selection, Optuna (Akiba et al. 2019) with the objective of maximising the AUROC was used to select the model's hyperparameters, and optimal alternative cut-off thresholds were determined from the preprocessing validation subset. Youden's Index (Youden 1950) was used to optimise the AUROC, while finding the point where precision, recall and the F1 score were equal was used as a stringent cut-off for reducing the false positive rate (FPR). Multiple thresholds allow for a simple traffic light explanation system to be presented to clinicians with regards to if a switch could be appropriate at a particular time. Where green represents switch, yellow represents potentially switch and red represents don't switch, as shown in Figure 5.8. Subsequently, once the features and models were finalised the unseen hold-out set was randomly split 10 times into stratified training, validation, and testing sets for evaluation. Specifically, 10 naive models based on the previously identified features and model hyperparameters were trained and the final performance of such models was evaluated. Synthetic minority oversampling technique (Chawla et al. 2002) was used during training to address label class imbalance. The Adam optimiser (Kingma et al. 2014) was used with binary cross entropy with logits loss. Training utilised 10 epochs, and the model with the greatest AUROC on the validation dataset was selected as the final model to obtain results on the unseen test set.

5.3.4 Model evaluation

Standard ML metrics were used to evaluate model performance. Specifically for the switch classification task the AUROC, accuracy, precision, true positive rate (TPR), FPR, F1 score, and Area Under the Precision Recall curve (AUPRC) were calculated. The details and formulas for these metrics are provided in the Appendix (Section A.1). The standard deviation was calculated to indicate the variation in the results. To provide a baseline for comparison two infection markers that are clearly defined within the latest guidelines (UK Health Security Agency 2022) were separately also used for predicting when switching could be appropriate in each patient. Specifically, the patient's temperature must have been between 36°C and 38°C for the past 24 hours and their National Early Warning Score (NEWS) must be decreasing, upon which a switch would then be suggested for the rest of that patient's stay. It was not possible to include all the aspects of the guidelines as many of them are ambiguous and not recorded within the data. However, it acts as a fair comparison to our models as it utilises similar patient data and actually contains additional information that is not fed into our models, such as the inspired O2 fraction. The model with the highest AUROC on the hold-out test set was selected as the final model and its respective hold-out split was used to break the distribution of labels and predictions down by IV treatment duration, to evaluate how predictions compare temporally to the real labels and discern when the model performed well versus poorly. To understand the value of the models' switch predictions and how they relate to patient outcomes, the difference in days between our predicted switch events and real switch events was calculated and the mean ICU length of stay (LOS) and mortality outcomes were taken. Furthermore, we analysed whether there was a variation in the remaining ICU LOS for patients who remained on IV versus those who switched on that day. This was done for dates with 2 to 7 days of previous IV treatment based on the dissimilarity between the model predictions and the labels on those days. For statistical analysis, the non-parametric Wilcoxon rank-sum (Mann-Whitney U) test with alpha set at 0.05, was used to determine whether the difference in means was statistically significant, given the non-normal data distribution. Models were evaluated using functions and metrics from the Scikit-learn and SciPy libraries (Pedregosa et al. 2011; Virtanen et al. 2020).

To further validate the findings, evaluations were performed on specific patient populations and infectious diseases within MIMIC. Obtaining target drug exposure against the pathogenic organism is important during antibiotic treatment and is often a concern when deciding to switch to oral administration (Broek et al. 2021). Antibiotics with incomplete oral absorption (bioavailability $<90\%$) were determined by consultation with a pharmacist and a literature search on PubMed, the Electronic Medicines Compendium, and UpToDate. The final list of antibiotics with incomplete oral absorption is shown in the Supplementary Table A.2. If patients have problems with enteral absorption, oral antibiotic therapy is less likely to be suitable (UK Health Security Agency 2022), so total parenteral nutrition was used as a proxy for those with poor oral absorption (malabsorption). Patients with UTIs, pneumonia, and sepsis were also selected because they were highly prevalent in the dataset, and UTIs and pneumonia are commonly treated with oral antibiotics but sepsis sees less oral use. The hospital ICD diagnostic codes were used to identify infection diagnosis. Note that infection types in MIMIC are linked to the hospital stay 'hadm_id' only and not the specific ICU stay 'stay_id' as diagnoses are only coded for billing purposes upon hospital discharge. eICU is a different dataset from MIMIC with a unique data distribution, distinct hospitals, and a separate patient population. These differences can often cause problems for machine learning models but allow us to validate our features and modelling approach on an external dataset. The best-performing short and long models from the MIMIC hold-out set were evaluated on data extracted from the eICU database via transfer learning to re-train and subsequently test the models. The same features and data processing pipeline was used for eICU and transfer learning utilised the same procedure as with evaluation on the MIMIC hold-out set, except the model parameters were initialised with the best performing final MIMIC trained models. eICU was selected as both it and MIMIC originate from the USA and it contained the data necessary for input into the models.

The best-performing final short and long models trained on the MIMIC hold-out set were used for fairness and interpretability research. Simplex (Crabbe et al. 2021) is as a post-hoc explanation methodology to extract similar patient examples from the training data, their importance, and the contribution of each feature for each example. It was used to pro-

vide data to help users interpret our models' predictions. To this extent, first, the corpus and test latent representations (hidden activations in the neural network's penultimate layer) are computed. SimplerEx was then fitted to minimise weights that mapped the corpus to the test latent representations and finally, the integrated Jacobian decomposition for a particular patient was calculated. Simpler patient examples, their importance, and feature contribution can then be provided alongside model predictions. To simplify visualizations and show only a couple of similar patients, examples with an importance greater than 0.1 were shown. This method was utilised as it is the cutting-edge of local, instance-based explanations and case-based approaches have previously seen utility in healthcare-focused decision support systems (Rawson et al. 2021). To assess model fairness, the demanding equalised odds (EO) metric was used, given we want to acknowledge and ideally minimise false positives as well as obtain equal performance across sensitive attribute classes. We defined that equalised odds (EO) was achieved for a given sensitive attribute group if the TPR was not less than 0.1 from the global average and the FPR was not greater than 0.1 from the global average. EO was assessed utilising the 1st threshold for the sensitive attributes age (grouped into brackets based on the nearest decade), sex, race, insurance, language, and marital status. Threshold optimisation (Hardt et al. 2016) was then employed to see if the fairness of the models could be improved. This method uses an optimisation algorithm to learn thresholds for each group within a sensitive attribute (e.g., male and female for sex) that satisfy a specified fairness constraint (e.g., equalised odds) while also maximising the overall utility of the model (e.g., through AUROC). The new learnt thresholds can be used at evaluation to convert a classifier's probability outputs into binary predictions that respect the fairness constraint. In this research, the postprocessing thresholdoptimizer method from the fairlearn Python package (Bird et al. 2020) was used, with the balanced accuracy objective and either the equalised odds, FPR parity or TPR parity constraint. The threshold optimisation method was chosen because we would have access to the sensitive attribute data at inference time (alternative methods such as fair representation learning avoid this), it can work with any model architecture, does not require re-training, and explicitly enforces fairness constraints.

5.3.5 Imperial College Healthcare NHS Trust

Once our approach to developing machine learning models for predicting IV-to-oral antibiotic switching was proven with publicly available datasets, we applied a similar methodological approach to data from Imperial College Healthcare Trust (ICHT), enabling switch predictions to be made for this local non-ICU patient population. Data was accessed through the ICHT Clinical Analytics, Research, and Evaluation (iCARE) environment, which contains de-identified real-world clinical data for patients with an inpatient encounter at ICHT. Patients were filtered to those who received oral treatment after IV antibiotics during their stay (not restricted to the ICU), with IV treatment restricted to < 8 days due to the exponential decrease in frequency of patients. 10 clinical variables based on the UKHSA IVOS criteria (UK Health Security Agency 2022) were extracted for each patient; heart rate, respiratory rate, temperature, SpO₂, systolic blood pressure, diastolic blood pressure, mean blood pressure, NEWS conscious level score, NEWS supplemental oxygen levels score, GCS. In addition, the patients' demographics (age, sex, ethnicity, index of multiple deprivation decile), type of infection (sepsis, pneumonia, UTI, cellulitis) and co-morbidities were collected. Given that this patient population was not in the ICU, vital measurements were more infrequent than MIMIC and eICU. To tackle these irregularly spaced time series, variables were aggregated by 6-hour windows (e.g., 12 am to 6 am) and forward filled, so there were 4 values for each variable per day per patient. The Catch22 methodology (Lubba et al. 2019) was then utilised as described in Section 5.3.2. However, in this case, to more closely align with the UKHSA IVOS criteria (UK Health Security Agency 2022) Catch22 was applied at 24 hours, 48 hours, and then daily thereafter using data up to 12 pm or 6 am depending on availability (as hospital ward rounds are normally conducted in the morning (NHS n.d.)). The final 253 vital features were selected based on removing those that had a low diversity of values. Specifically, features were removed if their unique values made up $\leq 10\%$ of their total values. This was done to help models focus on more informative features and reduce redundancy. The model architecture initially passes the vitals and demographics (including infection type) features into separate neural networks with patient co-morbidities being processed through a set transformer and SNOMED disease embeddings as described

in Chapter 3. The resulting embeddings are then concatenated before being passed into a final neural network model for prediction. An overview of the architecture of the final model is shown in Figure 5.2. 10-fold cross-validation was used to evaluate the model's performance. The optimal cut-off was determined with Youden's index (Youden 1950). Training utilised a batch size of 512, learning rate of 0.0001, Root Mean Squared Propagation for optimisation and binary cross entropy with logits loss. As with MIMIC and eICU, analysis was done on predictions by IV treatment length, in-patient hospital LOS and the subgroups' infection type and bad bioavailability. Furthermore, after this work was conducted, on the 26th of June 2023 iCARE released new data for patients admitted to ICHT before June 2023 (the original iCARE data only contained admissions before 2022). Therefore, the model could be run and prospectively evaluated on these new patients' data.

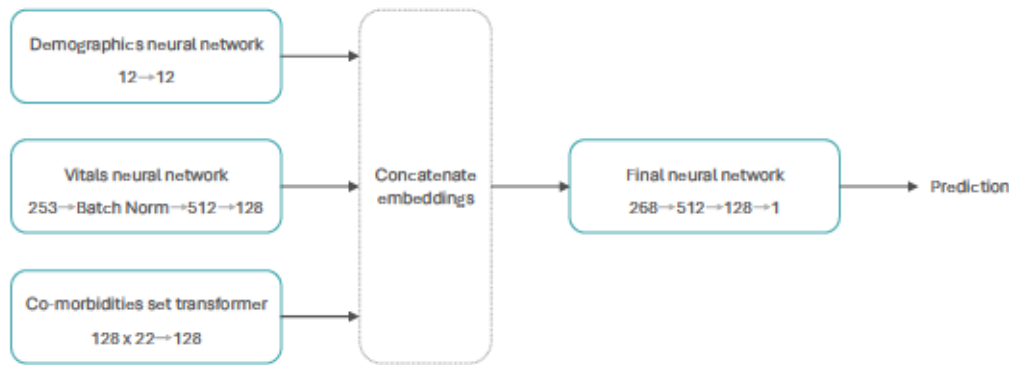


Figure 5.2: An overview of the structure of the model used to make IVOS predictions for patients in ICHT. Neural networks used the ReLU activation function and dropout. The set transformer used induced set attention followed by pooling by multihead attention (Lee et al. 2019), each with 4 heads and a hidden dimension of size 160, finally a linear decoder layer was applied.

5.3.5.1 Point prevalence survey

To prospectively validate the model, we worked with Mark Gilchrist and the antimicrobial pharmacy team at ICHT. Every year they conduct a point prevalence survey to assess antimicrobial prescribing within the trust. Outcomes include if prescriptions adhere to guidelines and are appropriately reviewed and stopped. One piece of information collected that is pertinent to this work is whether or not and when antimicrobial pharmacists think patients are suitable for IVOS. 32 unique patients on IVs were determined to be eligible for IVOS in the July 2024 point

prevalence survey. 20 patients (10 from Medicine and Integrated Care and 10 from Surgery and Cancer) were randomly selected from this list, due to time and resource constraints. 20 control patients who were also on IV antibiotics but were not deemed to be suitable for IVOS were also randomly selected (while controlling for patient speciality to account for any behavioural differences in prescribing). All random selection was done through a random number generator. Under an approved service evaluation project titled 'Evaluating the IV-to-oral antibiotic switching stewardship service and understanding the potential impact of a clinical decision support tool', data for these 40 patients associated with their inpatient stay during the point prevalence survey was extracted, processed and passed through the IVOS machine learning model trained on retrospective ICHT data, to prospectively evaluate its performance against antimicrobial pharmacists' opinions. The temporal relationship between the models' predictions, real-world actions, pharmacists' recommendations and IV treatment length was analysed, with the non-parametric Wilcoxon rank-sum test (due to the non-normal data distribution, alpha 0.05) used to compare the mean current antimicrobial treatment length between the route of administration groups.

5.4 Results

5.4.1 MIMIC and eICU

5.4.1.1 Data

8,694 unique ICU stays, were extracted from the MIMIC dataset (Goldberger et al. 2000; Johnson et al. 2021), with 1,668 from eICU (Goldberger et al. 2000; Pollard et al. 2018; Pollard et al. 2019). 10 clinical features were selected based on the UK antimicrobial IVOS criteria (UK Health Security Agency 2022). Transformation of these temporally dynamic values into time series features as detailed in Section 5.3.1 resulted in 960 unique features for each day of each patient's stay. Details of the MIMIC preprocessing, MIMIC hold-out, and eICU datasets are shown in Table 5.1. All the datasets are relatively equally balanced, however, the specific

antibiotic utilisation distribution varies between MIMIC and eICU. Furthermore, eICU represents a more unwell population given the higher proportion of life-threatening infections, such as sepsis.

Table 5.1: Dataset demographics and statistics.

Statistic		Dataset		
		MIMIC preprocessing	MIMIC hold-out	eICU
Number of stays		4,347	4,347	1,668
Age (mean)		65.30 (SD 15.16)	65.44 (SD 15.23)	64.74 (SD 15.91)
Length of stay (mean)		3.14 (SD 2.78)	3.12 (SD 2.71)	3.17 (SD 2.83)
Sex (%)	Male	58.82	58.92	50.30
	Female	41.18	41.08	49.70
Race (%)	White	67.97	68.09	78.77
	Black	9.55	9.92	15.08
	Unknown	10.28	9.35	3.50
	Other	6.22	5.48	1.39
	Hispanic	3.27	3.94	0.84
	Asian	2.34	2.94	0.42
	Native American	0.36	0.27	
Antimicrobial treatment length (mean)	Overall	3.34 (SD 2.16)	3.29 (SD 2.01)	2.97 (SD 1.94)
	IV	2.79 (SD 1.47)	2.75 (SD 1.46)	2.46 (SD 1.55)
	Oral	2.76 (SD 1.90)	2.69 (SD 1.81)	1.14 (SD 1.78)
Infection type (% most common shown)	UTI	65.10	64.69	10.19
	Pneumonia	26.30	26.62	31.71
	Sepsis	18.20	19.63	32.91
IV antibiotics (% those with a frequency of greater than 5% shown)	Vancomycin	35.96	35.56	41.98
	Cefepime	12.28	14.03	2.86
	Cefazolin	12.47	12.08	1.71
	Piperacillin-Tazobactam	9.35	8.42	8.92
	Ceftriaxone	7.82	8.04	6.10
	Levofloxacin	1.76	1.99	14.12
	Azithromycin	22.96	23.18	12.24
Oral antibiotics (% those with a frequency of greater than 5% shown)	Vancomycin	13.45	14.23	12.37
	Ciprofloxacin	12.77	12.14	-
	Levofloxacin	12.69	12.37	51.70
	Sulfameth/Trimethoprim	11.82	10.04	-
	Metronidazole	7.90	9.32	22.72
Microbiology (% those with a frequency of greater than 5% shown)	Positive growth	32.76	33.07	13.01
	<i>Escherichia coli</i>	6.86	6.51	-
	<i>Staphylococcus aureus</i>	6.69	7.33	-

Note that infection types in MIMIC are determined through 'hadm_id' and are not definitively linked to the antimicrobial switch of interest as diagnoses are only coded for billing purposes upon hospital discharge. This results in totals over 100% as patients have multiple infection episodes. Sex was determined based on the information contained within the dataset.

The ASI (Gerber et al. 2017) demonstrates if an antibiotic treatment regime shows broad or narrow activity. A statistically significant (p -value <0.01) difference was found between the ASI for IV and oral antibiotics upon switching (means of 8.25 and 5.89 respectively, medians of 8 and 5) through the non-parametric Wilcoxon rank-sum test (due to the non-normal data distribution). In addition, the majority of patients (70.03%) see a decrease in their treatment’s ASI and therefore a narrowing of activity upon switching, with a mean and median decrease of 23.04% and 27.27% respectively, although this was highly variable (Figure 5.3).

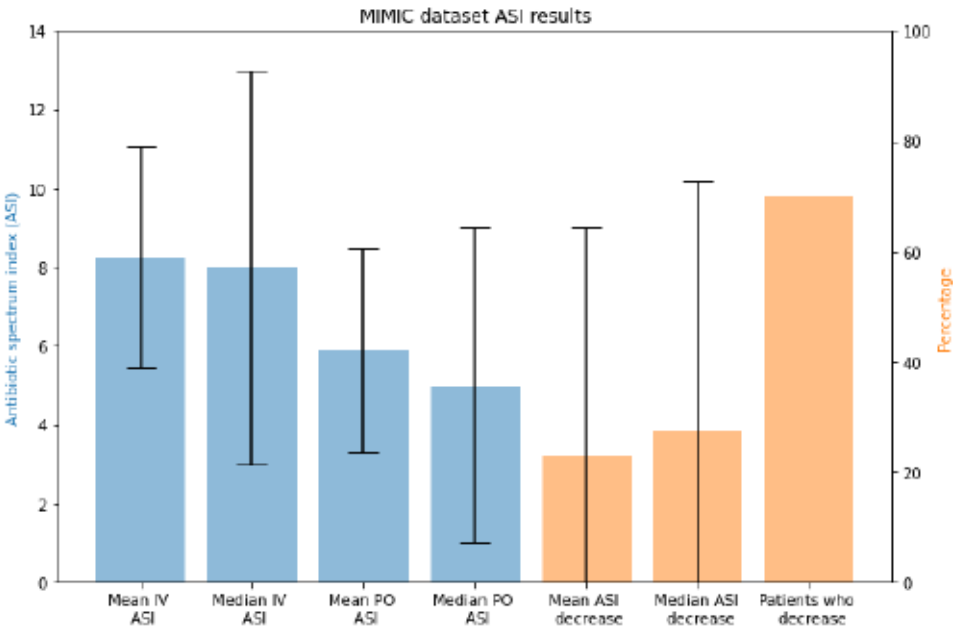


Figure 5.3: MIMIC antibiotic spectrum index results for the day before switching (IV) and the day after switching (oral). Error bars indicate the standard deviation.

5.4.1.2 Feature and model optimisation

The first large neural network trained on the preprocessing training subset achieved an AUROC of 0.76 on the preprocessing test subset. SHAP values (Lundberg et al. 2017) were calculated (Supplementary Figure A.6) with the top 98 features included in the genetic algorithm feature selection approach. The genetic algorithm produced two sets of features, one short set, containing only 5 features, and another longer set of 37. The short and long feature sets achieved an AUROC of 0.80 and 0.82 on the preprocessing test subset, respectively. The final features for each set are shown in Table 5.2 with the respective Catch22 time-series transformations

listed in Supplementary Table A.3. During hyperparameter optimisation, our objective was to find the most simple models whilst maintaining performance. For both feature sets, this was achieved with new less complex models being found to achieve the same AUROC. The final hyperparameters of each model are shown in Table 5.3. Finally, alternative cut-off thresholds were explored for both models to maximise the AUROC and minimise the FPR (Figure 5.4). This then allows a traffic light system to be employed at deployment for simplicity and interpretability (Figure 5.8). Youden's index (Youden 1950) which optimises the AUROC, found the 1st cut-off point of 0.54 and 0.52 for the short and long models, respectively. The point where precision, recall, and the F1 score were equal acted as a 2nd stringent threshold. This cut-off point was 0.74 for the short models and 0.79 for the long models, resulting in a lower AUROC (0.70 and 0.74 respectively) on the preprocessing test subset but a superior FPR (0.11 and 0.09 respectively versus 0.26 and 0.22 using Youden's threshold).

Table 5.2: Feature sets identified by the genetic algorithm. The first 5 were used within the short model while all the features were used in the long model. The short feature set contained the 1st, 2nd, 4th, 5th and 6th SHAP-ranked features, while the long feature set contained those same 5 alongside an additional 32. Details of the Catch22 transformation key can be found in the appendix (Table A.3.)

Clinical parameter	Catch22 transformation key	Calculated over the entire current stay or the given day	Calculated as the difference from the previous day	SHAP value
Systolic blood pressure	17	Entire current stay	×	2.27
Heart rate	17	Given day	×	2.05
Mean blood pressure	11	Entire current stay	×	1.62
O2 saturation pulseoxymetry	3	Entire current stay	×	1.38
GCS motor response	15	Entire current stay	×	1.37
Heart rate	17	Entire current stay	×	1.27
GCS verbal response	10	Entire current stay	×	1.16
Heart rate	14	Given day	×	1.15
O2 saturation pulseoxymetry	17	Given day	×	1.06
GCS motor response	10	Entire current stay	×	0.98
Respiratory rate	17	Entire current stay	×	0.92
Mean blood pressure	5	Entire current stay	×	0.91
Respiratory rate	6	Entire current stay	✓	0.80
Temperature	7	Entire current stay	✓	0.77
Systolic blood pressure	15	Entire current stay	×	0.77
Heart rate	9	Entire current stay	✓	0.77
Systolic blood pressure	5	Entire current stay	✓	0.75
Heart rate	3	Given day	×	0.73
O2 saturation pulseoxymetry	9	Given day	×	0.72
O2 saturation pulseoxymetry	2	Given day	×	0.69
GCS motor response	23	Entire current stay	×	0.69
GCS verbal response	7	Entire current stay	×	0.68
GCS motor response	13	Given day	×	0.63
Heart rate	16	Entire current stay	×	0.63
GCS motor response	0	Entire current stay	×	0.62
Temperature	10	Given day	✓	0.61
Heart rate	21	Given day	×	0.61
GCS motor response	22	Given day	×	0.58
Diastolic blood pressure	3	Entire current stay	×	0.56
GCS eye opening	21	Given day	×	0.55
Respiratory rate	12	Given day	✓	0.55
GCS motor response	8	Entire current stay	×	0.52
O2 saturation pulseoxymetry	1	Entire current stay	✓	0.52
Temperature	7	Entire current stay	×	0.52
Temperature	6	Entire current stay	✓	0.51
Respiratory rate	3	Given day	×	0.51
Heart rate	8	Given day	×	0.50

Table 5.3: Models' hyperparameters and optimisation options.

Hyperparameter	Short model	Long model	Optimisation options
Learning rate	0.001	0.001	0.1, 0.01, 0.001, 0.0001, 0.00001
Batch size	256	256	16, 32, 64, 128, 256, 512
Optimiser	Adam	Adam	Adam, RMSprop, SGD
Input dimension	5	37	–
Batch normalization	BatchNorm1d	BatchNorm1d	–
Number of hidden layers	2	3	1, 2, 3, 4, 5, 6, 7, 8, 9
Hidden dimension 1	512	64	2, 4, 8, 16, 32, 64, 128, 256, 512, 1024
Hidden dimension 2	16	512	2, 4, 8, 16, 32, 64, 128, 256, 512, 1024
Hidden dimension 3	–	256	2, 4, 8, 16, 32, 64, 128, 256, 512, 1024
Output dimension	1	1	–
Activation function	ReLU	ReLU	–
Dropout	0.4	0.5	0, 0.1, 0.2, 0.3, 0.4, 0.5

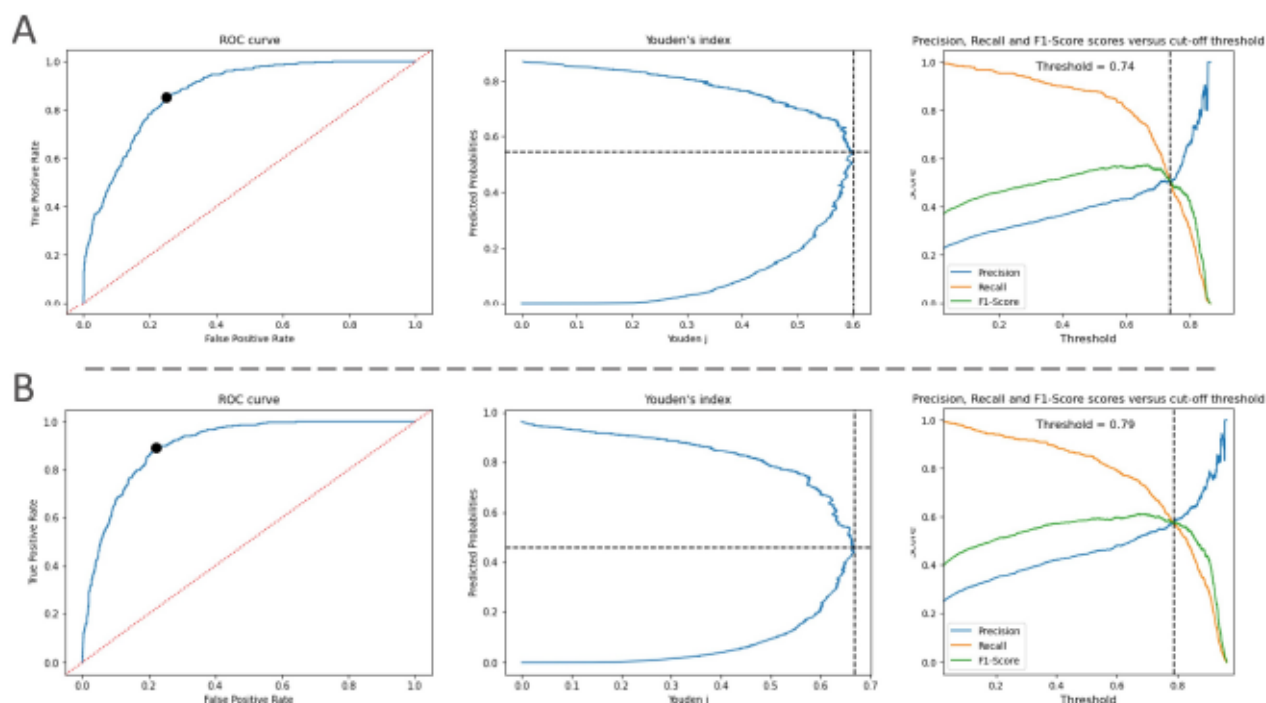


Figure 5.4: Preprocessing threshold optimisation graphs. ROC curve, Youden's index and Precision, Recall and F1-Score plots for the short model [A] (1st and 2nd threshold of 0.54 and 0.74 respectively) and the long model [B] (1st and 2nd threshold of 0.52 and 0.79 respectively).

5.4.1.3 Model evaluation

The final short models trained and tested on the MIMIC hold-out set obtained a mean AUROC of 0.78 (SD 0.02), FPR 0.25 (SD 0.02) with the 1st Youden threshold, and a mean AUROC of 0.69 (SD 0.03), FPR 0.10 (SD 0.02) with the 2nd threshold. Meanwhile, the final long models achieved a mean AUROC of 0.80 (SD 0.01), FPR 0.25 (SD 0.04) with the 1st cut-off, and a mean AUROC of 0.75 (SD 0.02), FPR 0.10 (SD 0.03) with the 2nd cut-off. Further evaluation metrics for each model and threshold can be found in Table 5.4. For comparison, a baseline using two clear infection markers (temperature and NEWS) from the latest guidelines (UK Health Security Agency 2022) obtained worse results with an AUROC of 0.66, accuracy of 0.61, TPR of 0.75, and FPR of 0.43. Predictions and labels broken down by IV treatment duration (Figure 5.5) shows that the majority of incorrect predictions occurred in the middle of IV treatment days when the models predicted to switch, but the real label indicated that the patient continued with IV. The short model predicted on average 70% and 38% of patients could switch earlier than they did with the 1st and 2nd thresholds, respectively. Arguably, the long model demonstrated a more balanced profile with switch predictions of 51%/28% early, 38%/41% agreement, and 11%/31% late with the 1st and 2nd thresholds, respectively. When the difference between the real and predicted switch event was minimal, the mean LOS outcomes of the patients were reduced (Figure 5.6). Furthermore, a statistically significant difference (Wilcoxon rank-sum test, alpha 0.05) was observed in the remaining LOS between those who received oral versus those who had IV treatment, with 2, 3, and 4 prior days of IV treatment (all p-values <0.01). No statistically significant differences were observed on the later days 5, 6, and 7 (Figure 5.7). No mortality differences were observed due to imbalanced data (Supplementary Table A.4).

Table 5.4: Evaluation results.

Dataset	Model	Threshold	Evaluation Metric					
			AUROC	Accuracy	Precision	F1 Score	AUPRC	TPR
MIMIC	Short	1st	0.78 (SD 0.02)	0.76 (SD 0.01)	0.39 (SD 0.02)	0.53 (SD 0.03)	0.35 (SD 0.03)	0.80 (SD 0.05)
		2nd	0.69 (SD 0.03)	0.83 (SD 0.01)	0.49 (SD 0.03)	0.48 (SD 0.04)	0.32 (SD 0.03)	0.48 (SD 0.06)
	Long	1st	0.80 (SD 0.01)	0.77 (SD 0.03)	0.41 (SD 0.04)	0.55 (SD 0.03)	0.37 (SD 0.03)	0.85 (SD 0.04)
		2nd	0.75 (SD 0.02)	0.85 (SD 0.02)	0.55 (SD 0.05)	0.57 (SD 0.03)	0.40 (SD 0.03)	0.61 (SD 0.06)
MIMIC Incomplete absorption	Short	1st	0.73 (SD 0.03)	0.72 (SD 0.03)	0.64 (SD 0.04)	0.71 (SD 0.03)	0.59 (SD 0.03)	0.79 (SD 0.06)
		2nd	0.67 (SD 0.05)	0.71 (SD 0.04)	0.74 (SD 0.06)	0.56 (SD 0.11)	0.57 (SD 0.05)	0.47 (SD 0.12)
	Long	1st	0.77 (SD 0.02)	0.76 (SD 0.03)	0.68 (SD 0.05)	0.74 (SD 0.03)	0.63 (SD 0.04)	0.82 (SD 0.06)
		2nd	0.73 (SD 0.03)	0.75 (SD 0.03)	0.79 (SD 0.09)	0.66 (SD 0.06)	0.63 (SD 0.05)	0.57 (SD 0.09)
MIMIC Mal-absorption	Short	1st	0.76 (SD 0.10)	0.64 (SD 0.17)	0.44 (SD 0.25)	0.57 (SD 0.23)	0.44 (SD 0.25)	1.00 (SD 0.00)
		2nd	0.75 (SD 0.11)	0.73 (SD 0.06)	0.41 (SD 0.20)	0.50 (SD 0.21)	0.39 (SD 0.15)	0.78 (SD 0.33)
	Long	1st	0.75 (SD 0.07)	0.65 (SD 0.10)	0.39 (SD 0.16)	0.53 (SD 0.17)	0.38 (SD 0.15)	0.93 (SD 0.13)
		2nd	0.71 (SD 0.16)	0.79 (SD 0.10)	0.57 (SD 0.35)	0.48 (SD 0.26)	0.45 (SD 0.24)	0.53 (SD 0.34)
MIMIC UTI	Short	1st	0.77 (SD 0.03)	0.71 (SD 0.03)	0.38 (SD 0.07)	0.52 (SD 0.07)	0.35 (SD 0.07)	0.87 (SD 0.05)
		2nd	0.74 (SD 0.04)	0.81 (SD 0.02)	0.49 (SD 0.08)	0.54 (SD 0.08)	0.38 (SD 0.08)	0.63 (SD 0.10)
	Long	1st	0.78 (SD 0.02)	0.72 (SD 0.03)	0.39 (SD 0.06)	0.54 (SD 0.06)	0.36 (SD 0.06)	0.87 (SD 0.04)
		2nd	0.77 (SD 0.04)	0.83 (SD 0.02)	0.54 (SD 0.09)	0.58 (SD 0.07)	0.42 (SD 0.07)	0.66 (SD 0.12)
MIMIC Pneumonia	Short	1st	0.76 (SD 0.03)	0.68 (SD 0.03)	0.28 (SD 0.05)	0.42 (SD 0.07)	0.26 (SD 0.05)	0.86 (SD 0.07)
		2nd	0.76 (SD 0.03)	0.82 (SD 0.02)	0.41 (SD 0.05)	0.51 (SD 0.04)	0.32 (SD 0.05)	0.68 (SD 0.07)
	Long	1st	0.77 (SD 0.02)	0.70 (SD 0.04)	0.30 (SD 0.06)	0.44 (SD 0.07)	0.28 (SD 0.05)	0.87 (SD 0.03)
		2nd	0.74 (SD 0.04)	0.82 (SD 0.03)	0.41 (SD 0.09)	0.48 (SD 0.08)	0.31 (SD 0.07)	0.62 (SD 0.10)
MIMIC Sepsis	Short	1st	0.82 (SD 0.05)	0.71 (SD 0.05)	0.38 (SD 0.15)	0.53 (SD 0.17)	0.38 (SD 0.15)	1.00 (SD 0.00)
		2nd	0.79 (SD 0.12)	0.82 (SD 0.08)	0.53 (SD 0.25)	0.56 (SD 0.20)	0.43 (SD 0.18)	0.76 (SD 0.27)
	Long	1st	0.77 (SD 0.07)	0.68 (SD 0.08)	0.35 (SD 0.14)	0.48 (SD 0.14)	0.33 (SD 0.13)	0.90 (SD 0.13)
		2nd	0.76 (SD 0.18)	0.80 (SD 0.10)	0.46 (SD 0.21)	0.52 (SD 0.22)	0.42 (SD 0.16)	0.68 (SD 0.33)
eICU	Short	1st	0.72 (SD 0.02)	0.75 (SD 0.03)	0.36 (SD 0.05)	0.47 (SD 0.05)	0.29 (SD 0.04)	0.68 (SD 0.06)
		2nd	0.65 (SD 0.05)	0.85 (SD 0.02)	0.60 (SD 0.12)	0.42 (SD 0.10)	0.31 (SD 0.06)	0.34 (SD 0.10)
	Long	1st	0.72 (SD 0.02)	0.75 (SD 0.02)	0.36 (SD 0.06)	0.46 (SD 0.04)	0.29 (SD 0.04)	0.67 (SD 0.07)
		2nd	0.64 (SD 0.06)	0.84 (SD 0.02)	0.48 (SD 0.17)	0.38 (SD 0.15)	0.28 (SD 0.07)	0.33 (SD 0.15)

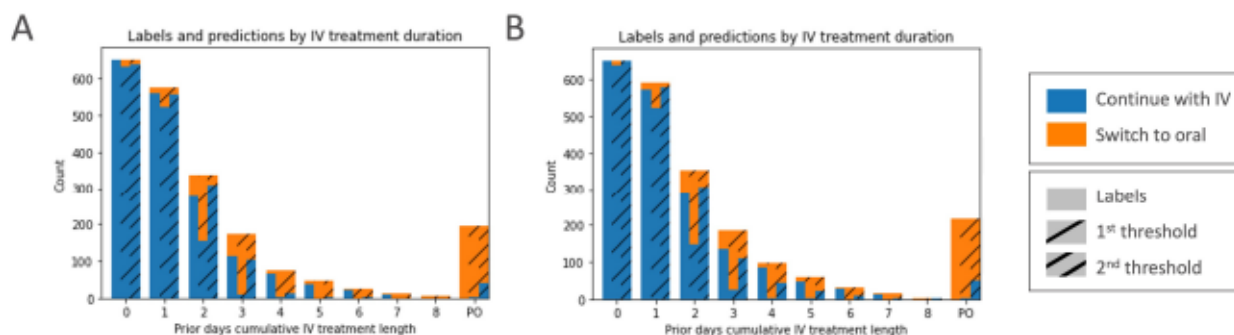


Figure 5.5: Labels and predictions by IV treatment duration. Plots for the short model [A] and the long model [B].

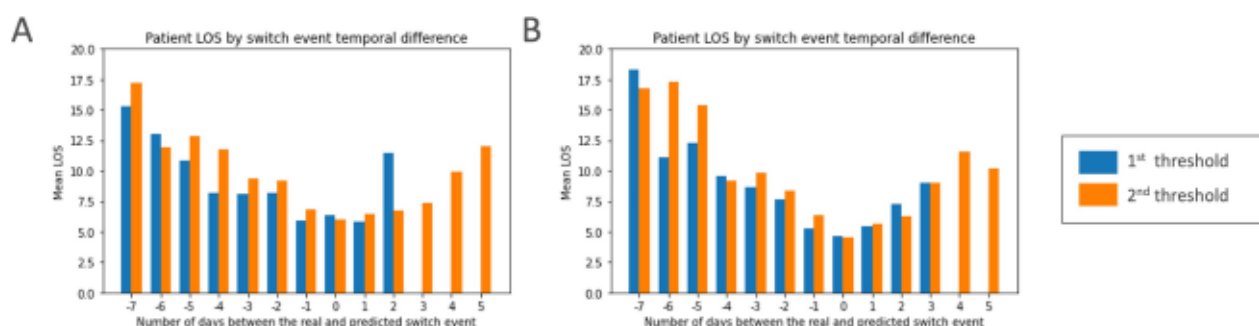


Figure 5.6: Mean MIMIC patient LOS outcomes by days between the real and predicted switch event. Plots for the short model [A] and the long model [B]. A negative number on the x axis indicates the predicted switch event was before the real switch event. The opposite is true for positive numbers, while 0 means they occurred on the same day.

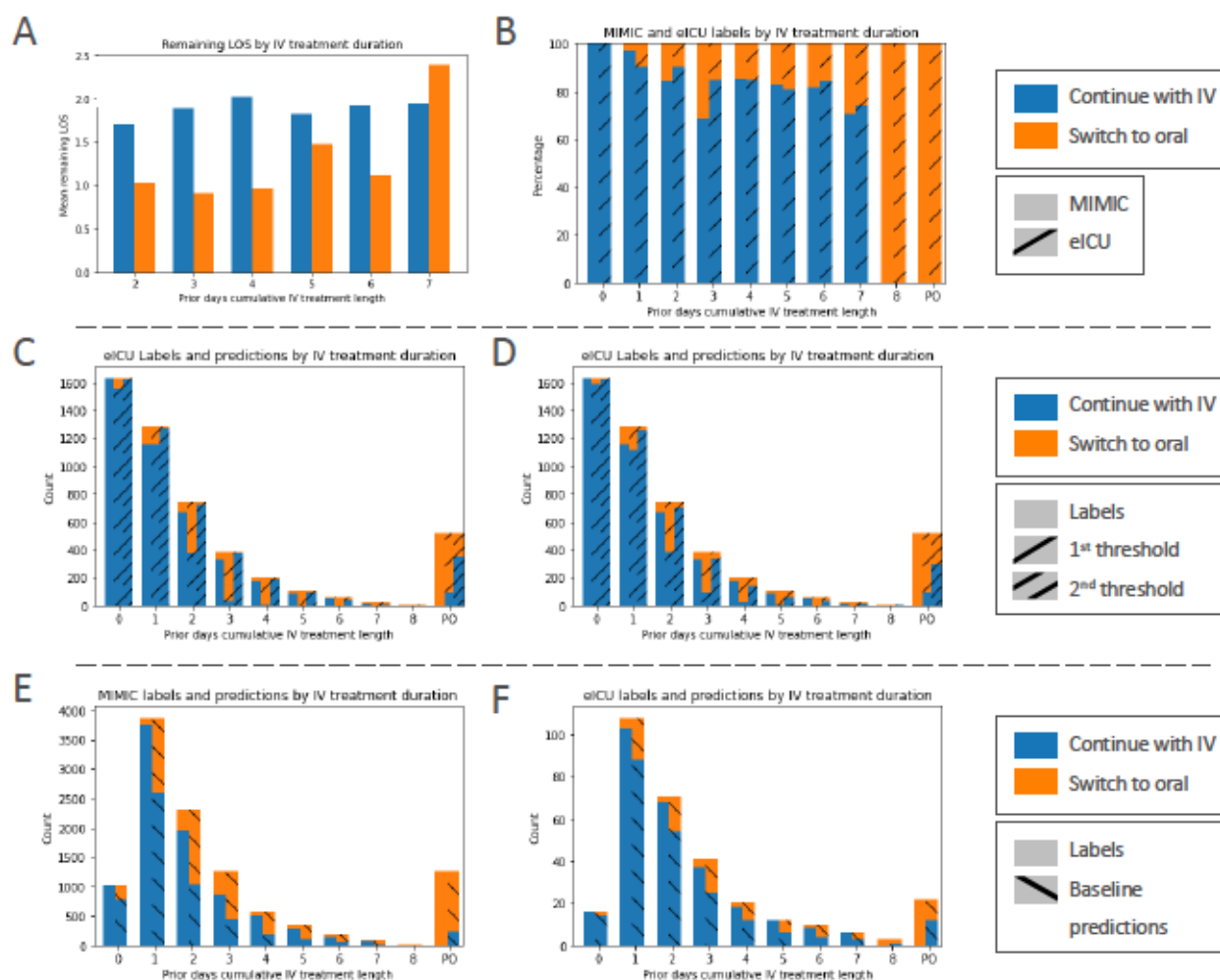


Figure 5.7: Labels, predictions, and LOS broken down by IV treatment duration. [A] demonstrates those patients on oral therapies on average have a shorter remaining LOS than those on IV. [B] shows switch labels for both the MIMIC and eICU datasets. They are generally consistent although eICU sees fewer patients switch on day 3. [C] and [D] are predictions for the transfer learning short and long models respectively on the eICU dataset. Models predict many more patients should not switch when compared to MIMIC results (Figure 5.5). [E] and [F] show how baseline predictions compare to the labels for both MIMIC and eICU.

Transfer learning on the eICU dataset resulted in a mean AUROC of 0.72 (SD 0.02), 0.65 (SD 0.05), 0.72 (SD 0.02), 0.64 (SD 0.06), and a FPR of 0.24 (SD 0.04), 0.05 (SD 0.02), 0.24 (SD 0.04) and 0.06 (SD 0.03) was obtained for the short and long models 1st and 2nd thresholds respectively (Table 5.4). Both models outperformed the eICU baseline, which obtained an AUROC of 0.55, accuracy of 0.67, TPR of 0.38, and FPR of 0.28.

For those patients in MIMIC who were on oral antibiotics with incomplete absorption, a mean AUROC of 0.73 (SD 0.03), 0.67 (SD 0.05), 0.77 (SD 0.02), 0.73 (SD 0.03), and a FPR of 0.33 (SD 0.06), 0.12 (SD 0.04), 0.28 (SD 0.07) and 0.12 (SD 0.07) was achieved for the short and long models 1st and 2nd cut-offs respectively (Table 5.4).

When tested on MIMIC patients with poor absorption, a mean AUROC of 0.76 (SD 0.10), 0.75 (SD 0.11), 0.75 (SD 0.07), 0.71 (SD 0.16), and a FPR of 0.48 (SD 0.20), 0.28 (SD 0.12), 0.43 (SD 0.14) and 0.12 (SD 0.12) was obtained for the short and long models 1st and 2nd thresholds respectively (Table 5.4).

Results were then examined for MIMIC patients with specific infections. For patients with a urinary tract infection (UTI) a mean AUROC of 0.77 (SD 0.03), 0.74 (SD 0.04), 0.78 (SD 0.02), 0.77 (SD 0.04), and an FPR of 0.33 (SD 0.03), 0.15 (SD 0.03), 0.31 (SD 0.04) and 0.13 (SD 0.05) was achieved for the short and long models 1st and 2nd cut-offs respectively (Table 5.4). When tested on patients with pneumonia a mean AUROC of 0.76 (SD 0.03), 0.76 (SD 0.03), 0.77 (SD 0.02), 0.74 (SD 0.04), and a FPR of 0.35 (SD 0.03), 0.16 (SD 0.04), 0.32 (SD 0.04) and 0.14 (SD 0.04) was obtained for the short and long models 1st and 2nd thresholds respectively (Table 5.4). Finally, for sepsis patients, a mean AUROC of 0.82 (SD 0.05), 0.79 (SD 0.12), 0.77 (SD 0.07), 0.76 (SD 0.18), and a FPR of 0.36 (SD 0.10), 0.17 (SD 0.09), 0.35 (SD 0.08) and 0.16 (SD 0.07) was achieved for the short and long models 1st and 2nd cut-offs respectively (Table 5.4).

5.4.1.4 Interpretability

Infectious disease clinicians were consulted and utilising the SimplEx (Crabbe et al. 2021) methodology and alternative cut-off thresholds informative visual representations were created. Figure 5.8 shows an example of these for short model predictions.

5.4.1.5 Fairness

Overall, the MIMIC models demonstrated EO across the majority of sensitive attribute groups. Table 5.5 shows the AUROC, TPR, and FPR for both short and long models by sensitive attribute group. The short model did not obtain EO for those in the age bracket of 90, of Native American descent, or with Medicaid insurance (Table 5.5). On the other hand, the long model only showed a discrepancy for patients in the age bracket of 30. For the short model, threshold optimisation (Hardt et al. 2016) with the TPR parity constraint enabled EO to be achieved for those in the age bracket of 90, while the EO constraint standardised performance across insurance groups (Figure 5.9, Supplementary Table A.5,). No constraint enabled the model to demonstrate EO for the native group. For the long model, the FPR parity constraint caused EO to be obtained for those in the age bracket of 30 (Figure 5.9, Supplementary Table A.6).

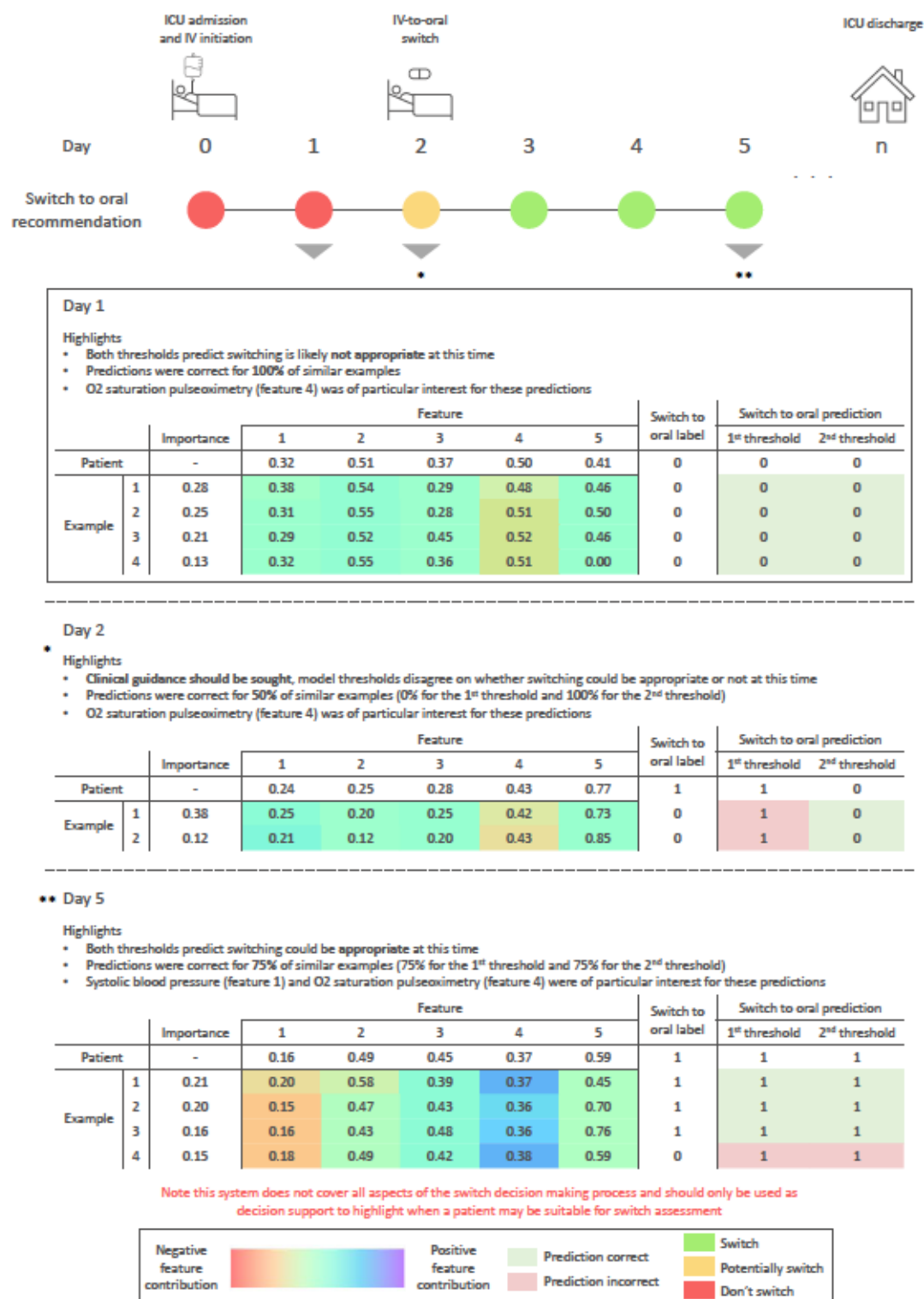


Figure 5.8: Example visual representation for a particular patient. 'Traffic light' suggestions are initially displayed in a temporal manner as the patient progresses. If required clinicians can obtain more information for any given day of the patient's stay. Simple textual descriptions are provided alongside detailed tabular graphics to maximise clarity. The text quickly gets across key points, while tables show similar example patients and their features, both with their relative importance and contribution respectively. Furthermore, switch labels and predictions across both thresholds are displayed. Note that the patient's label would obviously not be available during clinical use, but is shown here to be comprehensive. Finally, the limitations and use cases for the system are clearly labelled.

Table 5.5: Fairness results.

Sensitive attribute	Group	Short model			EO	Long model			EO
		AUROC	TPR	FPR		AUROC	TPR	FPR	
-	All patients	0.78 (SD 0.02)	0.80 (SD 0.05)	0.25 (SD 0.02)	-	0.80 (SD 0.01)	0.85 (SD 0.04)	0.25 (SD 0.04)	-
Sex	Female	0.74 (SD 0.18)	0.79 (SD 0.35)	0.29 (SD 0.13)	✓	0.80 (SD 0.15)	0.85 (SD 0.34)	0.25 (SD 0.10)	✓
	Male	0.80 (SD 0.08)	0.82 (SD 0.22)	0.23 (SD 0.06)	✓	0.80 (SD 0.16)	0.84 (SD 0.33)	0.23 (SD 0.08)	✓
Age	20	0.73 (SD 0.08)	0.74 (SD 0.15)	0.27 (SD 0.06)	✓	0.76 (SD 0.09)	0.77 (SD 0.16)	0.24 (SD 0.05)	✓
	30	0.80 (SD 0.02)	0.86 (SD 0.06)	0.26 (SD 0.04)	✓	0.72 (SD 0.03)	0.64 (SD 0.09)	0.20 (SD 0.05)	✗
	40	0.78 (SD 0.04)	0.81 (SD 0.08)	0.25 (SD 0.03)	✓	0.77 (SD 0.02)	0.80 (SD 0.06)	0.26 (SD 0.06)	✓
	50	0.76 (SD 0.04)	0.78 (SD 0.09)	0.25 (SD 0.04)	✓	0.80 (SD 0.04)	0.87 (SD 0.08)	0.26 (SD 0.05)	✓
	60	0.79 (SD 0.02)	0.82 (SD 0.04)	0.23 (SD 0.04)	✓	0.80 (SD 0.03)	0.84 (SD 0.03)	0.24 (SD 0.05)	✓
	70	0.73 (SD 0.08)	0.69 (SD 0.19)	0.23 (SD 0.07)	✓	0.81 (SD 0.06)	0.86 (SD 0.12)	0.23 (SD 0.05)	✓
	80	0.77 (SD 0.02)	0.81 (SD 0.04)	0.26 (SD 0.03)	✓	0.81 (SD 0.01)	0.85 (SD 0.06)	0.23 (SD 0.05)	✓
	90	0.78 (SD 0.03)	0.79 (SD 0.07)	0.23 (SD 0.02)	✗	0.78 (SD 0.02)	0.78 (SD 0.04)	0.22 (SD 0.04)	✓
	Asian	0.79 (SD 0.08)	0.83 (SD 0.12)	0.24 (SD 0.11)	✓	0.80 (SD 0.11)	0.84 (SD 0.18)	0.24 (SD 0.08)	✓
Race	Black	0.78 (SD 0.04)	0.83 (SD 0.07)	0.27 (SD 0.05)	✓	0.80 (SD 0.04)	0.85 (SD 0.07)	0.24 (SD 0.06)	✓
	Hispanic	0.80 (SD 0.07)	0.85 (SD 0.12)	0.25 (SD 0.08)	✓	0.80 (SD 0.08)	0.84 (SD 0.16)	0.25 (SD 0.08)	✓
	Native	0.78 (SD 0.17)	0.97 (SD 0.07)	0.43 (SD 0.35)	✗	0.82 (SD 0.13)	1.00 (SD 0.00)	0.35 (SD 0.23)	✓
	Other	0.76 (SD 0.06)	0.72 (SD 0.10)	0.19 (SD 0.05)	✓	0.79 (SD 0.07)	0.77 (SD 0.09)	0.20 (SD 0.09)	✓
	Unknown	0.79 (SD 0.05)	0.83 (SD 0.11)	0.25 (SD 0.03)	✓	0.82 (SD 0.03)	0.87 (SD 0.06)	0.23 (SD 0.05)	✓
	White	0.77 (SD 0.02)	0.79 (SD 0.06)	0.24 (SD 0.03)	✓	0.80 (SD 0.02)	0.84 (SD 0.04)	0.24 (SD 0.05)	✓
	Medicaid	0.72 (SD 0.07)	0.69 (SD 0.17)	0.26 (SD 0.06)	✗	0.76 (SD 0.08)	0.77 (SD 0.16)	0.26 (SD 0.05)	✓
Insurance	Medicare	0.78 (SD 0.03)	0.81 (SD 0.06)	0.25 (SD 0.02)	✓	0.81 (SD 0.02)	0.85 (SD 0.04)	0.24 (SD 0.05)	✓
	Other	0.78 (SD 0.02)	0.80 (SD 0.05)	0.24 (SD 0.03)	✓	0.80 (SD 0.02)	0.84 (SD 0.05)	0.23 (SD 0.04)	✓
Language	English	0.77 (SD 0.04)	0.79 (SD 0.09)	0.25 (SD 0.04)	✓	0.81 (SD 0.06)	0.85 (SD 0.11)	0.24 (SD 0.05)	✓
	Other	0.78 (SD 0.02)	0.80 (SD 0.05)	0.25 (SD 0.02)	✓	0.77 (SD 0.01)	0.78 (SD 0.04)	0.24 (SD 0.04)	✓
Marital status	Divorced	0.78 (SD 0.04)	0.80 (SD 0.10)	0.24 (SD 0.03)	✓	0.79 (SD 0.05)	0.82 (SD 0.09)	0.24 (SD 0.05)	✓
	Married	0.77 (SD 0.03)	0.77 (SD 0.06)	0.22 (SD 0.02)	✓	0.81 (SD 0.01)	0.83 (SD 0.05)	0.22 (SD 0.05)	✓
	Single	0.78 (SD 0.02)	0.84 (SD 0.05)	0.28 (SD 0.03)	✓	0.79 (SD 0.03)	0.84 (SD 0.06)	0.27 (SD 0.04)	✓
	Widowed	0.79 (SD 0.04)	0.82 (SD 0.08)	0.24 (SD 0.02)	✓	0.81 (SD 0.03)	0.85 (SD 0.07)	0.24 (SD 0.06)	✓
	Unknown	0.77 (SD 0.05)	0.83 (SD 0.09)	0.29 (SD 0.05)	✓	0.84 (SD 0.04)	0.93 (SD 0.06)	0.24 (SD 0.07)	✓

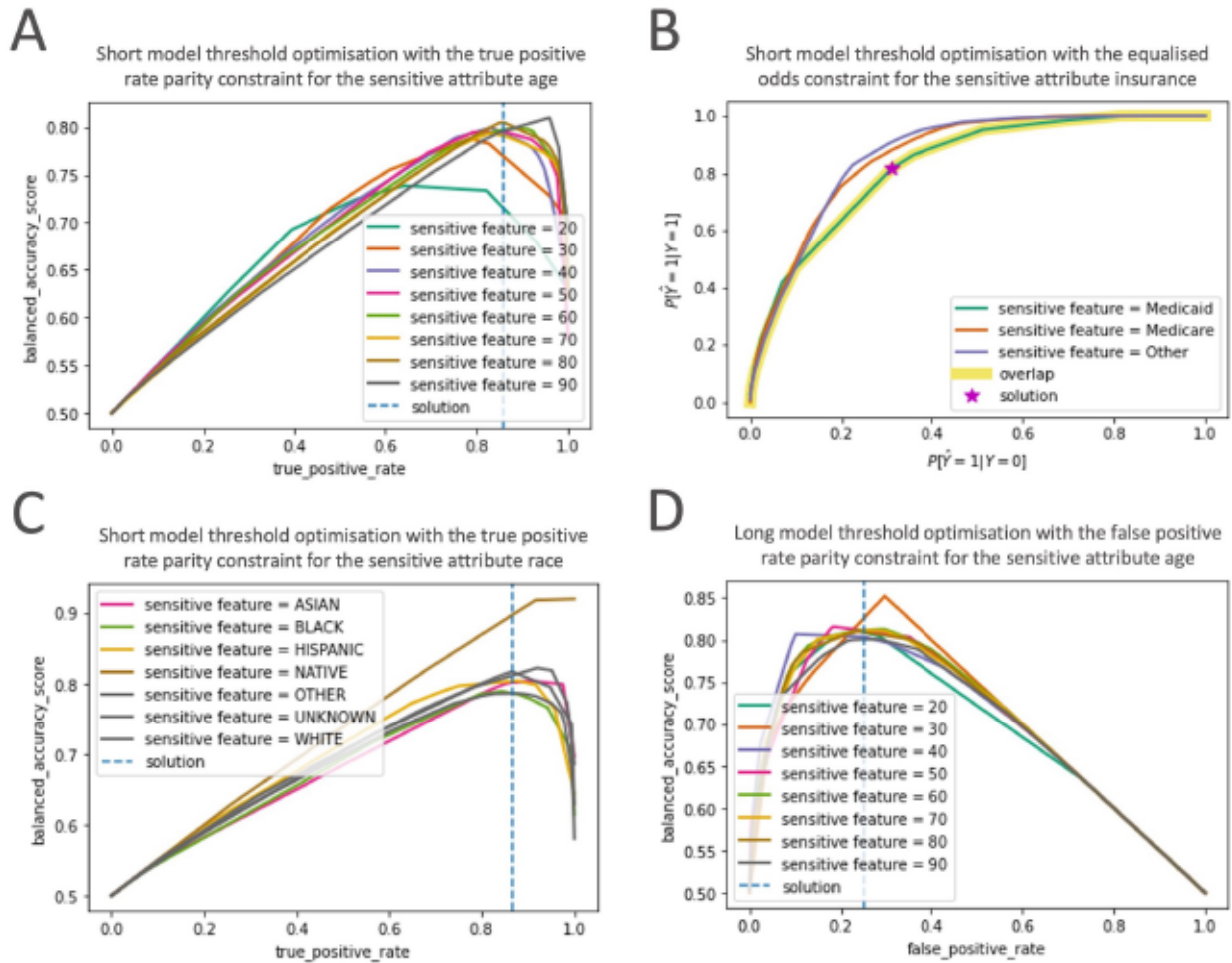


Figure 5.9: Sensitive attribute threshold optimisation graphs. [A] indicates that while fairness is now achieved by the short model for the 30-year-old age group, 20-year-olds suffer (Supplementary Table A.5). [B] demonstrates fair performance by the short model across insurance groups. [C] shows equalised odds cannot be achieved by the short model for the native population. [D] exhibits fair performance by the long model across age groups.

5.4.2 Imperial College Healthcare NHS Trust

5.4.2.1 Data

5,610 unique stays were extracted from ICHT with the prospective dataset containing 547 unique stays. Both datasets were relatively balanced with respect to demographics and type of infection, but patients in the prospective dataset were more co-morbid as shown in Table 5.6.

Table 5.6: ICHT demographics and statistics.

Statistic		Dataset	
		ICHT	ICHT Prospective
Number of stays		5610	547
Length of stay (mean, SD)		8.02 (8.69)	8.11 (8.12)
IV antimicrobial treatment length (mean, SD)		2.28 (1.73)	2.22 (1.78)
Age statistics	Mean	61.02	71.07
	Median	64.00	74.00
	Standard Deviation	20.62	16.62
Sex (%)	Female	60.39	52.48
	Male	39.61	47.52
Ethnicity (%)	White	43.49	56.13
	Other	34.20	22.85
	Black	11.36	11.61
	Asian	10.95	9.42
Infection type (%)	UTI	13.12	13.71
	Pneumonia	26.84	34.73
	Sepsis	31.79	37.84
	Cellulitis	4.66	7.68
Index of multiple deprivation decile statistics	Mean	4.39	3.89
	Median	4.00	3.00
	Mode	2.00	2.00
	Standard Deviation	2.27	2.11
Co-morbidities (%, those with a frequency of greater than 5% shown)	Hypertensive disorder	30.09	53.56
	Diabetes mellitus type 2	13.99	23.95
	Varicella	10.39	6.76
	Hypercholesterolemia	9.63	18.83
	Asthma	8.97	14.26
	Atrial fibrillation	8.22	9.69
	Chronic obstructive pulmonary disease	7.15	19.74
	Hypothyroidism	5.42	14.08
	Osteoarthritis	5.08	9.69

5.4.2.2 Model evaluation

The final model obtained a mean AUROC of 0.79 (SD 0.01), and FPR of 0.21 (SD 0.03) on the ICHT dataset. When evaluating by subgroups, similar performance was obtained across the board, with a slight drop in results for those on antibiotics with reduced bioavailability (Table 5.7). This performance was maintained in the prospective ICHT dataset with an AUROC of 0.77 and FPR of 0.20. Figure 5.10 shows the labels and predictions broken down by IV treatment duration and timeframe (24 hours, 48 hours or later). For the ICHT dataset, the model predicted switching was appropriate earlier in 48.39% of the cases, later 26.94%, and on the same day 24.66% of the time. In the prospective dataset similar results were observed with 50.83% earlier, 28.05% later, and 21.12% on the same day. Patients' LOS was reduced when model predictions aligned with the true switch event (Figure 5.11). Statistically significant differences (Wilcoxon rank-sum test, alpha 0.05) in remaining LOS were observed between those who received oral versus those who had IV treatment, with 1-5, and 7 prior days of IV treatment, as well as in all time frames (24 hours, 48 hours and other, Figure 5.10 [E, F], all p-values <0.01).

Table 5.7: ICHT evaluation results

Dataset	Evaluation Metric						
	AUROC	Accuracy	Precision	F1 Score	AUPRC	TPR	FPR
ICHT	0.79 (SD 0.01)	0.71 (SD 0.01)	0.82 (SD 0.01)	0.73 (SD 0.02)	0.84 (SD 0.01)	0.66 (SD 0.03)	0.21 (SD 0.03)
ICHT Incomplete absorption	0.76 (SD 0.09)	0.68 (SD 0.09)	0.72 (SD 0.24)	0.66 (SD 0.22)	0.81 (SD 0.08)	0.61 (SD 0.21)	0.21 (SD 0.08)
ICHT UTI	0.80 (SD 0.03)	0.72 (SD 0.02)	0.83 (SD 0.04)	0.75 (SD 0.02)	0.86 (SD 0.03)	0.68 (SD 0.04)	0.21 (SD 0.05)
ICHT Pneumonia	0.79 (SD 0.02)	0.72 (SD 0.02)	0.82 (SD 0.02)	0.74 (SD 0.04)	0.84 (SD 0.02)	0.68 (SD 0.06)	0.22 (SD 0.03)
ICHT Sepsis	0.81 (SD 0.05)	0.75 (SD 0.01)	0.81 (SD 0.02)	0.76 (SD 0.01)	0.84 (SD 0.02)	0.71 (SD 0.03)	0.21 (SD 0.03)
ICHT Prospective	0.77	0.70	0.77	0.68	0.68	0.61	0.20

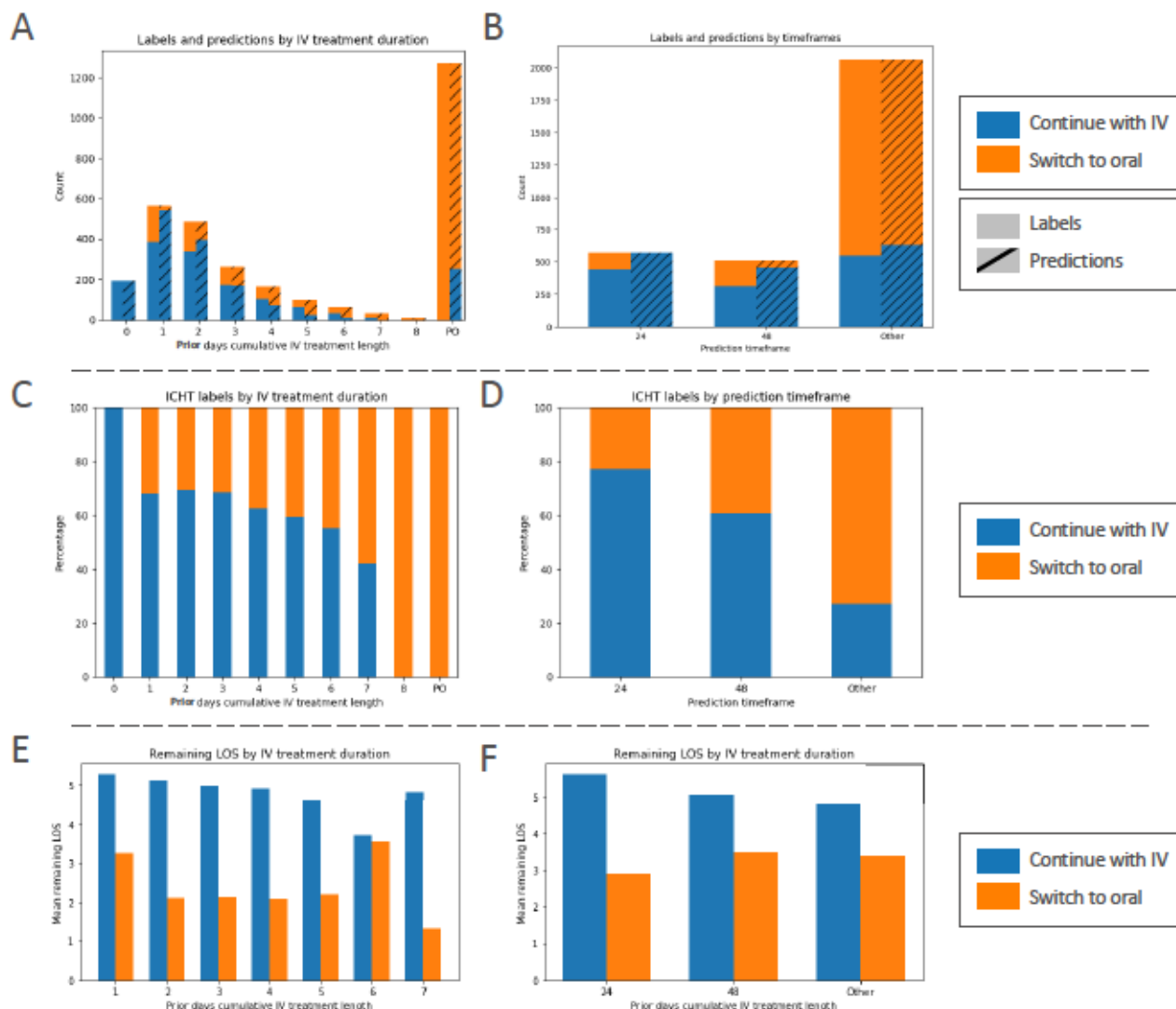


Figure 5.10: Labels, predictions, and LOS broken down by IV treatment duration. [A] and [B] show that predictions are generally temporally balanced with the labels, but that the model starts by predicting fewer patients should switch and over time progresses to predicting more should switch. [C] and [D] shows the distribution of switch labels for the ICHT dataset. [E] and [F] demonstrate those patients on oral therapies on average have a shorter remaining LOS than those on IV.

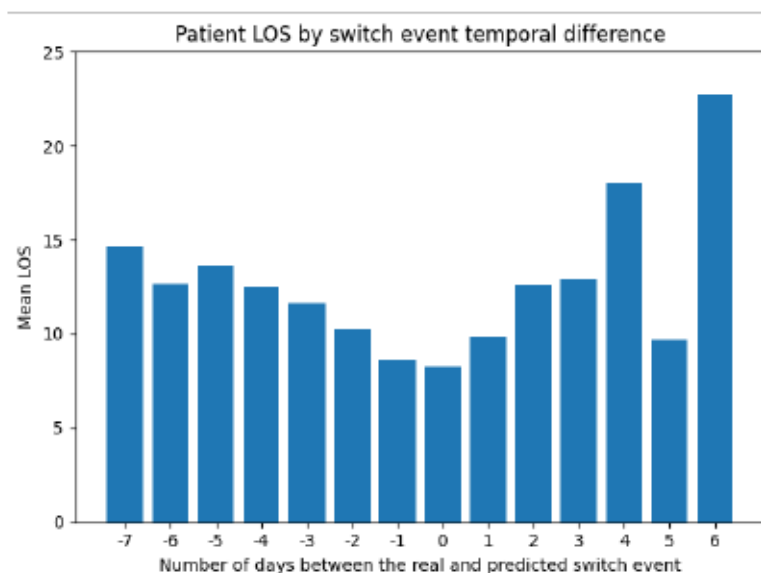


Figure 5.11: Mean ICHT patient LOS outcomes by days between the real and predicted switch event. A negative number on the x axis indicates the predicted switch event was before the real switch event. The opposite is true for positive numbers, while 0 means they occurred on the same day.

5.4.2.3 Prospective point prevalence survey evaluation

Information on the 40 patients selected from the point prevalence survey is shown in Table 5.8. The model achieved an AUROC of 0.68 against switch labels defined by antimicrobial pharmacists with a lower accuracy of 0.62 for the labels defined by the real prescribing actions. The full details of the results are shown in Table 5.9. When analysing the results by division, the model performed better in surgery and cancer than in medicine and integrated care.

Table 5.8: Characteristics of patients from the point prevalence survey

Characteristic	No. (%) of participants
Age statistics	
Mean	64.73
Median	66.50
Standard Deviation	19.67
Sex	
Male	21 (52.50%)
Female	19 (47.50%)
Ethnicity	
White	23 (57.50%)
Other	12 (30.00%)
Asian	3 (7.50%)
Black	2 (5.00%)
Index of multiple deprivation decile statistics	
Mean	4.25
Median	3.50
Standard Deviation	2.01
Division	
Medicine and Integrated Care	20 (50.00%)
Surgery and Cancer	20 (50.00%)
Speciality	
Elderly medicine	6 (15.00%)
General surgery	6 (15.00%)
Respiratory medicine	6 (15.00%)
Gastrointestinal surgery	4 (10.00%)
Neurology	2 (5.00%)
Renal nephrology	2 (5.00%)
Renal dialysis	2 (5.00%)
Hepatology	2 (5.00%)
Vascular	2 (5.00%)
Urology	2 (5.00%)
Oncology	2 (5.00%)
Ear, Nose, and Throat	2 (5.00%)
Major trauma	2 (5.00%)
Indication	
Gastrointestinal infections	13 (32.50%)
Respiratory Infections	9 (22.50%)
Genitourinary Infections	6 (15.00%)
Blood Infections	6 (15.00%)
Skin Infections	4 (10.00%)
Ear, Nose, and Throat Infections	2 (5.00%)

Table 5.9: Point prevalence survey model evaluation results

Dataset	Label	Evaluation Metric						
		AUROC	Accuracy	Precision	F1 Score	AUPRC	TPR	FPR
All	Pharmacists	0.68	0.70	0.48	0.55	0.41	0.64	0.28
	Real-world	-	0.62	-	-	-	-	0.38
Medicine and Integrated Care	Pharmacists	0.62	0.64	0.44	0.50	0.39	0.58	0.33
	Real-world	-	0.59	-	-	-	-	0.41
Surgery and Cancer	Pharmacists	0.75	0.76	0.54	0.61	0.46	0.71	0.22
	Real-world	-	0.65	-	-	-	-	0.35

Note many of the evaluation metrics cannot be calculated when using the real-world labels due to the lack of any actual IV-to-oral switching.

Overall, the model predicted to switch early for 37.50% of the patients, late for 30.00% and was the same for 32.50%, when analysing the temporal nature of the model's predictions versus when the pharmacists perceived the patients were suitable for switching. These metrics can be further broken down. For 17.50% the model predicted to switch before pharmacists did with an average temporal difference of 2.14 days earlier (SD 1.86). In 20.00% of cases, the model predicted to switch and the pharmacists had not up to that point in time with a minimum time difference of 2.12 days earlier (SD 1.89). 15.00% of the time the model predicted to switch after pharmacists with an average temporal difference of 1.5 days later (SD 0.54). For 15.00% of patients, pharmacists predicted to switch but the model did not up until the point of the data ending (minimum time difference 1.33 days later, SD 0.52). Finally, for one patient (2.50%) both the model and the pharmacists concluded that switching was appropriate on the same day, with 30.00% of patients not being perceived as suitable for oral treatment at any time by both the model and pharmacists.

The difference in days between when pharmacists deemed switching was appropriate (or the last day if they did not think switching was appropriate), the real-world actions and our model's predictions is shown in Figure 5.12. When considering all patients our model's

switch predictions are on average 0.38 days before the pharmacists (SD 1.84) with the real-world actions having a greater difference of 1.18 days after (SD 1.46). However, the absolute difference (ignoring negatives) was similar at 1.23 (SD 1.42) and 1.18 (SD 1.46) days respectively. When just considering those patients where at some point in time both the model and pharmacists perceive switching is appropriate this results in our model's switch predictions being on average 0.43 days before the pharmacists (SD 2.16) while real-world actions are 2.79 days after (SD 1.12). When taking the absolute values this results in 1.71 (SD 1.39) and 2.79 (SD 1.21) respectively. Finally, statistically significant differences (Wilcoxon rank-sum test, all p-values ≤ 0.01) were observed between the duration of IV treatment and switching being deemed appropriate by both the model and pharmacists (Figure 5.13).

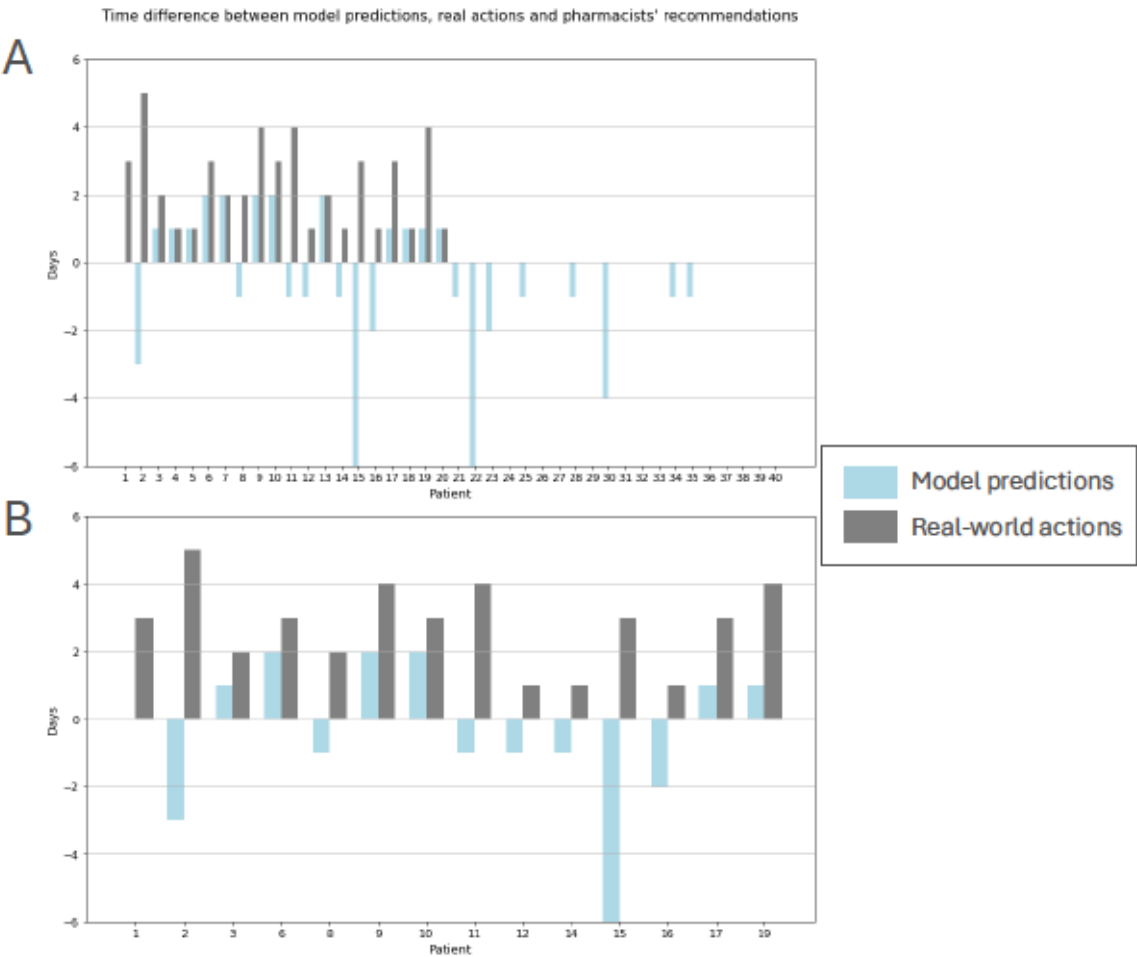


Figure 5.12: Time difference in days for model predictions and real-world actions when compared to pharmacists' recommendations. [A] shows all patients while [B] shows only those where at some point in time both the model and pharmacists perceived switching was appropriate. Note that in the cases where no switch prediction occurred during the patient's stay, 1 day was added to the calculation to account for the earliest they could possibly switch.

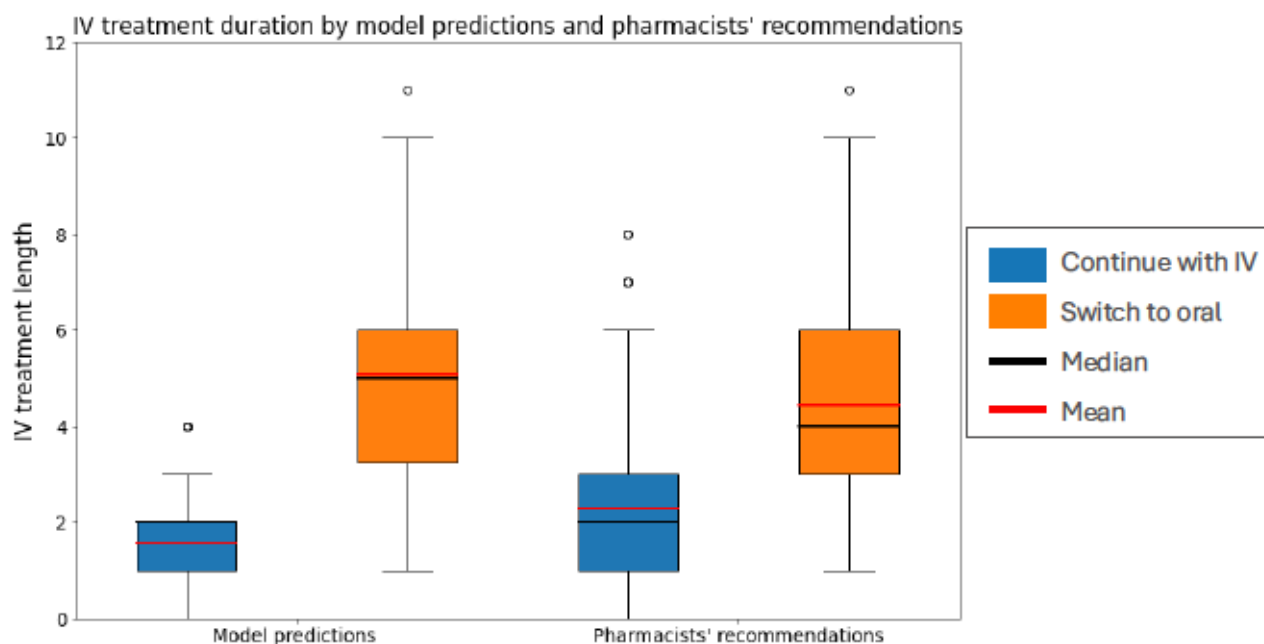


Figure 5.13: Box plot of IV treatment durations by model predictions and pharmacists' recommendations for switch and don't switch.

5.5 Discussion

We aimed to maximise the clinical utility of our IVOS-focused machine learning models. When working with MIMIC this was pursued by minimising complexity during feature selection and model development. Through the genetic algorithm, two feature sets of interest were identified. The short set used only five features but maintained performance, while the long set enabled slight improvements in the evaluation metrics. The two most important SHAP features utilised the same time series transformation (SB_MotifThree_quantile_hh) for systolic blood pressure over the whole ICU stay and heart rate over the current day respectively. This measure uses equiprobable binning to indicate the predictability of a time series. This is medically relevant to switching the administration route as clinicians look for vitals to stabilise before switching. Interestingly, the 3rd and 4th SHAP ranked features represent the same type of feature (IN_AutoMutualInfoStats_40_gaussian_fmfi calculated over the whole of the ICU stay) for two different clinical parameters, respiratory rate, and mean blood pressure. Furthermore, their feature values (shown in the Supplementary Figure A.6) are very similar, indicating why hav-

ing both features was probably redundant to the models. This similarity may be due to both of these variables being artificially controlled for some patients (e.g., through mechanical ventilation and blood pressure management). Other features in the short set also demonstrate clinical importance. For example, the first minimum of an O2 saturation pulseoximetry autocorrelation function (CO_FirstMin_ac) would indicate variable stability and hence clinical improvement or deterioration. A value indicating the importance of low frequencies in the GCS motor response (SP_Summaries_welch_rect_area_5_1), could show if the patient is retaining consciousness or not over a long period, which is often necessary for administering oral medication. Overall, these features combine to provide a comprehensive but succinct view of the general health status of the patient which can be used to determine if switching could be appropriate.

MIMIC results on specific infections and antibiotic characteristics demonstrate that the models have stable performance across numerous different patient groups. Particularly important is understanding when oral antibiotics with incomplete absorption can be used, given concerns surrounding achieving therapeutic concentrations (Broek et al. 2021). Our long model achieved an AUROC of 0.77 (SD 0.02) in this subpopulation. Furthermore, in conditions such as sepsis where patients are critically ill for prolonged periods and fewer oral therapies are used, our short model obtained an AUROC of 0.82 (SD 0.02), indicating that such a support system could be utilised in severe infections. Transfer learning results on the eICU dataset were stringent with regards to predicting when switching could be appropriate (Supplementary Figure 5.7), this is to be expected considering the patients in eICU are on average more severely unwell than in MIMIC (Table 5.1). Switching administration route is influenced by many behavioural factors that are not easily modelled. Given that eICU contains data from many different hospitals, the prescribing behaviour with regard to oral switching is likely much more heterogeneous than in MIMIC whose data are from a single institution. As such, the eICU model has to approximate many different behaviours, which results in varying performance across institutions (Supplementary Figure A.7), and likely causes it to be more stringent with regard to predicting a switch to optimise performance. Similar behaviour is observed with the baseline eICU results, which confirms that predicting the route of administration is a more challenging task in eICU when compared to MIMIC (AUROC of 0.66 and 0.55 respectively). Good predictive performance

(AUROC 0.79, FPR 0.21) was obtained when applying our methodological approach of IVOS prediction models to ICHT. This setting represents a very different patient population to MIMIC and eICU, given the patients are not from the USA or an ICU. Therefore, these results (Table 5.7) are very promising and highlight that our approach can be applied to different medical settings, including locally within London hospitals. Overall, the results demonstrate the feature sets identified are informative, the selected hyperparameters can model the underlying data and our methods and models are generalisable across datasets and healthcare environments.

MIMIC models demonstrated reasonably fair performance across all sensitive attribute groups. When equalised odds were not achieved, threshold optimisation (Hardt et al. 2016) was able to improve the results for a given group in all cases, except for that of the Native American group. This population was the most under-represented within the data with an average of only 11 patients in the test set, highlighting the need for further good quality real or synthetic data on minority populations. When threshold optimisation was undertaken a trade-off between groups in a sensitive attribute class was sometimes observed. For example, the TPR parity constraint on the short model achieved EO for those in the age bracket of 90 but caused the FPR of those in the minority group around 20 years old to increase from 0.29 to 0.61 (Supplementary Table A.5). This loss in performance for the 20-year-old group was also partially seen for the FPR parity constraint on the long model (Supplementary Table A.6). This shows the importance of balance when considering if a model is defined as fair or not, in particular for drastically different patient populations, such as 90 versus 20-year-olds. Prioritising one group or a sensitive attribute can hinder model performance in others. As such honest and decent precautions and analysis are needed to ensure algorithms are equal and reasonable without discrimination. Moreover, for antibiotic decision-making further ethical considerations need to be taken into account including the effect on other individuals outside of the patient being treated (Bolton et al. 2022b) as discussed in Section 2.4. This research demonstrates such CDSSs can be fair and help support ethical prescribing; however, further analysis and validation would certainly be required before deployment.

Two sets of features were used for the MIMIC and eICU models to evaluate the trade-off between simplicity and explainability versus performance which has been widely discussed in

the machine learning literature (Herm et al. 2022). Results show that the long model often demonstrates slightly superior performance to the short model. However, it is inherently more complex and in some scenarios such as in those with sepsis, it performs worse than the short model. More research is required, including understanding clinician opinions, to determine which model is most appropriate in specific circumstances. Alternative cut-off thresholds were also investigated for our binary classification task to maximise the AUROC and minimise the FPR. The results show that this was achieved for both the short and long models by fixing the thresholds from the preprocessing validation subset, with the 1st threshold achieving a reasonable AUROC and the 2nd threshold having a lower FPR, although as expected this comes at the expense of a worse AUROC score. Further research into subpopulations and other datasets could identify unfavourable IV-to-oral switch characteristics, such as individuals with abnormal pharmacokinetics or immunosuppression. Specific thresholding or separate models (Afrose et al. 2022) could then ensure that patients with such attributes require a larger output to be flagged as suitable for switching. However, combining this with alternative thresholds to ensure fairness can very quickly make CDSSs excessively complex, leading to misunderstanding, misuse, and reluctant adoption (Rawson et al. 2017; D'Hondt et al. 2022; Schouten et al. 2022). We envisage the two thresholds used in this work being utilised similar to a traffic light, whereby suggestions can be split into do not, potentially, or do switch based on the model's level of confidence (Figure 5.8). This type of structure is simple, familiar to individuals and should ensure along with interpretability methods that such a model acts as an appropriate CDSS and allows the end user to understand the output alongside other information in order to make the final decision. This research therefore strikes a practical balance between performance and usefulness for IV-to-oral switch decision support.

Explainability and interpretability have been highlighted as critical aspects of using machine learning models in the real-world (Barredo Arrieta et al. 2020; Cinà et al. 2022). To ensure that our MIMIC model and its outputs could be understood and interrogated, Simplex (Crabbe et al. 2021) was utilised and visual representations created (Figure 5.8). These visual summaries include a number of aspects that were noted as important for understanding by clinical colleagues. Firstly, textual descriptions enable key information to be conveyed quickly

and reduce the barrier to adoption through universal understanding. Secondly, related patient examples are shown and scored. Clinicians rely heavily on previous experience when making antibiotic treatment decisions (Smith et al. 2008); as such, showing historical examples and how they compare to the current patient of interest is perceived as appropriate. In conjunction, highlighting whether the model was correct on previous examples at each threshold provides some level of reassurance on how well the model performs on this type of patient and therefore if the predictions should be trusted or not. Finally, patient-specific feature contribution can be shown to illustrate how the model arrived at that conclusion. Figure 5.8 shows that while in many cases a clear switch decision is apparent, inherently some days (e.g., day 3) and patients present a particularly complex case. This reflects what is often seen in reality with decisions regarding antimicrobial switching not being clear-cut. By incorporating interpretability methods, models such as those developed in this research can become clinically useful CDSSs.

The objective of a CDSS to support IV-to-oral switch decision-making is to facilitate antimicrobial stewardship. MIMIC ASI results are in line with current literature indicating frequent oral prescribing may use less broad spectrum IV antibiotics overall and therefore could be beneficial from an AMR and HCAI perspective (Ray-Barruel et al. 2019; McCarthy et al. 2020). As such, this evidence supports the drive to maximise the use of oral therapies and alongside limited adoption (Hospenthal et al. 2020; Buis et al. 2022) highlights why a switch-focused CDSS may be useful. It is however notoriously difficult to discern the value of predictions from a CDSS. A retrospective analysis was conducted to understand how such switch models may benefit healthcare institutions and patients. Figures 5.5 and 5.10 show that for the first two days after starting IV treatment our models predict that the majority of patients should not switch. This is in line with the latest UK guidelines whereby the IV-to-oral switch should be considered daily after 48 hours (UK Health Security Agency 2022). However, in MIMIC and eICU for dates with 2 to 7 previous days of IV treatment, there develops a disconnect between the labels and model predictions when using the 1st more lenient threshold. These outputs indicate that by day 4 almost all patients could be suitable for switching to oral administration from a clinical parameter, health status perspective. For some patients, there will be risk factors beyond the model's input features that the clinician considered meaning they did not switch, but for others,

the clinician may have been unaware or neglected the decision meaning switching earlier may have been suitable. This aligns with recent public health data in the UK that showed that at the end of 2023 in England around 20% of patients were still receiving IV antibiotics past the point at which they met the IV-oral switch criteria (Office for Health Improvement and Disparities 2024). This prescribing behaviour is not seen with the higher thresholds or in the ICHT model implying that more input data, such as co-morbidities processed through the pipeline developed in Chapter 3, can help the model discern when edge cases should not switch. On a patient level results show that LOS is minimised when predictions and the true labels align (Figure 5.6 and 5.11), and upon switching patients usually see prompt discharge (Figure 5.7 and 5.10). LOS is an important outcome measure with reductions benefiting both patients from a comfort perspective and hospitals financially (NHS 2024). Our models may therefore be able to provide useful decision support by raising awareness of when switching could be suitable for a particular patient. Given that this decision is often neglected and postponed, such a CDSS may be able to promote switching when appropriate, which could potentially support efforts to stop AMR, prevent HCAs, and benefit patients.

Progressing health-focused AI technologies from research on retrospective datasets to being integrated and adopted within a hospital is notoriously difficult (Higgins et al. 2020; McCradden et al. 2020; Ahmed et al. 2023; Lambert et al. 2023; Olaye et al. 2023). One key step in this pathway is prospective validation of the model. To progress this research beyond retrospective analysis and embark on the translation journey, we conducted an initial prospective evaluation of the ICHT IVOS model through two methods. First by using the latest real-world clinical data from patients with inpatient encounter at ICHT available in iCARE (released after initial model development). The performance of the model on this dataset was promising being highly comparable to the original ICHT results (AUROC of 0.77 versus 0.79 respectively, Table 5.7). Secondly, we leveraged the point prevalence survey conducted in July 2024 by pharmacists at ICHT. Although the sample size of this analysis was small due to data extraction limitations, the presence of pharmacists' IVOS recommendations can act as a gold standard ground truth, enabling us to understand how the model could augment antimicrobial stewardship support. The model obtained reasonable performance compared to

the pharmacists' recommendations (AUROC 0.68, FPR 0.28, Table 5.9). However, these results were worse than the previous analysis discussed above, which was primarily driven by patients in medicine and integrated care and may be due to labels being defined in a different way (i.e., through antimicrobial pharmacists' opinions versus real actions). Poor performance was seen against the labels defined by real-world actions but this is likely due to the skewed population (where no IVOS actually occurred due to the nature of the point prevalence survey only examining IVOS for those on IVs) and limited sample size versus, for example, the ICHT prospective dataset. When analysing the temporal relationships our model is slightly closer to the gold standard pharmacists' recommendations than real-world actions with an average difference of 0.38 days versus 1.18 days, respectively. This though is driven by the presence of model predictions before and after, with absolute differences being similar overall (1.23 versus 1.18 respectively) but better than real-world actions for those cases where switching may have been appropriate (absolute difference 1.71 versus 2.79 respectively). In general, the model had a slight preference towards predicting switching earlier (37.50%), which aligns with the results observed in MIMIC (Figure 5.5). There were two outlier cases where the model predicted switching was appropriate much earlier than the pharmacists (Patients 15 and 22 in Figure 5.12). Interestingly both of these patients were highly multi-morbid with 8 and 10 co-morbidities, respectively. This potentially indicates a failure mode of the model where the presence of many co-morbidities causes a feature explosion where values grow excessively, leading to numerical instability and erroneous outputs. Overall, this initial approach towards prospectively validating the ICHT IVOS model was valuable, highlighting both its successes and pitfalls. Such evaluations are critical to enable AI CDSSs to progress towards delivering patient care.

To improve the clinical applicability of our models a number of logic based rules could be implemented. For example, if a patient has a certain type of infection, malabsorption, immunosuppression, has recently vomited, or could have compliance issues, an overriding rule based on the latest guidelines (UK Health Security Agency 2022) could suggest not to switch. Furthermore, the number of days of IV treatment should be highlighted alongside conditions, such as sepsis, in which extra care should be taken, as these factors influence switch decision-

making. If a patient is receiving an IV antibiotic and there is a similar oral version available this could be flagged alongside model outputs as a 'simple' switch. Moreover, given the potential comfort, workload, and discharge benefits when patients have no IV catheters, CDSSs should consider the wider patient treatment paradigm, and potentially further encourage switching when IV access is only for antibiotic treatment. Finally, to improve practice, it is important for clinicians to document when a switch occurred and why that decision was made. This ensures that in the future such individualised antibiotic decision-making can be data-driven based on real evidence, rather than decided by habit or general population evidence. By combining machine learning approaches with clinical logic, we can ensure patient safety while driving a positive change in antimicrobial utilisation.

There are limitations to this research study. Firstly, the use of historical patient data means that all of our model's predictions are based on historical prescribing practices. Due to concerns surrounding AMR, there has been a large amount of research on antibiotic prescribing in recent years (Holmes et al. 2016; Rawson et al. 2017; Tamma et al. 2019; Charani et al. 2021; Cantón et al. 2022) and hence it is plausible that our model's switch suggestions are 'out of date'. To combat this in the future gold standard ground truths such as the pharmacists' recommendations in the point prevalence survey could be used for training. Secondly, our model only analyses a snapshot of the patient and not all factors that are clinically used to assess the suitability of a patient to change (UK Health Security Agency 2022). As discussed in the methods, this is due to data challenges, but incorporating additional criteria into the model so that under certain circumstances a switch suggestion cannot be given is an avenue for future work. However, we believe by analysing and summarising multiple variables regarding the patient's clinical and infection status such a system could support switch decision-making with the final decision always made by the responsible clinician.

5.6 Conclusion

In this chapter, our research resulted in simple, fair, and interpretable models to predict when a patient could switch from IV-to-oral antibiotic treatment. The models use clinically relevant features and the results were generalisable across three unique datasets, with an initial prospective evaluation conducted. Such a CDSS could provide prescribers with individualised recommendations, act as an alert for when switching may be appropriate or support appropriate stewardship for the often large volume of patients in secondary care hospitals on antibiotics. Further research is needed to evaluate the safety of such a system. But given AMR, HCAs, and the interest in promoting oral therapies, this research holds great promise to provide clinically useful antimicrobial decision support.

References

1. Bolton, William J. (2023). *Personalising intravenous to oral antibiotic switch decision making through fair interpretable machine learning*. doi: 10.5281/zenodo.10251521. URL: https://github.com/WilliamBolton/iv_to_oral.
2. Bolton, William J. et al. (Jan. 2024). "Personalising intravenous to oral antibiotic switch decision making through fair interpretable machine learning". In: *Nature Communications* 15.1, p. 506. ISSN: 2041-1723. doi: 10.1038/s41467-024-44740-2. URL: <https://www.nature.com/articles/s41467-024-44740-2>.
3. Dyar, O. J. et al. (Nov. 2017). "What is antimicrobial stewardship?" In: *Clinical Microbiology and Infection* 23.11, pp. 793–798. ISSN: 1198-743X. doi: 10.1016/j.cmi.2017.08.026. URL: [https://www.clinicalmicrobiologyandinfection.com/article/S1198-743X\(17\)30489-5/fulltext](https://www.clinicalmicrobiologyandinfection.com/article/S1198-743X(17)30489-5/fulltext).
4. Cyriac, Jissa Maria and James, Emmanuel (2014). "Switch over from intravenous to oral therapy: A concise overview". In: *Journal of Pharmacology & Pharmacotherapeutics* 5.2, pp. 83–87. ISSN: 0976-500X. doi: 10.4103/0976-500X.130042. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4008927/>.
5. Oosterheert, Jan Jelrik et al. (Dec. 2006). "Effectiveness of early switch from intravenous to oral antibiotics in severe community acquired pneumonia: multicentre randomised trial". In: *BMJ* 333.7580, p. 1193. ISSN: 0959-8138, 1468-5833. doi: 10.1136/bmj.38993.560984.BE. URL: <https://www.bmj.com/content/333/7580/1193>.
6. UK Health Security Agency (Mar. 2015). *Antimicrobial stewardship: Start smart – then focus*. URL: <https://www.gov.uk/government/publications/antimicrobial-stewardship-start-smart-then-focus>.
7. Davar, Kusha et al. (Dec. 2022). "Can the Future of ID Escape the Inertial Dogma of Its Past? : The Exemplars of Shorter Is Better and Oral Is the New IV". In: *Open Forum Infectious Diseases*, ofac706. ISSN: 2328-8957. doi: 10.1093/ofid/ofac706. URL: <https://doi.org/10.1093/ofid/ofac706>.
8. Gilchrist, Mark et al. (May 2022). "Outpatient parenteral antimicrobial therapy (OPAT) in the UK: findings from the BSAC National Outcomes Registry (2015–19)". In: *Journal of Antimicrobial Chemotherapy* 77.5, pp. 1481–1490. ISSN: 0305-7453. doi: 10.1093/jac/dkac047. URL: <https://doi.org/10.1093/jac/dkac047>.
9. Iversen, Kasper et al. (Jan. 2019). "Partial Oral versus Intravenous Antibiotic Treatment of Endocarditis". In: *New England Journal of Medicine* 380.5, pp. 415–424. ISSN: 0028-4793. doi: 10.1056/NEJMoa1808312. URL: <https://doi.org/10.1056/NEJMoa1808312>.
10. Li, Ho-Kwong et al. (Jan. 2019). "Oral versus Intravenous Antibiotics for Bone and Joint Infection". In: *New England Journal of Medicine* 380.5, pp. 425–436. ISSN: 0028-4793. doi: 10.1056/NEJMoa1710926. URL: <https://doi.org/10.1056/NEJMoa1710926>.

11. Spellberg, Brad et al. (May 2020). "Evaluation of a Paradigm Shift From Intravenous Antibiotics to Oral Step-Down Therapy for the Treatment of Infective Endocarditis: A Narrative Review". In: *JAMA internal medicine* 180.5, pp. 769–777. ISSN: 2168-6114. doi: 10.1001/jamainternmed.2020.0555.
12. Platts, Stephen et al. (Nov. 2022). "Oral step-down for *Staphylococcus aureus* bacteraemia: An opportunity for antimicrobial stewardship?" In: *Clinical Infection in Practice* 16, p. 100202. ISSN: 2590-1702. doi: 10.1016/j.clinpr.2022.100202. URL: <https://www.sciencedirect.com/science/article/pii/S259017022200070X>.
13. Wald-Dickler, Noah et al. (Mar. 2022). "Oral Is the New IV. Challenging Decades of Blood and Bone Infection Dogma: A Systematic Review". In: *The American Journal of Medicine* 135.3, 369–379.e1. ISSN: 0002-9343, 1555-7162. doi: 10.1016/j.amjmed.2021.10.007. URL: [https://www.amjmed.com/article/S0002-9343\(21\)00699-9/fulltext](https://www.amjmed.com/article/S0002-9343(21)00699-9/fulltext).
14. Kaasch, Achim J. et al. (July 2023). *Early oral switch in low-risk Staphylococcus aureus bloodstream infection*. doi: 10.1101/2023.07.03.23291932. URL: <https://www.medrxiv.org/content/10.1101/2023.07.03.23291932v1>.
15. Ray-Barruel, Gillian et al. (Aug. 2019). "Effectiveness of insertion and maintenance bundles in preventing peripheral intravenous catheter-related complications and bloodstream infection in hospital patients: A systematic review". In: *Infection, Disease & Health* 24.3, pp. 152–168. ISSN: 2468-0451. doi: 10.1016/j.idh.2019.03.001. URL: <https://www.sciencedirect.com/science/article/pii/S2468045119300161>.
16. McMeekin, Nicola et al. (Jan. 2020). "Cost-effectiveness of oral versus intravenous antibiotics (OVIVA) in patients with bone and joint infection: evidence from a non-inferiority trial". In: *Wellcome Open Research* 4, p. 108. ISSN: 2398-502X. doi: 10.12688/wellcomeopenres.15314.4. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6944252/>.
17. UK Health Security Agency (Nov. 2022). *National antimicrobial intravenous-to-oral switch (IVOS) criteria for early switch*. URL: <https://www.gov.uk/government/publications/antimicrobial-intravenous-to-oral-switch-criteria-for-early-switch/national-antimicrobial-intravenous-to-oral-switch-ivos-criteria-for-early-switch>.
18. Shraytey, Zeina M, Rahal, Mohamad K, and Malaeb, Diana N (Dec. 2014). "Practice of switch from intravenous to oral antibiotics". In: *SpringerPlus* 3, p. 717. ISSN: 2193-1801. doi: 10.1186/2193-1801-3-717. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4320166/>.
19. Rawson, Timothy Miles et al. (Dec. 2016). "Mapping the decision pathways of acute infection management in secondary care among UK medical physicians: a qualitative study". In: *BMC Medicine* 14.1, p. 208. ISSN: 1741-7015. doi: 10.1186/s12916-016-0751-y. URL: <https://doi.org/10.1186/s12916-016-0751-y>.

20. Hospenthal, Duane R et al. (Dec. 2020). "Practice Patterns of Infectious Diseases Physicians in Transitioning From Intravenous to Oral Therapy in Patients With Bacteremia". In: *Open Forum Infectious Diseases* 7.12, ofz386. ISSN: 2328-8957. doi: 10.1093/ofid/ofz386. URL: <https://doi.org/10.1093/ofid/ofz386>.
21. Buis, D T P et al. (Oct. 2022). "Current clinical practice in antibiotic treatment of Staphylococcus aureus bacteraemia: results from a survey in five European countries". In: *Journal of Antimicrobial Chemotherapy* 77.10, pp. 2827–2834. ISSN: 0305-7453. doi: 10.1093/jac/dkac237. URL: <https://doi.org/10.1093/jac/dkac237>.
22. Pandolfo, Alyssa M. et al. (Mar. 2022). "Understanding decisions about antibiotic prescribing in ICU: an application of the Necessity Concerns Framework". In: *BMJ Quality & Safety* 31.3, pp. 199–210. ISSN: 2044-5415, 2044-5423. doi: 10.1136/bmjqs-2020-012479. URL: <https://qualitysafety.bmj.com/content/31/3/199>.
23. Hamilton, R. A. et al. (Aug. 2024). "Nurses' attitudes, behaviours, and enablers of intravenous to oral switching (IVOS) of antibiotics: a mixed-methods survey of nursing staff in secondary care hospitals across the Midlands region of England". In: *Journal of Hospital Infection* 150, pp. 9–16. ISSN: 0195-6701. doi: 10.1016/j.jhin.2024.05.007. URL: <https://www.sciencedirect.com/science/article/pii/S0195670124001774>.
24. Sendi, Parham et al. (Sept. 2023). "Early switch from intravenous to oral anti-microbial therapy in infectious diseases". In: *Clinical Microbiology and Infection* 29.9, pp. 1117–1119. ISSN: 1198-743X. doi: 10.1016/j.cmi.2023.06.021. URL: [https://www.clinicalmicrobiologyandinfection.com/article/S1198-743X\(23\)00299-9/fulltext](https://www.clinicalmicrobiologyandinfection.com/article/S1198-743X(23)00299-9/fulltext).
25. Rawson, T. M. et al. (Aug. 2017). "A systematic review of clinical decision support systems for antimicrobial management: are we failing to investigate these interventions appropriately?" In: *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 23.8, pp. 524–532. ISSN: 1469-0691. doi: 10.1016/j.cmi.2017.02.028.
26. Peiffer-Smadja, N. et al. (May 2020). "Machine learning for clinical decision support in infectious diseases: a narrative review of current applications". In: *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 26.5, pp. 584–595. ISSN: 1469-0691. doi: 10.1016/j.cmi.2019.09.009.
27. Bolton, William J. et al. (2022a). "Machine learning and synthetic outcome estimation for individualised antimicrobial cessation". In: *Frontiers in Digital Health* 4. ISSN: 2673-253X. URL: <https://www.frontiersin.org/articles/10.3389/fdgth.2022.997219>.
28. D'Hondt, Ellie et al. (Dec. 2022). "Identifying and evaluating barriers for the implementation of machine learning in the intensive care unit". In: *Communications Medicine* 2.1, pp. 1–12. ISSN: 2730-664X. doi: 10.1038/s43856-022-00225-1. URL: <https://www.nature.com/articles/s43856-022-00225-1>.

29. Goldberger, A. L. et al. (June 2000). "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals". In: *Circulation* 101.23, E215–220. ISSN: 1524-4539. doi: 10.1161/01.cir.101.23.e215.
30. Johnson, Alistair et al. (2021). *MIMIC-IV*. doi: 10.13026/s6n6-xd98. URL: <https://physionet.org/content/mimiciv/1.0/>.
31. Pollard, Tom J. et al. (Sept. 2018). "The eICU Collaborative Research Database, a freely available multi-center database for critical care research". In: *Scientific Data* 5.1, p. 180178. ISSN: 2052-4463. doi: 10.1038/sdata.2018.178. URL: <https://www.nature.com/articles/sdata2018178>.
32. Pollard, Tom Joseph et al. (2019). *The eICU Collaborative Research Database (version 2.0)*. PhysioNet. doi: 10.13026/C2WM1R. URL: <https://doi.org/10.13026/C2WM1R>.
33. Agency for Healthcare Research and Quality (2019). *Four Moments of Antibiotic Decision Making*. <https://www.ahrq.gov/antibiotic-use/acute-care/four-moments/index.html>.
34. Gerber, Jeffrey S. et al. (Aug. 2017). "Development and Application of an Antibiotic Spectrum Index for Benchmarking Antibiotic Selection Patterns Across Hospitals". In: *Infection Control & Hospital Epidemiology* 38.8, pp. 993–997. ISSN: 0899-823X, 1559-6834. doi: 10.1017/ice.2017.94. URL: <https://www.cambridge.org/core/journals/infection-control-and-hospital-epidemiology/article/development-and-application-of-an-antibiotic-spectrum-index-for-benchmarking-antibiotic-selection-patterns-across-hospitals/6CC7D60E4CBCBFF83CA52C37462CE84F>.
35. Korzilius, Julia W et al. (June 2023). "Oral antimicrobial agents in patients with short bowel syndrome: worth a try!" In: *Journal of Antimicrobial Chemotherapy*, dkad198. ISSN: 0305-7453. doi: 10.1093/jac/dkad198. URL: <https://doi.org/10.1093/jac/dkad198>.
36. Lubba, Carl H. et al. (Nov. 2019). "catch22: CANonical Time-series CHaracteristics". In: *Data Mining and Knowledge Discovery* 33.6, pp. 1821–1852. ISSN: 1573-756X. doi: 10.1007/s10618-019-00647-x. URL: <https://doi.org/10.1007/s10618-019-00647-x>.
37. Löning, Markus et al. (Sept. 2019). *sktime: A Unified Interface for Machine Learning with Time Series*. doi: 10.48550/arXiv.1909.07872. URL: <http://arxiv.org/abs/1909.07872>.
38. Löning, Markus et al. (Sept. 2022). *sktime/sktime: v0.13.4*. doi: 10.5281/zenodo.7117735. URL: <https://zenodo.org/record/7117735>.
39. Lundberg, Scott M and Lee, Su-In (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc. URL: <https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf>.

40. Vafaie, H. and De Jong, K. (Nov. 1992). "Genetic algorithms as a tool for feature selection in machine learning". In: *Proceedings Fourth International Conference on Tools with Artificial Intelligence TAI '92*, pp. 200–203. doi: 10.1109/TAI.1992.246402.
41. Paszke, Adam et al. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., pp. 8024–8035.
42. Akiba, Takuya et al. (2019). "Optuna: A Next-generation Hyperparameter Optimization Framework". In: *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
43. Youden, W. J. (Jan. 1950). "Index for rating diagnostic tests". In: *Cancer* 3.1, pp. 32–35. ISSN: 0008-543X. doi: 10.1002/1097-0142(1950)3:1<32::aid-cnrcr2820030106>3.0.co;2-3.
44. Chawla, N. V. et al. (June 2002). "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16, pp. 321–357. ISSN: 1076-9757. doi: 10.1613/jair.953. URL: <https://www.jair.org/index.php/jair/article/view/10302>.
45. Kingma, Diederik P. and Ba, Jimmy (2014). "Adam: A Method for Stochastic Optimization". In: URL: <http://arxiv.org/abs/1412.6980>.
46. Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.
47. Virtanen, Pauli et al. (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17, pp. 261–272. doi: 10.1038/s41592-019-0686-2.
48. Broek, Annemieke K. van den et al. (Mar. 2021). "Systematic review: the bioavailability of orally administered antibiotics during the initial phase of a systemic infection in non-ICU patients". In: *BMC Infectious Diseases* 21.1, p. 285. ISSN: 1471-2334. doi: 10.1186/s12879-021-05919-w. URL: <https://doi.org/10.1186/s12879-021-05919-w>.
49. Crabbe, Jonathan et al. (2021). "Explaining Latent Representations with a Corpus of Examples". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., pp. 12154–12166.
50. Rawson, Timothy M et al. (June 2021). "A Real-world Evaluation of a Case-based Reasoning Algorithm to Support Antimicrobial Prescribing Decisions in Acute Care". In: *Clinical Infectious Diseases* 72.12, pp. 2103–2111. ISSN: 1058-4838. doi: 10.1093/cid/ciaa383. URL: <https://doi.org/10.1093/cid/ciaa383>.
51. Hardt, Moritz, Price, Eric, and Srebro, Nathan (Oct. 2016). "Equality of Opportunity in Supervised Learning". In: *arXiv:1610.02413 [cs]*. URL: <http://arxiv.org/abs/1610.02413>.
52. Bird, Sarah et al. (May 2020). *Fairlearn: A toolkit for assessing and improving fairness in AI*. Tech. rep. MSR-TR-2020-32. Microsoft. URL: <https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/>.

53. NHS (n.d.). *A typical day on an NHS ward*. URL: <https://www.myhsc.co.uk/top-tip/a-typical-day-on-an-nhs-ward/>.
54. Lee, Juho et al. (May 2019). *Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks*. DOI: 10.48550/arXiv.1810.00825.
55. Bolton, William J. et al. (Nov. 2022b). "Developing moral AI to support decision-making about antimicrobial use". In: *Nature Machine Intelligence* 4.11, pp. 912–915. ISSN: 2522-5839. DOI: 10.1038/s42256-022-00558-5. URL: <https://www.nature.com/articles/s42256-022-00558-5>.
56. Herm, Lukas-Valentin et al. (June 2022). "Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability". In: *International Journal of Information Management*, p. 102538. ISSN: 0268-4012. DOI: 10.1016/j.ijinfomgt.2022.102538. URL: <https://www.sciencedirect.com/science/article/pii/S026840122200072X>.
57. Afroze, Sharmin et al. (Sept. 2022). "Subpopulation-specific machine learning prognosis for underrepresented patients with double prioritized bias correction". In: *Communications Medicine* 2.1, pp. 1–14. ISSN: 2730-664X. DOI: 10.1038/s43856-022-00165-w. URL: <https://www.nature.com/articles/s43856-022-00165-w>.
58. Schouten, Bo et al. (Dec. 2022). "Implementing artificial intelligence in clinical practice: a mixed-method study of barriers and facilitators". In: *Journal of Medical Artificial Intelligence* 5.0. ISSN: 2617-2496. DOI: 10.21037/jmai-22-71. URL: <https://jmai.amegroups.com/article/view/7519>.
59. Barredo Arrieta, Alejandro et al. (June 2020). "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58, pp. 82–115. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2019.12.012. URL: <https://www.sciencedirect.com/science/article/pii/S1566253519308103>.
60. Cinà, Giovanni et al. (June 2022). *Why we do need Explainable AI for Healthcare*. DOI: 10.48550/arXiv.2206.15363. URL: <http://arxiv.org/abs/2206.15363>.
61. Smith, Megan, Higgs, Joy, and Ellis, Elizabeth (2008). "Factors influencing clinical decision making". In: *Clinical reasoning in the health professions* 3, pp. 89–100.
62. McCarthy, Kate and Avent, Minyon (Apr. 2020). "Oral or intravenous antibiotics?" In: *Australian Prescriber* 43.2, pp. 45–48. ISSN: 0312-8008. DOI: 10.18773/austprescr.2020.008. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7186270/>.
63. Office for Health Improvement and Disparities (2024). *Public health profiles*. URL: <https://fingertips.phe.org.uk>.
64. NHS (2024). *NHS England – National Cost Collection for the NHS*. URL: <https://www.england.nhs.uk/costing-in-the-nhs/national-cost-collection/>.

65. Higgins, David and Madai, Vince I. (2020). "From Bit to Bedside: A Practical Framework for Artificial Intelligence Product Development in Healthcare". In: *Advanced Intelligent Systems* 2.10, p. 2000052. ISSN: 2640-4567. doi: 10.1002/aisy.202000052. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/aisy.202000052>.
66. McCradden, Melissa D., Stephenson, Elizabeth A., and Anderson, James A. (Sept. 2020). "Clinical research underlies ethical integration of healthcare artificial intelligence". In: *Nature Medicine* 26.9, pp. 1325–1326. ISSN: 1546-170X. doi: 10.1038/s41591-020-1035-9. URL: <https://www.nature.com/articles/s41591-020-1035-9>.
67. Ahmed, Molla Imaduddin et al. (Oct. 2023). "A Systematic Review of the Barriers to the Implementation of Artificial Intelligence in Healthcare". In: *Cureus* 15.10, e46454. ISSN: 2168-8184. doi: 10.7759/cureus.46454. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10623210/>.
68. Lambert, Sophie Isabelle et al. (June 2023). "An integrative review on the acceptance of artificial intelligence among healthcare professionals in hospitals". In: *npj Digital Medicine* 6.1, pp. 1–14. ISSN: 2398-6352. doi: 10.1038/s41746-023-00852-5. URL: <https://www.nature.com/articles/s41746-023-00852-5>.
69. Olaye, Iredia M. and Seixas, Azizi A. (May 2023). "The Gap Between AI and Bedside: Participatory Workshop on the Barriers to the Integration, Translation, and Adoption of Digital Health Care and AI Startup Technology Into Clinical Practice". In: *Journal of Medical Internet Research* 25.1, e32962. doi: 10.2196/32962. URL: <https://www.jmir.org/2023/1/e32962>.
70. Holmes, Alison H et al. (Jan. 2016). "Understanding the mechanisms and drivers of antimicrobial resistance". In: *The Lancet* 387.10014, pp. 176–187. ISSN: 0140-6736. doi: 10.1016/S0140-6736(15)00473-0. URL: <http://www.sciencedirect.com/science/article/pii/S0140673615004730>.
71. Tamma, Pranita D., Miller, Melissa A., and Cosgrove, Sara E. (Jan. 2019). "Rethinking How Antibiotics Are Prescribed: Incorporating the 4 Moments of Antibiotic Decision Making Into Clinical Practice". In: *JAMA* 321.2, pp. 139–140. ISSN: 0098-7484. doi: 10.1001/jama.2018.19509. URL: <https://doi.org/10.1001/jama.2018.19509>.
72. Charani, Esmita et al. (Aug. 2021). "Optimising antimicrobial use in humans – review of current evidence and an interdisciplinary consensus on key priorities for research". In: *The Lancet Regional Health – Europe* 7, p. 100161. ISSN: 2666-7762. doi: 10.1016/j.lanepe.2021.100161. URL: <https://www.sciencedirect.com/science/article/pii/S2666776221001381>.
73. Cantón, Rafael et al. (Sept. 2022). "Relevance of the Consensus Principles for Appropriate Antibiotic Prescribing in 2022". In: *Journal of Antimicrobial Chemotherapy* 77.Supplement_1, pp. i2–i9. ISSN: 0305-7453. doi: 10.1093/jac/dkac211. URL: <https://doi.org/10.1093/jac/dkac211>.

Chapter 6

Clinician Evaluation

Abstract

Challenges exist when translating Artificial intelligence (AI)-driven clinical decision support systems (CDSSs) from the research setting into healthcare systems. Particularly in infectious diseases where behaviour, culture, uncertainty, and the frequent lack of ground truth enhance the complexity of medical decision-making. AI CDSSs hold promise to optimise prescribing and tackle antimicrobial resistance. In our previous research, we developed recommendation systems to support intravenous (IV)-to-oral antibiotic switching (Chapter 5). Here, we conducted a clinician evaluation of this technology to understand user perceptions and how it influences decision-making.

A randomised clinical vignette study was used with semi-structured interviews, demographics, and surveys to triangulate findings. Each vignette contained a patient currently on IV antibiotics, with clinicians ($n=42$) asked to decide if they were suitable for IV-to-oral switching (IVOS) or not. Different interventions, standard of care (SOC) through the UK Health Security Agency (UKHSA) IVOS guidelines or SOC alongside CDSS predictions and their explanations were experienced by participants for each vignette across two arms. AI model explanations included similar patients, feature importance and free text insights. This design allowed us to understand the impact of the AI system, its explanations, and test nuances such as if healthcare professionals can detect when the AI predictions are clearly wrong.

In most cases, decisions were equivalent between interventions with clinicians able to identify and ignore incorrect advice. The presence of CDSS predictions caused a statistically significant increase in clinicians not switching for 1 scenario ($p\text{-value} < 0.01$). The same trend was seen in other scenarios in subgroups including females and when controlling for sex

and grade. 9% of the participants used AI explanations when available. Completion times were equivalent between the SOC and CDSS interventions. The web app AI-CDSS developed obtained a good system usability survey score of 72.32 (out of 100) and a perceived usefulness of 3.59 (out of 5).

This AI CDSS has the potential to support antimicrobial stewardship with a greater influence on clinicians when it recommends not to switch. How useful clinicians find a system affects its impact with AI explanations seeming to have little effect. Trust in such a system is mainly built through clinical evidence with ease of use being essential for adoption.

This chapter’s research is currently out for peer review at the Lancet Digital Health. Furthermore, it was presented at the International Congress on Infectious Diseases (ICID) 2024. An overview of the research study conducted in this chapter is shown in Figure 6.1.

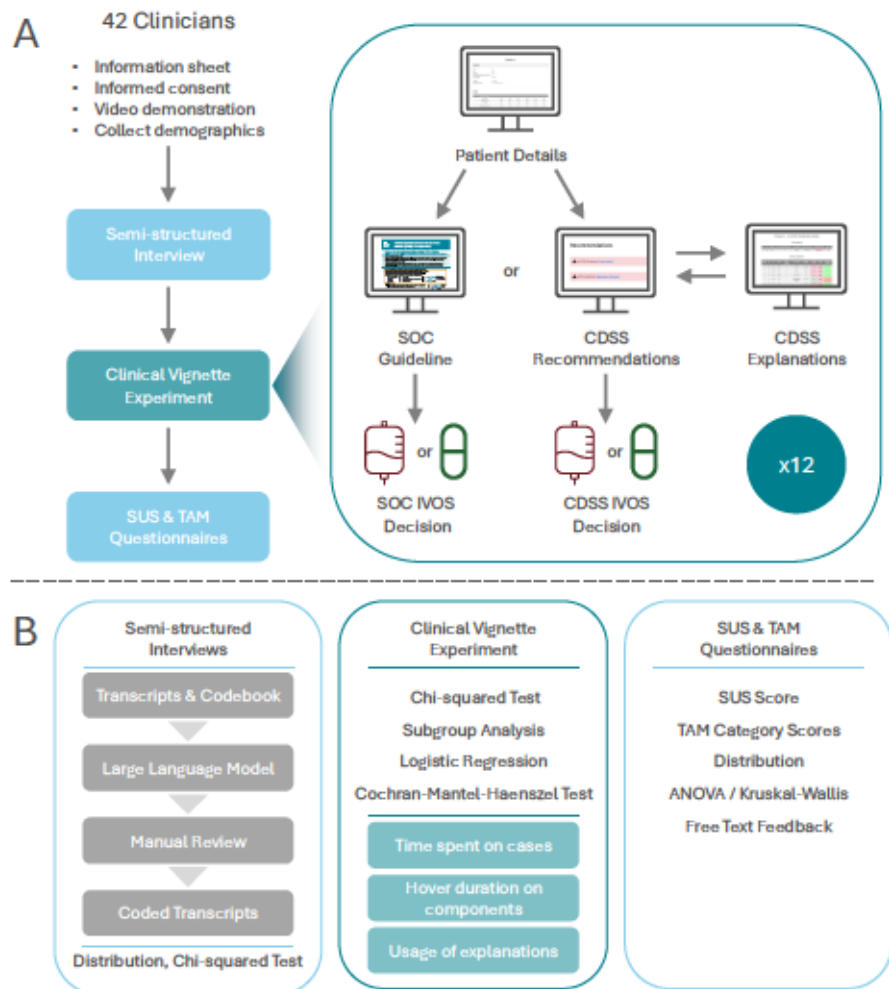


Figure 6.1: Overview of the experimental protocol conducted with clinicians in this study [A] and the subsequent analysis undertaken [B].

6.1 Chapter objectives

- *Develop a web app that incorporates an AI CDSS for IVOS*
- *Conduct a study with clinicians consisting of interviews, questionnaires and a case vignette experiment to discern the ability of the AI CDSS to influence antibiotic prescribing decisions in a controlled, simulated environment*
- *Understand user perceptions, barriers for implementation and adoption, as well as where such technology could fit in the clinical pathway*

6.2 Introduction

AI-based CDSSs have promised to revolutionise healthcare (Alowais et al. 2023). However, a valley of death exists, with few technologies truly translated, integrated and adopted (Higgins et al. 2020; McCradden et al. 2020; Ahmed et al. 2023; Lambert et al. 2023; Olaye et al. 2023). This is particularly true for systems that aim to improve antibiotic prescribing in order to tackle the growing problem of antimicrobial resistance (Peiffer-Smadja et al. 2020; Tokgöz et al. 2024). This speciality has numerous complexities compared to areas where AI uptake has been more successful to date, such as radiology or stroke diagnosis (Mello-Thoms et al. 2023; NICE 2024). For example, there is no clear gold standard ground truth, decisions are often made in uncertain scenarios with a lack of information, behavioural factors such as risk appetite play a key role in decision-making, clinical practice is heterogeneous, and a balance is needed between the benefits and risks to an individual versus those of wider society (Wojcik et al. 2021; Bolton et al. 2022; Tarrant et al. 2022; Poluektova et al. 2023).

To cross the translation gap, robust evaluations and clinical evidence are necessary (Kelly et al. 2019; Sendak et al. 2020; Vasey et al. 2022; White et al. 2023). Numerous aspects

of AI CDSSs must be considered, including who is the end user, what is the intended use, and how and when predictions are presented. Furthermore, the potential impact of the tool and any explanation must be understood. In Chapter 5, we developed an AI-driven CDSS to predict based on routinely collected clinical information when a patient could be suitable for switching from intravenous (IV) to oral antibiotics (Bolton et al. 2024). Some of these models utilise the co-morbidity representation pipeline created in Chapter 3. In this chapter, we evaluated this technology using a randomised clinical vignette web app-based study alongside semi-structured interviews and surveys. The aim was to understanding the impact of the AI system, and its explanations on decisions, as well as participants' perspectives and how such technology could fit into a clinical pathway. This study provides further evidence of the potential of this AI IVOS technology and advances the literature on the appropriate design, utilisation, and clinician evaluation of AI CDSSs.

6.3 Methods

6.3.1 Study and participants

Healthcare professionals regularly involved in antibiotic prescribing were recruited through personal networks and the British Infection Association general email list between the beginning of May and the end of July 2024. Participants received an information sheet on the study before informed consent was obtained and a time to complete the study was scheduled. The study was split into three phases; a pre-intervention semi-structured interview, the clinical vignette experiment, and questionnaires. All parts of the study were completed in one 60 to 90-minute online meeting. The final two stages of the research study were conducted through a web application created using Django, HTML, and JavaScript. Before the case vignette experiment and questionnaires, clinicians' medical speciality, grade, age (grouped by decade), sex, and familiarity with AI (collectively referred to as demographics) were collected through the web app. This study was approved by the Research Governance and Integrity Team (RGIT) at Imperial College London (ICREC reference 6880677).

6.3.2 Semi-structured interviews

The objective of the semi-structured interviews was to understand the current antibiotic prescribing of clinicians and their perception of AI. A topic guide (overview included in the appendix at Page 240) containing example questions was developed in collaboration with clinicians and pharmacists to ensure that all relevant points were covered during the discussions. This included their current antimicrobial prescribing practices, factors that influence their decision-making, perceptions on AI, and what would convince them to consistently use an AI-driven CDSS. Interviews were recorded and transcribed through Zoom. To convert raw qualitative data into quantitative data, a content analysis coding process was used (Erlingsson et al. 2017). A codebook (Table 6.3) was deductively developed based on the topic guide to facilitate the coding process, with inductive processes to adapt and adjust codes as needed during analysis. Recent research has highlighted the ability of language models to assist with qualitative research (Xiao et al. 2023; Dunivin 2024; Tai et al. 2024). To this extent, GPT-4o (version released on the 13th of May 2024) was used to fill in the code book for each transcript. Coding each transcript involved assigning a binary 1 or 0 to each code depending on whether the participant mentioned that particular aspect during the interview. Prompt engineering was initially used to improve the Cohen-Kappa agreement score between a language model output and a manually coded transcript, to ensure that the language model coding reflected our perspectives, the Cohen-Kappa score was increased from 0.38 (fair agreement) to 0.72 which is considered substantial agreement. The final prompt can be viewed in the appendix at Page 241 and involved few shot in-context learning (i.e., including some examples in the prompt) as well as providing the language model with the codebook. In addition to prompt engineering, each CSV output produced by the language model was cross-checked by a human researcher against the original transcript to ensure consistency and prevent incorrect coding. This language model-powered coding method was utilised to accelerate the research, given manual coding is a time consuming process. Once coding was completed frequencies were calculated and the Chi-Square test ($\alpha=0.05$) was used to understand if there were any differences in perspectives raised during the interviews across participant demographics.

6.3.3 Clinical vignette experiment

Table 6.1: Details of the experimental setup and case vignettes.

Patient	Intervention		AI CDSS recommendation	Guideline CDSS recommendation	Ground Truth	Data	Group
	Arm A	Arm B					
1	CDSS	SOC	Don't switch	Don't switch	Don't switch	ICHT	AI correct
2	SOC	CDSS	Switch	Switch	Switch	ICHT	AI correct
3	SOC	CDSS	Switch	Don't switch	Don't switch	ICHT	AI incorrect
4	CDSS	SOC	Don't switch	Switch	Switch	ICHT	AI incorrect
5	SOC	CDSS	Switch	Don't switch	–	ICHT	Uncertain
6	CDSS	SOC	Don't switch	Switch	–	ICHT	Uncertain
7	CDSS	SOC	Don't switch	Don't switch	Didn't switch	MIMIC	AI correct
8	SOC	CDSS	Switch	Switch	Switched	MIMIC	AI correct
9	SOC	CDSS	Potentially switch	Switch	Didn't switch	MIMIC	Uncertain
10	CDSS	SOC	Don't switch	Switch	Switched	MIMIC	AI incorrect
11	CDSS	SOC	Switch	Don't switch	Didn't switch	MIMIC	AI incorrect
12	SOC	CDSS	Potentially switch	Don't switch	Switched	MIMIC	Uncertain

The objective of the clinical vignette experiment was to determine whether an AI CDSS can influence IV-to-oral switch decision-making. 12 clinical vignettes were used given the time available with each participant. Each vignette involved a patient currently on IV antibiotics. 6 were simulated cases based on patients' real data distribution in Imperial College Healthcare NHS Trust (ICHT). The Imperial Clinical Analytics, Research and Evaluation (iCARE) environment was used to access de-identified, real-world, clinical data for those with inpatient encounter at ICHT. The design of these cases was informed by expert clinicians to focus on scenarios they regularly encounter and investigate the research hypothesis. The other 6 were real patient cases randomly extracted from MIMIC-IV a large, publicly available, de-identified, real-world, clinical dataset (Goldberger et al. 2000; Johnson et al. 2022). Details of the cases are provided in Table 6.1. The ground truth for each vignette was defined as the real action undertaken for MIMIC cases and was determined by expert clinical collaborators for the ICHT vignettes. For each of the ICHT and MIMIC cases we specified that for 2 the AI CDSS switch predictions would be correct relative to the ground truth, 2 were incorrect, and 2 were uncer-

tain. For each of those three groups in one case the AI model would recommend switch or potentially switch while on the other it would recommend don't switch. All clinical vignettes were completed in the same order by all participants. A web app was developed to present the case vignettes and surveys to participants as well as record their results and actions. This approach was chosen given its scalability and because it simulates what interacting with the CDSS would be like in a real-world environment. Developing the web app involved creating a Django web app with different models that define data structures and handle interactions with databases. The key models in this application were the user (i.e., the clinician completing the study) and the patient which represented the different case vignettes. Numerous views were used to process users' requests, retrieve data from models, and pass it on to templates. These HTML templates presented the information to clinicians and used JavaScript to handle some user interactions. Finally, CSS was used for styling. Pilot testing with clinicians enabled for the design and functionality of the web app to be iteratively build based on end-users feedback. This ensured the web app was accessible and the study could be completed by clinicians with a range of digital proficiency. Figure 6.2 shows example screenshots of case vignettes in the web app.

An illustration of the study design is shown in Figure 6.1. For each case vignette, participants were asked to determine whether the patient in question was suitable for stepping down from IV-to-oral antibiotic treatment. They were given a binary 'Switch' or 'Don't switch' selection box as well as a free text box to explain their reasoning or add any clarifications. This binary option does not encompass all of the possible permutations of clinical care seen for real patients but it allows us to focus this study on IV-to-oral antibiotic switching and disentangle the uncertainty in antimicrobial switch decision-making which is frequently seen. Participants were told to only select 'Switch' if they would change that patient's treatment to oral antibiotics or if they would stop antibiotics completely (another form of de-escalation, which would be written down in the free text box). Furthermore, it was noted that none of the patients have an infection that requires special consideration, their enteral route was not compromised (unless the Glasgow Coma Score said otherwise), and an oral option was available. This was to clarify commonly requested information and highlight that in this exercise we were focused

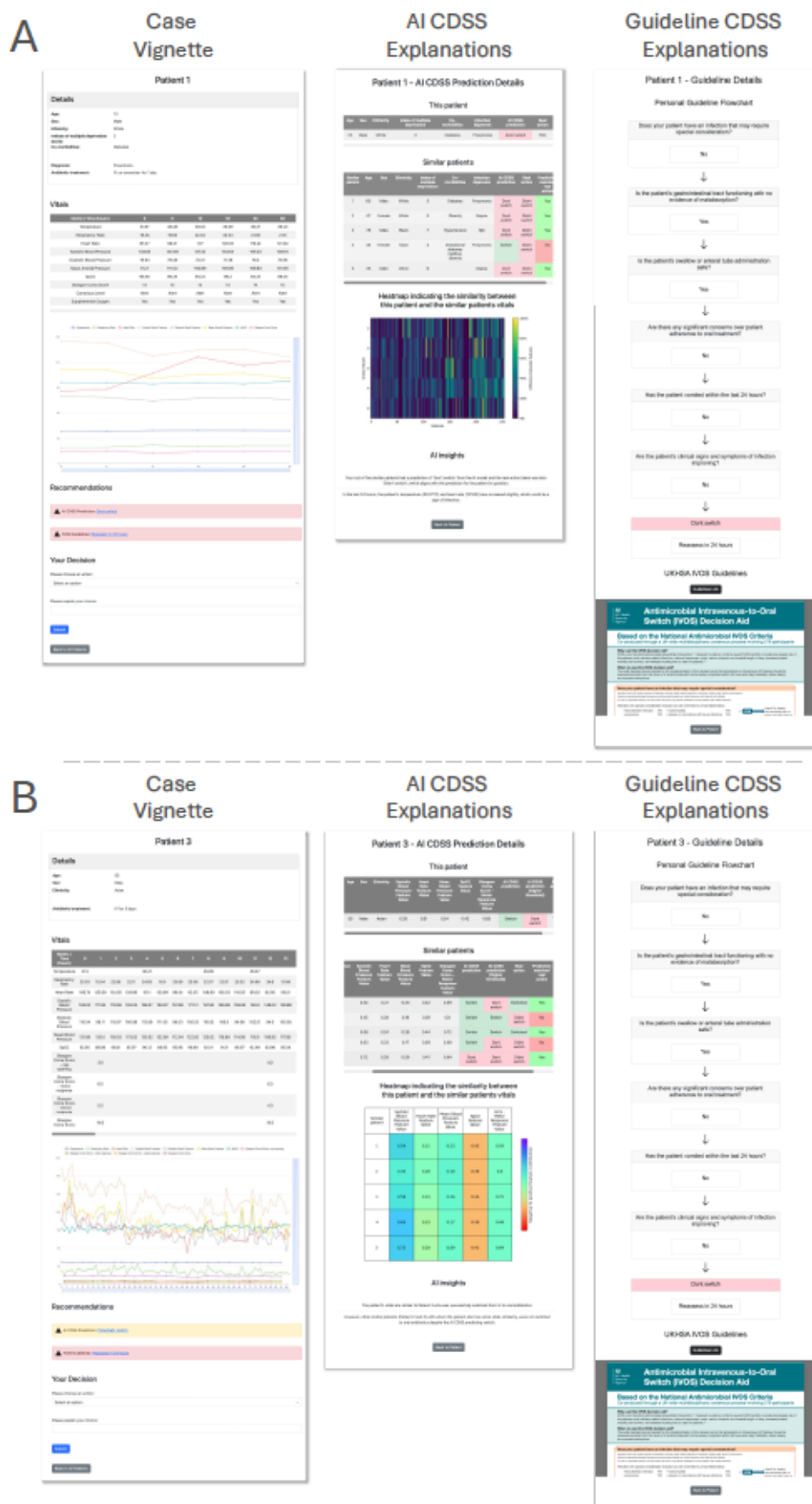


Figure 6.2: Screenshots of the web app used for the case vignette experiment. A is a simulated patient from ICHT while B is from MIMIC. Both examples are from the video demonstration used to inform participants on how to complete the study.

on understanding from a clinical sign and symptom and infection marker perspective if it was appropriate to switch that patient to oral antibiotics. Before beginning the experiment, the task at hand was explained to the participants and a demonstration of the exercise and the web app was given through a video and three example clinical vignettes. Finally, any questions the subjects had were answered to ensure that they fully understood the experiment.

Participants were randomly assigned to two experimental arms (A or B). Arms influenced what a participant viewed and had access to for a particular vignette (Table 6.1). First, standard of care (SOC) information acted as our control. Subjects could view patient data, an embedded PDF of the UK Health Security Agency (UKHSA) IV-to-oral switching (IVOS) decision aid, and they had a link to the UKHSA IVOS website (UK Health Security Agency 2023). Providing access to the AI-CDSS acted as our intervention. Subjects had access to all the information available in standard of care, plus the CDSS's recommendations (from an AI model and guideline-based system using the UKHSA IVOS criteria, these were shown below the standard of care information on the web page) and their explanations (such as similar patients, feature importance, and free text insights) through a separate link as detailed in Section 6.3.3.1. For each arm, each vignette was assigned to either the SOC or CDSS intervention. The two arms were opposites of each other in terms of the intervention for a particular case vignette, with each having 6 decision support (3 for ICHT cases and 3 for MIMIC cases) and 6 SOC interventions (3 for ICHT cases and 3 for MIMIC cases); this ensured the amount of data collected for each intervention across vignettes was equal and allows for estimation of between arm variability by controlling for the case. Data displayed for the ICHT cases included the patient's age, sex, ethnicity, indices of multiple deprivation decile, co-morbidities, diagnosis, antibiotic treatment, duration of treatment, temperature, respiratory rate, heart rate, systolic blood pressure, diastolic blood pressure, mean blood pressure, peripheral oxygen saturation (SpO₂), Glasgow Coma Score, conscious level and oxygen status. MIMIC cases displayed similar data but did not include the indices of multiple deprivation decile, co-morbidities, diagnosis, antibiotic treatment, consciousness level or oxygen status. Demographic and infection information was displayed at the top of the web page with clinical data that varied over time displayed below in both a table and an interactive graph. This data was provided in multiple formats to

investigate preferences and key drivers of decision-making, while visualisations were designed to mimic how real UK clinicians often encounter patient data in their electronic health records. Data collected for each participant on each vignette included: intervention experienced, binary switch choice ('Switch' or 'Don't switch'), free text explanation, mouse hover time on each element, URL clicks (for example, to see AI explanations) and time elapsed. Overall, this design allowed us to understand the impact of the AI system, its explanations and test nuances such as if healthcare professionals can detect and ignore when the AI recommendations are wrong.

For statistical analysis, all demographics were converted into categorical variables. Age was grouped by decade, sex by male and female, medical speciality by infectious diseases, microbiology, pharmacist, and other. Grade was grouped into consultant and other. Finally, familiarity with AI was assessed on a 5-point Likert scale. Any missing values were imputed with the median. All alphas were set at 0.05. Switch decisions were analysed using the Chi-square test (with Yates correction) to assess if there were statistically significant differences in choices between those who experienced SOC versus the decision support system intervention for each case vignette. Subgroup analysis was also conducted using this method for each demographic and based on responses to the questionnaires. The Cochran-Mantel-Haenszel test was additionally used to determine if significant associations existed between interventions and decisions while controlling for a third demographic variable. Logistic regression quantified the effect of the intervention experienced on participants' switch choice. The variability in decisions was analysed through the Shannon entropy. This method results in a value between 0 and 1, where a higher value indicates more diversity in decisions. The hover durations over various components within the experiment were aggregated to act as a proxy for which aspects had a greater influence on clinicians' decision-making. Time spent on each case vignette was combined across participants with the mean and standard deviations calculated. This was also done according to the intervention subjects experienced, using the paired t-test (after meeting parametric requirements) to see if there were differences in means. Finally, the percentage of participants that clicked on URLs such as the AI explanations was determined at any time and across all vignettes.

6.3.3.1 Clinical decision support systems

Two very similar AI clinical decision support systems were used to provide IVOS recommendations. Both were described in chapter 5, with the case vignettes utilising real data from MIMIC using the MIMIC trained model (Bolton et al. 2024) and the simulated ICHT cases using the ICHT model that was prospectively evaluated. For the MIMIC model, the *simplex* (Crabbe et al. 2021) method enabled similar patients to be identified and displayed, as well as their feature contribution. For the ICHT model explanations, the top 5 most similar patients in the training set were extracted from the embedding space using k-dimensional trees with a feature similarity plot indicating the similarities and differences between those similar patients and the patient in question across the feature set. For all vignettes free text insights were generated using the Gemini Pro multi-modal language model (Team et al. 2024)(version active on the 27th of March 2024). The prompts used can be found in the appendix (Pages 242 and 243); each textual prompt was accompanied by a screenshot of the AI explanation for that patient and their vitals data in CSV format. These bullet points highlighted key aspects of the patient's data and AI models and were included to understand clinician perceptions toward textual explanations. In the decision support intervention alongside AI model predictions, a switch recommendation was also given by a strict rule-based model using the UKHSA IVOS criteria. This was added to understand clinician perceptions of a decision support system based on approved guidelines as opposed to AI. The guideline CDSS aligned with the AI recommendations in those vignettes where the AI CDSS prediction was correct but disagreed in the other groups. For explainability, a flow diagram indicating which criteria failed was shown, as well as an embedded PDF of the UKHSA IVOS decision aid and a link to the UKHSA IVOS website (UK Health Security Agency 2023). All explainability aspects for AI CDSS and guideline CDSS were located on separate web pages from the main vignette page, where data and decisions could be viewed and made for a particular patient. This aimed to prevent information overload (Laker et al. 2018; Phillips-Wren et al. 2020) and ensured that we could reliably track when participants wanted explanations.

6.3.4 Questionnaires

The final part of the study involved subjects completing the System Usability Scale (SUS, (Lewis 2018)) and Technology Acceptance Model (TAM, (Holden et al. 2010)) questionnaires. The purpose of the SUS survey is to assess participants' satisfaction with the application and interface, while the more nuanced TAM questions were designed to understand the acceptance of AI technology by clinicians and their opinions of the AI CDSS. Both surveys' questions used a 5-point Likert scale (Likert 1932). The SUS survey was analysed as is standard to return a SUS score where the average score is 68 (out of 100) and anything above that is considered good. The ANOVA (analysis of variance) and Kruskal-Wallis tests were used to assess if there were statistically significant differences between the SUS scores for different demographic groups, depending on whether the data met the parametric assumptions of homogeneity of variance and normality which were determined using Levene's and Shapiro-Wilk tests, respectively. All alphas were set at 0.05. The TAM survey contained 28 questions and was developed based on the TAM 2 methodology and similar research (Holden et al. 2010; Ketikidis et al. 2012; Buenestado et al. 2013). Some questions were assigned to categories common to the TAM framework; 12 for Perceived-Usefulness (PU), 2 for Perceived-Ease-of-Use (PEOU, fewer were included here due to the SUS survey assessing the same concept), 2 for Self-Efficacy (SE), 2 for Subjective-Norm (SN), 3 for Behavior-Change (BC). Others were independent questions designed to assess a particular aspect from TAM, evaluate a certain perspective or the influence of AI for antimicrobial decision-making. The internal consistency of each category was measured with Cronbach's alpha. The mean and standard deviation were calculated for each category and individual question, where a higher score indicates a more favourable result, with ANOVA or Kruskal-Wallis test used to assess any differences between demographic groups as for the SUS survey. At the end of the surveys, a free text box was used to gather any final perspectives that the participants wanted to share.

6.4 Results

6.4.1 Study and participants

59 clinicians were directly contacted or responded to outreach emails, with 42 individuals from 23 different hospitals in the UK participating in the study (Table 6.2). Their mean age was 42 years (SD 8.84) with 19 (45%) female subjects and 26 (62%) consultants. Most of the participants (n=14, 33%) specialised in infectious diseases, with 12 (29%) microbiologists also completing the study. The rest of the subjects were pharmacists (n=6, 14%), had no speciality or were from another area in medicine (called 'other' moving forward, n=10, 24%). Figure 6.3 shows participants' familiarity with AI assessed on a 5-point Likert scale, most were slightly familiar (n=17, 45%) with 3 (8%) very familiar.

Table 6.2: Participants' demographic information

Demographic	No. (%) of participants
Grade	
Consultant	26 (61.90%)
Other	16 (38.10%)
Medical Speciality	
Infectious Diseases	14 (33.33%)
Microbiology	12 (28.57%)
Pharmacist	6 (14.29%)
Other	10 (23.81%)
Sex	
Male	23 (54.76%)
Female	19 (45.24%)
Age	
20s	4 (9.52%)
30s	17 (40.48%)
40s	12 (28.57%)
50s	6 (14.29%)
60s	2 (4.76%)
Prefer not to say	1 (2.38%)
Age Statistics	
Mean	42
Median	37
Standard Deviation	8.84
AI Familiarity	
Not familiar	9 (23.68%)
Slightly familiar	17 (44.74%)
Moderately familiar	9 (23.68%)
Very familiar	3 (7.89%)

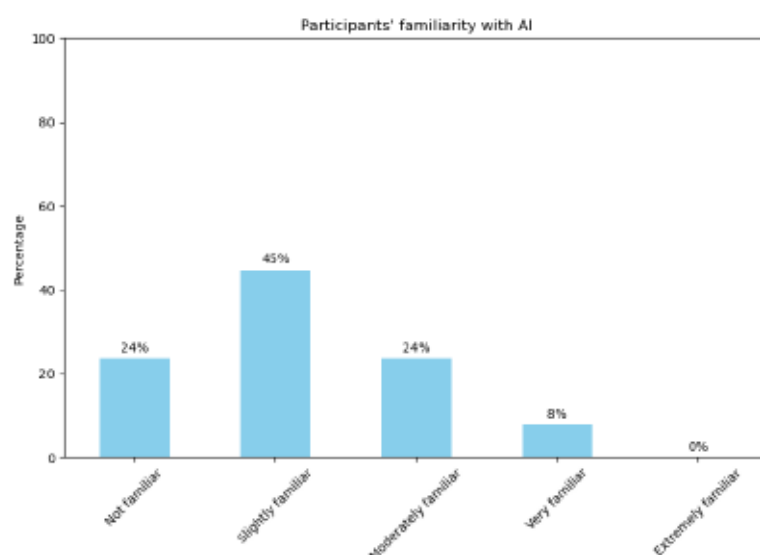


Figure 6.3: Distribution of participants' familiarity with AI.

6.4.2 Semi-structured interviews

Full details of the results from coding interview transcripts can be seen in Table 6.3 with the distribution of some key results shown in Figure 6.4. 64% of participants directly treat patients with antibiotics, 57% give prescribing advice to other clinicians, with a third involved in shaping treatment or stewardship guidelines. A statistically significant difference in responsibilities was observed across medical specialities (chi2 statistic=18.54, p-value=0.01, Supplementary Figure A.8) with fewer clinicians from other non-infection focused specialities being involved in guideline generation as would be expected. In addition, fewer junior participants noted that they had helped generate guidelines (n=2/16, 13% versus n=12/26, 46% consultants) although this was not statistically significant (chi2 statistic=3.48, p-value=0.18).

There was disagreement among the respondents about how individualised their current antimicrobial prescribing is, with one respondent noting 'we take patient-specific factors into account, but a lot of our choice is driven by what's in the guidelines'. A range of current technologies used to support prescribing were reported, including Microguide and flagging systems; however, 31% noted that they have not used any technology. The majority of participants (62%) were optimistic about the use of AI, with clinical evidence (69%) being the biggest driver behind trust, while ease of use (52%) was noted as an important factor influencing whether such an

AI system is actually used. Several barriers to adoption were highlighted though including challenges with changing behaviour (50%) and the technological infrastructure of hospitals not being ready for AI (45%).

No other statistically significant differences were observed from the interviews across demographics, which may be due to the sample size of our study or the subjective nature of qualitative research. However, other interesting differences observed (Supplementary Table A.7) include for behavioural factors older participants noting people are cautious with respect to antimicrobial prescribing, while more junior doctors point out the 'conflict that exists, which is the benefit to the system versus the benefit to the individual' (p -value=0.15) (Bolton et al. 2022). The aspects of antimicrobial prescribing that subjects would like AI to support align with their specialities (p -value=0.12). Pharmacists noted stewardship-focused applications such as IVOS ($n=5/6$, 83%), microbiologists wanted to stratify patients' risk ($n=4/12$, 33%) to identify those they should help, while infectious disease doctors discussed that AI could support optimal treatment duration ($n=7/14$, 50%) and 'other' specialities wanted assistance with selecting the appropriate antimicrobial ($n=3/10$, 30%). Finally, those with a medical speciality of 'other' think stakeholder support is important to ensure appropriate usage ($n=2/10$, 20%, p -value=0.07) implying that they want the wider hospital including patients to support any tool.

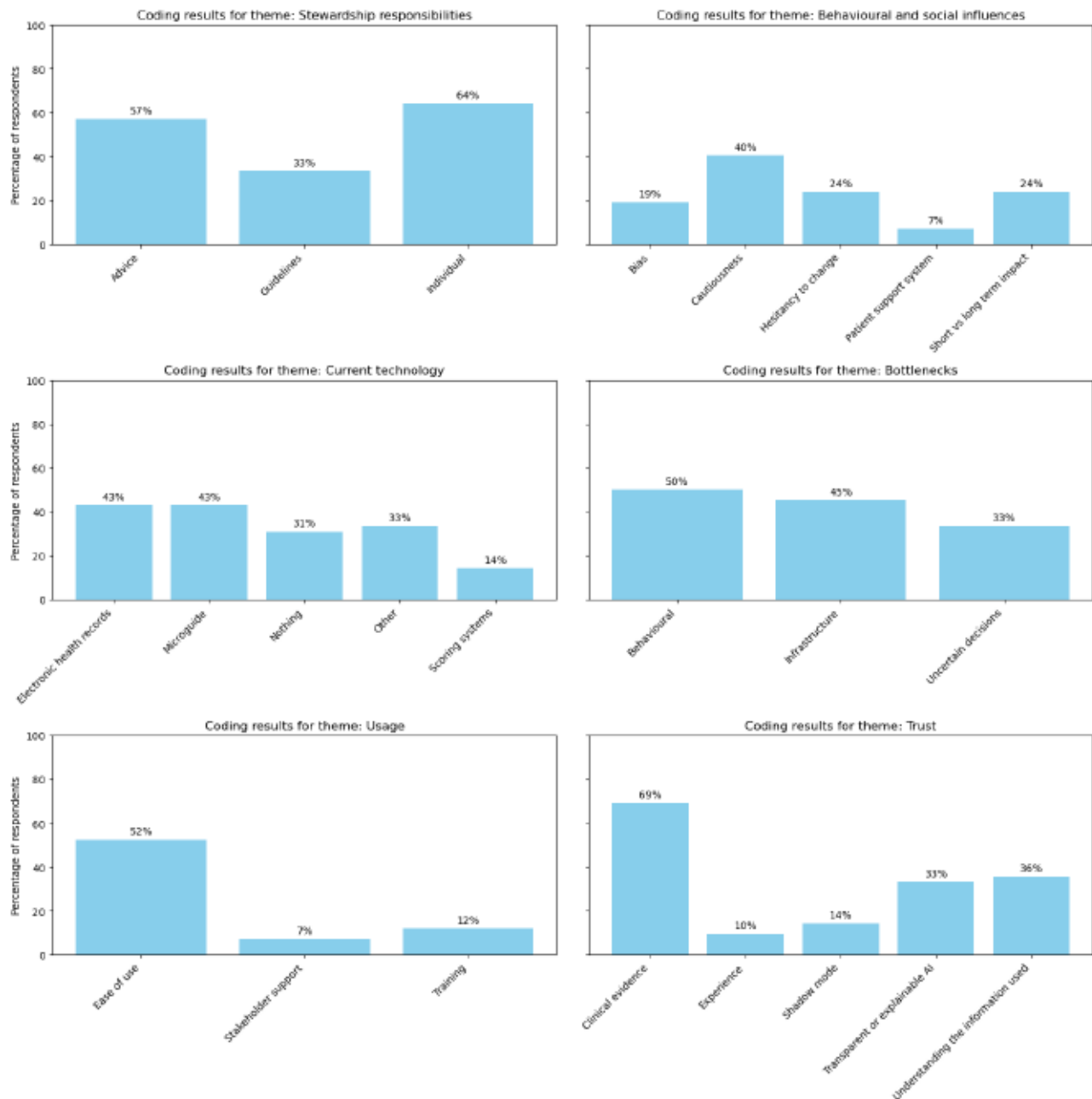


Figure 6.4: Distribution of codes for key themes explored in semi-structured interviews.

Table 6.3: Results from coding semi-structured interviews.

Theme	Code	Count (%)
Stewardship responsibilities	Advice	24 (57.14%)
Stewardship responsibilities	Guidelines	14 (33.33%)
Stewardship responsibilities	Individual	27 (64.29%)
Treatment individualisation	Not very individualised	15 (35.71%)
Treatment individualisation	Quite individualised	22 (52.38%)
Behavioural and social influences	Bias	8 (19.05%)
Behavioural and social influences	Cautiousness	17 (40.48%)
Behavioural and social influences	Hesitancy to change	10 (23.81%)
Behavioural and social influences	Patient support system	3 (7.14%)
Behavioural and social influences	Short vs long term impact	10 (23.81%)
Current technology	Electronic health records	18 (42.86%)
Current technology	Microguide	18 (42.86%)
Current technology	Nothing	13 (30.95%)
Current technology	Other	14 (33.33%)
Current technology	Scoring systems	6 (14.29%)
View on AI	Optimistic	26 (61.90%)
View on AI	Pessimistic	1 (2.38%)
Problem for AI	Antimicrobial selection	11 (26.19%)
Problem for AI	Iv to oral switch	17 (40.48%)
Problem for AI	Risk classification	7 (16.67%)
Problem for AI	Treatment duration	12 (28.57%)
Value of AI	Economic benefit	8 (19.05%)
Value of AI	Optimised prescribing	19 (45.24%)
Value of AI	Time saving	14 (33.33%)
Usage	Ease of use	22 (52.38%)
Usage	Stakeholder support	3 (7.14%)
Usage	Training	5 (11.90%)
Trust	Clinical evidence	29 (69.05%)
Trust	Experience	4 (9.52%)
Trust	Shadow mode	6 (14.29%)
Trust	Transparent or explainable AI	14 (33.33%)
Trust	Understanding the information used	15 (35.71%)
Outcomes	Amount of antimicrobial use	16 (38.10%)
Outcomes	Cost savings	8 (19.05%)
Outcomes	Length of stay	10 (23.81%)
Outcomes	Mortality	6 (14.29%)
Outcomes	Patient related outcomes	2 (4.76%)
Outcomes	Side effects	5 (11.90%)
Outcomes	Treatment duration	8 (19.05%)
Outcomes	Treatment failure rate	5 (11.90%)
Bottlenecks	Behavioural	21 (50.00%)
Bottlenecks	Infrastructure	19 (45.24%)
Bottlenecks	Uncertain decisions	14 (33.33%)

6.4.3 Clinical vignette experiment

For all but one case vignette, no statistically significant differences were observed in participants' switch choice based on the intervention they experienced (Table 6.4). This difference was observed for patient 7 (chi2 statistic=7.73, p-value=<0.01, Figure 6.5). In this case, the majority of participants with SOC chose to switch (n=16/21, 76%), while those who had access to the CDSS intervention chose to not switch (n=15/21, 71%). The recommendation of both AI and guideline CDSS in this scenario was don't switch which aligned with the ground truth action from the MIMIC dataset. The coefficient of the intervention was -2.08 for patient 7's statistically significant (p-value=<0.01) logistic regression model. This means that the odds of the switch decision being 1 are about 12% of the odds when the intervention is SOC and, hence, for this patient if the intervention changed from SOC to CDSS, the odds of the respondent switching decreased by approximately 88%.

Subgroup analysis on demographics revealed that a statistically significant change in choices was observed for females on patient 10 (chi2 statistic=4.02, p-value=0.04, Supplementary Table A.8). Those who viewed the CDSS (AI recommended don't switch and guideline recommended switch) chose to not switch (n=6/8, 75%) more often than those who did not have access to the CDSSs (n=2/11, 18%). Furthermore, the Cochran-Mantel-Haenszel test indicated that the proportions of switch decisions were statistically different from each other for patient 6 when controlling for sex (p-value=0.04) and grade (p-value=0.04), and patient 7 when controlling for each demographic (p-value=<0.01, Supplementary Table A.9). No statistical differences were observed in the switch choices when analysing by whether subjects reported that using the AI CDSS changed their decision-making. However, some significant differences were seen for patients 7 and 10 when stratifying for those who responded positively to some PU, PEOU, and the intention to use TAM questions (Supplementary Table A.10).

Table 6.4: All participants' switch choices for each patient (case vignette) by the intervention they experienced (CDSS or SOC). Chi-square statistical analysis results are also shown.

Patient	View	Switch Choice	Count	Chi-squared statistic	p-value
1	CDSS	Don't switch	21	0.0	1.00
	SOC	Don't switch	21		
2	CDSS	Switch	21	0.52	0.47
	SOC	Switch	19		
3	CDSS	Don't switch	2	0.2	0.66
		Switch	18		
	SOC	Don't switch	4		
		Switch	15		
4	CDSS	Don't switch	6	0.52	0.47
		Switch	19		
	SOC	Switch	2		
5	CDSS	Switch	21	0.52	0.47
		Don't switch	19		
	SOC	Don't switch	2		
6	CDSS	Switch	17	2.49	0.12
		Don't switch	4		
	SOC	Switch	21		
7	CDSS	Don't switch	15	7.73	<0.01
		Switch	6		
	SOC	Switch	16		
		Don't switch	5		
8	CDSS	Switch	13	0.0	1.00
		Don't switch	8		
	SOC	Switch	12		
9	CDSS	Don't switch	9	0.0	1.00
		Switch	19		
	SOC	Switch	2		
		Don't switch	18		
10	CDSS	Don't switch	3	2.39	0.12
		Switch	13		
	SOC	Switch	8		
		Don't switch	14		
11	CDSS	Don't switch	7	0.0	1.00
		Switch	11		
	SOC	Switch	10		
		Don't switch	11		
12	CDSS	Don't switch	10	0.41	0.52
		Switch	15		
	SOC	Don't switch	6		
		Switch	12		
		Don't switch	9		
		Switch	9		

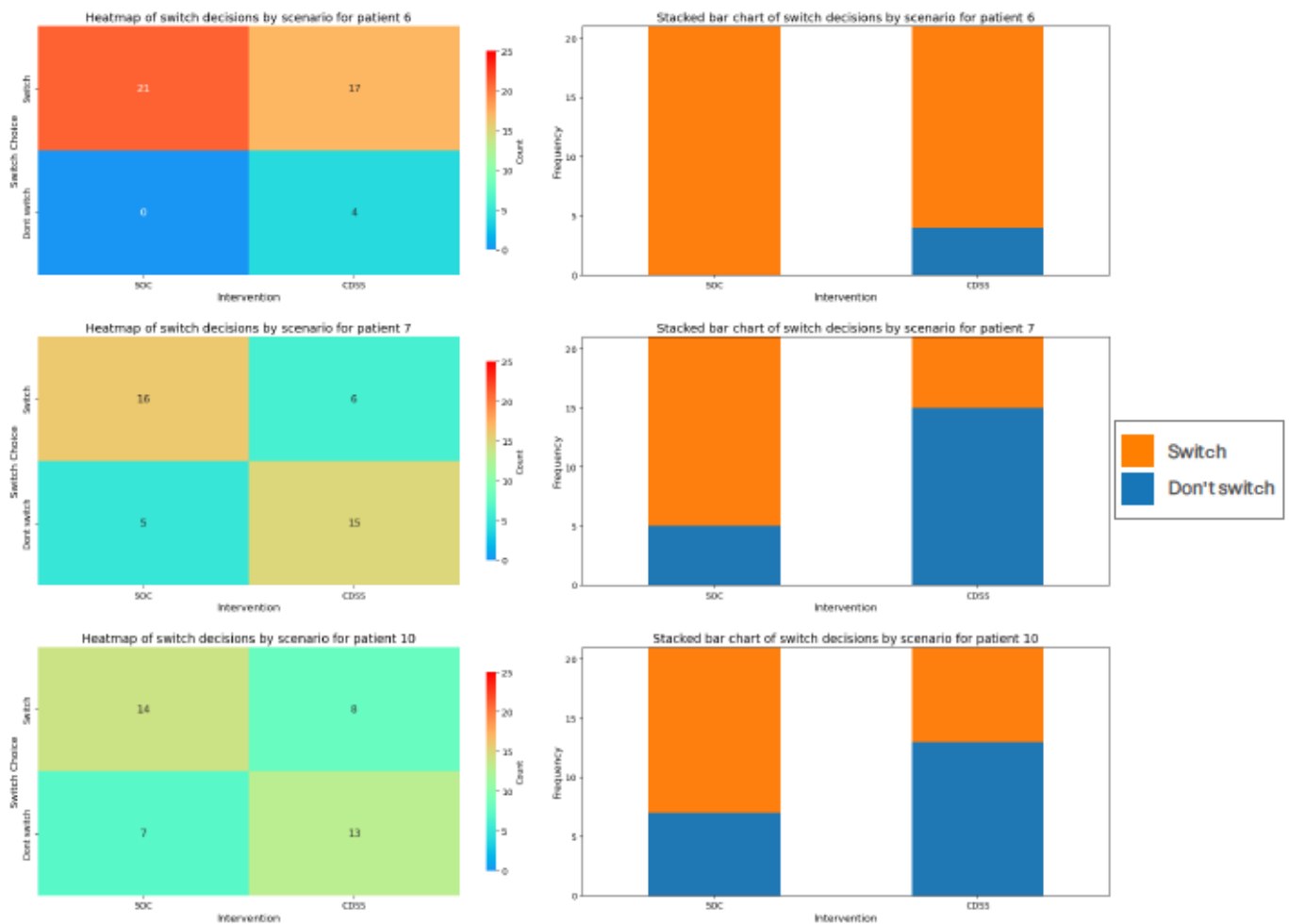


Figure 6.5: Participants' switch decisions for patients 6, 7 and 10 by intervention.

A high degree of variability was observed in the decisions for many patients regardless of the intervention experienced (Supplementary Table A.11). In particular, three cases (7, 10 and 11) had almost equal disagreement between clinicians with an entropy of 1. Two other MIMIC patients, 8 and 12, caused large disagreements with an entropy of 0.97 and 0.94 respectively. Perfect agreement was only observed for patient 1, who had a temperature over 39 degrees Celsius and was designed to be a simple case to gently introduce participants to the task.

The mean time to complete a case vignette was 148.50 seconds (SD 26.10). No significant differences in completion time were observed between the SOC and CDSS interventions. Hover duration (Supplementary Table A.13) was greatest for the table containing clinical data followed by the graph that displayed the same information with 19% of participants interacting with it. 55% of the participants viewed the AI CDSS explanations at any time, but across all case

vignettes where decision support was available, this was reduced to 9%. Similarly, 29% of the subjects clicked for an explanation of the guideline CDSS but across all examples, this URL was only clicked 5% of the time. The mean time to look at the AI and guideline CDSS explanations was 59.82 seconds and 64.52 seconds, respectively, which accounted for on average 24.31% and 25.45% of the time spent on the case. 12% visited the UKHSA website at any time; however, across all patients this only occurred in 1% of the cases. The percentage of participants who requested explanations by decision support system and patients is shown in Figure 6.6. No differences were observed between demographics.

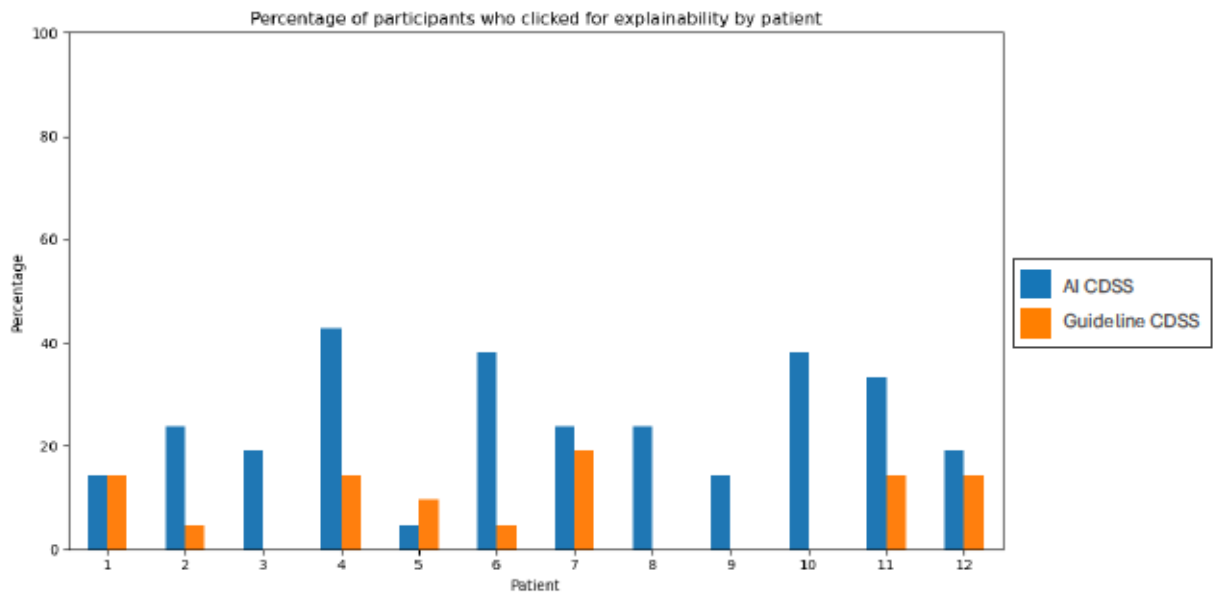


Figure 6.6: AI and guideline CDSS explainability click-through rate for each case vignette.

Analysing the free text explanations (Supplementary Table A.12) it is clear that the variables that most influenced clinicians' switch decisions were the temperature (14.43%), Glasgow coma score (14.43%), and heart rate (7.33%). 8.25% of don't switch decisions mentioned febrile (1.61% for switch choices) with 4.50% of switch choices including afebrile (0% for don't switch). Other medical terms mentioned included infection (11.68%) and antibiotics (7.33%) as should be expected, with the terms stable (23.47%) and improving (9.65%) often used when a switch decision was chosen. Finally, in 20.26% of the switch decisions, subjects mentioned that the patient could likely stop antibiotics. This was driven by patients 5 and 6 where almost half (42.5% and 44.74% respectively) of those who chose to switch said stopping would be appropriate.

6.4.4 Questionnaires

A SUS score of 72.32 (SD 8.79) was obtained for the web apps AI-CDSS which is considered good (Supplementary Table A.14). No statistically significant differences were observed in the SUS score between demographics. In the TAM questionnaire, a Cronbach alpha of 0.84 (Good), 0.52 (Poor), 0.51 (Poor), 0.07 (Unacceptable) and 0.40 (Unacceptable) were obtained for the group's PU, PEOU, SE, SN, and BC, respectively. Given the unacceptable scores for SN and BC, their questions were evaluated separately. All questions were rated on a scale from 1 to 5. The mean score for PU, PEOU and SE was 3.59 (SD 0.31), 3.83 (SD 0.20), and 4.05 (SD 0.05), respectively. Table 6.5 shows the results for all TAM questions, with Figure 6.7 presenting the distribution of responses to some key questions. Most participants reported that they are comfortable with AI in healthcare (62%), with their subjective norm (for either patients or people important to them) or perception of prestige showing little impact on their use of AI technology. Although responses were variable, half of the clinicians disagreed that their decisions were changed by the AI CDSS, which matches data from the clinical vignette experiment, where differences were not observed across interventions for most cases. However, their use of the AI CDSS was voluntary (100%), they were reassured when their clinical decision aligned with the AI CDSS recommendation (78%) and investigated further when it disagreed (74%), implying that the CDSS did have some influence on their practice. 64% agreed that the AI CDSS could be seamlessly integrated into their current clinical workflow, but there was disagreement about whether or not healthcare institutions have the infrastructure to support such systems, which likely reflects the diverse range of hospitals represented in this study. Finally, 79% of the participants said that they would intend to use the AI system if they had access, which was supported by a descriptive norm question in which 83% noted that some or most of their colleagues would use it.

Statistically significant differences (Supplementary Table A.15 and Supplementary Figure A.9) were observed for SE by AI familiarity (p -value=0.04) and based on grade for whether or not participants reported that the AI CDSS changed their decision-making (p -value=0.03), with more junior clinicians noting a larger change in behaviour. Furthermore, sex (p -value=0.02)

and medical speciality ($p\text{-value}=\leq 0.01$) had a significant influence on whether respondents reported that their hospital had the infrastructure necessary for AI. Females and those with a speciality of other were generally more optimistic while microbiologists disagreed. Finally, sex also had an influence on if participants thought people important to them think they should use AI technology ($p\text{-value}=0.05$).

Participants provided their comments and feedback in the final free text box. In general, clinicians were excited by the potential of the technology but noted that further development and testing were required. This would include accounting for a range of additional information such as examinations, microbiology results, radiology, or blood markers such as C-reactive protein (CRP) or procalcitonin (PCT). Such information would enable for a more comprehensive view of the patient and their state to determine if they are ready for stepping down antimicrobial therapy. Indeed during the case vignette exercise for 29.38% of the don't switch choices participants noted in the free text explanation that they would need more information to switch (Supplementary Table A.12). Participants also recommended that the most useful application of the AI CDSS would be to empower 'junior non-specialist antibiotic prescribers' to switch, or to reduce the burden on pharmacists and microbiologists by highlighting those that are suitable for switching and more complex cases for them to investigate further.

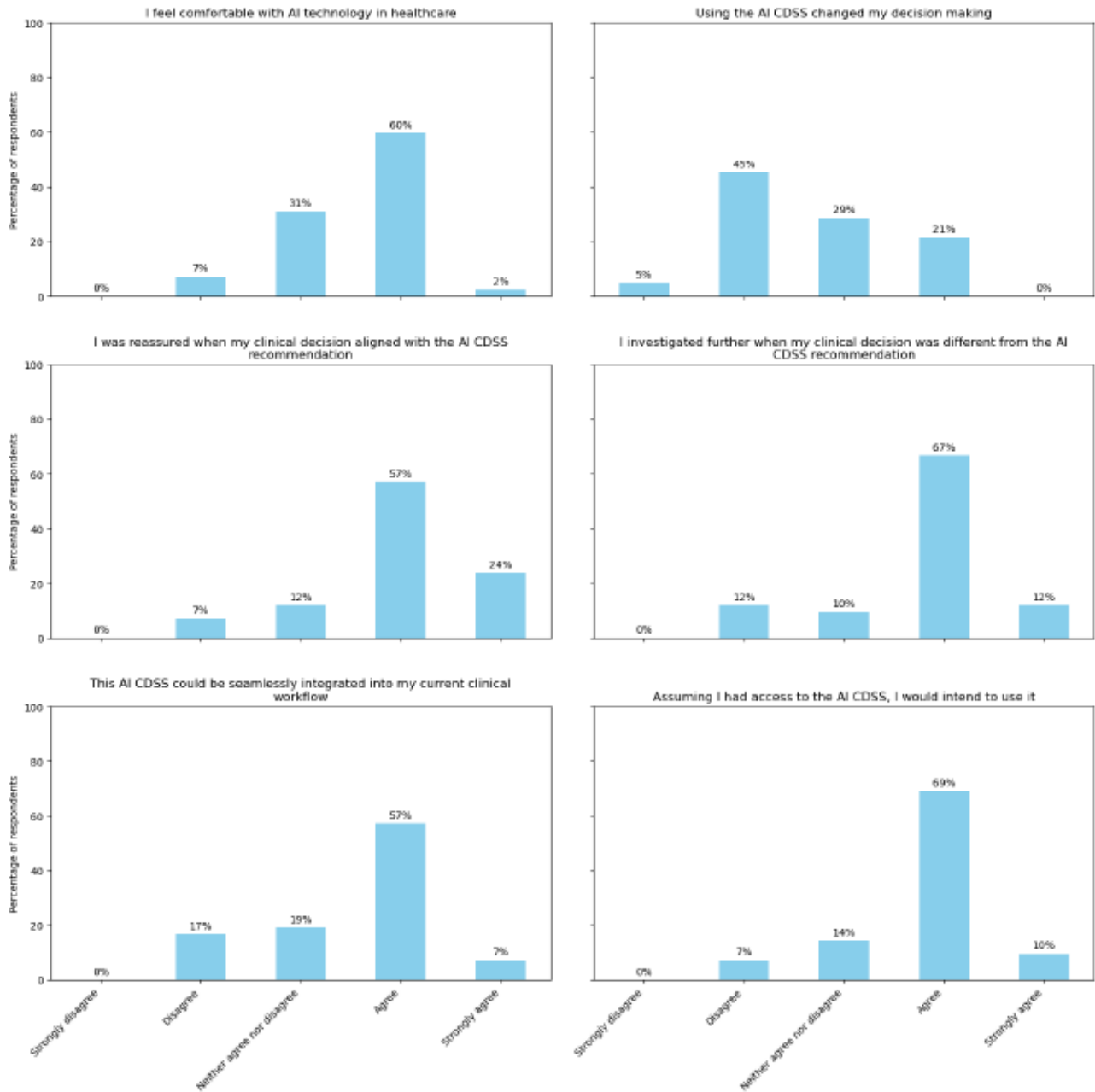


Figure 6.7: Result distribution for key questions asked in the TAM questionnaire.

Table 6.5: Technology Acceptance Model (TAM) questionnaire results. Note that the questions with a * were added after 7 clinicians had completed the study.

TAM Category	TAM Question	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
Perceived usefulness	I liked using the AI CDSS	0 (0%)	0 (0%)	13 (31%)	26 (62%)	3 (7%)
	I preferred making IV-to-oral switch decisions with the AI CDSS versus without	0 (0%)	13 (31%)	11 (26%)	16 (38%)	2 (5%)
	In my job using this AI CDSS would be relevant	0 (0%)	3 (7%)	7 (17%)	24 (57%)	8 (19%)
	The data presented in this study is important for delivering care and prescribing antibiotics	1 (2%)	3 (7%)	6 (14%)	22 (52%)	10 (24%)
	I would find the AI CDSS useful in my job	0 (0%)	4 (10%)	5 (12%)	26 (62%)	7 (17%)
	Using this AI CDSS in my job would increase my productivity	0 (0%)	9 (21%)	12 (29%)	15 (36%)	6 (14%)
	Using this AI CDSS would make it easier to do my job	1 (2%)	7 (17%)	10 (24%)	18 (43%)	6 (14%)
	This AI CDSS could improve the quality of care that I deliver	0 (0%)	3 (7%)	9 (21%)	26 (62%)	4 (10%)
	I preferred using this interface for antibiotic decision-making over software I currently use in clinical practice	1 (2%)	10 (24%)	16 (38%)	13 (31%)	2 (5%)
	The quality of the AI CDSS predictions was high	0 (0%)	3 (7%)	22 (52%)	17 (40%)	0 (0%)
	Details such as similar patients and their features improved my trust in the AI CDSS	0 (0%)	8 (19%)	11 (26%)	22 (52%)	1 (2%)
	Overall, I think this AI CDSS has the potential to be a useful tool for antibiotic intravenous-to-oral switch decision-making	1 (2%)	3 (7%)	1 (2%)	25 (60%)	12 (29%)
Perceived ease of use	The interface was pleasant with information clearly visible	0 (0%)	7 (17%)	4 (10%)	26 (62%)	5 (12%)
	It was easy to understand recommendations given by the AI CDSS	0 (0%)	1 (2%)	6 (14%)	28 (67%)	7 (17%)
Self efficacy	It was easy to get acquainted using the AI	0 (0%)	0 (0%)	3 (7%)	35 (83%)	4 (10%)
	Training on the AI CDSS was sufficient	0 (0%)	0 (0%)	6 (14%)	26 (62%)	10 (24%)
Subjective norm	Patients will welcome that I use this AI technology	0 (0%)	1 (2%)	29 (69%)	11 (26%)	1 (2%)
	People who are important to me think I should use AI technology	1 (2%)	4 (10%)	26 (62%)	11 (26%)	0 (0%)
Behavior change	Using the AI CDSS changed my decision-making	2 (5%)	19 (45%)	12 (29%)	9 (21%)	0 (0%)
	I was reassured when my clinical decision aligned with the AI CDSS recommendation*	0 (0.0%)	3 (9%)	5 (14%)	17 (49%)	10 (29%)
	I investigated further when my clinical decision was different from the AI CDSS recommendation*	0 (0%)	5 (14%)	4 (11%)	21 (60%)	5 (14%)
Voluntariness	My use of the AI CDSS was voluntary	0 (0%)	0 (0%)	0 (0%)	26 (62%)	16 (38%)
Prestige	People in my organization involved in AI research have more prestige than those who do not	2 (5%)	15 (36%)	15 (36%)	10 (24%)	0 (0%)
Workflow	This AI CDSS could be seamlessly integrated into my current clinical workflow	0 (0%)	7 (17%)	8 (19%)	24 (57%)	3 (7%)
Infrastructure	I think my healthcare institution has the necessary infrastructure to support this AI CDSS	3 (7%)	10 (24%)	15 (36%)	13 (31%)	1 (2%)
AI attitude	I feel comfortable with AI technology in healthcare	0 (0%)	3 (7%)	13 (31%)	25 (60%)	1 (2%)
Intention	Assuming I had access to the AI CDSS, I would intend to use it	0 (0%)	3 (7%)	6 (14%)	29 (69%)	4 (10%)
		None of them	Few of them	Some of them	Most of them	All of them
Descriptive norm	How many of your colleagues would use this AI CDSS if it were implemented	0 (0%)	7 (17%)	24 (57%)	11 (26%)	0 (0%)

6.5 Discussion

The main objective of this study was to discern clinical perceptions of an antibiotic prescribing focused AI CDSS and to understand if and how it can influence IV-to-oral switch decision-making. For most vignettes, no differences were observed between interventions. This is to be expected in patients 1 and 2 where the switch decision was designed to be obvious and the AI CDSS recommendation correct. For those cases where the AI prediction was incorrect (3, 4, 10 and 11), this result highlights that clinicians could correctly identify this and ignore the support, which differs from evidence from some previous studies (Gaube et al. 2021; Jacobs et al. 2021; Jabbour et al. 2023). Accounting for subgroup analysis, statistically significant differences in switch choices were observed in 3 patients (6, 7, and 10). We hypothesised that differences would be observed in those case vignettes where the switch decision was uncertain. This was true for patient 6 but cases 7 and 10 were instead where the AI prediction was correct and incorrect relative to the MIMIC ground truth, respectively. Interestingly, in all of these cases, the AI CDSS suggested that not switching would be appropriate, and the shift in decisions relative to those who experienced SOC reflected this. This observation implies that clinicians are persuaded to take the option that is perceived as safer and not switch when the AI system supports this, regardless of their initial inclination. This matches the culture of cautiousness and hesitancy to change reported by 40% and 24% of the respondents in the interviews. The significant difference in decisions observed for patient 7 for those randomised to use SOC versus CDSS is likely extended by the guideline CDSS also recommending don't switch, this hints that the alignment of multiple decision support systems may have a cumulative effect. Respondents often reported that the AI CDSS did not change their decision-making (50%), but no statistical differences were observed in switch choices based on this factor, implying that self-reporting may not be a valid metric in this setting (Nagendran et al. 2023). Rather, significant differences were observed for some PU, PEOU and intention to use TAM questions, highlighting that assessing how useful a tool was provides a better gauge of its potential impact. SUS and TAM questionnaire results were promising indicating that the AI CDSS was useful (PU of 3.59, 79% intention to use), usable (SUS score of 72.32, PEOU of 3.83, SE of

4.05) and appropriate (64% fit into current workflow). Respondents were reassured when the AI recommendation aligned with their clinical decision (78%), this aligns with previous research which has shown that expert clinicians are more confident when AI advice is given (Gaubert et al. 2023). Overall, the results indicate that the AI CDSS was well perceived and that its main influence is in directing clinicians towards not switching.

The influence of including or excluding explanations for AI CDSSs has previously been studied, with numerous publications indicating that explainable AI often has limited impact (Ghassemi et al. 2021; Shafit et al. 2022; Jabbour et al. 2023; Nagendran et al. 2023). In this research, we aimed to understand the extent to which AI explanations were used when available. Approximately half of the subjects (55%) viewed the AI explanations at least once during their participation. However, across all cases where decision support was available, explanations were only used 9% of the time, with around a quarter of the duration of a case vignette being dedicated to looking at the AI explanations. Even when selecting those who saw the AI CDSS and their decision aligned with its recommendation for patients 6, 7 and 10 (where a difference in decision-making was observed), between 20% and 40% viewed the AI explanations. This observation suggests that the persuasive nature of the explanation was not what changed their decision-making, but rather the presence of the AI recommendation itself. No correlation was observed between participants' familiarity with AI and if they were comfortable with AI in healthcare, also indicating that understanding the technical details of an AI system is not a key driver behind adoption (Henry et al. 2022). On the other hand, 33% of the subjects mentioned explainable AI in their interview (Figure 6.4), with one respondent noting "if you don't explain it to somebody on the front line, they are going to get frustrated, and all they will do is just start ignoring the system". Time is precious when making clinical decisions though, with the same respondent noting "I don't have time to look up loads of other stuff" and that free text explanations should just be incorporated directly into the patient's clinical notes, which is supported by the limited graph interaction (19%) or external website viewing (1%) seen in this study. Together, the evidence from this research agrees with previous studies indicating that explanations are likely to have little impact on the influence of an AI system for healthcare tasks. Rather, it seems, as demonstrated by the interview results and the statistical

differences in decision-making based on the TAM categories of PU, PEOU, and intention to use (Supplementary Table A.10), whether an individual trusts the system or finds it useful plays a larger role.

AI explanations were viewed more often than those for the guideline-based CDSS (Figure 6.6). These explanations may have been sought out of curiosity, given that many of the clinicians involved in the study were familiar with the UKHSA IVOS guidelines, although if this were true, we would expect to see higher usage of the AI explanations in the first few case vignettes, which we do not observe. Conversely, if the use of explanations indicates a lack of trust, then clinicians were certainly more wary of the AI CDSS versus one based on more traditional logic applicable to healthcare guidelines. Regarding the aspects of AI explainability that participants used, displaying similar patients and summarizing information in free text warrant further research, with participants noting they would like to know the outcome of similar historical patients as well if the AI model was correct or not. The AI explanation was used the most for patient 4 (43%, Figure 6.6). In this vignette the AI recommendation was don't switch; however, this was incorrect relative to the ground truth. Almost all participants correctly ignored this advice and continued with their decision to switch. However, the fact that almost half of the clinicians looked into the AI explanation here, as opposed to only 19% for patient 3 which is the opposite of this case, supports our insight that subjects were most influenced by the AI CDSS when it suggested not to switch, even though in this example switching was quite obviously the appropriate decision. This also aligns with a question from the TAM survey in which participants reported that they investigated further when their clinical decision was different from the AI CDSS recommendation (79%). Patient 4 was one of the cases in this study designed to test if intentionally incorrect AI recommendations could be ignored. Contrary to previous studies (Gaube et al. 2021; Jacobs et al. 2021; Jabbour et al. 2023), in this research incorrect AI recommendations were largely ignored, particularly for patients 3 and 4 (Table 6.4). Although, some changes in decision-making were observed for patient 10 as previously discussed, suggesting that while clinicians were generally able to disregard incorrect AI recommendations, further refinement in AI explainability and research into its influence in these situations is needed.

Challenges exist when trying to implement CDSSs into healthcare settings. Universal bottlenecks highlighted by this study include unsuitable infrastructure (45%) which can be made additionally complex due to variation between hospitals, a lack of funding, and clinical data being non-interoperable. This last point is very pertinent for systems tackling infectious diseases and antimicrobial stewardship as associated decisions are highly complex (Rawson et al. 2016). Despite significant research efforts (Charani et al. 2021; Rawson et al. 2021), what is best for an individual patient is often unknown with one participant saying 'I see very little evidence for an awful lot of what we do'. We tried to disentangle some of this grey area for IVOS decision-making in this study by including complicated cases where the optimal decision was unknown. Interestingly, for most of these particular cases (5, 6 and 9) the majority of participants chose one action, this new 'ground truth' data gathered in this study could potentially act as useful information to educate non-experts on optimal stewardship practices. Across many cases though a high variability in IVOS decisions was seen regardless of the intervention experienced. This was particularly true for the MIMIC cases, which excluding patient 9 all had a Shannon entropy ≥ 0.94 (Supplementary Table A.11). This is likely due to the reduced clinical information available to participants in these cases, but such situations are common in clinical practice (Wojcik et al. 2021), highlighting a key barrier when applying AI to infectious disease. If there is a lack of scientific understanding or information as to support optimal decision-making, then heterogeneity in practice can prevent AI models trained on such data from converging effectively. Current AI systems are not yet at the capabilities of conducting scientific research themselves (Lu et al. 2024; Messeri et al. 2024), and we know that on a larger scale neural network-based models do not learn or process information in the same way as human brains (Cevora 2019; Song et al. 2024). Hence, when trained to tackle highly variable healthcare problems, AI outputs can often come into contradiction with many clinicians' mental model of what is considered optimal care, leading to a breakdown in confidence and such systems being rejected by the user. One interviewee stated, 'I think the challenge will come if it is suggesting you do something very opposed to your normal practice'. The broader interview results show key bottlenecks in translating such AI systems are the uncertainty of antimicrobial stewardship decisions (33%) and the challenge of getting clinicians to change their existing

prescribing behaviour (50%, Figure 6.4). The ability for infectious diseases to affect others and the development of AMR also raises challenges including balancing the needs of an individual versus wider society (Bolton et al. 2022, Section 2.4), and updating the system temporally and geographically to match changes in pathogens and their resistance patterns. How this latter point can be addressed while still meeting regulatory requirements is an outstanding question. Ultimately, even if uncertainty exists tools that help antimicrobial prescribing and are supported by evidence can still provide a benefit to clinicians, patients, and help tackle AMR. However, infectious diseases pose unique hurdles for AI CDSSs, with one participant stating 'the layers of complexity on an individual basis is why microbiology as a service exists'.

Behavioural, social and cultural factors play an important role in antimicrobial prescribing (Charani et al. 2011; Charani et al. 2014; Lorencatto et al. 2018; Donisi et al. 2019; Dionisio et al. 2023; Poluektova et al. 2023). Interviewees noted that prescribers are often biased toward particular antimicrobial agents based on their lived experience (19%), which can lead to cautious prescribing (40%) that unnecessarily extends IV durations. All of this changes by speciality and seniority, with junior colleagues being heavily influenced by their more senior peers (24%). Participants mentioned that there is high individual variability with the frequent rotation of junior doctors and the presence of temporary staff meaning stewardship education must be continuously restarted. Another factor that influences adherence to stewardship practices is resources including the presence of pharmacists and microbiologists on wards. When present, such experts can provide regular advice and steer clinicians towards optimal prescribing. Unfortunately, in resource-constrained environments such experts are frequently absent, leading to the potential for an AI system to augment this support. However, this also gives rise to the barrier that new systems can be overwhelming from a user perspective when healthcare systems are stretched. To ensure adoption data showing the safety and benefit of an AI CDSS is necessary (69%), with clinicians wanting some understanding of how the system was devised and the information it uses (36%). Regarding outcome measures patient outcomes were most important (Kelly et al. 2019), but participants noted that showing a benefit in, for example, mortality or treatment failure rate is going to be difficult for a stewardship-focused CDSS. Instead, demonstrating non-inferiority for patient outcomes and a benefit elsewhere, such as

in the amount of antimicrobials used (38%), development of resistance, or with cost savings (19%) through clinical time or bed days saved (24%), was considered an appropriate approach. Indeed beyond optimizing prescribing (45%), many participants noted the potential time (33%) and economic (19%) benefits of such an AI tool, although this would need to be proven in a clinical trial. Trust is also built over time through experience with a system (10%) (Henry et al. 2022), with subjects noting a 'shadow mode' (14%) where the tool is first run in the background without influencing decisions could be useful. In our results, there was little evidence that other stakeholders (e.g., colleagues, patients) have a big impact on a clinician's perception of an AI system (Table 6.3), which reflects some previous research (Nadri et al. 2018). However, it is likely that regardless of this on a wider scale, an organisation's ethos, support, training (12%) and social norm feedback would be essential for driving behaviour change (Huang et al. 2023; Tokgöz et al. 2023; Zeng et al. 2023; Tokgöz et al. 2024). Indeed, the UK's regulation of software as a medical device supports such systems by ensuring they clearly define when and how they should be used (MHRA 2023; MHRA 2024). Finally, as with any software, maintenance and contingency plans for when things go wrong are important. All together, setting up the appropriate support around an AI CDSS and gathering robust data are important to drive behavioural changes in traditionally heterogeneous stewardship practices.

Limitations exist for this study. Firstly, whilst a large and diverse population participated, our recruitment primarily relied on those interested to respond to our emails. This inherently biases our study population towards those who are interested in research and AI technology, although as shown in Figure 6.3 this does not mean they were familiar with AI. Secondly, the clinical vignette experiment in this research was conducted through a web app which, despite using simulated and real patient cases, is inherently different from a clinical setting. Concordant with this is the fact that despite each case containing a large amount of infection-relevant data (as noted by the participants in Table 6.5), not all the information often present in real clinical practice could be included, due to limitations with the datasets. A number of participants noted that they would request additional information to confirm their decisions. The case vignettes given to participants were chosen randomly from MIMIC or simulated from ICHT data according to our criteria, as detailed in Subsection 6.3.3. 12 was chosen

as an appropriate number of cases to cover a range of different scenarios, gather enough data, and match participants' time restrictions (45 minutes were allocated for this exercise, giving 3 minutes and 45 seconds per case). However, this number does mean that we were unable to include vignettes on all potential use cases of the IVOS CDSS, or thoroughly investigate the differential impact of AI versus guideline recommendations. Finally, detailed data was not collected on the internal decision-making processes of the participants during the clinical vignette experiment. This means that when analysing the results some level of inference must be applied, which is supported by data from the semi-structured interviews, questionnaires and mouse movement. Future work could use more formal approaches such as 'thinking-aloud' methodologies (Eccles et al. 2017) or eye tracking (Nagendran et al. 2024).

6.6 Conclusion

Findings from this research can benefit the wider AI CDSS ecosystem. In particular, our work suggests such systems can often have unexpected impacts on decision-making, AI explanations are likely to be underutilised and assessing how useful an individual found a system is a good proxy for its influence. Furthermore, clinical evidence to build trust and the system being easy to use are essential factors to ensure adoption. Specifically for this IVOS-focused AI CDSS, it seems there are two use cases where it provides the most utility. First, by empowering clinicians who are not experts in antimicrobial stewardship, particularly in settings with limited infection expertise, with one participant saying "if junior doctors were able to be more independent prescribers, as a result of this tool, then that would be a good thing". Secondly, as a risk stratification tool for pharmacists and microbiologists so that they can focus their time on treating complicated patients. Indeed, both can support optimal antimicrobial prescribing and try to drive the cultural and behavioural change needed at the system level to combat AMR. Prospective clinical trials are required though to gather safety and value evidence. AI systems have the potential to support uncertain clinical decision-making. The great complexity of medicine though necessitates research to understand their impact, perception and interaction with existing healthcare dynamics, ultimately bridging the translation gap to benefit patients.

References

1. Alowais, Shuroug A. et al. (Sept. 2023). "Revolutionizing healthcare: the role of artificial intelligence in clinical practice". In: *BMC Medical Education* 23.1, p. 689. ISSN: 1472-6920. doi: 10.1186/s12909-023-04698-z. URL: <https://doi.org/10.1186/s12909-023-04698-z>.
2. Higgins, David and Madai, Vince I. (2020). "From Bit to Bedside: A Practical Framework for Artificial Intelligence Product Development in Healthcare". In: *Advanced Intelligent Systems* 2.10, p. 2000052. ISSN: 2640-4567. doi: 10.1002/aisy.202000052. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/aisy.202000052>.
3. McCradden, Melissa D., Stephenson, Elizabeth A., and Anderson, James A. (Sept. 2020). "Clinical research underlies ethical integration of healthcare artificial intelligence". In: *Nature Medicine* 26.9, pp. 1325–1326. ISSN: 1546-170X. doi: 10.1038/s41591-020-1035-9. URL: <https://www.nature.com/articles/s41591-020-1035-9>.
4. Ahmed, Molla Imaduddin et al. (Oct. 2023). "A Systematic Review of the Barriers to the Implementation of Artificial Intelligence in Healthcare". In: *Cureus* 15.10, e46454. ISSN: 2168-8184. doi: 10.7759/cureus.46454. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10623210/>.
5. Lambert, Sophie Isabelle et al. (June 2023). "An integrative review on the acceptance of artificial intelligence among healthcare professionals in hospitals". In: *npj Digital Medicine* 6.1, pp. 1–14. ISSN: 2398-6352. doi: 10.1038/s41746-023-00852-5. URL: <https://www.nature.com/articles/s41746-023-00852-5>.
6. Olaye, Iredia M. and Seixas, Azizi A. (May 2023). "The Gap Between AI and Bedside: Participatory Workshop on the Barriers to the Integration, Translation, and Adoption of Digital Health Care and AI Startup Technology Into Clinical Practice". In: *Journal of Medical Internet Research* 25.1, e32962. doi: 10.2196/32962. URL: <https://www.jmir.org/2023/1/e32962>.
7. Peiffer-Smadja, N. et al. (May 2020). "Machine learning for clinical decision support in infectious diseases: a narrative review of current applications". In: *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 26.5, pp. 584–595. ISSN: 1469-0691. doi: 10.1016/j.cmi.2019.09.009.
8. Tokgöz, Pinar et al. (Apr. 2024). "Decision support systems for antibiotic prescription in hospitals: a survey with hospital managers on factors for implementation". In: *BMC Medical Informatics and Decision Making* 24.1, p. 96. ISSN: 1472-6947. doi: 10.1186/s12911-024-02490-7. URL: <https://doi.org/10.1186/s12911-024-02490-7>.

9. Mello-Thoms, Claudia and Mello, Carlos A B (Oct. 2023). "Clinical applications of artificial intelligence in radiology". In: *British Journal of Radiology* 96.1150, p. 20221031. ISSN: 0007-1285. doi: 10.1259/bjr.20221031. URL: <https://doi.org/10.1259/bjr.20221031>.
10. NICE (Jan. 2024). *Overview | Artificial intelligence (AI)-derived software to help clinical decision making in stroke | Guidance*. URL: <https://www.nice.org.uk/guidance/dg57>.
11. Wojcik, Gosha et al. (July 2021). "Understanding the complexities of antibiotic prescribing behaviour in acute hospitals: a systematic review and meta-ethnography". In: *Archives of Public Health* 79.1, p. 134. ISSN: 2049-3258. doi: 10.1186/s13690-021-00624-1. URL: <https://doi.org/10.1186/s13690-021-00624-1>.
12. Bolton, William J. et al. (Nov. 2022). "Developing moral AI to support decision-making about antimicrobial use". In: *Nature Machine Intelligence* 4.11, pp. 912-915. ISSN: 2522-5839. doi: 10.1038/s42256-022-00558-5. URL: <https://www.nature.com/articles/s42256-022-00558-5>.
13. Tarrant, Carolyn and Krockow, Eva M. (Mar. 2022). "Antibiotic overuse: managing uncertainty and mitigating against overtreatment". In: *BMJ Quality & Safety* 31.3, pp. 163-167. ISSN: 2044-5415, 2044-5423. doi: 10.1136/bmjqs-2021-013615. URL: <https://qualitysafety.bmj.com/content/31/3/163>.
14. Poluektova, Olga et al. (June 2023). "A scoping review and behavioural analysis of factors underlying overuse of antimicrobials". In: *JAC-Antimicrobial Resistance* 5.3, dlad043. ISSN: 2632-1823. doi: 10.1093/jacamr/dlad043. URL: <https://doi.org/10.1093/jacamr/dlad043>.
15. Kelly, Christopher J. et al. (Oct. 2019). "Key challenges for delivering clinical impact with artificial intelligence". In: *BMC Medicine* 17.1, p. 195. ISSN: 1741-7015. doi: 10.1186/s12916-019-1426-2. URL: <https://doi.org/10.1186/s12916-019-1426-2>.
16. Sendak, Mark P. et al. (Jan. 2020). "A Path for Translation of Machine Learning Products into Healthcare Delivery". In: URL: <https://www.emjreviews.com/innovations/article/a-path-for-translation-of-machine-learning-products-into-healthcare-delivery/>.
17. Vasey, Baptiste et al. (May 2022). "Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI". In: *Nature Medicine* 28.5, pp. 924-933. ISSN: 1546-170X. doi: 10.1038/s41591-022-01772-9. URL: <https://www.nature.com/articles/s41591-022-01772-9>.
18. White, Nicole M et al. (June 2023). "Evaluating the costs and consequences of computerized clinical decision support systems in hospitals: a scoping review and recommendations for future practice". In: *Journal of the American Medical Informatics Association* 30.6, pp. 1205-1218. ISSN: 1527-974X. doi: 10.1093/jamia/ocad040. URL: <https://doi.org/10.1093/jamia/ocad040>.

19. Bolton, William J. et al. (Jan. 2024). "Personalising intravenous to oral antibiotic switch decision making through fair interpretable machine learning". In: *Nature Communications* 15.1, p. 506. ISSN: 2041-1723. DOI: 10.1038/s41467-024-44740-2. URL: <https://www.nature.com/articles/s41467-024-44740-2>.
20. Erlingsson, Christen and Brysiewicz, Petra (Sept. 2017). "A hands-on guide to doing content analysis". In: *African Journal of Emergency Medicine* 7.3, pp. 93–99. ISSN: 2211-419X. DOI: 10.1016/j.afjem.2017.08.001. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6234169/>.
21. Xiao, Ziang et al. (Apr. 2023). *Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding*. DOI: 10.1145/3581754.3584136. URL: <http://arxiv.org/abs/2304.10548>.
22. Dunivin, Zackary Okun (Feb. 2024). *Scalable Qualitative Coding with LLMs: Chain-of-Thought Reasoning Matches Human Performance in Some Hermeneutic Tasks*. DOI: 10.48550/arXiv.2401.15170. URL: <http://arxiv.org/abs/2401.15170>.
23. Tai, Robert H. et al. (Jan. 2024). "An Examination of the Use of Large Language Models to Aid Analysis of Textual Data". In: *International Journal of Qualitative Methods* 23, p. 16094069241231168. ISSN: 1609-4069. DOI: 10.1177/16094069241231168. URL: <https://doi.org/10.1177/16094069241231168>.
24. Goldberger, A. L. et al. (June 2000). "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals". In: *Circulation* 101.23, E215–220. ISSN: 1524-4539. DOI: 10.1161/01.cir.101.23.e215.
25. Johnson, Alistair et al. (2022). *MIMIC-IV*. DOI: 10.13026/s6n6-xd98. URL: <https://physionet.org/content/mimiciv/2.0/>.
26. UK Health Security Agency (Jan. 2023). *National antimicrobial intravenous-to-oral switch (IVOS) criteria for early switch*. URL: <https://www.gov.uk/government/publications/antimicrobial-intravenous-to-oral-switch-criteria-for-early-switch/national-antimicrobial-intravenous-to-oral-switch-ivos-criteria-for-early-switch>.
27. Crabbé, Jonathan et al. (2021). "Explaining Latent Representations with a Corpus of Examples". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., pp. 12154–12166.
28. Team, Gemini et al. (June 2024). *Gemini: A Family of Highly Capable Multimodal Models*. DOI: 10.48550/arXiv.2312.11805. URL: <http://arxiv.org/abs/2312.11805>.
29. Laker, Lauren F. et al. (Dec. 2018). "Quality and Efficiency of the Clinical Decision-Making Process: Information Overload and Emphasis Framing". In: *Production and Operations Management* 27.12, pp. 2213–2225. ISSN: 1059-1478. DOI: 10.1111/poms.12777. URL: <https://doi.org/10.1111/poms.12777>.

30. Phillips-Wren, Gloria and Adya, Monica (Aug. 2020). "Decision making under stress: the role of information overload, time pressure, complexity, and uncertainty". In: *Journal of Decision Systems* 29.sup1, pp. 213–225. ISSN: 1246-0125. doi: 10.1080/12460125.2020.1768680. URL: <https://www.tandfonline.com/doi/full/10.1080/12460125.2020.1768680>.
31. Lewis, James R. (July 2018). "The System Usability Scale: Past, Present, and Future". In: *International Journal of Human–Computer Interaction* 347, pp. 577–590. ISSN: 1044-7318. doi: 10.1080/10447318.2018.1455307. URL: <https://doi.org/10.1080/10447318.2018.1455307>.
32. Holden, Richard J. and Karsh, Ben-Tzion (Feb. 2010). "The Technology Acceptance Model: Its past and its future in health care". In: *Journal of Biomedical Informatics* 43.1, pp. 159–172. ISSN: 1532-0464. doi: 10.1016/j.jbi.2009.07.002. URL: <https://www.sciencedirect.com/science/article/pii/S1532046409000963>.
33. Likert, R. (1932). "A technique for the measurement of attitudes." In: *Archives of Psychology* 22 140, pp. 55–55.
34. Ketikidis, Panayiotis et al. (June 2012). "Acceptance of health information technology in health professionals: An application of the revised technology acceptance model". In: *Health Informatics Journal* 18.2, pp. 124–134. ISSN: 1460-4582. doi: 10.1177/1460458211435425. URL: <https://doi.org/10.1177/1460458211435425>.
35. Buenestado, David et al. (Feb. 2013). "Evaluating Acceptance and User Experience of a Guideline-based Clinical Decision Support System Execution Platform". In: *Journal of Medical Systems* 37.2, p. 9910. ISSN: 1573-689X. doi: 10.1007/s10916-012-9910-7. URL: <https://doi.org/10.1007/s10916-012-9910-7>.
36. Gaube, Susanne et al. (Feb. 2021). "Do as AI say: susceptibility in deployment of clinical decision-aids". In: *npj Digital Medicine* 4.1, pp. 1–8. ISSN: 2398-6352. doi: 10.1038/s41746-021-00385-9. URL: <https://www.nature.com/articles/s41746-021-00385-9>.
37. Jacobs, Maia et al. (Feb. 2021). "How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection". In: *Translational Psychiatry* 11.1, pp. 1–9. ISSN: 2158-3188. doi: 10.1038/s41398-021-01224-x. URL: <https://www.nature.com/articles/s41398-021-01224-x>.
38. Jabbour, Sarah et al. (Dec. 2023). "Measuring the Impact of AI in the Diagnosis of Hospitalized Patients: A Randomized Clinical Vignette Survey Study". In: *JAMA* 330.23, pp. 2275–2284. ISSN: 0098-7484. doi: 10.1001/jama.2023.22295. URL: <https://doi.org/10.1001/jama.2023.22295>.
39. Nagendran, Myura et al. (Nov. 2023). "Quantifying the impact of AI recommendations with explanations on prescription decision making". In: *npj Digital Medicine* 6.1, pp. 1–7. ISSN: 2398-6352. doi: 10.1038/s41746-023-00955-z. URL: <https://www.nature.com/articles/s41746-023-00955-z>.

40. Gaube, Susanne et al. (Jan. 2023). "Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays". In: *Scientific Reports* 13.1, p. 1383. ISSN: 2045-2322. doi: 10.1038/s41598-023-28633-w. URL: <https://www.nature.com/articles/s41598-023-28633-w>.
41. Ghassemi, Marzyeh, Oakden-Rayner, Luke, and Beam, Andrew L. (Nov. 2021). "The false hope of current approaches to explainable artificial intelligence in health care". In: *The Lancet Digital Health* 3.11, e745–e750. ISSN: 2589-7500. doi: 10.1016/S2589-7500(21)00208-9. URL: [https://www.thelancet.com/journals/landig/article/PIIS2589-7500\(21\)00208-9/fulltext](https://www.thelancet.com/journals/landig/article/PIIS2589-7500(21)00208-9/fulltext).
42. Shafti, Ali et al. (Feb. 2022). *The Response Shift Paradigm to Quantify Human Trust in AI Recommendations*. doi: 10.48550/arXiv.2202.08979. URL: <http://arxiv.org/abs/2202.08979>.
43. Henry, Katharine E. et al. (July 2022). "Human–machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system". In: *npj Digital Medicine* 5.1, pp. 1–6. ISSN: 2398-6352. doi: 10.1038/s41746-022-00597-7. URL: <https://www.nature.com/articles/s41746-022-00597-7>.
44. Rawson, Timothy Miles et al. (Dec. 2016). "Mapping the decision pathways of acute infection management in secondary care among UK medical physicians: a qualitative study". In: *BMC Medicine* 14.1, p. 208. ISSN: 1741-7015. doi: 10.1186/s12916-016-0751-y. URL: <https://doi.org/10.1186/s12916-016-0751-y>.
45. Charani, Esmita et al. (Aug. 2021). "Optimising antimicrobial use in humans – review of current evidence and an interdisciplinary consensus on key priorities for research". In: *The Lancet Regional Health – Europe* 7, p. 100161. ISSN: 2666-7762. doi: 10.1016/j.lanepe.2021.100161. URL: <https://www.sciencedirect.com/science/article/pii/S2666776221001381>.
46. Rawson, Timothy M. et al. (Dec. 2021). "Optimizing antimicrobial use: challenges, advances and opportunities". In: *Nature Reviews Microbiology* 19.12, pp. 747–758. ISSN: 1740-1534. doi: 10.1038/s41579-021-00578-9. URL: <https://www.nature.com/articles/s41579-021-00578-9>.
47. Lu, Chris et al. (Aug. 2024). *The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery*. doi: 10.48550/arXiv.2408.06292. URL: <http://arxiv.org/abs/2408.06292>.
48. Messeri, Lisa and Crockett, M. J. (Mar. 2024). "Artificial intelligence and illusions of understanding in scientific research". In: *Nature* 627.8002, pp. 49–58. ISSN: 1476-4687. doi: 10.1038/s41586-024-07146-0. URL: <https://www.nature.com/articles/s41586-024-07146-0>.
49. Cevora, George (May 2019). *The relationship between Biological and Artificial Intelligence*. doi: 10.48550/arXiv.1905.00547. URL: <http://arxiv.org/abs/1905.00547>.
50. Song, Yuhang et al. (Feb. 2024). "Inferring neural activity before plasticity as a foundation for learning beyond backpropagation". In: *Nature Neuroscience* 27.2, pp. 348–358. ISSN: 1546-1726. doi: 10.1038/s41593-023-01514-1. URL: <https://www.nature.com/articles/s41593-023-01514-1>.

51. Charani, Esmita et al. (2011). "Behavior change strategies to influence antimicrobial prescribing in acute care: a systematic review". In: *Clinical Infectious Diseases* 53.7, pp. 651–662.
52. Charani, Esmita, Castro-Sánchez, Enrique, and Holmes, Alison (2014). "The role of behavior change in antimicrobial stewardship." In.
53. Lorencatto, Fabiana et al. (Oct. 2018). "Driving sustainable change in antimicrobial prescribing practice: how can social and behavioural sciences help?" In: *Journal of Antimicrobial Chemotherapy* 73.10, pp. 2613–2624. ISSN: 0305-7453. DOI: 10.1093/jac/dky222. URL: <https://doi.org/10.1093/jac/dky222>.
54. Donisi, V et al. (Oct. 2019). "Emotional, cognitive and social factors of antimicrobial prescribing: can antimicrobial stewardship intervention be effective without addressing psycho-social factors?" In: *Journal of Antimicrobial Chemotherapy* 74.10, pp. 2844–2847. ISSN: 0305-7453. DOI: 10.1093/jac/dkz308. URL: <https://doi.org/10.1093/jac/dkz308>.
55. Dionisio, Francisco, Baquero, Fernando, and Fuertes, Marina (June 2023). "Psychological and cultural factors influencing antibiotic prescription". In: *Trends in Microbiology* 31.6, pp. 559–570. ISSN: 0966-842X, 1878-4380. DOI: 10.1016/j.tim.2022.12.010. URL: [https://www.cell.com/trends/microbiology/abstract/S0966-842X\(22\)00346-8](https://www.cell.com/trends/microbiology/abstract/S0966-842X(22)00346-8).
56. Nadri, Hamed et al. (Apr. 2018). "Factors Affecting Acceptance of Hospital Information Systems Based on Extended Technology Acceptance Model: A Case Study in Three Paraclinical Departments". In: *Applied Clinical Informatics* 9.2, pp. 238–247. ISSN: 1869-0327. DOI: 10.1055/s-0038-1641595. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884692/>.
57. Huang, Zhilian et al. (Dec. 2023). "Are physicians ready for precision antibiotic prescribing? A qualitative analysis of the acceptance of artificial intelligence-enabled clinical decision support systems in India and Singapore". In: *Journal of Global Antimicrobial Resistance* 35, pp. 76–85. ISSN: 2213-7165. DOI: 10.1016/j.jgar.2023.08.016. URL: <https://www.sciencedirect.com/science/article/pii/S2213716523001406>.
58. Tokgöz, Pinar, Hafner, Jessica, and Dockweiler, Christoph (Feb. 2023). "Factors influencing the implementation of decision support systems for antibiotic prescription in hospitals: a systematic review". In: *BMC Medical Informatics and Decision Making* 23.1, p. 27. ISSN: 1472-6947. DOI: 10.1186/s12911-023-02124-4. URL: <https://doi.org/10.1186/s12911-023-02124-4>.
59. Zeng, Yingchao et al. (May 2023). "Effects of social norm feedback on antibiotic prescribing and its characteristics in behaviour change techniques: a mixed-methods systematic review". In: *The Lancet Infectious Diseases* 23.5, e175–e184. ISSN: 1473-3099, 1474-4457. DOI: 10.1016/S1473-3099(22)00720-4. URL: [https://www.thelancet.com/journals/laninf/article/PIIS1473-3099\(22\)00720-4/fulltext#seccestitle160](https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(22)00720-4/fulltext#seccestitle160).

60. MHRA (Mar. 2023). *Crafting an intended purpose in the context of software as a medical device (SaMD)*. URL: <https://www.gov.uk/government/publications/crafting-an-intended-purpose-in-the-context-of-software-as-a-medical-device-samd/crafting-an-intended-purpose-in-the-context-of-software-as-a-medical-device-samd>.
61. MHRA (Apr. 2024). *Impact of AI on the regulation of medical products*. URL: <https://www.gov.uk/government/publications/impact-of-ai-on-the-regulation-of-medical-products/impact-of-ai-on-the-regulation-of-medical-products>.
62. Eccles, David W. and Arsal, Güler (Aug. 2017). "The think aloud method: what is it and how do I use it?" In: *Qualitative Research in Sport, Exercise and Health* 9.4, pp. 514–531. ISSN: 2159-676X. DOI: 10.1080/2159676X.2017.1331501. URL: <https://doi.org/10.1080/2159676X.2017.1331501>.
63. Nagendran, Myura et al. (Aug. 2024). "Eye tracking insights into physician behaviour with safe and unsafe explainable AI recommendations". In: *npj Digital Medicine* 7.1, pp. 1–10. ISSN: 2398-6352. DOI: 10.1038/s41746-024-01200-x. URL: <https://www.nature.com/articles/s41746-024-01200-x>.

Chapter 7

Prospective Outlook and Conclusion

7.1 Contributions and takeaways

In this thesis, I aimed to address the research hypothesis:

*Artificial intelligence-based clinical decision support systems
utilising routinely collected electronic health record data
can support antibiotic stewardship*

The work described in the previous chapters addresses this through the development and evaluation of AI models that support various aspects of antimicrobial stewardship that had not been previously explored. Furthermore, novel AI methodologies have been presented including segregating the embedding spaces of an LSTM autoencoder for counterfactual prediction, a pipeline for embedding patients and their co-morbidities, and a generalisable preprocessing and evaluation framework for IVOS models. This research advances the body of literature on AI CDSSs, in particular those for antimicrobial prescribing, by providing evidence that routinely collected data is informative for models aimed at stewardship-focused tasks, and their clinical perception and potential impact can be evaluated. The work conducted in this thesis was supported by an interdisciplinary team that included clinicians from many specialities, pharmacists, and AI experts. Throughout my PhD, I have ensured that my work focused on

tackling real-world problems that clinicians currently experience with antimicrobial prescribing. As shown in Chapter 6 Table 6.3 many subjects in the clinician evaluation study noted the decision to switch or stop antimicrobials is an area they believe AI could benefit. The research presented in this thesis serves as an example of how to design, develop, and assess an AI CDSS for infectious diseases. This work contributes to and advances AI CDSSs for antimicrobial optimisation, by providing evidence that AI models can obtain good performance at these tasks, that clinicians are receptive to decision support, and that, under certain circumstances, this technology can influence decision-making.

Key takeaways from this thesis are as follows:

- *Research into optimising antibiotic de-escalation and stewardship decision-making was lacking. This thesis presents AI models that can support antimicrobial IV-to-oral switching and cessation through simple, fair, interpretable prediction-focused algorithms or by providing prescribers with an understanding of the potential impact of their actions.*
- *Leveraging medical knowledge for the extraction and representation of clinical data such as patients' co-morbidities can support AI applications in healthcare.*
- *Antimicrobial prescribing decision-making is morally and technically complex and affected by behavioural, cultural and societal factors. By applying AI in a fair way it can support moral decision-making and help address ethical questions such as balancing the needs of an individual patient and those of wider and future society.*
- *Real-world research including app demos, case vignettes, interviews, and surveys enables an understanding of an AI system's potential impact, perception and interaction with existing healthcare practices. AI CDSSs can often have unexpected and nuanced impacts on decision-making, with how useful an individual found a system, a good proxy for its influence. Such clinician evaluations act as promising steps towards crossing the translation gap and benefiting patients.*

- *Interpretability is important for understanding when and how a model makes a mistake during development, but explainability at deployment is unlikely to be frequently utilised in healthcare. Rather, ensuring that the user has enough knowledge and information so that they are comfortable using and interacting with the system is essential. Human-AI interaction can be thought of as communicating with someone who speaks a different language, by finding a common ground you can understand each other to a limited level and reach the desired outcome.*
- *Evidence of safety and benefit is essential to build clinical trust in AI CDSSs, with ease of use being important to ensure adoption. Further work to understand and streamline the translation and integration of AI CDSSs in high-consequence and uncertain decision-making scenarios in healthcare is required.*

7.2 Future work and progressing AI CDSSs in healthcare

The field of AI CDSSs has expanded rapidly in recent years, with success stories being seen, including companies raising significant capital and numerous medical devices gaining regulatory approval (FDA 2024). The benefit to the end users – clinicians, hospitals and patients though remains to be seen. Domains such as infectious diseases have seen limited translation and adoption due to numerous technical, clinical, and commercial challenges (Howard et al. 2024a). Such factors were explored in more detail in Chapter 6 and include uncertainty, lack of a gold standard ground truth, behavioural factors, heterogeneous clinical practice, temporal and geographical variation, and the effect of AMR on wider society (Bolton et al. 2022). From an AI perspective to date machine learning (ML) and deep learning (DL) techniques have seen most success in modalities with structure (e.g., images and text). From large amounts of data, models can learn this structure and become useful for related objective tasks. Humans and their behaviour however are inherently stochastic (Glimcher 2005), meaning learning this from data is non-trivial. Advancements in AI methods to learn in a sample efficient manner from messy data are therefore needed with recent works using imitation learning to mimic

human actions (Pearce et al. 2023). From a clinical and commercial perspective, acceptance by clinicians, direct image access bypassing electronic medical records, and a clear return on investment are some of the key reasons the vast majority of current regulatory approvals focus on radiology. Such progress can be achieved in the field of infectious disease but the high degree of complexity necessitates a nuanced approach.

Future work should aim to progress the research discussed in this thesis towards a real-world CDSS that can help clinicians and their patients. To this extent, the prototype created for clinician evaluation (Chapter 6) needs to be turned into a fully functioning app with a simple user interface, where the design is directed by clinicians as end users, as well as by patients and experts in the user interface of medical software applications. In addition, infrastructure needs to be established to effectively and securely deploy software, and pass patient data from hospital systems through AI algorithms in real-time. How to best present AI outputs to clinicians and drive a change in prescribing behaviour through an AI tool is an important question. The co-design of systems and nudge theory can assist CDSS adoption, transparency and engagement (Rawson et al. 2019a; Sant’Anna et al. 2021), while AI literacy training, understanding what behaviour change is needed and how the tool fits into clinical workflows is vital to ensure impact and appropriate patient and clinician engagement. Indeed, results from Chapters 5 and 6 potentially come into conflict here given temporal analysis consistently showed the AI models were predicting switching was appropriate slightly earlier, while the clinical vignette experiment showed clinicians were more influenced by the CDSS when it recommended don’t switch. Further work is therefore needed to find out how to balance these contrasting behaviours, meet user expectations, and understand how AI CDSSs interplay with the existing social aspects of antimicrobial prescribing. For the cessation model described in Chapter 4 research would need to discern users’ perceptions and preferences with regards to reviving information on the potential impact of a decision versus a recommended action. The prospective evaluation and translation of this type of AI model would not be trivial given the counterfactual scenario never occurs. The next key milestone in the translation of this technology would be to continuously prospectively evaluate the performance of an AI CDSS system on real-world data. This would likely be done in a ‘shadow mode’ initially to prevent the

CDSS from over-influencing decisions but enable key stakeholders to become familiar with the technology. This evaluation would not only assess the predictive performance of the algorithms (e.g., AUROC) versus the current standard of care and expert opinion, but also understand the impact of the CDSS on decisions (as in Chapter 6), the amount of antimicrobials used, and patient outcomes. Such work would demonstrate the clinical value and safety of the AI system, with health economic modelling also necessary to show potential cost savings at the system level. Interestingly while economic analysis shows AI has great potential for cost savings in healthcare (Khanna et al. 2022), how freed-up time would be best utilised is an outstanding question with evidence showing it is unlikely it would all be directed towards more patient care (Moulds et al. 2024). Once the evidence has been gathered, an approved body can be engaged to obtain regulatory approval. In this case, in the UK, such technology would likely involve becoming a Class IIa certified software as a medical device, complying with the NHS Digital Technology Assessment Criteria and potentially some future version of the EU AI Act. Looking towards commercialisation, intellectual property must be identified and protected and a go-to-market strategy created to define the software's intended use and price. After implementation, continuous monitoring should be undertaken (Warraich et al. 2024) to prevent performance drift, further understand the tool's behavioural influence and see if this changes temporally.

From research conducted in Chapter 6 we distilled that the two most promising uses of our antimicrobial stewardship-focused CDSSs are:

1. *Supporting junior doctors or nurses in making day-to-day individualised decisions for their patients*
2. *Helping expert pharmacists, microbiologists or infectious disease physicians focus on the more complex patient cases*

Exactly which of these areas should be pursued initially and how the technology would be positioned within the care pathway requires further investigation. The first approach certainly holds the greatest potential to drive a lasting change in antimicrobial prescribing and benefit AMR, given the widespread decision-making of such individuals within institutions such

as the NHS. Nurses play a vital role in prescriptions and could benefit the most from actions such as early switching, and so should be more involved in future work. However, this approach also poses a challenge as ensuring appropriate use and consistent adoption across different departments and specialities is not trivial. The second option therefore may be a more suitable initial option given the focused nature of such individuals' work means getting them on board to use the technology may be easier, and there is a greater chance of reimbursement due to the potential time savings. The Black Country Pathology Service reported at the European Society of Clinical Microbiology and Infectious Diseases Global 2024 that an AI system for microbiology interpretation resulted in a 95% time reduction due to the number of cultures that could be reviewed at once. Such a benefit could likely be seen by stewardship focused tools by helping experts to review a large numbers of patients in a shorter time. More behavioural research is needed though to support translation (Charani et al. 2014; Jermutus et al. 2022). This is particularly true for antimicrobial prescribing where in our research (Chapter 6) many influencing factors such as cautiousness (40%) were elucidated and participants noted that simple approaches such as encouraging that review and stop dates are included in antimicrobial prescriptions and creating competitive dynamics between teams, can often be most influential in changing stewardship behaviour. Future work should aim to understand how existing biases and peers influence the use of an AI CDSS to ensure appropriate adoption. Human-AI interaction is a significant field of research in itself (Nazar et al. 2021), but by extending this to, and gathering data on patient-clinician-AI situations, we can ensure that technology is appropriately developed and deployed. Ultimately science enables humans to pursue a greater understanding of the world which can lead to improved healthcare, and technology is a tool that can support that journey. Humans, however, are emotional and social beings shaped by evolution. Primal instincts such as being part of a group and afraid of change exist and often drive behaviour. The Covid-19 pandemic made the impact of infectious diseases tangible to people around the world. There is therefore an understanding that problems such as AMR need to be tackled, but where people's values lie (e.g., the value of their own health versus wider society (Bolton et al. 2022)) needs to be understood. By aligning with these values, technology such as AI CDSSs can have a real positive impact on the health of our society.

In this thesis, the work presented focuses primarily on antimicrobial switching and cessation. Other aspects of stewardship which require further data-driven research include side effects, antibiotic re-treatment, dosing and duration. I supervised two master's projects during my PhD focusing on using ML and DL to predict antibiotic side effects and re-treatment (work not presented in this thesis). Both these students successfully completed their projects and obtained excellent grades. Work towards optimising antimicrobial dosing is one of the key research objectives for the NIHR HPRU in HCAI and AMR at Imperial College London, with associated work underway (Rawson et al. 2019b; Rawson et al. 2021). Treatment duration is inherently linked to switching and stopping but other than clinical studies (Hanretty et al. 2018; Royer et al. 2018; Spellberg et al. 2019; Wald-Dickler et al. 2019; Yahav et al. 2019) approaches towards applying AI directly to this aspect of stewardship and understanding an appropriate length of antimicrobial therapy are missing. As discussed in Chapter 6 the lack of complete scientific understanding and the fact that current clinical data is collected for the provision of care and not for research means that it is difficult to design AI models to augment prescribing practices. The complexity of medical decision-making therefore often goes beyond what can be captured in current datasets. Work on improving clinical datasets has already begun including linking primary and secondary care records, setting up secure data and trusted research environments, using genetic data, creating datasets that use structured and unstructured data (Fleming et al. 2023), and developing platforms such as OpenSAFELY. For antibiotic stewardship, in particular, a problem lies in defining a gold-standard ground truth. Expert guidelines can be used but lead to data leakage, with alternatives including human annotation (e.g., centaurlabs annotator platform) or even large language models (LLMs) (Zheng et al. 2023; Bolton et al. 2024). But all of these methods suffer if there is a lack of fundamental understanding. Therefore, more scientific research to deepen our knowledge of the interaction between an antimicrobial, pathogen, and host is essential with concurrent work curating high-quality research-focused datasets.

To mitigate some of the issues discussed in the previous paragraph, alternative approaches towards AI for antibiotic stewardship could be explored. In Chapter 3 we developed a pipeline to represent and embed medical knowledge on co-morbidities and how they are

related. In Chapter 5 we demonstrated that this methodology can be interoperated into AI systems. In the future, further work on the representation and incorporation of external knowledge and information will enable AI-CDSSs to overcome limitations with their training datasets and better align with real-world expectations. Such work has been seen in the fields of graphs and natural language processing (Lecue 2020; Safavi et al. 2021; Liu et al. 2023), but more healthcare-focused research is needed. Self-supervised learning has seen great success with large transformer-based models learning language (Naveed et al. 2024). By applying self-supervised learning to clinical data a model can learn a meaningful patient representation which can be applied to downstream tasks. Examples of training such a model include using data augmentations and a contrastive loss, masked data prediction, or next event prediction (Kraljevic et al. 2023). Another approach would be to use reinforcement learning, which has seen success in similar uncertain scenarios such as supporting the treatment of sepsis (Komorowski et al. 2018). Here a state and action space would be defined using the clinical data and potential decisions respectively. The reward function could then be designed based on factors such as clinical improvement, patient outcomes, length of stay, side effects, antimicrobial re-treatment, and the development of resistance, to balance the benefit of a prescription to an individual versus wider society. Problems that may be encountered here include the temporal effects of decisions meaning rewards may not be immediately apparent, and finding a balance between exploration and exploitation. Ultimately the goal of an AI approach towards antibiotic stewardship should be to understand the impact of any prescription at any time, for the individual patient, the hospital and wider society with regards to AMR, so prescribing can be optimised and ethical as discussed in Section 2.4. The research presented in this thesis begins to address this through AI-CDSSs that support an understanding of the likelihood of effectiveness of a prescription. But further work is needed on other aspects such as the chance of resistance developing. Once obtained, such parameters could be utilised in methods such as Pareto Front Optimisation (Kang et al. 2024) or Robust decision-making (Lempert 2019) to decipher across multiple dimensions and under uncertainty whether or not an antimicrobial prescription or stewardship action is appropriate or not for that patient at that time. By creating models that estimate the dynamic effects of antimicrobial resistance on multiple stakeholders

(Howard et al. 2024b) we can hope to disentangle uncertainty and tackle the complex global problem of AMR.

Given the significant recent advancements in foundational models and LLMs, it is evident they will play a big role in the future of AI CDSSs in healthcare, whether involved in inference or simply supporting AI-human interactions. One outstanding problem with such models is 'hallucinations' where they fabricate false information. This can be intrinsic where it directly contradicts another source or, more importantly, extrinsic where information cannot be verified. Several approaches have been developed to mitigate hallucinations (Tonmoy et al. 2024) including using Retrieval Augmented Generation (Lewis et al. 2021), Knowledge Graphs (Pan et al. 2024), Chain-of-Thought (Wei et al. 2023), verifiers (Zheng et al. 2023), or evaluation frameworks (Bolton et al. 2024), but no method is full proof given the stochastic nature of LLMs. The explainability of such models may therefore be more important than the results in Chapter 6 for more simple models elude to. However, as models get larger understanding their outputs becomes increasingly complex, reasoning becomes less faithful (Lanham et al. 2023), and we currently have a limited understanding of how machines 'think' (Anthropic 2024). Fairness will also need to be ensured with research demonstrating post-processing methods such as those used in Chapter 5 are often the best approach (Cruz et al. 2023). Reasoning and planning are important for clinical decision-making (Sandhu et al. 2006). Significant work is aiming to make such models better at these tasks (OpenAI 2024; Wang et al. 2024), with agents able to undertake autonomous actions being the goal of many researchers. Such agents are unlikely to be fully autonomous for key clinical decisions in the near future but advancements in world models, long-term memory, and applications to narrow focused tasks will surely benefit healthcare. Indeed work on building foundational models for health data is ongoing (Abbaspourazad et al. 2024). In the future, such models may be beneficial for a wide range of healthcare applications and adapted to new tasks simply by fine-tuning the model head. Foundational models therefore will play a growing role in AI CDSSs moving forwards, but exactly how they are combined with more traditional methods to best tackle multi-dimensional healthcare problems such as antimicrobial prescribing and AMR remains unanswered.

7.3 Conclusion

This thesis presents research on artificial intelligence-driven clinical decision support systems for antibiotic optimisation. Novel artificial intelligence methodologies have been demonstrated with the work exploring aspects of antibiotic prescribing under-represented in the literature but impactful with regard to a clinical team, their patient, and antimicrobial resistance. Specifically IV-to-oral switching and cessation, challenging real-world decisions faced by clinicians when conducting antimicrobial stewardship. Research spans from fundamental data analysis and model development to deploying artificial intelligence-based decision support applications with clinicians. This provides a basis for the design, development and evaluation of such systems as well as learnings applicable to antibiotic optimisation and artificial intelligence-based clinical decision support systems more broadly. Antimicrobial decisions are technically, ethically, and behaviourally complex but results from this thesis show such technology holds great promise to support optimised and individualised antibiotic prescribing decisions.

References

1. FDA (July 2024). *Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices*. URL: <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-ai-ml-enabled-medical-devices>.
2. Howard, Alex et al. (Jan. 2024a). "Antimicrobial learning systems: an implementation blueprint for artificial intelligence to tackle antimicrobial resistance". In: *The Lancet Digital Health* 6.1, e79–e86. ISSN: 2589-7500. DOI: 10.1016/S2589-7500(23)00221-2. URL: [https://www.thelancet.com/journals/landig/article/PIIS2589-7500\(23\)00221-2/fulltext](https://www.thelancet.com/journals/landig/article/PIIS2589-7500(23)00221-2/fulltext).
3. Bolton, William J. et al. (Nov. 2022). "Developing moral AI to support decision-making about antimicrobial use". In: *Nature Machine Intelligence* 4.11, pp. 912–915. ISSN: 2522-5839. DOI: 10.1038/s42256-022-00558-5. URL: <https://www.nature.com/articles/s42256-022-00558-5>.
4. Glimcher, Paul W. (2005). "Indeterminacy in brain and behavior". In: *Annual Review of Psychology* 56, pp. 25–56. ISSN: 0066-4308. DOI: 10.1146/annurev.psych.55.090902.141429.
5. Pearce, Tim et al. (Mar. 2023). *Imitating Human Behaviour with Diffusion Models*. DOI: 10.48550/arXiv.2301.10677. URL: <http://arxiv.org/abs/2301.10677>.
6. Rawson, Timothy M. et al. (June 2019a). "Artificial intelligence can improve decision-making in infection management". In: *Nature Human Behaviour* 3.6. ISSN: 2397-3374. DOI: 10.1038/s41562-019-0583-9. URL: <https://www.nature.com/articles/s41562-019-0583-9>.
7. Sant'Anna, Anita, Vilhelmsson, Andreas, and Wolf, Axel (June 2021). "Nudging healthcare professionals in clinical settings: a scoping review of the literature". In: *BMC Health Services Research* 21.1, p. 543. ISSN: 1472-6963. DOI: 10.1186/s12913-021-06496-z. URL: <https://doi.org/10.1186/s12913-021-06496-z>.
8. Khanna, Narendra N. et al. (Dec. 2022). "Economics of Artificial Intelligence in Healthcare: Diagnosis vs. Treatment". In: *Healthcare* 10.12, p. 2493. ISSN: 2227-9032. DOI: 10.3390/healthcare10122493. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9777836/>.
9. Moulds, Alison and Horton, Tim (2024). *How would clinicians use time freed up by technology? – The Health Foundation*. URL: <https://www.health.org.uk/publications/long-reads/how-would-clinicians-use-time-freed-up-by-technology>.
10. Warraich, Haider J., Tazbaz, Troy, and Califf, Robert M. (Oct. 2024). "FDA Perspective on the Regulation of Artificial Intelligence in Health Care and Biomedicine". In: *JAMA*. ISSN: 0098-7484. DOI: 10.1001/jama.2024.21451. URL: <https://doi.org/10.1001/jama.2024.21451>.

11. Charani, Esmita, Castro-Sánchez, Enrique, and Holmes, Alison (June 2014). "The Role of Behavior Change in Antimicrobial Stewardship". In: *Infectious Disease Clinics of North America*. Antimicrobial Stewardship 28.2, pp. 169–175. ISSN: 0891-5520. doi: 10.1016/j.idc.2014.01.004. URL: <https://www.sciencedirect.com/science/article/pii/S0891552014000087>.
12. Jermutus, Eva et al. (2022). "Influences on user trust in healthcare artificial intelligence: A systematic review". In: *Wellcome Open Research* 7.65, p. 65.
13. Nazar, Mobeen et al. (2021). "A Systematic Review of Human–Computer Interaction and Explainable Artificial Intelligence in Healthcare With Artificial Intelligence Techniques". In: *IEEE Access* 9, pp. 153316–153348. ISSN: 2169-3536. doi: 10.1109/ACCESS.2021.3127881. URL: <https://ieeexplore.ieee.org/abstract/document/9614151>.
14. Rawson, Timothy M et al. (Jan. 2019b). "Public acceptability of computer-controlled antibiotic management: An exploration of automated dosing and opportunities for implementation". In: *Journal of Infection* 78.1, pp. 75–86. ISSN: 0163-4453. doi: 10.1016/j.jinf.2018.08.005. URL: <https://www.sciencedirect.com/science/article/pii/S0163445318302500>.
15. Rawson, Timothy M. et al. (Dec. 2021). "Optimizing antimicrobial use: challenges, advances and opportunities". In: *Nature Reviews Microbiology* 19.12, pp. 747–758. ISSN: 1740-1534. doi: 10.1038/s41579-021-00578-9. URL: <https://www.nature.com/articles/s41579-021-00578-9>.
16. Hanretty, Alexandra M. and Gallagher, Jason C. (2018). "Shortened Courses of Antibiotics for Bacterial Infections: A Systematic Review of Randomized Controlled Trials". In: *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* 38.6, pp. 674–687. ISSN: 1875-9114. doi: 10.1002/phar.2118. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/phar.2118>.
17. Royer, Stephanie et al. (May 2018). "Shorter Versus Longer Courses of Antibiotics for Infection in Hospitalized Patients: A Systematic Review and Meta-Analysis". In: *Journal of hospital medicine* 13.5, pp. 336–342. ISSN: 1553-5592. doi: 10.12788/jhm.2905. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5945333/>.
18. Spellberg, Brad and Rice, Louis B. (Aug. 2019). "Duration of Antibiotic Therapy: Shorter Is Better". In: *Annals of internal medicine* 171.3, pp. 210–211. ISSN: 0003-4819. doi: 10.7326/M19-1509. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6736742/>.
19. Wald-Dickler, Noah and Spellberg, Brad (Oct. 2019). "Short-course Antibiotic Therapy—Replacing Constantine Units With "Shorter Is Better"". In: *Clinical Infectious Diseases* 69.9, pp. 1476–1479. ISSN: 1058-4838. doi: 10.1093/cid/ciy1134. URL: <https://doi.org/10.1093/cid/ciy1134>.

20. Yahav, Dafna et al. (Sept. 2019). "Seven Versus 14 Days of Antibiotic Therapy for Uncomplicated Gram-negative Bacteremia: A Noninferiority Randomized Controlled Trial". In: *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 69:7, pp. 1091–1098. ISSN: 1537-6591. doi: 10.1093/cid/ciy1054.
21. Fleming, Scott L. et al. (Dec. 2023). *MedAlign: A Clinician-Generated Dataset for Instruction Following with Electronic Medical Records*. doi: 10.48550/arXiv.2308.14089. URL: <http://arxiv.org/abs/2308.14089>.
22. Zheng, Lianmin et al. (Dec. 2023). *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. doi: 10.48550/arXiv.2306.05685. URL: <http://arxiv.org/abs/2306.05685>.
23. Bolton, William James et al. (Mar. 2024). *RAmBLA: A Framework for Evaluating the Reliability of LLMs as Assistants in the Biomedical Domain*. doi: 10.48550/arXiv.2403.14578. URL: <http://arxiv.org/abs/2403.14578>.
24. Lecue, Freddy (Jan. 2020). "On the role of knowledge graphs in explainable AI". In: *Semantic Web* 11:1, pp. 41–51. ISSN: 1570-0844. doi: 10.3233/SW-190374. URL: <https://content.iospress.com/articles/semantic-web/sw190374>.
25. Safavi, Tara and Koutra, Danai (Sept. 2021). *Relational World Knowledge Representation in Contextual Language Models: A Review*. doi: 10.48550/arXiv.2104.05837. URL: <http://arxiv.org/abs/2104.05837>.
26. Liu, Zhiyuan, Lin, Yankai, and Sun, Maosong, eds. (2023). *Representation Learning for Natural Language Processing*. Springer Nature. doi: 10.1007/978-981-99-1600-9. URL: <https://library.oapen.org/handle/20.500.12657/76271>.
27. Naveed, Humza et al. (Apr. 2024). *A Comprehensive Overview of Large Language Models*. doi: 10.48550/arXiv.2307.06435. URL: <http://arxiv.org/abs/2307.06435>.
28. Kraljevic, Zeljko et al. (Jan. 2023). *Foresight – Generative Pretrained Transformer (GPT) for Modelling of Patient Timelines using EHRs*. doi: 10.48550/arXiv.2212.08072. URL: <http://arxiv.org/abs/2212.08072>.
29. Komorowski, Matthieu et al. (Nov. 2018). "The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care". In: *Nature Medicine* 24:11, pp. 1716–1720. ISSN: 1546-170X. doi: 10.1038/s41591-018-0213-5. URL: <https://www.nature.com/articles/s41591-018-0213-5>.
30. Kang, Shida, Li, Kaiwen, and Wang, Rui (Sept. 2024). "A survey on pareto front learning for multi-objective optimization". In: *Journal of Membrane Computing*. ISSN: 2523-8914. doi: 10.1007/s41965-024-00170-z. URL: <https://doi.org/10.1007/s41965-024-00170-z>.

31. Lempert, Robert J (2019). "Robust decision making (RDM)". In: *Decision making under deep uncertainty: From theory to practice*, pp. 23–51.
32. Howard, Alex et al. (Jan. 2024b). "Antimicrobial treatment imprecision: an outcome-based model to close the data-to-action loop". In: *The Lancet Infectious Diseases* 24.1, e47–e58. ISSN: 1473-3099. DOI: 10.1016/S1473-3099(23)00367-5. URL: <https://www.sciencedirect.com/science/article/pii/S1473309923003675>.
33. Tonmoy, S. M. Towhidul Islam et al. (Jan. 2024). *A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models*. DOI: 10.48550/arXiv.2401.01313. URL: <http://arxiv.org/abs/2401.01313>.
34. Lewis, Patrick et al. (Apr. 2021). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. DOI: 10.48550/arXiv.2005.11401. URL: <http://arxiv.org/abs/2005.11401>.
35. Pan, Shirui et al. (July 2024). "Unifying Large Language Models and Knowledge Graphs: A Roadmap". In: *IEEE Transactions on Knowledge and Data Engineering* 36.7, pp. 3580–3599. ISSN: 1041-4347, 1558-2191, 2326-3865. DOI: 10.1109/TKDE.2024.3352100. URL: <http://arxiv.org/abs/2306.08302>.
36. Wei, Jason et al. (Jan. 2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. DOI: 10.48550/arXiv.2201.11903. URL: <http://arxiv.org/abs/2201.11903>.
37. Lanham, Tamera et al. (July 2023). *Measuring Faithfulness in Chain-of-Thought Reasoning*. DOI: 10.48550/arXiv.2307.13702. URL: <http://arxiv.org/abs/2307.13702>.
38. Anthropic (May 2024). *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet*.
39. Cruz, André and Hardt, Moritz (Oct. 2023). "Unprocessing Seven Years of Algorithmic Fairness". In: URL: <https://openreview.net/forum?id=jr03SfWsBS>.
40. Sandhu, Harpartap et al. (Dec. 2006). "Clinical Decisionmaking: Opening the Black Box of Cognitive Reasoning". In: *Annals of Emergency Medicine* 48.6, pp. 713–719. ISSN: 0196-0644. DOI: 10.1016/j.annemergmed.2006.03.011. URL: <https://www.sciencedirect.com/science/article/pii/S0196064406004082>.
41. OpenAI (Sept. 2024). *Learning to Reason with LLMs*. URL: <https://openai.com/index/learning-to-reason-with-llms/>.
42. Wang, Lei et al. (Mar. 2024). "A survey on large language model based autonomous agents". In: *Frontiers of Computer Science* 18.6, p. 186345. ISSN: 2095-2236. DOI: 10.1007/s11704-024-40231-1. URL: <https://doi.org/10.1007/s11704-024-40231-1>.
43. Abbaspourazad, Salar et al. (Mar. 2024). *Large-scale Training of Foundation Models for Wearable Biosignals*. DOI: 10.48550/arXiv.2312.05409. URL: <http://arxiv.org/abs/2312.05409>.

Appendix

A.1 Evaluation metric formulas

Area Under the Receiver Operating Characteristic Curve (AUROC)

The AUROC is the area under the Receiver Operating Characteristic (ROC) curve.

$$\text{AUROC} = \int_0^1 \text{TPR}(\text{FPR}) d\text{FPR}$$

Accuracy

Accuracy is the proportion of correct predictions.

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

Precision

Precision measures the proportion of correctly predicted positive observations.

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

Recall

Recall measures the proportion of actual positives that were correctly predicted.

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

F1 Score

The F1 score is the harmonic mean of precision and recall.

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

Area Under the Precision-Recall Curve (AUPRC)

The AUPRC is the area under the Precision-Recall (PR) curve.

$$\text{AUPRC} = \int_0^1 \text{Precision}(\text{Recall}) d\text{Recall}$$

True Positive Rate (TPR)

True Positive Rate is the same as recall.

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

False Positive Rate (FPR)

False Positive Rate measures the proportion of negatives incorrectly classified as positives.

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

Root Mean Squared Error (RMSE)

RMSE measures the square root of the average squared difference between predicted and actual values.

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

TP (True Positives) = The number of positive cases correctly identified as positive.

FP (False Positives) = The number of negative cases incorrectly identified as positive.

FN (False Negatives) = The number of positive cases incorrectly identified as negative.

TN (True Negatives) = The number of negative cases correctly identified as negative.

A.2 Chapter 3

Table A.1: Co-morbidities for the patient in question and for those patients identified as most similar by each method for each question in the multi-morbid patient similarity survey.

Question	Patient	Co-morbid patient embeddings	One hot encodings	Rocheteau similarity score
1	Hypertensive disorder, Hypercholesterolemia and Neoplasm of kidney	Hypertensive disorder, Neoplasm of kidney	Hypertensive disorder, Hypercholesterolemia, Insulin treated type 2 diabetes mellitus, Neoplasm of kidney	Hypertensive disorder, Diabetes mellitus type 2, Hypercholesterolemia, Osteoarthritis, Atrial fibrillation, Mixed anxiety and depressive disorder, Malignant tumor of prostate, Neoplasm of kidney
2	Asthma, Depressive disorder, Hypothyroidism and Fibromyalgia	Asthma, Depressive disorder, Hypothyroidism, Asperger	Asthma, Depressive disorder, Fibromyalgia	Asthma, Hypertensive disorder, Diabetes mellitus type 2, Gastroesophageal reflux disease, Cerebrovascular accident, Mixed anxiety and depressive disorder, Hyperlipidemia, Fibromyalgia
3	Ischemic heart disease, Hearing loss, Dementia, Essential hypertension, Alzheimer's disease and Diverticulosis of sigmoid colon	Hypertensive disorder, Myocardial infarction, Cataract, Dementia, Alzheimer's disease	Hearing loss, Dementia	Hypertensive disorder, Hypercholesterolemia, Transient ischemic attack, Osteoarthritis, Diverticular disease, Benign prostatic hyperplasia, Myocardial infarction, Alzheimer's disease, Retention of urine, Small vessel cerebrovascular disease
4	Hypertensive disorder, Varicella, Polycystic ovaries and Scoliosis deformity of spine	Hypertensive disorder, Varicella, Polycystic ovaries	Varicella, Polycystic ovaries, Scoliosis deformity of spine	
5	Asthma, Hemorrhoids, Colitis, Viral hepatitis type B, Gastritis, Human immunodeficiency virus infection, Supraventricular tachycardia, Non-alcoholic fatty liver, Portal hypertension and Hemophilia	Asthma, Osteoporosis, Hypertensive disorder, Hypercholesterolemia, Vitamin D deficiency, Depressive disorder, Hemorrhoids, Atrial fibrillation, Chronic obstructive lung disease, Anemia, Rheumatoid arthritis, Inflammatory disease of liver, Dysphagia, Steatosis of liver, Aortic valve regurgitation	Asthma, Depressive disorder, Hemorrhoids, Human immunodeficiency virus infection	Hemorrhoids, Atrial fibrillation, Hypothyroidism, Viral hepatitis type B
6	Hypertensive disorder, Diabetes mellitus type 2, Hypercholesterolemia, Vitamin D deficiency, Benign prostatic hyperplasia, Angina, Old myocardial infarction and Chronic ischemic heart disease	Hypertensive disorder, Diabetes mellitus type 2, Ischemic heart disease, Angina	Hypertensive disorder, Diabetes mellitus type 2, Hypercholesterolemia, Vitamin D deficiency	Hypertensive disorder, Diabetes mellitus type 2, Benign prostatic hyperplasia, Gastroesophageal reflux disease, Iron deficiency anemia, Acute non-ST segment elevation myocardial infarction, Angina, Heart failure, Constipation, Hyperlipidemia, Retention of urine, Bilateral cataracts, Heart failure with normal ejection fraction
7	Varicella and Rheumatoid arthritis with arteritis	Varicella, Rheumatoid arthritis	Rheumatoid arthritis with arteritis	Asthma, Varicella, Transient ischemic attack, Cerebrovascular accident, Fibromyositis, Rheumatoid arthritis with aortitis, Seizures in response to acute event
8	Hypertensive disorder, Varicella, Gestational diabetes mellitus and Pre-eclampsia	Varicella, Gestational diabetes mellitus, Pre-eclampsia, Pregnancy-induced hypertension	Hypertensive disorder, Varicella, Gestational diabetes mellitus	Varicella, Gestational diabetes mellitus, Pre-eclampsia
9	Asthma, Hypertensive disorder, Osteoarthritis, Type 2 diabetes, Hypercholesterolemia, Anemia, Gastroesophageal reflux disease and Hypothyroidism	Asthma, Hypertensive disorder, Osteoarthritis, Diabetes, Hyperlipidemia, Anemia, Gastroesophageal reflux disease, Obstructive sleep apnea	Asthma, Hypertensive disorder, Osteoarthritis, Type 2 diabetes, Hypercholesterolemia	Asthma, Hypertensive disorder, Rheumatoid arthritis, Diabetes, Hypercholesterolemia, Anemia, Gastroesophageal reflux disease, Hypothyroidism, Coronary arteriosclerosis, Pulmonary embolism, Chronic kidney disease
10	Osteoarthritis and Alcoholism	Osteoarthritis, Alcohol dependence	Osteoarthritis, Alcoholism, Peripheral nerve entrapment	

Patient co-morbidity similarity questionnaire

Thank you for taking the time to complete our survey. For each question please select the patient (A, B or C) you believe is the most similar to the patient in question with regards to their co-morbidities.

Please note that these co-morbidities were extracted from SNOMED codes and may be historic diagnosis.

[Sign in to Google](#) to save your progress. [Learn more](#)

* Indicates required question

Patient in question has: Hypertensive disorder, Hypercholesterolemia and Neoplasm of kidney *

☐ Patient A: Hypertensive disorder, Neoplasm of kidney

☐ Patient B: Hypertensive disorder, Hypercholesterolemia, Insulin treated type 2 diabetes mellitus, Neoplasm of kidney

☐ Patient C: Hypertensive disorder, Diabetes mellitus type 2, Hypercholesterolemia, Osteoarthritis, Atrial fibrillation, Mixed anxiety and depressive disorder, Malignant tumor of prostate, Neoplasm of kidney

Any comments or questions on this case?

Your answer

Figure A.1: Screenshot of a section from the multi-morbid patient similarity survey. Note that the context, instructions, and patient similarity questions given to participants were asked in a plain text format so as not to introduce any biases.

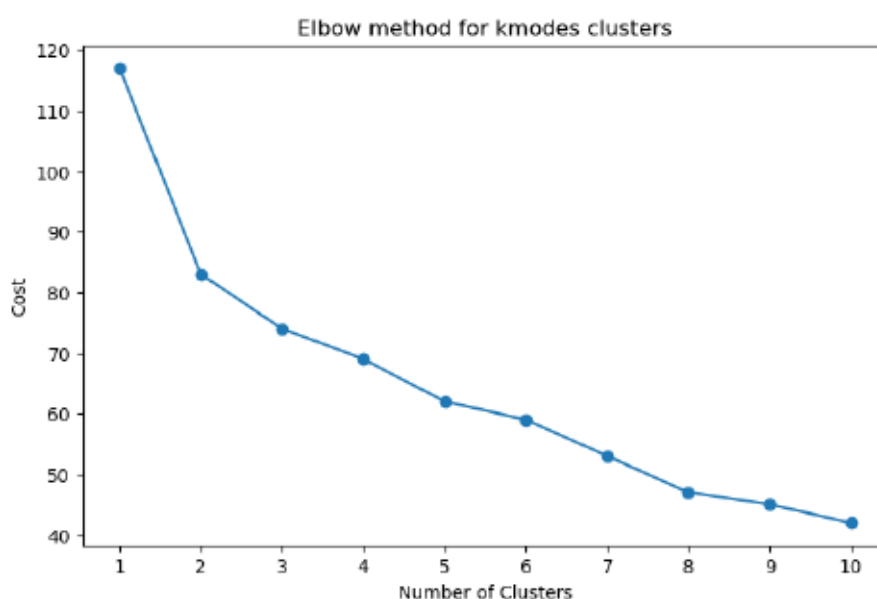


Figure A.2: Results for the elbow method on K-modes clustering of the survey data. Note that the infection point occurs at two clusters.

A.3 Chapter 4

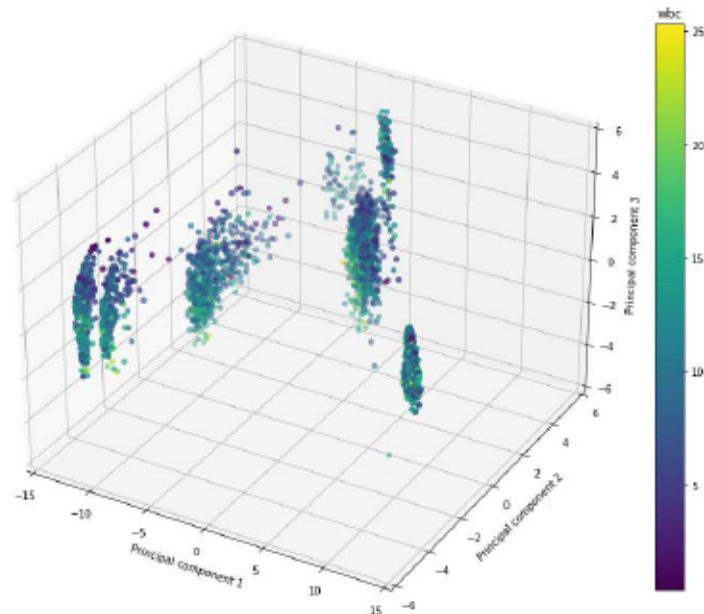


Figure A.3: Principal component analysis visualisation of all the embeddings for WBC estimation within the unseen test set.

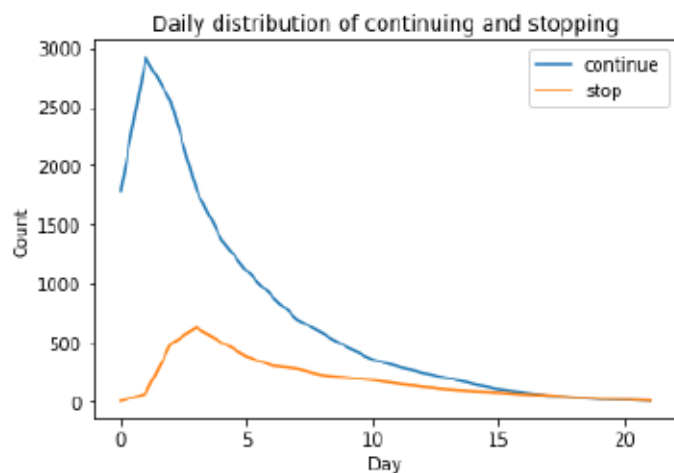


Figure A.4: Distribution of continuing and stopping antibiotic treatment by day within the test dataset. Excluding the particular patient i whose outcomes are being investigated, those who continue on a given day are within the continue donor pool, while those who stop on that day are within the stop donor pool.

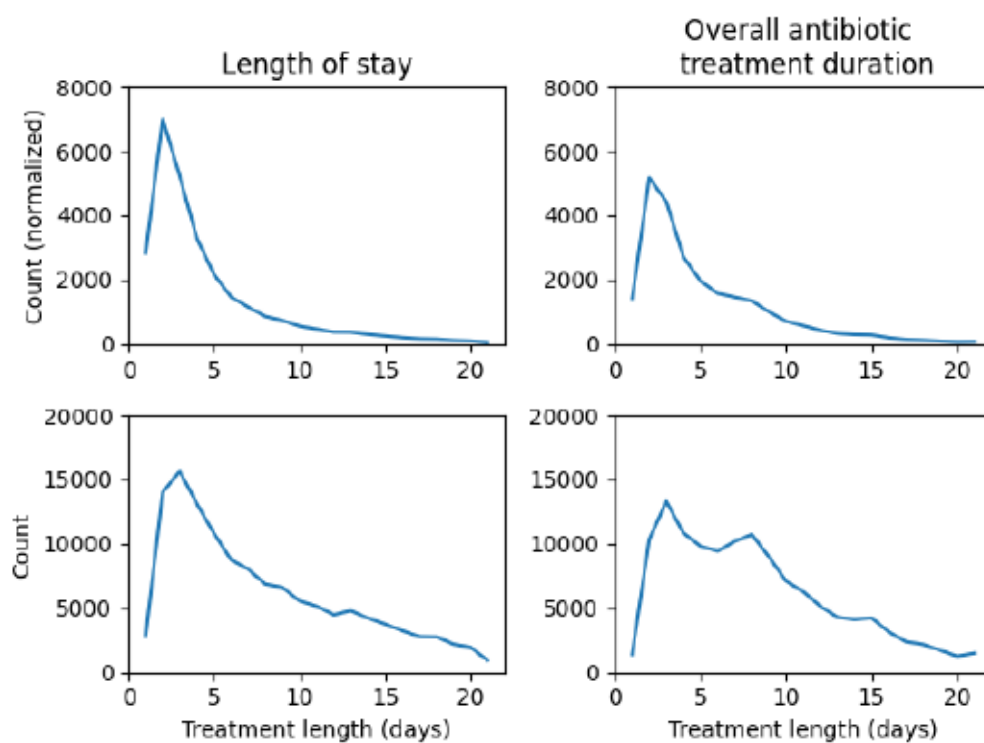


Figure A.5: Overall antibiotic treatment duration and length of stay for patients. The normalised count represents the number of distinct occurrences for that specific day, while the raw count represents the number of days of data available.

A.4 Chapter 5

Table A.2: Antibiotics selected for incomplete oral absorption analysis. * Selected due to nonlinear absorption.

Antibiotic	Oral bioavailability (%)
Amoxicillin	70
Ampicillin	50
Augmentin	70
Azithromycin	37
Cefpodoxime	50
Ciprofloxacin	70
Clarithromycin	50
Clindamycin*	90
DiCLOXacillin	49
Erythromycin	32
Flucloxacillin	50
Neomycin	3
Nitrofurantoin	80
Penicillin	60
Tetracycline	60

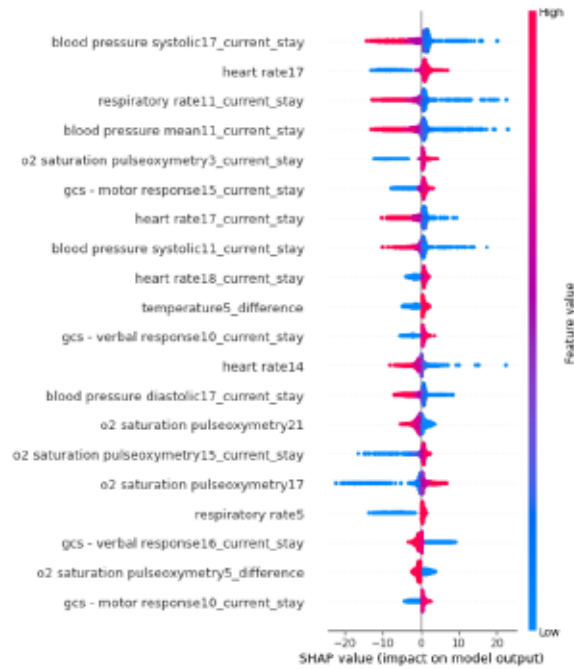


Figure A.6: SHAP and feature value distribution for the 20 most important features. Notice how 'respiratory rate11_current_stay' and 'blood pressure mean11_current_stay' (features of importance 3 and 4 respectively) are very similar implying redundancy.

Table A.3: Catch22 time-series transformation key. Note that the mean and standard deviation were also included. More information on these transformations can be found at <https://feature-based-time-series-analys.gitbook.io/catch22-features/>.

Key	Corresponding transformation
0	DN_HistogramMode_5
1	DN_HistogramMode_10
2	CO_f1ecac
3	CO_FirstMin_ac
4	CO_HistogramAMI_even_2_5
5	CO_trev_1_num
6	MD_hrv_classic_pnn40
7	SB_BinaryStats_mean_longstretch1
8	SB_TransitionMatrix_3ac_sumdiagcov
9	PD_PeriodicityWang_th0_01
10	CO_Embed2_Dist_tau_d_expfit_meandiff
11	IN_AutoMutualInfoStats_40_gaussian_fmfi
12	FC_LocalSimple_mean1_taresrat
13	DN_OutlierInclude_p_001_mdrmd
14	DN_OutlierInclude_n_001_mdrmd
15	SP_Summaries_welch_rect_area_5_1
16	SB_BinaryStats_diff_longstretch0
17	SB_MotifThree_quantile_hh
18	SC_FluctAnal_2_rsrangefit_50_1_logi_prop_r1
19	SC_FluctAnal_2_dfa_50_1_2_logi_prop_r1
20	SP_Summaries_welch_rect_centroid
21	FC_LocalSimple_mean3_stderr
22	Mean
23	Standard deviation

Table A.4: Patient mean mortality by switch event temporal difference.

	Model	Short		Long	
	Threshold	1st	2nd	1st	2nd
Number of days between the real and predicted switch event	-7	0.00	0.00	0.00	0.00
	-6	0.50	0.00	0.00	0.00
	-5	0.00	0.50	0.40	0.00
	-4	0.17	0.00	0.00	0.33
	-3	0.43	0.17	0.00	0.00
	-2	0.11	0.38	0.09	0.17
	-1	0.04	0.29	0.00	0.00
	0	0.07	0.00	0.05	0.02
	1	0.00	0.00	0.00	0.05
	2	0.00	0.10	0.00	0.00
	3	0.00	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.00

Table A.5: Threshold optimisation fairness results for the short model.

Sensitive attribute	Group	Original results				Threshold optimisation results			
		AUROC	TPR	FPR	EO	AUROC	TPR	FPR	EO
Age	20	0.73	0.74	0.27	✓	0.63	0.86	0.61	✗
	30	0.80	0.86	0.26	✓	0.72	0.73	0.28	✓
	40	0.78	0.81	0.25	✓	0.76	0.82	0.31	✓
	50	0.76	0.78	0.25	✓	0.81	0.86	0.25	✓
	60	0.79	0.82	0.23	✓	0.79	0.87	0.29	✓
	70	0.73	0.69	0.23	✓	0.78	0.87	0.31	✓
	80	0.77	0.81	0.26	✓	0.80	0.87	0.26	✓
	90	0.78	0.79	0.23	✗	0.78	0.86	0.3	✓
Race	Asian	0.79	0.83	0.24	✓	0.71	0.81	0.38	✓
	Black	0.78	0.83	0.27	✓	0.79	0.86	0.28	✓
	Hispanic	0.80	0.85	0.25	✓	0.76	0.84	0.31	✓
	Native	0.78	0.97	0.43	✗	0.75	0.93	0.43	✗
	Other	0.76	0.72	0.19	✓	0.78	0.84	0.29	✓
	Unknown	0.79	0.83	0.25	✓	0.81	0.86	0.23	✓
	White	0.77	0.79	0.24	✓	0.78	0.87	0.31	✓
Insurance	Medicaid	0.72	0.69	0.26	✗	0.74	0.82	0.34	✓
	Medicare	0.78	0.81	0.25	✓	0.77	0.88	0.33	✓
	Other	0.78	0.80	0.24	✓	0.79	0.9	0.33	✓

Table A.6: Threshold optimisation fairness results for the long model.

Sensitive attribute	Group	Original results				Threshold optimisation results			
		AUROC	TPR	FPR	EO	AUROC	TPR	FPR	EO
Age	20	0.76	0.77	0.24	✓	0.69	0.63	0.24	✗
	30	0.74	0.64	0.20	✗	0.78	0.88	0.31	✓
	40	0.77	0.8	0.26	✓	0.76	0.76	0.23	✓
	50	0.8	0.87	0.26	✓	0.81	0.88	0.25	✓
	60	0.8	0.84	0.24	✓	0.81	0.86	0.24	✓
	70	0.81	0.86	0.23	✓	0.81	0.88	0.25	✓
	80	0.81	0.85	0.23	✓	0.82	0.89	0.25	✓
	90	0.78	0.78	0.22	✓	0.78	0.81	0.25	✓

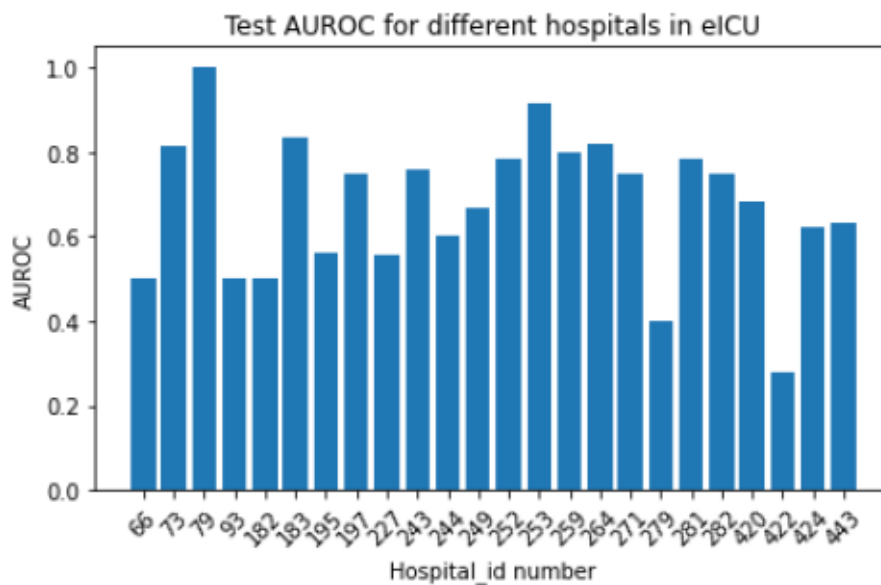


Figure A.7: eICU test results broken down by hospitals.

A.5 Chapter 6

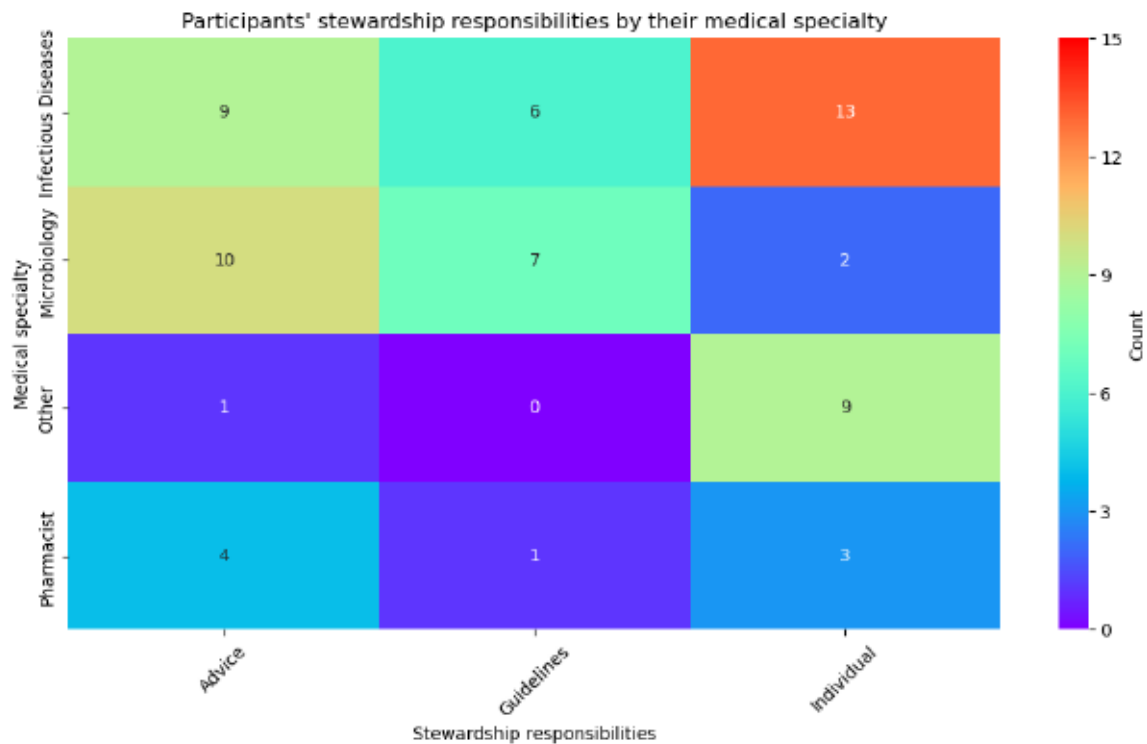


Figure A.8: Participants' stewardship responsibilities by their medical speciality. Note that those with a medical speciality of 'other' are not involved in guideline creation.

Table A.7: Chi-squared differences in interview coding results across demographics with a p-value below 0.20.

Theme	Demographic	Chi-squared statistic	p-value
Stewardship responsibilities	Medical speciality	18.55	<0.01
Stewardship responsibilities	Grade	3.48	0.18
Behavioural and social influences	Age	21.88	0.15
Problem for AI	Medical speciality	13.97	0.12
Usage	Medical speciality	11.57	0.07

Table A.8: Statistically significant differences in decision-making observed for patients when stratifying by participant demographics.

Patient	Demographic	Group	View	Switch Choice	Count	Chi-squared statistic	p-value
7	Grade	Consultant	CDSS	Don't switch	10	5.75	0.02
				Switch	4		
			SOC	Don't switch	2		
				Switch	10		
7	AI familiarity	Slightly familiar	CDSS	Don't switch	8	4.74	0.03
				Switch	2		
			SOC	Don't switch	1		
				Switch	6		
10	Sex	Female	CDSS	Don't switch	6	4.02	0.04
				Switch	2		
			SOC	Don't switch	2		
				switch	9		

Table A.9: Statistically significant differences in decision-making observed through the Cochran-Mantel-Haenszel (CMH) test when controlling for participant demographics.

Patient	Demographic	CMH statistic	p-value
6	Sex	4.42	0.04
6	Grade	4.42	0.04
7	Age	8.79	<0.01
7	Sex	9.55	<0.01
7	Medical speciality	9.55	<0.01
7	Grade	9.55	<0.01
7	AI familiarity	7.06	<0.01

Table A.10: Statistically significant differences in decision-making observed for patients when stratifying by TAM subgroups.

Patient	TAM Category	TAM Question	Subgroup	Chi-squared statistic	p-value
7	Perceived usefulness	I feel comfortable with AI technology in healthcare	Agree	5.53	0.02
		In my job using this AI CDSS would be relevant	Strongly agree	4.30	0.04
		I would find the AI CDSS useful in my job	Agree	5.54	0.02
		This AI CDSS could improve the quality of care that I deliver	Agree	7.43	<0.01
		I preferred using this interface for antibiotic decision-making over software I currently use in clinical practice	Neither agree nor disagree	4.15	0.04
		The quality of the AI CDSS predictions was high	Agree	6.87	<0.01
		Details such as similar patients and their features improved my trust in the AI CDSS	Agree	8.98	<0.01
		Overall, I think this AI CDSS has the potential to be a useful tool for antibiotic intravenous-to-oral switch decision-making	Agree	4.87	0.03
	Perceived ease of use	The interface was pleasant with information clearly visible	Agree	12.54	<0.01
		It was easy to understand recommendations given by the AI CDSS	Agree	7.22	<0.01
	Self efficacy	It was easy to get acquainted using the AI	Agree	8.24	<0.01
	Subjective norm	Patients will welcome that I use this AI technology	Agree	4.65	0.03
10	Voluntariness	My use of the AI CDSS was voluntary	Strongly agree	6.67	<0.01
	Workflow	This AI CDSS could be seamlessly integrated into my current clinical workflow	Agree	6.04	0.01
	Intention	Assuming I had access to the AI CDSS, I would intend to use it	Agree	5.80	0.02
	Descriptive norm	How many of your colleagues would use this AI CDSS if it were implemented	Neither agree nor disagree	4.29	0.04
	Perceived usefulness	I liked using the AI CDSS	Agree	4.84	0.03
		I preferred making IV-to-oral switch decisions with the AI CDSS versus without	Agree	4.06	0.04
		In my job using this AI CDSS would be relevant	Agree	5.53	0.02
		I would find the AI CDSS useful in my job	Agree	5.54	0.02
		The quality of the AI CDSS predictions was high	Agree	10.02	<0.01
		Details such as similar patients and their features improved my trust in the AI CDSS	Agree	8.98	<0.01
		Overall, I think this AI CDSS has the potential to be a useful tool for antibiotic intravenous-to-oral switch decision-making	Strongly agree	5.19	0.02
	Perceived ease of use	The interface was pleasant with information clearly visible	Agree	3.87	0.05
		It was easy to understand recommendations given by the AI CDSS	Agree	5.17	0.02
	Behavior change	I was reassured when my clinical decision aligned with the AI CDSS recommendation	Agree	10.02	<0.01
	Intention	Assuming I had access to the AI CDSS, I would intend to use it	Agree	7.60	<0.01

Table A.11: Shannon entropy values for the switch decisions made by participants for each patient. Note due to two options (i.e., switch or don't switch) the minimum and maximum possible Shannon entropy values are 0 and 1 respectively.

Patient	Shannon entropy
1	0.00
2	0.28
3	0.78
4	0.28
5	0.28
6	0.45
7	1.00
8	0.97
9	0.53
10	1.00
11	1.00
12	0.94

Table A.12: Frequency analysis of clinicians' free text switch choice explanations. Please see the associated code for all the terms used when searching for those with a *. Searches were not case-sensitive.

Theme	Term	Overall frequency (%)	Don't switch frequency (%)	Switch frequency (%)
More information	Information*	23.76	20.26	29.38
Stop antibiotics	Stop	12.48	20.26	0.00
Clinical parameters	Temperature*	9.70	14.43	6.75
	Respiratory Rate*	3.56	6.7	1.61
	Heart Rate*	7.33	13.4	3.54
	Mean Arterial Pressure*	1.39	3.09	0.32
	SpO2*	2.38	4.64	0.96
	Glasgow Coma Score*	7.92	14.43	3.86
Other terms	Sepsis	1.98	3.09	1.29
	Afebrile	2.77	0.00	4.50
	Febrile	4.16	8.25	1.61
	Stable	15.84	3.61	23.47
	Apyrexial	1.58	0.0	2.57
	Clinical improvement	2.18	1.03	2.89
	Infection	11.68	8.76	13.5
	Pyrexia	1.39	2.58	0.64
	Hypotensive	1.19	2.06	0.64
	Improving	7.33	3.61	9.65
	Antibiotics	7.33	1.55	10.93

Table A.13: Hover duration for different components of the vignettes across all cases. Note that the low hover time for CDSS recommendations is because they were buttons only clicked on to access explanations.

Case vignette component	Total hover duration (s)
Vitals table	7453.28
Vitals graph	5231.41
UKHSA IVOS PDF	1349.90
Patient infection details and demographics	1246.19
Similar patients (explanation)	540.54
Guideline CDSS flowchart (explanation)	212.88
LLM free text insights (explanation)	194.28
Feature importance (explanation)	173.52
AI CDSS Patient table (explanation)	171.10
Guideline CDSS recommendation	145.85
AI CDSS recommendation	116.19

Table A.14: System Usability Scale (SUS) questionnaire results.

Question	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
I think that I would like to use this system frequently	1 (2%)	4 (10%)	12 (29%)	23 (56%)	1 (2%)
I found the system unnecessarily complex	6 (15%)	27 (66%)	5 (12%)	3 (7%)	0 (0%)
I thought the system was easy to use	0 (0%)	2 (5%)	4 (10%)	31 (76%)	4 (10%)
I think that I would need the support of a technical person to be able to use this system	13 (32%)	24 (59%)	3 (7%)	1 (2%)	0 (0%)
I found the various functions in this system were well integrated	0 (0%)	4 (10%)	6 (15%)	26 (63%)	5 (12%)
I thought there was too much inconsistency in this system	5 (12%)	27 (66%)	6 (15%)	2 (5%)	1 (2%)
I would imagine that most people would learn to use this system very quickly	0 (0%)	0 (0%)	3 (7%)	28 (68%)	10 (24%)
I found the system very cumbersome to use	7 (17%)	27 (66%)	5 (12%)	2 (5%)	0 (0%)
I felt very confident using the system	0 (0%)	2 (5%)	11 (27%)	23 (56%)	5 (12%)
I needed to learn a lot of things before I could get going with this system	7 (17%)	29 (71%)	4 (10%)	1 (2%)	0 (0%)

Table A.15: Statistically significant differences observed in the Technology Acceptance Model (TAM) questionnaire by patient demographics.

TAM Category / Question	Demographic	Statistical test	Statistic	p-value
Self efficacy	AI familiarity	ANOVA	3.18	0.04
People who are important to me think I should use AI technology	Sex	Kruskal-Wallis	3.91	0.05
Using the AI CDSS changed my decision-making	Grade	ANOVA	4.97	0.03
I think my healthcare institution has the necessary infrastructure to support this AI CDSS	Sex	ANOVA	6.42	0.02
I think my healthcare institution has the necessary infrastructure to support this AI CDSS	Medical speciality	ANOVA	6.18	<0.01

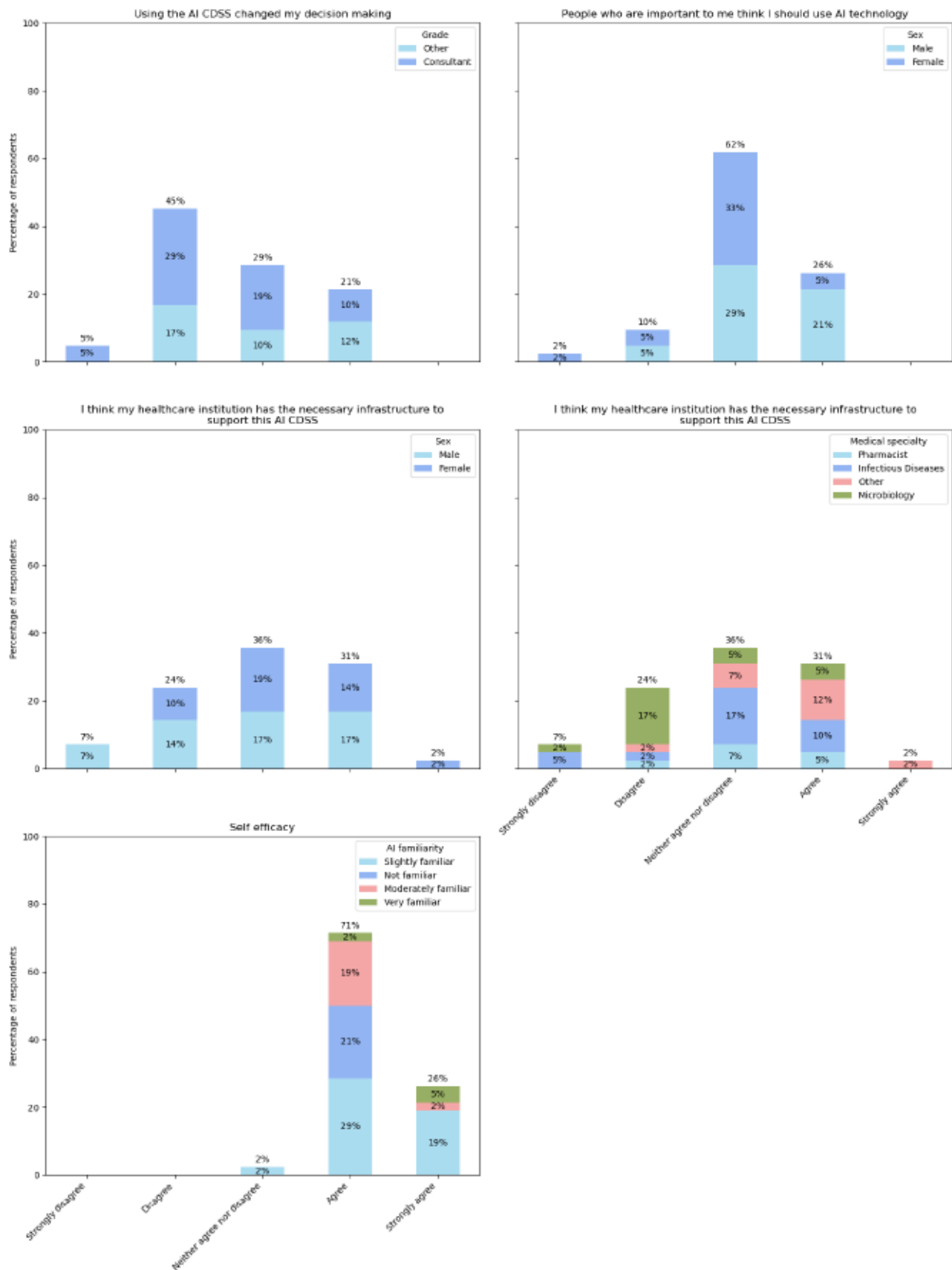


Figure A.9: Distribution of TAM results by patient subgroups for statistically significant differences.

Questions included in the topic guide for the semi-structured interviews.

Please describe your professional experience and your role within the organization where you currently work?

PROBE: • Position • speciality area • Organization

What types of decisions do you make regarding antibiotic prescribing/stewardship?

PROBE: • Initiation, selection of antimicrobial therapy, choice of route, switching, de-escalation, dosage, review, duration and cessation of treatment • Four Moments of Antibiotic decision-making – infection, empiric therapy, cultures, de-escalation, duration, cessation

What factors are key in influencing these decisions?

PROBE: • Data drivers: Pathogen/ Infectious species, Gram status of bacteria, Microbiology susceptibility results, Biomarker results, Length of patient stay, Allergy status, Current prescriptions, Drug-drug interactions, Previous procedures, Diagnoses, Co-morbidities, Patient medical history, Experience with prior patients, Local resistance rates, Guidelines (local, national, international), Policy, Age, Gender, Ethnicity • Outcome drivers: Mortality, 30-day Mortality, Hospital Mortality, Length of hospital stay, Length of ICU stay, 30-day re-admission, Antibiotic re-admission, Treatment response, Adverse events, Disease symptoms, Financial cost, Antibiotic free days, Avoiding antibiotic resistance, Avoiding antibiotic resistance

How often do you make these decisions?

PROBE: • Do you review the need to continue the prescribed antibiotic regimen daily? • Do you consider patient antibiotic switching/cessation daily?

How much of this decision-making is currently supported by innovative technology?

PROBE: • Type of technology • How the technology is used • Clinical decision support systems • AI

How much of this decision-making do you feel requires innovation?

PROBE: • What do you perceive such technology will look like? – EHR, App etc

What challenges, if any, do you face in implementing antibiotic stewardship practices/ programmes?

PROBE: • Most difficult part?

What additional information or support would you like to make antibiotic prescribing/ stewardship easier?

PROBE: • Resource needs • Innovation needs

Currently how individualized is antibiotic prescribing/ stewardship decision-making for a particular patient?

PROBE: • Do they need additional information for individualized decisions?

How do you expect this to change in the future (the next 3-5 years)?

What behavioural/ social aspects impact how you use antibiotics? Please explain.

PROBE: • Fear of HCAI or outbreaks in ICU/ward • Acceptance rate of failure/death in ICU

What is your awareness and perception of artificial intelligence (AI), technology?

PROBE: • What do you think it is/ does • Is AI technology currently impacting your work in clinical practice? • Have you previously used a digital health tool to help antibiotic prescribing/stewardship decisions?

To what degree are you optimistic or pessimistic about the use of such technology in healthcare, specifically with regards to infectious diseases, antibiotic prescribing, and stewardship?

To what degree do you believe an AI decision support system could help you optimize antibiotic use?

What factors would determine the level of trust you have in using an AI system to guide your antibiotic prescribing?

PROBE: • What would need to be shown/ done for you to trust such an AI system? • Do they have any major concerns regarding using AI?

What would convince you to consistently use an AI decision support system for antibiotic optimization?

PROBE: • Clinical evidence/ perception • Understanding how such a system works (Transparency) • Understand the reasons behind an output (Explainability) • Ease of use • Behavioural components – Fear of HCAI etc. • Organisational ethos • Organisational ethos/ management pressure • Perspective of people who are important to you • Guideline change

How do you expect AI CDSS/ digital health tool for antibiotic prescribing/stewardship to deliver value?

PROBE: • Increase the standard of care for its target patient population • Decrease time required for effective decision-making • Tackle AMR • Improve hard patient outcomes (mortality, LOS etc) • Decrease antibiotic prescribing

From your experience what is the biggest bottleneck between AI research and an actual decision aid/ product in use in real clinical practice aiming to improve antibiotic use?

PROBE: • Why does this bottleneck exist • What can be done about it

Do you have any final insights or recommendations?

System and user prompt templates used to code interview transcripts. Note the codebook and transcript were also included in the query.

System prompt:

You are a highly intelligent language model specialized in qualitative research analysis. Your task is to analyze interview transcripts and populate a qualitative research codebook in CSV format. The codebook contains the following columns: **theme**, **theme_definition**, **theme_notes**, **code**, **code_definition**, **example**, **code_notes**, **exists_in_text**, **text_snippet**. You must accurately determine if each code is explicitly referenced in the text and mark 'exists_in_text' with 1 if the code is directly referenced and 0 if it is not mentioned. Ensure that your determination is strict; only exact matches or direct paraphrases should be marked as 1. If referenced, and 'exists_in_text' is 1, include an extract of the transcript where the code is mentioned in the 'text_snippet' column. If you identify a relevant concept or topic that is not currently in the codebook, add a new row with appropriate details. Additionally, provide a very brief summary of key points and nuanced views from the respondent in bullet points. These bullet points do not need to directly align with the themes in the codebook but should rather highlight views the participant expressed that are most interesting and novel. Always provide the updated codebook before the bullet points. Be detailed and clear in your assessment to align closely with the provided instructions.

User prompt:

Below is an interview transcript. The questions and responses are transcribed verbatim. Analyze the transcript and fill in the qualitative research codebook based on the provided CSV structure. Please return the full codebook with 'exists_in_text' and 'text_snippet' completed. Ensure that 'exists_in_text' is marked as 1 only when the code is specifically mentioned or directly paraphrased in the text. Make sure the 1 and 0 are in the same column ('exists_in_text') in the csv. If you identify new relevant concepts or topics not currently in the codebook, add new rows with appropriate information. Additionally, provide a summary of the key highlights and nuanced views from the respondent in bullet points.

Below is an example of analyzing a snippet from a transcript and subsequently completing the codebook:

“Transcript Snippet:”

“Bolton, William: What outcomes would you want to be measured in a clinical trial evaluating such an AI clinical decision support system for antimicrobial prescribing?
Respondent: I think an easy metric would be the minimization of intravenous use, while not compromising pathogen directed therapy. It would also be great to an improvement in patient survival but I imagine that would be difficult to show.”

“Relevant Codes within the Codebook Completed:”

“theme,theme_definition,theme_notes,code,code_definition,example,code_notes,exists_in_text,text_snippet

Outcomes,The outcomes or endpoints the participant feels are key to demonstrate when evaluating an AI clinical decision support system for antibiotic prescribing and stewardship,Amount of antimicrobial use,Volume of antimicrobials used across the healthcare insitution,A reduction in antimicrobials use, while not compromising patients treatment,1,minimization of intravenous use, while not compromising pathogen directed therapy Outcomes,The outcomes or endpoints the participant feels are key to demonstrate when evaluating an AI clinical decision support system for antibiotic prescribing and stewardship,Mortality,Patient mortality or death,If you're AI algorithm showed that it could improve mortality. I think that would be very likely to be adopted. But even non inferiority with a benefit elsewhere would be good,1,an improvement in patient survival but I imagine that would be difficult to show Outcomes,The outcomes or endpoints the participant feels are key to demonstrate when evaluating an AI clinical decision support system for antibiotic prescribing and stewardship,Length of stay,Inpatient length of stay in the hospital,You might be able to show something like less time in the hospital, that would be very beneficial,0,”

Below is an example of analyzing a snippet from a transcript and subsequently updating the code book with a new row:

“Transcript Snippet:”

“Bolton, William: Are there any behavioural or social aspects that influence antimicrobial prescribing? Respondent: Yes, there are definitely social aspects that influence prescribing. For instance when choosing, if a patient would be suitable for outpatient antibiotics. That depends a lot on whether they can cope with it with the self-administration and also switching from an IV to an oral depends a lot on the patient support system, and how confident we are they could seek help if needed.”

“Revised Codebook Update:”

“theme,theme_definition,theme_notes,code,code_definition,example,code_notes,exists_in_text,text_snippet

Behavioural and social influences,Description of the behaviour and social factors that influence antimicrobial prescribing,Patient support system,The availability and reliability of a patients support system,Switching from an iv to an oral depends a lot on the patient and their support system and if we think they will take the medication,1,clinical risk of switching from an iv to an oral depends a lot on the patient support system”

Key points to remember: 1. “Explicit Reference Requirement”: Ensure that 'exists_in_text' is marked as 1 strictly when codes match specific words or direct paraphrases in the text. 2. “Addition of New Codes”: Add new rows/codes when a new concept or relevant topic is identified, to encourage comprehensive coverage. 3. “Snippets and Examples”: Use accurate snippets from the text to illustrate the references made, helping associate the codes directly with the interview text.###

Prompt template used to generate free text insights for the ICHT cases. Note a screenshot of the other explainable elements was also included in the query.

Title:

Explaining an AI Clinical Decision Support Systems (CDSS) Predictions for Antibiotic Stewardship

Scenario:

An inpatient is in hospital and is receiving intravenous (IV) antibiotic treatment. An AI-model based clinical decision support system (CDSS) can help clinicians in the important decision on if and when to switch the patient from IV-to-oral antibiotics. The model takes a patient's demographics including their co-morbidities and their vitals as input features. If the AI CDSS prediction is 'Switch' the model believes changing the antibiotic administration route from IV-to-oral is appropriate. Alternatively, if the AI CDSS prediction is 'Don't switch' the model believes changing the antibiotic administration route from IV-to-oral is not appropriate, and that IV treatment should be continued. To help clinicians understand why the AI model predicts a patient should or should not switch they can view more details as shown in the attached image.

Task:

You are an expert in both the medical domain and AI explainability and interpretability. Your task is to provide an overall summary in the form of a couple of bullet points outlining why the AI CDSS made its prediction. You will be provided with CSV data containing the patients' vitals since antibiotic initiation. The attached image presents the explainability and interpretability information available to clinicians. This includes a table containing the patients' demographics and clinical information, the model's prediction is shown under the 'AI CDSS prediction' column. A table containing historical similar patient cases is also shown with their demographics, clinical information, models predictions shown under the 'AI CDSS prediction' column, the real switch action that was taken is shown under the 'Real action' column, if the prediction was correct and aligns with the real action a yes is indicated in the 'Prediction correct' column otherwise it is No. Finally, a similarity plot based on the patients' features derived from their vitals compares the similar patients to the patient in question (this was plotted via imshow in matplotlib, blue indicates greater similarity between the similar patient and the patient in question while yellow indicates less similarity).

Instructions:

1. Review the data and image: - Familiarise yourself with the provided CSV data containing the patients' vitals during their stay. - Examine the attached image, which encompasses demographics, clinical information, AI CDSS prediction, neighbouring patient profiles, predicted and real actions, and a feature similarity plot.
2. Patient Information: - Review the demographics and clinical details of the patient in question. - Identify key factors relevant to the decision-making process, such as age, sex, underlying health conditions, infection, and vitals.
3. Patient data: - Analyse the csv data provided and identify key trends that could link to the AI models prediction. - In particular focus on the last 24-hour period.
4. Similar Patient Profiles: - Evaluate the demographics and clinical details of similar patients retrieved from the embedding space. - Compare the real actions taken for these similar patients with the actions predicted by the AI model.
5. Feature Similarity Plot: - Interpret the vitals feature similarity plot provided, indicating the degree of resemblance between the patient in question and neighbouring patients. - Remember that for each feature on the x-axis blue indicates greater feature similarity between the similar patient (on the y-axis) and the patient in question, while yellow indicates less similarity, ensuring you understand this distinction. - Discuss how the plot reflects the decision-making process and influences the AI model's prediction.
6. Prediction Explanation: - Provide a comprehensive rationale for the AI model's prediction regarding the transition from IV-to-oral antibiotic therapy or the continuation of the current IV therapy. - Focus on logical reasoning over prior knowledge and consider all the information provided including the similarities and differences between the patient in question and the neighbouring patients in terms of demographics and clinical features.

Example Bullet Points (note the context and meaning of these should be ignored, they are only intended to provide an example of how to structure a potential output): - The patient's temperature has shown an increasing trend over the past 24 hours, which could indicate a worsening infection. - For four out of the five similar patients the AI model predicted that they should not switch from IV-to-oral antibiotics, which aligns with the prediction for the patient in question. - The patient's vitals show a greater similarity to patients for whom the AI model predicted that IV treatment should be switched.

Additional Guidance: - Utilise your medical expertise and understanding of AI explainability to interpret the information and situation provided. - Consider the complex and critical nature of medical decision-making for antibiotic stewardship. - Emphasise the potential significance of each factor in the AI CDSS prediction. - Ensure clarity, conciseness, coherence, and relevance in the explanation provided. - Do not make assumptions outside what the data indicates.

CSV Data: {INSERT_PATIENTS_DATA}

Conclusion:

To repeat your task is to analyse the provided CSV data and image, summarise the information and provide guidance on why the AI CDSS made its prediction in a few bullet points, do not provide any other text. Please provide at least provide one bullet point related to the csv data and one that comments on the similar patients, feel free to include additional points as necessary based on your analysis.

Prompt template used to generate free text insights for the MIMIC cases. Note a screenshot of the other explainable elements was also included in the query.

Title:

Explaining an AI Clinical Decision Support Systems (CDSS) Predictions for Antibiotic Stewardship

Scenario:

An inpatient is in hospital and is receiving intravenous (IV) antibiotic treatment. An AI-model based clinical decision support system (CDSS) can help clinicians in the important decision on if and when to switch the patient from IV-to-oral antibiotics. The model has 5 input features that are numerical values between 0 and 1 derived from the patients' vitals time series data during their stay. Two thresholds can be used to transform the output of the model into a binary prediction, the first maximised the area under the receiver operating characteristic in the validation set while the second higher threshold minimised the false positive rate. If the AI CDSS prediction is 'Switch' using both thresholds the model believes changing the antibiotic administration route from IV-to-oral is appropriate. If the model prediction is outputting 'Switch' but the second higher threshold outputs 'Don't switch' the AI CDSS prediction is 'Potentially switch'. Alternatively, if the AI CDSS prediction is 'Don't switch' for both the model believes changing the antibiotic administration route from IV-to-oral is not appropriate, and that IV treatment should be continued. To help clinicians understand why the AI model predicts a patient should or should not switch they can view more details as shown in the attached image.

Task:

You are an expert in both the medical domain and AI explainability and interpretability. Your task is to provide an overall summary in the form of a couple of bullet points outlining why the AI CDSS made its prediction. You will be provided with CSV data containing the patients' vitals since antibiotic initiation. The attached image presents the explainability and interpretability information available to clinicians. This includes a table containing the patients' demographics, features, and the two model predictions shown under the 'AI CDSS prediction' and 'AI CDSS prediction (higher threshold)' columns. A table containing historical similar patient cases is also shown with their demographics, the models two predictions are again shown under the 'AI CDSS prediction' and 'AI CDSS prediction (higher threshold)' columns, the real switch action that was taken is shown under the 'Real action' column, if the prediction was correct and aligns with the real action a yes is indicated in the 'Prediction correct' column otherwise it is No, if the higher threshold prediction was correct and aligns with the real action a yes is indicated in the 'Higher threshold prediction correct' column otherwise it is No. Finally, a feature contribution table indicates for the similar patients how each feature contributed towards the prediction. A red/orange colour indicates it led towards a more negative prediction (i.e., 'Don't switch'), while blue indicates a contribution towards the prediction 'Switch'. Green or teal is neutral and should be ignored.

Instructions:

1. Review the data and image: - Familiarise yourself with the provided CSV data containing the patients' vitals during their stay. - Examine the attached image, which encompasses demographics, features, clinical information, AI CDSS prediction, neighbouring patient profiles, predicted and real actions, and a feature contribution table.
2. Patient Information: - Analyse the demographics, features and clinical details of the patient in question. - Identify key factors relevant to the decision-making process, such as age, sex, features, and vitals.
3. Patient data: - Analyse the csv data provided and identify key trends that could link to the AI models prediction. - In particular focus on the last 24-hour period.
4. Similar Patient Profiles: - Evaluate the demographics, and features of similar patients. - Compare the real actions taken for these similar patients with the actions predicted by the AI model.
5. Feature Contribution Table: - Interpret the feature contribution table provided, indicating the impact of the features on the AI predictions for the neighbouring patients. - Remember that a red/orange colour indicates it led towards a more negative prediction (i.e., 'Don't switch'), while blue indicates a contribution towards the prediction 'Switch'. Green or teal is neutral and should be ignored. - Discuss how the plot reflects the decision-making process and influences the AI model's prediction.
6. Prediction Explanation: - Provide a comprehensive rationale for the AI model's prediction regarding the transition from IV-to-oral antibiotic therapy or the continuation of the current IV therapy. - Focus on logical reasoning over prior knowledge and consider all the information provided including the similarities and differences between the patient in question and the neighbouring patients in terms of demographics, features, and predictions.

Additional Guidance: - Utilise your medical expertise and understanding of AI explainability to interpret the information and situation provided. - Consider the complex and critical nature of medical decision-making for antibiotic stewardship. - Emphasise the potential significance of each factor in the AI CDSS prediction. - Ensure clarity, conciseness, coherence, and relevance in the explanation provided. - Do not make assumptions outside what the data indicates.

CSV Data: {INSERT_PATIENTS_DATA}

Conclusion:

To repeat your task is to analyse the provided CSV data and image, summarise the information and provide guidance on why the AI CDSS made its prediction in a few bullet points, do not provide any other text. Please provide at least provide one bullet point related to the csv data and one that comments on the similar patients, feel free to include additional points as necessary based on your analysis.

Publications

B.1 Journal articles

———— Chapter 2 ————

Bolton, W.J., Badea, C., Georgiou, P., Holmes, A. and Rawson, T.M., 2022. Developing moral AI to support decision-making about antimicrobial use. *Nature Machine Intelligence*, 4(11), pp.912-915.

———— Chapter 3 ————

Bolton, W.J., Georgiou, P., Holmes, A. and Rawson, T.M., 2024. Co-morbidity Representation in Artificial Intelligence: Tapping into Unused Clinical Knowledge. In *AI for Health Equity and Fairness: Leveraging AI to Address Social Determinants of Health* (pp. 173-196). Springer Nature Switzerland.

———— Chapter 4 ————

Bolton, W.J., Rawson, T.M., Hernandez, B., Wilson, R., Antcliffe, D., Georgiou, P. and Holmes, A.H., 2022. Machine learning and synthetic outcome estimation for individualised antimicrobial cessation. *Frontiers in Digital Health*, 4, p.997219.

———— Chapter 5 ————

Bolton, W.J., Wilson, R., Gilchrist, M., Georgiou, P., Holmes, A. and Rawson, T.M., 2024. Personalising intravenous to oral antibiotic switch decision-making through fair interpretable machine learning. *Nature communications*, 15(1), p.506.

Bolton, W.J., Wilson, R., Gilchrist, M., Georgiou, P., Holmes, A. and Rawson, T.M., 2025. The impact of artificial intelligence-driven decision support on uncertain antimicrobial prescribing: a randomized multi-method study. Undergoing peer review at the Lancet Digital Health.

B.2 Other articles

Bolton, W.J., Poyiadzi, R., Morrell, E., Bueno, G.V.B.G. and Goetz, L., RAMBLA: A Framework for Evaluating the Reliability of LLMs as Assistants in the Biomedical Domain. In ICLR 2024 Workshop on Reliable and Responsible Foundation Models.

B.3 Conference abstracts and posters

Bolton, W.J., Rawson, T.M., Hernandez, B., Wilson, R., Antcliffe, D., Georgiou, P. and Holmes, A.H., 2022. Predicting mortality and length of stay outcomes for individual patients receiving antibiotics using a recurrent neural network.

Bolton, W.J., Rawson, T.M., Hernandez, B., Wilson, R., Antcliffe, D., Georgiou, P. and Holmes, A.H., 2022. Recognising the inherent diversity in patients with infections and the challenge of fairness in AI systems.

Bolton, W.J., Rawson, T.M., Hernandez, B., Wilson, R., Antcliffe, D., Georgiou, P. and Holmes, A.H., 2022. Exploring the impact of obesity on infection related outcomes in a large ICU dataset.

Bolton, W.J., Wilson, R., Gilchrist, M., Georgiou, P., Holmes, A. and Rawson, T.M., 2023. Machine learning for individualised antibiotic intravenous-to-oral decision-making.

Bolton, W.J., Georgiou, P., Holmes, A. and Rawson, T.M., 2023. Exploring the impact of co-morbidities on empiric treatment of urinary tract infections and antimicrobial resistance in a large dataset. Won a top rated poster prize.

Bolton, W.J., Wilson, R., Gilchrist, M., Georgiou, P., Holmes, A. and Rawson, T.M., 2024. Machine learning based individualised intravenous-to-oral antibiotic switch decision support.

Bolton, W.J., Georgiou, P., Holmes, A. and Rawson, T.M., 2024. Machine learning for antibiotic optimisation in multi-morbid patients.

———— AAI ————

Bolton, W.J., Georgiou, P., Holmes, A. and Rawson, T.M., 2024. Co-morbidity Representation in Artificial Intelligence: Tapping into Unused Clinical Knowledge. In AAI 2024 Workshop on Health Intelligence.

———— ICLR ————

Bolton, W.J., Poyiadzi, R., Morrell, E., Bueno, G.V.B.G. and Goetz, L., RAMBLA: A Framework for Evaluating the Reliability of LLMs as Assistants in the Biomedical Domain. In ICLR 2024 Workshop on Reliable and Responsible Foundation Models. Work conducted during an internship at GSK.ai.

B.4 Conference oral presentations

———— ECCMID 2023 ————

Bolton, W.J., Rawson, T.M., Hernandez, B., Wilson, R., Antcliffe, D., Georgiou, P. and Holmes, A.H. Machine learning based white blood count estimation for individualised antimicrobial cessation.

———— AAI Health Intelligence Workshop 2024 ————

Bolton, W.J., Georgiou, P., Holmes, A. and Rawson, T.M. Co-morbidity Representation in Artificial Intelligence: Tapping into Unused Clinical Knowledge.

Bolton, W.J., Wilson, R., Gilchrist, M., Georgiou, P., Holmes, A. and Rawson, T.M. Prospective and End User Evaluation of an AI-based Clinical Decision Support System for Individualized Antimicrobial Stewardship.

Bolton, W.J., Wilson, R., Gilchrist, M., Georgiou, P., Holmes, A. and Rawson, T.M. Prospective evaluation of a machine learning-based decision support system for intravenous-to-oral antibiotic switching.

B.5 Additional selected activities

- *Developed a zine in collaboration with Damien Ming, Bernard Hernandez, Holly Worrall, Anjana Nair, and Charlie Delikan, illustrating important aspects behind the development and use of AI systems, 2022.*
- *Gave a talk on ethical and responsible AI for CAMO-Net UK, December 2023.*
- *Presented my PhD research at an event exploring the impact of AI in research and practice on AMR, February 2024. Co-organized by Imperial College London's Institute of Infection, the AI Network, the Centre of Antimicrobial Optimisation, and the NIHR Health Protection Research Unit for Healthcare-Associated Infections and Antimicrobial Resistance.*
- *Delivered an invited talk on my PhD research at the 2024 Advancing Data Technologies to Corner AMR (ADTCA) Conference, June 2024.*
- *Led a workshop on the design, evaluation and integration of AI decision support in healthcare for CAMO-Net UK, June 2024. It included giving a talk on translating AI, web-app demos, interactive questionnaires and breakout exercises.*
- *Accepted into the Conception X and AI SuperConnector programmes to explore entrepreneurship and the commercialisation of AI research into ethical and impactful ventures, 2023-2025.*

Code and data

The computer code used in this research is available at <https://github.com/WilliamBolton>

The MIMIC-IV dataset can be found at <https://physionet.org/content/mimiciv/2.0/> and the eICU dataset at <https://physionet.org/content/eicu-crd/2.0/>. Both are accessible once you are a credentialed user on Physionet, have completed the required training and signed the appropriate data use agreement. Data from Imperial College Healthcare NHS Trust can be accessed via contacting the iCARE team and completing the steps described at <https://www.imperial.ac.uk/medicine/research-and-impact/groups/icare/icare-facility/information-for-researchers/>.

Glossary

AAAI	Association for the Advancement of Artificial Intelligence
AI	Artificial Intelligence
AMR	Antimicrobial Resistance
ANOVA	Analysis Of Variance
AUPRC	Area Under the Precision Recall curve
AUROC	Area Under the Receiver Operating Characteristic curve
ASI	Antibiotic Spectrum Index
BC	Behaviour-Change
BCE	Binary Cross Entropy
BIDMC	Beth Israel Deaconess Medical Center
CAMO	Centre for Antimicrobial Optimisation
Catch22	Canonical Time-series Characteristics 22
CDSS	Clinical Decision Support System
Chi2	Chi-squared
CI	Confidence Interval
CMH	Cochran-Mantel-Haenszel
CRP	C-Reactive Protein
csv	Comma-separated values
DL	Deep Learning
ECCMID	European Congress of Clinical Microbiology and Infectious Diseases
ED	Emergency Department
EHR	Electronic Health Record
eICU	eICU Collaborative Research Database
EO	Equalised Odds
ESCMID	European Society of Clinical Microbiology and Infectious Diseases
EU	European Union
FPR	False Positive Rate
GCS	Glasgow Coma Scale
GMC	General Medical Council
GPT	Generative Pre-Trained Transformer
HCAIs	Healthcare-Associated Infections
HPRU	Health Protection Research Unit
ICD	International Classification of Diseases
ICHT	Imperial College Healthcare NHS Trusts
iCARE	Imperial Clinical Analytics Research and Evaluation

ICID	International Congress on Infectious Diseases
ICREC	Imperial College Research Ethics Committee
ICU	Intensive Care Unit
IV	Intravenous
IVOS	Intravenous-to-Oral Switch
LLM	Large Language Model
LOS	Length Of Stay
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MIMIC-IV	Medical Information Mart for Intensive Care IV
ML	Machine Learning
MRSA	Methicillin-resistant Staphylococcus aureus
NICE	National Institute for Health and Care Excellence
NHS	National Health Service
NIHR	National Institute for Health Research
PCT	Procalcitonin
PEOU	Perceived-Ease-of-Use
p-value	Probability Value
PU	Perceived-Usefulness
ReLU	Rectified Linear Unit
RGIT	Research Governance and Integrity Team
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
SD	Standard Deviation
SE	Self-Efficacy
SHAP	SHapley Additive exPlanations
SGD	Stochastic gradient descent
SMOTE	Synthetic Minority Oversampling Technique
SN	Subjective-Norm
SNOMED CT	Systematized Nomenclature of Medicine Clinical Terms
SOC	Standard of Care
SpO2	saturation of peripheral oxygen
SUS	System Usability Scale
TAM	Technology Acceptance Model
t-SNE	t-distributed Stochastic Neighbour Embedding
TPR	True Positive Rate
UK	United Kingdom
UKHSA	UK Health Security Agency
UMLS	Unified Medical Language System
UTI	Urinary Tract Infection
WBC	White Blood Cell Count

Ad Astra