



**Reducing Post-Operative Infections: The Development and
Validation of Artificial Intelligence-Predictive Model to Inform
Shared Decision Making.**

Volume 1 of 1

By Nehal Abdelkhalek Mohamed Hassan

**Submitted for the degree of Doctor of Philosophy to Newcastle University, Faculty of
Medical Sciences (Population Health Sciences Institute), School of Pharmacy.**

January 2024

Abstract:

Background: Healthcare systems worldwide generate sizeable patient-related health data. There is growing interest amongst clinicians and healthcare staff in how this can be used to support patient care. One example is the use of predictive analytics in determining the risk of developing a particular complication, such as an infection post-surgery.

Objective: To develop an artificial intelligence (AI) model to predict the likelihood of post-operative infection in surgical patients, while also exploring clinicians' and patients' perceptions on using AI decision support tools more broadly to inform shared decision making.

Methods: This PhD programme of work involved a number of different stages. The literature was systematically reviewed for AI models that could inform clinical decision making with regard to post-surgical infection, and a candidate list of positive predictor variables extracted. A prognostic AI-model was developed to predict the risk of infection, and any inherent biases identified. Another systematic review was conducted to understand how clinicians and patients perceive using AI decision aids in shared decision making, and semi-structured interviews were carried out with clinicians to explore how to improve the clinical utility of AI decision support tools.

Results: Nine steps were identified for developing AI-predictive models; the first six steps were applied in the development and evaluation of our model. Nineteen predictors were used. The ensemble model displayed high performance in training (sensitivity: 85.3%, specificity: 74.6%, AUC: 88.6%) and internal validation (sensitivity: 96.9%, specificity: 74.1%, AUC: 85.5%). Patients and clinicians raised concerns about AI model interfaces, in general, and their impact on clinical/patient conversations. Several suggestions were made on how to improve the model's clinical application.

Conclusion: This study provided a deeper understanding of how AI-predictive models can guide shared decision making. Future work should concentrate on improving the user inclusivity of these tools and reducing the risk of inherent bias that could potentially mislead clinical decision making.

Acknowledgements:

A PhD is a long learning journey. This journey could not have been accomplished without the support and guidance of those who contribute along the way. Here, I extend my heartfelt gratitude to the individuals who have been instrumental in shaping my PhD experience:

First and foremost, my deep appreciation goes to my supervisors, Prof. Sarah Slight and Dr. Robert Slight. Their unwavering assistance, encouragement, and dedication began well before my PhD commenced, with early stages of my scholarship application. Their guidance propelled me towards numerous achievements and milestones, and as an international student, their welcoming and supportive manner made me feel right at home. I consider myself fortunate to have had such outstanding supervisors by my side.

I extend my sincere thanks to my examiners, Prof. Céire Costelloe, Professor of Health Informatics at Imperial College London and Dr. Richard McNally, Reader in Epidemiology at the Population Health Sciences Institute, Newcastle University. Their investment of time and expertise to provide constructive feedback on my work is deeply appreciated.

To Dr. Clare Tolley, Lecturer of Pharmacy Practice, School of Pharmacy, Newcastle University I am grateful for your support and advice at various stages of my PhD as an advisor on my annual progress review panel. Your significant input on the qualitative aspects of my work has been invaluable.

I express my gratitude to our collaborators; Prof. Graham Morgan, Professor of Computing at Newcastle University, Dr. Daniel Weiland, Consultant Microbiologist at Newcastle Upon Tyne Foundation NHS Trust, Prof. Akke Vellinga, Professor of Epidemiology at University College Dublin, Ireland, and Prof. David Bates, Professor of Medicine and Public Health at Harvard Medical School, United States for their expertise and support from the inception of this programme of work until the final stages of my thesis.

I would also like to thank my colleague, Eng. Hazem Mamdouh, a Machine learning Engineer and Technical lead at Research Development International (RDI), Egypt, for training me on Python programming language and providing advice and feedback throughout the process of writing code scripts.

I was privileged by collaborating with Suzy Gallier, Alexander Topham, Sarah Lim and Prof. Elizabeth Sapey from the PIONEER Research Hub. Their assistance in establishing the trusted

research environment at Newcastle University and their support in accessing and analysing the Sepsis Phenotype data has been a rewarding experience.

I am also grateful to Dr. Richard Hethrington, Senior Lecturer at the FMS Graduate School for his guidance through the PGR development programme at Newcastle University. His insightful feedback on my work presented during the doctoral prize competition and STEM for Britain, has been truly appreciated.

To Anna Robinson, my PhD buddy, and a Lecturer of Pharmacy Practice. Your friendship and companionship throughout this journey have been indispensable.

Last but not the least, I extend heartfelt thanks to my family in Egypt for their constant presence and support. My husband and daughter, Fathy and Alia, deserve special mention for their unwavering encouragement. I hope that Alia will learn the lesson from witnessing my PhD journey; learning is a lifelong endeavour, that is only achieved by passion, prayers, and perseverance.

Disclaimer:

Chat GPT version 3.5 was used as a proof-reading tool for some of the chapters in this thesis. All suggested changes were reviewed by the student and supervisors and some incorporated.

Publications:

- **Hassan N**, Slight RD, Bimpong K, Bates DW, Weiland D, Vellinga A, Morgan G, Slight SP. What are clinicians' and patients' perceptions on using artificial intelligence decision aids to inform shared decision making? A systematic review (Research article). [Under review by *Nature Digital Medicine – NPJ*]
- **Hassan N**, Slight R, Morgan G, Bates DW, Gallier S, Sapey E, Slight S. Road map for clinicians to develop and evaluate AI predictive models to inform clinical decision making. *BMJ Health & Care Informatics*. 2023;30(1). (Research article), [doi:10.1136/bmjhci-2023-100784](https://doi.org/10.1136/bmjhci-2023-100784) [Impact Factor: 4.1]
- **Hassan N**, Slight R, Weiland D, Vellinga A, Morgan G, Aboushareb F, Slight SP. Preventing sepsis; how can artificial intelligence inform the clinical decision making process? A systematic review. *International Journal of Medical Informatics*. 2021 Apr 10;104457. (Research article) <https://doi.org/10.1016/j.ijmedinf.2021.104457> [Impact Factor: 4.9]

Published Conference Papers and Proceedings:

- **Hassan N**, Slight R, Weiland D, Morgan G, Vellinga A, Mamdouh H, Gallier S, Sapey E, Slight S. Developing an AI-predictive model for predicting the likelihood of post-operative infection in surgical patients. *International Journal of Pharmacy Practice*. 2023 Dec 1;31(Supplement_2):ii22-3.
- Wilson S, Tolley C, Mc Ardle R, Beswick E, Lawson L, Slight R, Slight S, **Hassan N**. Strategies to develop, design, implement and support digital health equity in key underserved groups: protocol for a systematic review. In Alzheimer's Association International Conference 2023 Jul 16. ALZ. [Conference paper, Published]
- Rawi AA, **Hassan N**, Mc Ardle R, Slight S. Using artificial intelligence (AI) to predict the conversion of cognitively unimpaired individuals to mild cognitive impairment. In Alzheimer's Association International Conference 2023 Jul 16. ALZ. [Conference paper, Published]
- **Hassan N**, Slight R, Slight S. Healthcare staff perceptions on using artificial intelligence predictive tools: a qualitative study *BMJ Health & Care Informatics* 2022;29: [doi: 10.1136/bmjhci-2022-FCIASC.7](https://doi.org/10.1136/bmjhci-2022-FCIASC.7) [Impact Factor: 4.1]
- **Hassan N**, Slight RD, Weiland D, Vellinga A, Morgan G, Mamdouh H, Slight SP. AI-based Sepsis prediction in Surgical Practice; Lessons learnt from Model Development. AMIA annual Conference 2022 [Conference paper, Accepted]
- **Hassan N**, Slight RD, Bimpong K, Weiland D, Vellinga A, Morgan G, Slight SP. Clinicians' and patients' perceptions of the use of artificial intelligence decision aids to inform shared decision making: a systematic review. *The Lancet*. 2021

Nov 1;398:S80. (Conference abstract) [https://doi.org/10.1016/S0140-6736\(21\)02623-4](https://doi.org/10.1016/S0140-6736(21)02623-4) [Impact Factor: 168.9]

- **N Hassan**, R Slight, D Weiland, A Vellinga, G Morgan, F Aboushareb, S P Slight, Predicting infection and sepsis; what predictors have been used to train machine learning algorithms? A systematic review, *International Journal of Pharmacy Practice*, Volume 29, Issue Supplement_1, April 2021, Page i18, (Conference abstract) <https://doi.org/10.1093/ijpp/riab016.022> [Impact Factor: 1.8]
- **N Hassan**, R Slight, S P Slight, A survey of the knowledge and attitudes of Egyptian healthcare professionals towards the application of Health Information Technology to optimize antibiotic therapy, *International Journal of Pharmacy Practice*, Volume 29, Issue Supplement_1, April 2021, Page i28, (Conference abstract) <https://doi.org/10.1093/ijpp/riab015.033> [Impact Factor: 1.8]

Conferences and Speeches (Oral presentations):

- Podium presentation in the **Faculty of Clinical Informatics** annual conference, United Kingdom
- Oral Presentation in the **CiiEM Summer Research Summit**, Portugal
- Podium presentation in the **Faculty of Clinical Informatics** annual conference, United Kingdom
- Lightning talk at for the **UK Public Health Sciences** conference, United Kingdom.
- Podium presentation in **Faculty of Clinical Informatics** annual conference, United Kingdom
- Oral Presentation in **HSRPP 2021 Conference, University of Reading**, United Kingdom.
- Oral Presentation for **INSIGHTS Doctoral Thesis public lecture prize, Newcastle University**, United Kingdom
- Oral presentation for members of **both houses of parliament in STEM for Britain 2021**, United Kingdom
- **Three-minute thesis** competition, **Newcastle University**, United Kingdom

Conferences and Speeches (Poster presentations):

- Poster presentation at the **Royal Pharmaceutical Society 2023**, United Kingdom.
- Poster presentation at **American Medical Informatics Association (AMIA)** annual conference 2022, United States.
- Poster presentation at **Faculty of Clinical Informatics annual conference 2022**, United Kingdom.
- Two poster presentations at **Great North Pharmacy Research Conference 2021**, United Kingdom.
- Poster presentation at **HSRPP 2021 Conference**, University of Reading, United Kingdom
- Poster presentation at **National OPAT conference 2019** of the British Society of Antimicrobial Chemotherapy (BSAC), United Kingdom

- Poster presentation at **North East post-graduate conference (NEPG)** 2019, United Kingdom

Awards and Milestones during the PhD programme of work:

- Best e.paper oral presentation at the **Faculty of Clinical Informatics annual conference 2021 and 2023**
- Selected to join **Conception X** training (worth 5000 GBP funding by Newcastle University), which aims at transforming PhD researchers in deep tech. into venture scientists.
- Associate fellowship of the **Higher Education Academy**, UK.
- Winning **Insights doctoral thesis prize** by public votes, Faculty of Medical Sciences, Newcastle University
- One of the best 10 research projects for early career researchers in the UK at **STEM for Britain** Competition and presented PhD programme of work virtually to both **Houses of Parliament** in the UK.

Table of Contents

TABLE OF CONTENTS.....	VII
LIST OF TABLES	X
LIST OF FIGURES.....	X
LIST OF ABBREVIATIONS	XI
CHAPTER 1. INTRODUCTION.	1
1.1.ANTIMICROBIAL RESISTANCE - A GLOBAL HEALTH CRISIS	1
1.2.CONTRIBUTING FACTORS TO THE INCREASED RATES OF AMR	2
1.2.1. <i>The development of new antibiotics</i>	2
1.3.IMPACT OF AMR ON THE NATIONAL HEALTH SERVICE	3
1.3.1. <i>Cost to lives and the economy</i>	3
1.3.2. <i>Resistant infections and sepsis</i>	5
1.3.3. <i>Risk factors for Developing Infection and Sepsis</i>	6
1.4.TACKLING AMR – THE UK’S FIVE-YEAR NATIONAL ACTION PLAN	6
1.5.THE USE OF ARTIFICIAL INTELLIGENCE TO INFORM HEALTHCARE DECISIONS	6
1.6.AI-PREDICTIVE MODELS SUPPORTING SDM	7
1.7.THE USE OF AI-PREDICTIVE MODELS IN SURGICAL PRACTICE	9
1.8.CHAPTER SUMMARY	11
CHAPTER 2. AIM AND OBJECTIVES OF THE PHD PROGRAMME OF WORK.	12
2.1. AIM	12
2.1.2. <i>Objectives</i>	12
2.2. METHODOLOGICAL APPROACH	13
CHAPTER 3. THE STEPS INVOLVED IN DEVELOPING AN AI-PREDICTIVE MODEL TO INFORM SHARED DECISION MAKING.....	14
3.1. INTRODUCTION.....	14
3.2. BACKGROUND	14
3.2.1. <i>Clarifying the clinical question or outcome(s) of interest (output)</i>	15
3.2.2. <i>Identifying appropriate predictors (feature selection)</i>	16
3.2.3. <i>Choosing relevant datasets</i>	17
3.2.4. <i>Developing the AI-predictive model</i>	18
3.2.5. <i>Validating and testing the AI-predictive model</i>	20
3.2.6. <i>Presenting and interpreting the model prediction(s)</i>	21
3.2.7. <i>Licensing the AI-predictive model</i>	21
3.2.8. <i>Maintaining the AI-predictive model</i>	22
3.2.9. <i>On-going evaluation of the impact of the AI-predictive model</i>	22
3.3. CHAPTER SUMMARY	24
CHAPTER 4. IDENTIFYING THE OPTIMAL SET OF PREDICTORS USED TO TRAIN MACHINE LEARNING ALGORITHMS TO PREDICT THE LIKELIHOOD OF INFECTION AND SUBSEQUENT SEPSIS.	25
4.1. INTRODUCTION.....	25
4.2. BACKGROUND	25
4.3. METHODS.....	26
4.3.1. <i>Eligibility Criteria</i>	26
4.3.2. <i>Information sources and search strategy</i>	27
4.3.3 <i>Study selection</i>	27
4.3.4 <i>Data collection and analysis</i>	28
4.4. RESULTS.....	33
4.4.1 <i>Definitions</i>	33
4.4.2 <i>Predictors</i>	34

4.4.3 Machine learning algorithms	38
4.4.4 Prediction timeframe of the algorithms	42
4.5. DISCUSSION.....	42
4.6. CHAPTER SUMMARY	45
CHAPTER 5. THE DEVELOPMENT OF AN AI MODEL TO PREDICT THE LIKELIHOOD OF POST- OPERATIVE INFECTION IN SURGICAL PATIENTS (AIPIS).....	46
5.1. INTRODUCTION	46
5.2. BACKGROUND	46
5.3. AIM.....	47
5.3.1. Objectives	47
5.4. STUDY DESIGN AND METHODOLOGY	47
5.4.1. Clarifying the outcome of interest.....	47
5.4.2. Identifying appropriate predictors (feature selection)	50
5.4.3. Choosing relevant dataset.....	51
5.4.4. AI model development	52
5.4.5. Validating the AI-predictive model.....	54
5.5. RESULTS	55
5.5.1. Descriptive analysis	55
5.5.2. Predictors	59
5.5.3. Model performance.....	60
5.5.4. Presenting the model predictions and building the model interface	60
5.6. LIMITATIONS OF THE MODEL DEVELOPMENT	67
5.7. CHAPTER SUMMARY	68
CHAPTER 6. THE POTENTIAL FOR BIAS DURING THE DEVELOPMENT AND INTERNAL VALIDATION OF AI-BASED DECISION SUPPORT MODELS.....	70
6.1. INTRODUCTION	70
6.2. BACKGROUND	70
6.3. DIFFERENT TYPES OF BIAS.....	71
6.3.1. Bias occurring during development of AI models	71
6.3.2. Bias occurring during validation and testing of AI-predictive models.....	73
6.3.3. Bias occurring during the application of AI-predictive models.....	74
6.4. TOOLS TO ASSESS BIAS	75
6.5. GOVERNMENTAL GUIDANCE ON THE DETECTION AND MITIGATION OF BIAS IN AI TECHNOLOGIES	76
6.6. CHAPTER SUMMARY	77
CHAPTER 7. WHAT ARE CLINICIANS’ AND PATIENTS’ PERCEPTIONS ON USING ARTIFICIAL INTELLIGENCE DECISION AIDS TO INFORM SHARED DECISION MAKING?; A SYSTEMATIC REVIEW	78
7.1. INTRODUCTION	78
7.2. BACKGROUND	78
7.3. METHODS.....	79
7.3.1. Eligibility Criteria	79
7.3.2. Information sources and search strategy.....	79
7.3.3. Study selection	79
7.3.4. Data extraction and analysis.....	80
7.3.5. Quality assessment.....	81
7.4. RESULTS	81
7.4.1. Screening and prevention (risk prediction).....	87
7.4.2. Prognosis.....	89
7.4.3. Treatment.....	90
7.5. DISCUSSION	91
7.6. CHAPTER SUMMARY	94
CHAPTER 8. “THIS ISN'T TO REPLACE YOU, THIS ISN'T TO TELL YOU WHAT TO DO, THIS IS TO HELP YOU MAKE DECISIONS, SO YOU ARE MAKING THE DECISIONS WITH YOUR PATIENTS”;. 95	

CLINICIANS' PERCEPTIONS AROUND THE USE OF AI-BASED TECHNOLOGIES TO INFORM SHARED DECISION MAKING.....	95
8.1. INTRODUCTION.....	95
8.2. BACKGROUND	95
8.3. METHODS	96
8.3.1. <i>Participant recruitment</i>	96
8.3.2. <i>Data collection</i>	96
8.3.3. <i>Data Collection and analysis</i>	97
8.4. RESULTS.....	97
8.4.1. <i>Descriptive analysis</i>	97
8.4.2. <i>Theme 1: Participants' understanding of AI</i>	98
8.4.3. <i>Theme 2: Trustworthiness of the tool</i>	100
8.4.4. <i>Theme 3: Usability Features</i>	102
8.4.5. <i>Theme 4: Barriers to the adoption and use of AI tools in clinical practice</i>	103
8.4.6. <i>Theme 5: Future potential uses of AI</i>	106
8.5. DISCUSSION	108
8.6. CHAPTER SUMMARY	111
CHAPTER 9. FUTURE DIRECTIONS AND RECOMMENDATIONS.....	112
9.1. INTRODUCTION	112
9.2. KEY FINDINGS FROM DIFFERENT AREAS OF THIS WORK	112
9.3. STRENGTHS AND WEAKNESSES OF THE DIFFERENT AREAS OF WORK	114
9.4. AI-PREDICTIVE MODELS DEVELOPMENT; ACCESSING OF DATABASES THROUGH TRE	115
9.5. NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE (NICE) FRAMEWORK FOR DIGITAL HEALTH TECHNOLOGIES	116
9.6. EXTERNAL VALIDATION	119
9.7. EXPLORING HOW TO BEST PRESENT THE MODEL PREDICTIONS/OUTPUTS TO CLINICAL USERS	119
9.8. DESIGNING A FUTURE IMPACT STUDY	119
9.9. LICENSING AND MAINTENANCE.....	120
9.10. RECOMMENDATIONS	120
9.11. IMPACT OF COVID ON THE PROGRAMME OF WORK.....	121
9.12. CONCLUDING REMARKS	121
CHAPTER 10. APPENDICES	123
APPENDIX A	123
APPENDIX B.....	124
APPENDIX C.....	125
APPENDIX D	126
APPENDIX E.....	127
APPENDIX F	165
APPENDIX G	169
APPENDIX H	171
APPENDIX I.....	203
APPENDIX J	204
APPENDIX K.....	205
APPENDIX L.....	207
APPENDIX M	209
APPENDIX N	210
CHAPTER 11. REFERENCES.	211

List of Tables

Table 1. Newcastle-Ottawa Scale for Quality assessment of the included studies:.....	30
Table 2. Description of sepsis predictors in the included studies:	35
Table 3. Machine learning algorithms, predictive power and sepsis definition used:	39
Table 4. Descriptive analysis for categorical variables of the 2716 included patients:	55
Table 5. Descriptive analysis for continuous variables of the 2716 included patients:	57
Table 6. Performance measures of Ensemble and best models during training and internal validation:.....	60
Table 7. Description of AI decision aids:	82
Table 8. Descriptive analysis:	98
Table 9. Risk assessment of the AI model to predict the likelihood of post-operative infections in surgical patients:	118

List of Figures

Figure 1. Predicted mortality from AMR in 2050:[5].....	1
Figure 2. Rates of admission and readmission due to surgical site infections in England between 2017 and 2022:[20].....	4
Figure 3. Admission and Readmission trends of surgical site infections in England from 2012 to 2022 by micro-organism for superficial infections of all surgical types:[20]	5
Figure 4. Admission and Readmission trends of surgical site infections in England from 2012 to 2022 by micro-organism for deep, organ or space infections of all surgical types:[20]	5
Figure 5. Different definitions of sepsis and their related clinical predictors:.....	16
Figure 6. An illustration of a proposed two-level decision tree for sepsis prediction:	19
Figure 7. PRISMA chart for the search strategy and review process for the included studies:28	
Figure 8. Flow chart of the classification of patients in infected and non-infected groups:	49
Figure 9. A prototype of a possible model interface:.....	61
Figure 10. Model Output and recommendation:	62
Figure 11. Selected features using the RFECV without clinical input:	66
Figure 12. Changes in Python script to include clinical input:	67
Figure 13. Selected predictors after using a blend of clinical input and machine learning algorithms:	67
Figure 14. Different categories of Bias in AI models:[189]	71
Figure 15. Search strategy PRISMA flowchart:	80
Figure 16. Role of AI tools to guide SDM:.....	87
Figure 17. The blended approach to enhance the chances of successful clinical application of AI-predictive models:	114
Figure 18. Demonstrates different tiers of DHTs based on their function:[270].....	117

List of Abbreviations

Abbreviation	
AMR	Antimicrobial Resistance
WHO	World Health Organisation
GDP	Gross Domestic Product
DHSC	Department of Health and Social Care
UK	United Kingdom
NHS	National Health Services
ESPAUR	English Surveillance Programme for Antimicrobial Utilisation and Resistance
E.coli	Escherichia coli
SSI	Surgical Site Infection
CABG	Coronary Artery Bypass Grafting
MRSA	Methicillin Resistant Staphylococcus Aureus
MSSA	Methicillin Sensitive Staphylococcus Aureus
COVID-19	Coronavirus disease
AI	Artificial Intelligence
CDSS	Clinical Decision Support Systems
SDM	Shared Decision Making
NICE	National Institute for health and Care Excellence
ACS-SRC	American College of Surgeons Surgical Risk Calculator
ASA	American Society of Anaesthesiologists
POSSUM	Physiologic and Operative Severity Score for the Enumeration of Mortality and Morbidity
US	United States
AUC	Area Under the Curve
MIMIC III	Medical Information Mart for Intensive Care-III
BIDMC	Beth Israel Deaconess Medical Centre
ACS NSQIP	American College of Surgeons National Surgical Quality Improvement Program
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis
CONSORT-AI	Consolidated Standards of Reporting Trials in Artificial Intelligence
T2DM	Type 2 Diabetes Mellitus
SOFA	Sequential Organ Failure Assessment
SIRS	Systemic Inflammatory Response Syndrome
EHRs	Electronic Health Records
HRA	Health Research Authority
HDR	Health Data Research
GDPR	General Data Protection Regulation
NDOO	National Data Opt-Out
TRE	Trusted Research Environment
WBC	White blood cell
CRP	C-reactive protein
ROC	Receiver Operating Characteristic curve
AUROC	Area Under the Receiver Operating Characteristic Curve

MHRA	Medicines and Healthcare products Regulatory Agency
UKCA	United Kingdom Conformity Assessed
CE	Conformité Européenne
EU MDR	EU Medical Device Regulation
US FDA	United States Food and Drug Administration
MRC	Medical Research Council
ICU	Intensive Care Unit
PROSPERO	Prospective Register of Systematic Reviews
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
CINAHL	Cumulative Index of Nursing and Allied Health Literature
ICD	International Classification of the Disease
eGFR	estimated Glomerular Filtration Rate
DBP	Diastolic Blood Pressure
INR	international normalized ratio
SBP	Systolic blood pressure
AIS	Abbreviated Injury Scale
ISS	Injury Severity Score
GCS	Glasgow coma scale
SO2	Oxygen saturation
PaCO2	Partial pressure of carbon dioxide
MAP	Mean arterial pressure
SpO2	Partial oxygen pressure
AKI	Acute kidney injury
RR	Respiratory rate
HR	Heart rate
NEWS	National Early Warning Score
FiO2	Fraction of inspired oxygen
PTT	Prothrombin time
AST	Aspartate aminotransferase
BUN	Blood urea nitrogen
BNP	Atrial natriuretic peptide
CRP	C-reactive protein
HCO3	Bicarbonate
CCI	Charleston comorbidity index
UOP	urine output
ASA	American Society of Anaesthesiologists score
BMI	Body mass index
Hct	Haematocrit
hsCRP	Highly sensitive C-reactive protein
AF	Atrial fibrillation,
CAD	Coronary Artery Disease
DVT	Deep vein thrombosis
COPD	Chronic Obstructive Pulmonary Disease
MI	Myocardial infarction
PAD	Peripheral artery disease
ACR	Albumin-creatinine ratio
RNN	recurrent neural network

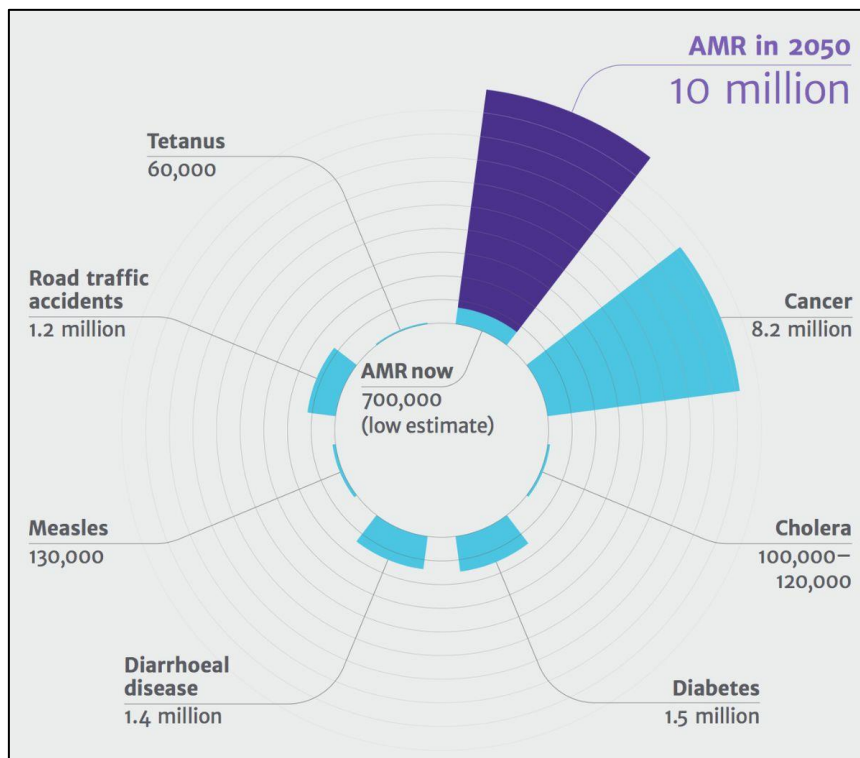
SMV	support vector machines
LASSO	least absolute shrinkage and selection operator
CARS	Computer-aided risk of sepsis model
AISE	Artificial Intelligence Sepsis Expert
CSRS	Composite sepsis risk score
MVA	Multivariable adjusted analyses
SRS	Sepsis risk score
SN	Sensitivity
SP	Specificity
PPV	Positive predictive value,
NPV	Negative predictive value
AIPIS	Artificial Intelligence to Predict Post-operative Infections in Surgery
HCAIs	Healthcare-associated Infections
PHE SMI I	Public Health England Standards for Microbiology Investigations
SNOMED-CT	Systematized Nomenclature of Medicine-Clinical Terms
RFECV	Recursive Feature Elimination with Cross-Validation
HDRUK	health data research UK
SQL	Structured Query Language
KNN	K-nearest neighbours
VS Code	Visual Studio Code
RF	Random Forest
PROBAST	Prediction model Risk of Bias Assessment Tool
MDR	Multi-drug Resistance
ITU	Intensive Therapy Unit
AVPU	Alert, Voice, Pain, Unresponsive
CK	Creatinine Kinase
PPIE	Public and Patients Involvement and Engagement
DVT	Deep Vein Thrombosis
ROB	Risk of Bias
QUIPS	Quality In Prognosis Studies
ENTREQ	ENhancing Transparency in REporting the synthesis of Qualitative research
CASP	Critical Appraisal Skills Programme
DTP	Direct-to-patient
AFSDM	Atrial Fibrillation Shared Decision Making
SCDA	Statin Choice Decision Aid
DMCDA	Diabetes Medication Choice Decision Aid
CPM	Clinical prediction models
WHAT/CDS	Women's Health Assessment Tool/Clinical Decision Support
EBDiMS	Evidence-Based Decision Support Tool in Multiple Sclerosis'
PSA	prostate-specific antigen
iPDA	integrated patient decision aid
DARTs	Decision Analysis in Routine Treatment Study
DHT	Digital Health Technologies
NPT	Normalisation Process Theory
NuTh	Newcastle Upon Tyne Hospitals Foundation Trust

Chapter 1. Introduction.

1.1. Antimicrobial Resistance - a Global Health Crisis

In 2019, antimicrobial resistance (AMR) was classified as one of the top ten global health problems by the World Health Organisation (WHO).[1] Antimicrobial resistance has been defined as the genetic change of microorganisms (bacteria, fungi, virus or parasite) making them non-responsive to antimicrobials, leading to infections that are more difficult to treat.[1] Antimicrobial resistance increases morbidity and mortality of hospitalised patients as well as those treated outside hospitals.[2] A systematic review assessed the global impact of AMR and found that approximately 4.95 million patients' deaths were associated with it in 2019; 1.27 million of these had AMR recorded as the direct cause of death.[3] Looking forward, it has been estimated that approximately 10 million people could die annually of AMR by 2050.[4] This figure exceeds the predicted mortality risk from conditions such as cancer (8.2 million), diabetes (1.5 million), and road traffic accidents (1.2 million), as shown in Figure 1.[5] Furthermore, the World Bank estimated that AMR could lead to a drop in the global gross domestic product (GDP) by 1.1–3.8% by 2050 if AMR is not curtailed.[6,7]

Figure 1. Predicted mortality from AMR in 2050:[5]



1.2.Contributing Factors to the Increased Rates of AMR

There are multiple contributing factors to the increased rates of AMR in hospitals, one of which is the misuse and overuse of antibiotics. Antibiotic misuse is defined as the inappropriate use of antibiotics which encompasses the antibiotic chosen for a particular indication and the duration of time over which it has been prescribed.[8] Inappropriate use of broad-spectrum antibiotics and incomplete antibiotic courses (i.e., stopping earlier than prescribed) have been associated with increased AMR.[9] Consequently, it is important that the right antibiotic is administered at the right time and for the correct duration. However, there has also been national and international antibiotic shortages of antibiotics that may, in part, be due to the complexity of supply chains. When antibiotics are unavailable, healthcare professionals may resort to less optimal treatments, with poorer clinical outcomes and increasing the risk of AMR. The highest rates of death from AMR have been observed in developing countries, where there have been challenges in accessing the right antibiotics as early as possible.[10] There is also a significant decline in the development of new antibiotics with different mechanisms of action.

1.2.1. *The development of new antibiotics*

The Department of Health and Social Care (DHSC) has increased the funding for new antibiotic research in 2015-2016 from £11.6 million to £30.6 million.[7] However, the United Kingdom (UK) Parliament also highlighted a number of challenges with this type of research, including how only six pharmaceutical companies in the UK are currently engaged in clinical trials in this area.[7] It has been reported that the development of a new antibiotic is a significantly time-consuming and costly process, with approximately 10 to 15 years of research needed to produce one new antibiotic.[11] Even with the successful development and deployment of a new antibiotic into the pharmaceutical market, clinical guidelines support restricting their use so as to preserve their effectiveness and prevent resistance. While this approach is clinically reasonable, it may not align with the pharmaceutical business strategy.[11] Consequently, there has been a "*discovery void*" over the last three decades in antibiotic research;[7] a situation that has been described by the Health and Social Care Committee to the UK House of Commons in 2018 as a "*pharmaceutical market failure*".[7]

1.3.Impact of AMR on the National Health Service

1.3.1. Cost to lives and the economy

Patients admitted to hospitals can often have multiple medical conditions that put them at an increased risk of infection and worsen their prognosis when infected; surgical patients are at a particularly higher risk due to the invasive nature of the procedures.[12] The rising rates of resistant infections and the declining number of novel antibiotics with new mechanisms of action has increased the risks associated with surgical procedures.[13] Outbreaks of nosocomial (hospital-acquired) infections have led to the cancellation of elective surgeries and an increase in waiting lists for patients, placing an additional burden on National Health Service (NHS) costs and workforce.[14] Consequently, infection prevention is key and viewed as the most effective approach to reducing the likelihood of nosocomial infections.[15]

The UK National Risk Register highlighted the current cost of AMR as 2,000 lives and £95 million per year from the NHS budget (2020).[16] A recent report from the English Surveillance Programme for Antimicrobial Utilisation and Resistance (ESPAUR) described a 10.8% increase in hospital admissions in England for bacteraemia (bacteria reaching the blood stream) and fungemia (fungi reaching the blood stream) between 2017 and 2021, with an additional 2% mortality risk in patients with resistant organisms.[17] There was a significant increase (2.5%, p-value <0.001) in the rates of resistant Enterococcal infections in particular (e.g., gram positive bacteria commonly causing urinary tract infection, bacteraemia, and infective endocarditis, and less often meningitis and intra-abdominal infections) during this period.[17] The ESPAUR report also revealed that London had one of the highest rates of AMR per 100,000 population, followed by the North West.[17] Another ecological study conducted in the UK demonstrated an increasing trend of resistant *Escherichia coli* (*E.coli*) infections in England between 2013 to 2018, with an observed increase in the resistance rates of both community and hospital acquired bacteraemia, and called for action to help tackle this.[18] The authors also reported a significant increase (p-value <0.05) in the risk of resistance to at least one of eleven antibiotics among patients above the age of 65.[18] Another study also described how patients with resistant *E.coli* infection were more likely to have prolonged hospitalisation, resulting in an estimated annual cost increase of £420 per infected patient.[19]

The UK Health Security Agency published a report summarising data that had been submitted by 184 NHS hospitals and independent treatment centres for 115,796 surgical procedures. It identified 968 surgical site infections (SSI) across 17 surgical categories between April 2021 and March 2022.[20] During this particular year, there was also an increase in SSI rate in six

types of surgeries including hip replacement (up to 0.44%), knee replacement (0.35%), reduction of long bone fracture (0.54%), Coronary Artery Bypass Grafting (CABG) (2.7%), cardiac (non-CABG) (1.4%) and breast surgery (0.6%). Figure 2 below describes the rates of admission and readmission due to SSI in England between 2017 and 2022, the highest rates being for abdominal surgeries.[20]

Figure 2. Rates of admission and readmission due to surgical site infections in England between 2017 and 2022:[20]

Surgical category	Number of participating hospitals	Number of operations	Inpatient and readmission				Inpatient only		
			Number of SSIs	SSI incidence (%)	95% CI	Median time to infection (days)	Number SSIs	Incidence density (per 1,000 patient days)	95% CI
Abdominal hysterectomy	14	1,429	25	1.7	(1.1-2.6)	11	6	1.3	(0.5-2.9)
Bile duct, liver or pancreatic surgery	4	773	119	15.4	(12.9-18.1)	8	103	11.9	(9.7-14.4)
Breast surgery	33	14,451	108	0.7	(0.6-0.9)	20	5	0.5	(0.1-1.1)
Cardiac surgery (non-CABG)	13	17,693	220	1.2	(1.1-1.4)	15	153	0.8	(0.6-0.9)
Cholecystectomy	3	268	26	9.7	(6.4-13.9)	7	22	9.2	(5.8-13.9)
Coronary artery bypass graft	16	27,424	722	2.6	(2.4-2.8)	15	371	1.6	(1.4-1.7)
Cranial surgery	8	7,200	101	1.4	(1.1-1.7)	16	35	0.6	(0.4-0.8)
Gastric surgery	11	1,275	24	1.9	(1.2-2.8)	10	20	2.0	(1.2-3.0)
Hip replacement	184	175,167	817	0.5	(0.4-0.5)	21	218	0.3	(0.2-0.3)
Knee replacement	177	180,071	716	0.4	(0.4-0.4)	23	106	0.1	(0.1-0.2)
Large bowel surgery	35	10,844	930	8.6	(8.1-9.1)	8	798	7.4	(6.9-7.9)
Limb amputation	11	785	23	2.9	(1.9-4.4)	12	18	1.4	(0.8-2.2)
Reduction of long bone fracture	40	14,456	98	0.7	(0.6-0.8)	20	49	0.4	(0.3-0.5)
Repair of neck of femur	114	91,811	729	0.8	(0.7-0.9)	18	436	0.3	(0.3-0.3)
Small bowel surgery	17	2,322	172	7.4	(6.4-8.5)	7	149	5.3	(4.5-6.2)
Spinal surgery	25	30,586	386	1.3	(1.1-1.4)	15	130	0.8	(0.6-0.9)
Vascular surgery	11	3,971	87	2.2	(1.8-2.7)	14	39	1.3	(0.9-1.8)

This surveillance report also showed an increasing trend in the incidence of hospital admission and readmission (2020/2021 and 2021/2022) with resistant micro-organisms like Methicillin Resistant Staphylococcus Aureus (MRSA), Methicillin Sensitive Staphylococcus Aureus (MSSA), Staphylococcus Aureus, and Enterococcus in *superficial* infections (defined as infections in the outer layers of the skin). See Figure 3 for further details. There was also an increasing trend in admission and readmission with *resistant* micro-organisms like MRSA and MSSA, and Staphylococcus Aureus for *deep, organ and space* infections (defined as infections deep into the muscles and tissues surrounding them). See figure 4 for further details.[20]

Figure 3. Admission and Readmission trends of surgical site infections in England from 2012 to 2022 by micro-organism for superficial infections of all surgical types:[20]

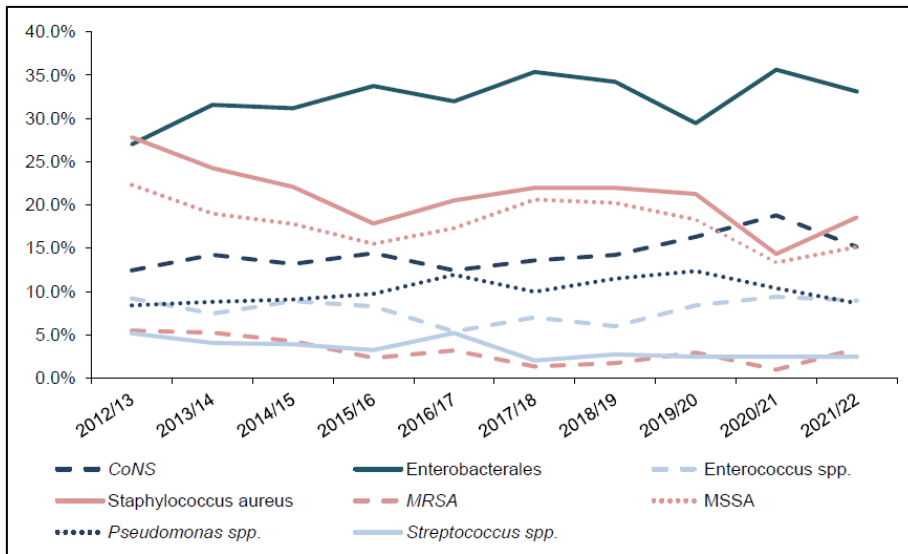
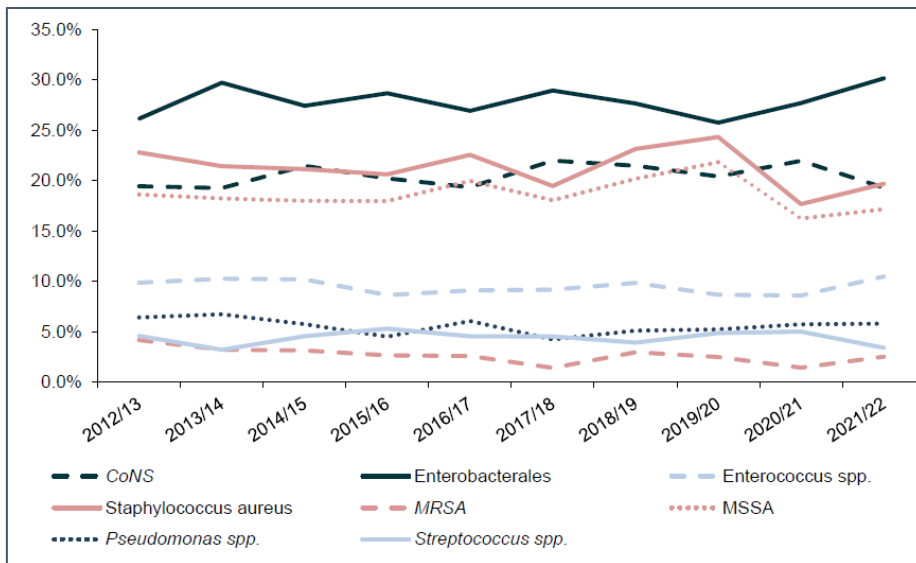


Figure 4. Admission and Readmission trends of surgical site infections in England from 2012 to 2022 by micro-organism for deep, organ or space infections of all surgical types:[20]



1.3.2. Resistant infections and sepsis

A serious complication of a resistant infection can be Sepsis. Sepsis is defined as life threatening syndrome that results as a response of the immune system to infection disseminating through blood stream to other organs leading to multiorgan dysfunction.[21] Sepsis is a rapidly progressive condition; patients can deteriorate within a few hours and require intensive treatment. In Europe, approximately 10% of hospitalised patients develop sepsis, and have a mortality rate of 38%.[22] In the UK, sepsis is reported to cause more deaths than bowel, breast, and prostate cancer combined, with an estimated 48,000 patients affected annually.[23] These

numbers were recorded pre the Coronavirus (COVID-19) pandemic, but are likely to have been higher during the pandemic.[23] Financially, sepsis costs the UK economy around £16 billion annually.[24] It is estimated that sepsis prevention measures could save the NHS, more specifically, approximately £170 million annually; a saving of more than £1 million for an average-sized hospital.[25]

1.3.3. Risk factors for Developing Infection and Sepsis

Some patients are at higher risk than others of developing infection and subsequent sepsis. They may have chronic conditions such as cardiovascular, renal, endocrine, obesity, autoimmune diseases, and/or cancer, which increases their vulnerability to resistant infections and their risk of hospitalisation.[26] For example, obese patients are at higher risk of infection as obesity is linked to altered immune responses and chronic inflammation that can increase risk of infection.[27] The elderly and infants are both at higher risk of sepsis due to their weakened or vulnerable immune systems.[28] Immunocompromised patients are also at higher risk due to their immune systems being suppressed.[21]

1.4.Tackling AMR – the UK’s Five-year National Action Plan

In addition to the funding for drug discovery, the UK government implemented a 5-year action plan (2019-2024) to address AMR.[29] This plan focused on reducing unnecessary antibiotic use, optimising antibiotic use, and developing innovative methods to tackle AMR. From a healthcare perspective, the government is aiming for a 15% overall reduction in the national use of antibiotics by 2024, with a targeted 25% antibiotic use reduction in primary care, and a 10% reduction in hospitals.[29] The term ‘antimicrobial stewardship’ was used to refer to an evidence-based strategy that promotes the appropriate and optimal use of antimicrobials in prevention, diagnosis and treatment of infections. As part of this 5-year action plan, the NHS has increased the deployment of clinical pharmacists for better implementation of antimicrobial stewardship programmes in care homes, primary care and hospitals.[29] The NHS is also putting a strong emphasis on the documentation and use of data, including prescribing data on the electronic systems to track compliance and antibiotics prescribing rate, and assess the impact of changes in antibiotic prescribing.[29]

1.5.The Use of Artificial Intelligence to Inform Healthcare Decisions

Learning from patterns in healthcare data can inform strategies to prevent infections in the future. One powerful tool to learn from the data and inform clinical decisions around disease prevention is artificial intelligence (AI). Artificial intelligence is defined as the ability of

machines (i.e. computers) to perform functions that normally human can do, in a more efficient, accurate, and faster manner.[30] Machines learn from the data using machine learning, a sub-speciality of AI that involves developing correlations and patterns between different variable(s) and predicted outputs, producing a machine learning model that can predict a particular output in a future, unseen dataset.[31] Radiology has seen the highest levels of AI advancement compared to other fields of medicine. Artificial intelligence has played a significant role in improving the quality of radiographic scans, providing accurate and precise measures for tumours, and guiding the early detection of cancer.[32] One example is the development of AI-assisted tools to improve the detection of breast cancer lesions in some cases where the sensitivity of conventional mammograms is limited.[33] Robotic surgery is another area that has evolved, where machines, under the control and supervision of human surgeons, can perform specific surgical/interventional techniques with high accuracy and minimal invasiveness compared to traditional techniques, reducing intervention-related complications such as infection, prolonged healing, and post-operative pain.[34] Artificial intelligence can also support clinical decision making through different types of clinical decision support systems (CDSS), such as AI-predictive models that provide clinicians with recommendations that can reduce the occurrence of a particular condition or a complication based on an estimated risk.[35] Improving the clinical utility of AI-predictive models will involve taking into consideration how the model is developed and can support shared decision making (SDM) process, including the involvement of patients in their healthcare decisions.

1.6.AI-predictive Models Supporting SDM

Shared decision making is defined as a “collaborative process” that involves both patient and clinician to reach a decision related to patient’s health including choice of treatment or consenting to carry out a procedure or an investigation, with the support of clinicians.[36] The National Institute for health and Care Excellence (NICE) have endorsed the benefits of SDM, which involves presenting patients with evidence-based information, giving them the opportunity to discuss this information with their healthcare providers, and helping them to reach a decision that is best for them. This is particularly important when considering different treatment options (surgical vs non-surgical) and obtaining consent. Patients need to understand information that is “*individualised*” to them and “*accurate*” to make informed decisions.[37] Shared-decision making support tools can come in different forms, some of which provide predictions (i.e. risk prediction tools) and, in this case, can be developed by either statistical approaches or AI-based techniques. The use of AI-predictive models versus traditional risk calculators has been discussed in a recent review on the use of AI-predictive models in surgical

practice.[38] The review stated that traditional risk scores for predicting post-operative morbidity such as the American College of Surgeons Surgical Risk Calculator (ACS-SRC), the American Society of Anaesthesiologists (ASA) score, and the Physiologic and Operative Severity Score for the Enumeration of Mortality and Morbidity (POSSUM) are all developed using Cox proportional hazards regression analysis or logistic regression models.[38] These models are limited by the risk of over or underestimation of risk, as they usually depend on variables that have statistically significant effects on outcomes.[38] This means that some important factors that can clinically influence the outcome may be excluded.[38] Furthermore, conventional models often hypothesise a linear correlation between the variables and the outcome of interest, without considering more complex interactions.[38] Additionally, limitations with regression analysis including overfitting and multicollinearity can hinder the assessment of large number of variables, and as a result, current models are usually restricted to relatively small number of variables.[39,40]

Artificial intelligence predictive models usually demonstrate better predictive performance compared to conventional logistic regression statistical models.[38,39,41-43] This improvement could be attributed to the ability of machine learning to evaluate a larger number of variables compared to statistical models, which helps the identification of weak predictors and interactions among predictors thereby enhancing prediction accuracy.[39] Developing non-linear models using multiple data sources, for example, diagnosis, treatment, and laboratory investigations, has shown higher performance compared to statistical logistic regression for predicting post-operative outcomes.[44-47] This was demonstrated in a recent AI-predictive model for post-operative complications following mastectomy and abdominal wall reconstruction surgeries, where the predictive performance was compared between a model developed using machine learning and another one using multivariable logistic regression.[42,43] In these examples, machine learning identified more predictors for post-operative complications, and exhibited higher predictive performance than multivariable logistic regression.[42,43] Moreover, machine learning possesses the capability for incremental learning, allowing it to continuously improve with the addition of new data.[48,49] This has led to the description of statistically developed predictive models as “static”, while machine learning models are “dynamic and continuously improve over time”.[38]

The difference between using AI techniques (i.e. machine learning) and statistical techniques (i.e. regression methods) have also been argued in a paper published by Bazdoc *et al.* in *Nature*. [50] The authors stated that “Statistics draws population inferences from a sample, and machine learning finds generalisable predictive patterns”.[50] This inference means that a

statistical approach can be used to test a particular hypothesis, so that it can find correlations between variables that are statistically true, not based on confounders.[50] Machine learning is more focused on predictions of an output (i.e., the likelihood of a disease to occur), through learning from a dataset with the output, so that it can be predicted in a new dataset. Additionally, machine learning can be trained on dataset even without an output available (i.e. unsupervised learning). These factors make machine learning perform better when dealing with “wide data” with a large number of variables to develop prediction.[50] For example, an early model was developed in the Pentagon in the United States (US) to predict cardiac complications (cardiac death and acute non-fatal myocardial infarction before hospital discharge) after non-cardiac surgery.[51] Both statistical logistic regression and machine learning (neural network) were used to develop the model and the performance of machine learning algorithm (neural network) was compared to statistical logistic regression in both models. The neural network model demonstrated significantly (p-value <0.01) higher specificity (increasing from 82% to 96%) compared to statistical logistic regression, and a higher positive predictive value with neural network (50%) compared to statistical logistic regression (18%). The authors concluded that AI techniques could be a “*useful alternative to conventional statistical logistic regression*”.[51]

1.7.The Use of AI-predictive Models in Surgical Practice

Artificial intelligence predictive models have been used in surgical practice to predict the likelihood of surgical complications. An AI-predictive model has been developed in lung resection surgery to predict the risk of cardiopulmonary complications.[52] Through 1360 patients and 23.3% exposure to cardiopulmonary complications, the model demonstrated high predictive performance, with an Area Under the Curve (AUC) of 0.75.[52] Another study developed an AI-predictive model for the risk of post-operative SSI in posterior spinal fusions surgery through 3034 patients in the training dataset, demonstrating an AUC of 0.775.[53]

As mentioned, surgical patients are at an increased risk of infection and subsequent sepsis. AI-predictive tools can be developed to predict the likelihood of infections based on an individual's specific risk factors, prior to surgery.[34] By leveraging these AI tools, clinicians can make informed decisions about whether to postpone the surgery or make changes to the management of these patients prior to surgery, such as tighter control of their blood pressure or blood glucose levels.[35] Numerous studies have been conducted to develop AI tools that can predict the likelihood of sepsis. However, the majority of these studies have focused either on assessing the lifetime risk of sepsis in community patients,[54] or on evaluating the risk of sepsis in medical or critically ill patients.[55-59]

In surgical setting, Sood et al. study that developed an AI model to predict sepsis within the first 30 days following major elective oncology surgery.[60] The study included 2,954 patients who experienced post-operative sepsis among nine of the most common types of oncology surgeries. The model demonstrated a satisfactory performance, with an (AUC) of 0.717 for sepsis prediction. However, the model had several limitations. Firstly, it did not consider socioeconomic factors that could have influenced the risk of infection and subsequent sepsis, nor did it incorporate cancer grades of the included patients, which questions how the diagnosis of cancer was defined in the database, and also challenges the correlation between the clinical variables over different cancer grades severities. Secondly, the model was only trained on cancer patients, making it difficult to generalise the algorithm to other surgical patients.[60] A second model was developed to predict post-operative sepsis in any type of elective surgery, utilising the Medical Information Mart for Intensive Care-III (MIMIC III) database, which comprised of over 40,000 critically ill patients from Beth Israel Deaconess Medical Centre (BIDMC) in Boston, Massachusetts, between 2001 and 2012.[61] This model also had limitations including treating all elective surgical procedures similarly in terms of the risk of infection and sepsis, while the risk cannot be generalised to all elective procedures. Although the study demonstrated good performance (AUC = 0.835), the model was developed using the XGboost model, which is not commonly employed in medical applications and has been associated with risk of overfitting.[61]

In the case of appendectomy, an AI-predictive model for post-operative sepsis was developed following appendectomy elective surgery using the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) database.[62] The model exhibited favourable performance with an AUC of 0.72. However, the authors did not specify which clinical definition of infection or sepsis was used. This is really important as it can potentially impact the accuracy of the developed model. Additionally, the incidence of post-operative sepsis in appendectomy is relatively low (1.2%),[63] raising concerns about the clinical utility of the developed model.

It should be noted in all these models, that the focus was always on elective surgery rather than emergency surgery. Emergency surgery is defined as an invasive procedure that is done within 24 hours of hospitalisation to manage an acute threat to life, such as trauma and fractures surgery.[64] This acute condition puts the body systems in stress, especially central organs (i.e. heart, lungs, kidneys, liver) which increases the risk of the patient to be in a hemodynamically unstable condition. Furthermore, with emergency surgery, it is more common to have a dirty or contaminated class of wound incisions, also it is more likely to have prolonged operative

duration and having a high American Society of Anaesthesiologists (ASA) score (a scoring system used preoperatively to predict surgical risk).[65] All these factors contributed to higher risk of post-operative infection following an emergency surgery compared to elective surgery.

1.8.Chapter Summary

This Chapter presented an overview of the global threat of AMR and the contributing factors to the progression of this threat. It also described how AMR could negatively impact the NHS in terms of monetary cost and mortality, in addition to the national plans to tackle AMR in the UK. This Chapter covered the different factors that would increase an individual's risk of infection and subsequent sepsis, particularly with hospitalisation, and how the risk is further increased with surgical procedures. It highlighted how AI-predictive models could help support SDM around infection prediction, which can subsequently reduce the incidence of life-threatening sepsis and indirectly reduce the burden of AMR.

This PhD programme of work attempts to address these knowledge gaps by providing a deeper understanding of how AI-predictive models can guide SDM, through developing an AI-predictive model to predict the likelihood of post-operative infection in elective surgery setting and exploring the users perceptions on using AI decision support tools more broadly to inform SDM. The next chapter of the thesis describes the aims and objectives of this PhD programme of work.

Chapter 2. Aim and Objectives of the PhD programme of Work.

In Chapter One, the researcher discussed the burden of AMR globally and the factors contributing to AMR, and how it impacts on the NHS. The researcher also explained how infection prevention needs to be prioritised. Artificial Intelligence can potentially optimise the use of healthcare data to inform SDM around infection prevention, with special emphasis on surgical practice due to higher risk of infection. Here, in this chapter, the researcher defines the aim and objectives, and the research methodology used in this study. This programme of work encompasses a number of different stages, the methods for which are described in each chapter.

2.1. Aim

To develop and internally validate an AI model to predict the likelihood of post-operative infection in surgical patients, while also exploring clinicians' and patients' perceptions on using AI-based decision support tools more broadly to inform SDM.

2.1.2. Objectives

- 1- To describe the key stages in developing and evaluating a predictive AI model, recognising the challenges that clinicians and/or researchers might face at each stage and provide practical tips to help manage them. (Chapter Three)
- 2- To conduct a systematic review of the literature to identify the optimal set of predictors used to train machine learning algorithms to predict the likelihood of an infection and subsequent sepsis. (Chapter Four)
- 3- To obtain the insights of patients and members of the public on developing an AI-predictive model to predict the likelihood of post-operative infections in elective surgery. (Chapter Five)
- 4- To develop and internally validate a machine learning model to predict the likelihood of post-operative infection in elective surgical patients to support SDM. (Chapter Five)
- 5- To identify the different types of bias that developers need to consider when developing and deploying an AI model. (Chapter Six)
- 6- To understand clinicians' and patients' perspectives on using AI decision aids to inform SDM at different stages of the patient journey and in various clinical settings. (Chapter Seven)

- 7- To explore the perceptions of clinicians from one English NHS Trust regarding the use of AI-based tools to inform SDM in the clinical environment. (Chapter Eight)

2.2. Methodological Approach

A mixed-method approach was used to meet the objectives set out above. This involved a descriptive piece detailing the key stages of AI model development, two systematic literature reviews, one quantitative study to develop and internally validate an AI-predictive model, and a qualitative study to explore healthcare professionals' perceptions about using AI tools more broadly in the clinical environment. The researcher was interested in predicting the likelihood of post-operative infections in elective surgical setting through the use of AI-predictive models in preoperative assessment stage. The researcher was also focused on exploring the key factors that can improve the clinical utility of these models and how to improve the users' (clinicians and patients) experience of using these models to better inform the SDM, thus promoting personalised healthcare as outlined by NICE.

Chapter 3. The Steps Involved in Developing an AI-Predictive Model to Inform Shared Decision Making.

3.1. Introduction

In order to develop and internally validate an AI model to predict the likelihood of post-operative infection in surgical patients, it is important to understand the key stages involved and also the challenges that researchers might face at each stage. In this Chapter, we will detail these key stages, which was informed from previous predictive models development and the researcher own experiences.

This descriptive piece has been published in an international peer-reviewed open access journal: Hassan N, Slight R, Morgan G, Bates DW, Gallier S, Sapey E, Slight S. *Road map for clinicians to develop and evaluate AI predictive models to inform clinical decision making*. BMJ Health & Care Informatics. 2023;30(1), doi:10.1114/ bmjhci-2023-100784 (*Appendix A*). The findings from this chapter have also been virtually presented at an international research conference: CiiEM Summer Research Summit in Portugal.

3.2. Background

Healthcare systems worldwide generate enormous amounts of patient-related health data, much of which is electronic in developed countries. There is growing interest amongst clinicians and healthcare staff in how they could use these data to support patient care.[66] Much of medicine is about anticipating and reducing risk, based on current and historical experiences. Predictive analytics in healthcare can help determine the risk of a patient developing a particular condition or complication, which can inform the SDM process between clinicians and patients and improve patient satisfaction with their overall medical care.[67-72] With the new era of AI, clinical prediction tools can help personalise treatment and management decisions.

The Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) framework was published to guide developing multivariate predictive models,[73] outlining what should be reported (e.g., data sources, modelling techniques) when written up for publication.[74] However, a recent systematic review highlighted how the reporting of these models has still been rather poor since its publication.[75] TRIPOD also only focused on regression-based prediction models (although it can be applied to AI-generated approaches) and highlighted the need for more “*practical methods*” for developing models more commonly used in healthcare (i.e., supervised learning techniques).[76] The Consolidated Standards of Reporting Trials (CONSORT-AI) guidelines were published in 2020 to help

readers conceive studies with AI interventions; however, there was limited guidance on how these AI-predictive models could be developed and usefully applied in clinical practice;[77] clinicians have sought further information on this.[66,78] Even if a newly developed AI model has a good predictive performance, this does not guarantee that it will be used in clinical practice or enhance clinical decision making, let alone bring about improvements in health outcomes.[79] The quality criteria important for evaluating AI-predictive models were described in a recent scoping review; however, little information was provided on how such tools affect the clinical routine of physicians, which may vary per physician.[80]

The nine stages for developing and evaluating predictive AI-models:

1. Clarifying the clinical question or outcome(s) of interest (output).
2. Identifying appropriate predictors (features selection).
3. Choosing relevant datasets.
4. Developing the AI-predictive model.
5. Validating and testing the developed model.
6. Presenting and interpreting the model prediction(s).
7. Licensing the AI-predictive model.
8. Maintaining the AI-predictive model.
9. On-going evaluation of the impact of the AI-predictive model.

It is vital to seek the input of a multidisciplinary team early when developing AI-predictive models. This includes clinical specialists when deciding on how the model could potentially enhance clinical decision making, and computing scientists when selecting the most appropriate algorithm(s).[81] Patients and providers should also be involved in the decision making if the recommendations will be presented to them, including what, how and when information might be usefully presented (i.e., content and alerts).[68,72,82] Taking each of these stages in turn:

3.2.1. Clarifying the clinical question or outcome(s) of interest (output)

The clinical question or outcome(s) of interest should be clearly defined from the onset. An example of a clinical question might be ‘*what is the likelihood of a patient developing type 2 diabetes mellitus (T2DM)?*’ with a view to modifying some of the patient’s potential risk factors through lifestyle changes and/or prescribing medication.[83] It is important to consider how we define T2DM here. Kopitar et al. defined it as the presence of a fasting plasma glucose level of 6.1 mmol/L or higher, without having diabetes symptoms.[83] This definition makes the model

a *prognostic* rather than *diagnostic* predictive model, given that it focuses on predicting a future health outcome. It is worth mentioning that this definition varies from those presented in different clinical guidelines,[83] and can also change over time, highlighting the importance of model upgrading and maintenance. Another example of a clinical question could be ‘*what is the likelihood of a patient developing an infection and subsequent sepsis as an inpatient?*’. Again, multiple definitions of sepsis could be used,[84-86] each varying in how closely aligned it is with the systemic effects of sepsis syndrome (see Figure 5).[84,85] However, the choice of definition here is critical as it can directly influence the model performance measures, particularly specificity, which I will discuss later.[87]

Figure 5. Different definitions of sepsis and their related clinical predictors:

Sepsis-1 definition (1991)*	Sepsis-2 definition (2001)**	Sepsis-3 definition (2016)***
≥ two systemic inflammatory response syndrome (SIRS) criteria: • fever >38.0°C, • hypothermia <36.0°C, • tachycardia >90 beats/min, • tachypnoea >20 breaths/min, • leucocytosis >12*10⁹/l or leucopenia <4*10⁹/l.	≥ two SIRS criteria <i>plus</i> a proven or suspected infection (as per the specific type of infection e.g., symptoms and CSF analysis (in the case of meningitis)) <i>plus</i> persistent hypotension (mean arterial pressure <60 mm Hg, systolic blood pressure <90 mm Hg, or a decrease in systolic blood pressure of ≥40 mm Hg) despite adequate fluid resuscitation.	≥ two points on Sequential Organ Failure Assessment (SOFA) score, which includes dysfunction of the following systems: • respiratory, • cardiovascular, • hepatic, • renal, • coagulation, • neurological.

* Note that SIRS criteria are non-specific to the type of infection.
 ** Note that suspected infection became a requirement to define sepsis.
 *** Note that clinical parameters are more specific to the systemic mechanism of sepsis.

3.2.2. Identifying appropriate predictors (feature selection)

The second step involves identifying appropriate clinical variables or predictors (features) related to the outcome of interest. Thus, if we take our sepsis-3 definition (Figure 5), for example, the next question relates to “*what clinical variables should we use for predicting an infection and subsequent sepsis?*”. These clinical predictors again will depend on whether you want to develop a *prognostic* predictive model (which predicts the likelihood of sepsis occurring in a patient before the systemic inflammation process begins),[23] or a *diagnostic* predictive model (which detects the likelihood of sepsis early in a patient but after the inflammation process has already begun).[58] A review of the medical literature can help identify potential predictors that might be worth considering; 194 clinical predictors have been previously used to train machine learning algorithms for sepsis prediction, 13 of which were used across all 17 newly developed algorithms.[87] These thirteen predictors contained a blend of non-modifiable (e.g., age, gender) and modifiable (e.g., blood glucose levels, blood pressure) predictors, the latter potentially increasing the applicability of the AI-predictive model in clinical practice.[87] It is important to consider here how these predictors have been defined

and selected in previous studies, their source (i.e., retrospective or real-time data), and whether any were excluded, thus recognising any inherent bias.[26,79] In terms of predictor type, numerical predictors should be given preference over categorical predictors, whenever possible.[73,88-90] A classic example is blood pressure, which can be recorded as a numerical (e.g., 108 mmHg) or categorical (e.g., high, normal, low) value. The latter assumes that a patient with systolic blood pressure 108 mmHg has the same level of hypotension as another patient with systolic blood pressure below 90 mmHg, which is more characteristic of sepsis. In the T2DM example mentioned above, Kopitar et al. screened the electronic health records (EHRs) of patients who went on to develop T2DM to identify potentially modifiable (e.g., total cholesterol) and non-modifiable (e.g., age) predictors.[83] EHR data can also provide an opportunity to explore variables with predictive potentials that might not have been previously considered. These EHR data could be particularly valuable when the model under development is intended to be used in the same clinical environment, as it is likely to increase applicability and reduce the risks of over/underfitting, discussed later.[83]

After selecting potential clinical predictors, these are then correlated to the model's outcome of interest (output) using either statistical methods or machine learning techniques.[91] Some predictors are likely to have a strong correlation to the output but may be more suitable for a *diagnostic* rather than a *prognostic* predictive model. For example, the SOFA score, which reflects multi-organ dysfunction, will have a strong correlation with the diagnosis of sepsis and would be more suitable for developing a *diagnostic* predictive model, whereas lipid profile will have a strong correlation to the prognosis of diabetes and would be more suitable for developing a *prognostic* predictive model; this is because patients with established diabetes are likely to have hypercholesterolemia.[83] the researcher suggested using a '*blended approach*' for predictor selection, where the predictors are correlated to the model's output, and clinical input is also obtained on the choice to support its clinical application.[84,87,92]

3.2.3. Choosing relevant datasets

The existence, choice, and access to relevant datasets often represent a limiting step for developing predictive AI models.[66,93] Health datasets in the UK are held by thousands of different organisations and so it can be difficult for clinicians, researchers and innovators to discover what datasets already exists.[94-96] Developers should first look at the relevance, data size and diversity of potential datasets; the proposed dataset should ideally have a good representation of the targeted population where the AI model is intended to be used to reduce the risk of inherent bias.[97] If the key outcome(s) of interest is unidentified, developers may have to decide how these available variables are used to define the key outcome. Researchers

and innovators can search and request access to UK health-related datasets through ‘the Gateway’, a common entry point established by the Health Research Authority (HRA) for nine UK-based health data research (HDR) hubs across the country.[95] These hubs include DATAMIND (mental health data), PIONEER (acute care data) and Discover-Now (primary care data), the latter being one of the largest primary care datasets in Europe. The UK HDR Alliance is also an independent alliance of leading healthcare and research organisations united to establish best practice for the ethical use of UK health data for research at scale.[96] In the UK, patients' information is protected by the General data protection regulation (GDPR) and patients can refuse to give their permission for their confidential data to be used through the national data opt-out (NDOO) service. De-identification can be challenging, specifically with demographic variables, some of which can be important predictors when training the model. Removing them can potentially risk the efficiency of the model performance. A trusted research environment (TRE) with anonymised patient data can be prepared for the clinician or researcher, once all the necessary ethical approvals have been obtained and the required training on data use and security completed.[98-101] Alternatively, data can be processed in a safe environment either at a hospital or University site; however, checks will need to be made on the safety of these environments and these data not approved for release if they do not meet the HDR UK 5 Safes (safe people, safe projects, safe settings, safe outputs and safe data).[96] The diabetes risk prediction model mentioned above was developed using anonymised data collected from ten diabetes screening clinics that were then pooled in a single database.[83] Internationally, the MIMIC database has clinical information from more than 40,000 patients admitted to critical care units at one tertiary centre (BIDMC, USA). Similarly, healthcare professionals can freely access the dataset after completing appropriate data use and security training and signing a data usage agreement.[98,102] An important consideration is how these data have been collected and recorded. Numerical variables in the chosen dataset should ideally be collected and recorded in a *synchronised* way.[99] The MIMIC database developers recognised this as a potential limitation of their dataset, with vital signs like heart rate and blood pressure recorded at different time points, thus potentially impacting the accuracy of the model.[98]

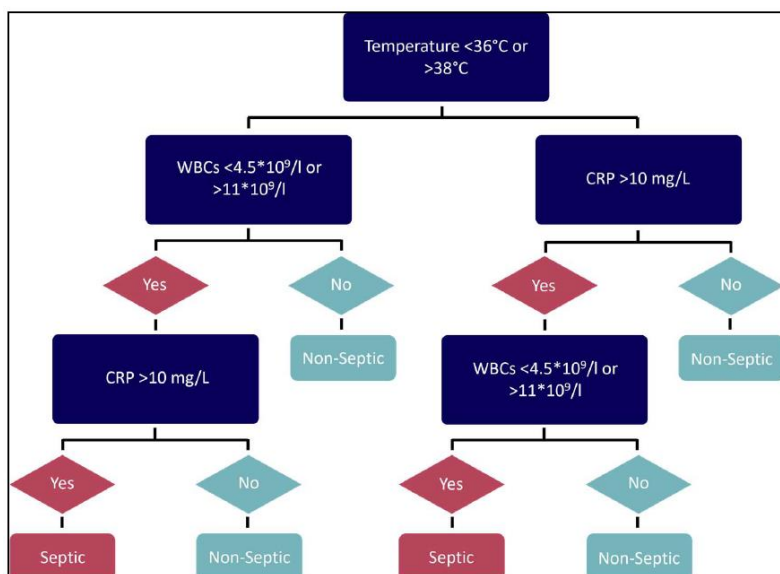
3.2.4. *Developing the AI-predictive model*

There are four major types of machine learning algorithms: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.[103] The choice of machine learning algorithm will depend on some factors, including the outcome of interest (i.e., numerical or discrete value); the number of predictors; the ‘shape’ of the dataset (i.e., size,

completeness, uniformity); and the performance measures of the algorithm (i.e., sensitivity, specificity, accuracy, AUC).[92] In the case of the latter, a number of algorithms may need to be tried first before finally deciding on the most suitable one or combination (ensemble model).[103] Supervised learning is commonly used for *predictive* models and can be sub-classified into regression (i.e., numerical output) or classification (i.e., discrete output) algorithms.[104] The higher the number of predictors used, the more computational power needed to train the model and the higher the potential risk of overfitting.[104] An overfitting model is a model that has high accuracy during the training phase, but lower accuracy during the validation and testing phase; potential ways to overcome this are described below. [89,104,105] It is important to bear in mind, however, that strong computational correlations are primarily dependent on the entry values (e.g., non-extreme vs extreme) and amount of missing data. Missing data can be potentially managed by statistical methods (i.e., multiple imputations) or machine learning algorithms (i.e., random forest classifier), the choice of which will usually depend on the type and extent of missing information.[50,103]

Random forest represents an example of a supervised machine learning algorithm.[106] Decision trees are the building blocks of random forest models. Figure 6 illustrates a very simple decision tree that uses a person's temperature (Celsius), white blood cell (WBC) count and C-reactive protein (CRP) to predict whether a patient is likely to develop sepsis or not. The decision around which predictor should be on the top of the decision tree (root node) depends on how well each of the individual predictors do at separating patients with or without sepsis. The gini impurity is a mathematical calculation that can help inform this decision.

Figure 6. An illustration of a proposed two-level decision tree for sepsis prediction:



The random forest technique involves developing a predictive AI model using a large number of decision trees (64 to 126 trees);[106] the higher the number of decision trees, the lower the risk of overfitting and the higher the predictive accuracy.[103,107] However, random forests with a very high number of decision trees (>118) can potentially increase the risk of overfitting and reduced performance.[106] The Gradient boosting algorithm can help overcome this overfitting problem by sequentially adding decision trees to the algorithm, creating a complex ensemble of decision trees.[108] Deep learning and artificial neural networks can provide higher performance levels than other conventional machine learning techniques. These networks act as a net of neurons that can identify patterns and correlations in a dataset so that the model can self-learn from these patterns. The “deep” refers to the depth of layers in a neural network, and the performance measures of a deep learning model are directly correlated to the data size (i.e., the larger the dataset, the better the model performance). However, this can be challenging with rare diseases.[104,107,109]

Python is one of the most common programming languages for developing AI-predictive models and is freely available.[110] It contains hundreds of codes for different algorithms and computational functions. After importing the dataset into a programming software, you usually divide it into two portions, one for training the algorithm (70%) and the other for internal validation (30%).[103,105] As described in Stage 2 above, each predictor is then correlated to the outcome of interest (feature selection) using the training set and the performance measures of the algorithm calculated. This includes the specificity, sensitivity, receiver operating characteristic curve (ROC) and the area under the receiver operating characteristic curve (AUROC curve). The AUROC curve measures the distinctive ability of the algorithm to predict the outcome, with a value of >0.9 considered excellent.[87,111] AI systems learn to make decisions based on these training data, which may reflect human biases or social inequities, even if predictors such as race or gender have been removed.[112] It is beneficial to have the input of a programming specialist when preparing/revising the codes and judging the performance measures of any resulting models.

3.2.5. *Validating and testing the AI-predictive model*

After developing the model, its predictive accuracy is usually re-assessed using a validation dataset (internal validation) and again in a completely new, unseen dataset (i.e., externally validated), ideally from another site. This comparison of performance measures is important for evaluating the risk of over/underfitting and widening the generalisability of the model, considering the diversity and representation of the patient population.[113] The testing phase usually involves running the model in a silent clinical environment, where the output is not

shared with clinicians but compared to conventional clinical judgement and diagnosis. The T2DM prediction model was tested in a silent clinical environment over six months to assess its performance, before ‘going live’ to support clinical decision making.[83] It is important to recognise that not all data are equal in terms of quality; laboratory values may be coded differently or missing for some or all of an entire predictor in validation and training datasets. Complete case analysis is a method that can be used to handle missing data and involves removing all missing patient cases; however, this requires having a large sample size and may introduce selection bias. Alternatively, mean imputation can be used for missing numerical predictors, but can be sensitive to outliers (i.e., extreme values).[114]

3.2.6. *Presenting and interpreting the model prediction(s)*

How the model prediction(s) is presented to target users (patients/clinicians) and whether it is accompanied by a recommendation are important considerations. The predicted probability (output) can be presented to users without any corresponding recommendations; this *assistive* presentation format gives clinicians the flexibility to combine these predictions with clinical judgment.[115,116] In contrast, a *directive* prediction model provides the physician with a recommendation in addition to the predicted probability; this, in turn, can potentially increase the ease of use of the AI prediction model, especially if integrated into the electronic ordering system.[117] Providing confidence intervals around the output can also be useful.[118] Clinicians should be informed of the underlying assumptions of the model, including which predictors were included and why, any inherent bias (e.g., if groups are over- or under-represented in the training data), and how patients with specific outcome risk profiles might be affected by different recommendations.[79] For example, the inclusion of health costs as a proxy for health needs could potentially introduce racial bias, as less money is spent on Black patients who have the same level of need in the US; in other words the algorithm could falsely conclude that Black patients are healthier than equally sick White patients.[119] There is some evidence that clinicians in English-speaking countries have felt more legally supported when using decision support tools because they can provide documented evidence for the rationale behind their decisions.[120] Chua et al. proposed an AI-human interface, where clinicians identify which patients might be eligible to use the tool, and the algorithm identifies (more accurately) which patients have serious illness communication needs and promotes upstream data collection.[72]

3.2.7. *Licensing the AI-predictive model*

In the UK, AI-based tools are classified as medical devices and therefore need approval by the Medicines and Healthcare products Regulatory Agency (MHRA). Before Brexit, approved

tools required either the "United Kingdom Conformity Assessed" (UKCA) or "*Conformité Européenne*" (CE) logo to be marketed in Europe.[121] However, from July 2023, only tools with the UKCA logo will be allowed to be marketed in the UK.[122] In Europe, AI-based software and tools are regulated by the EU Medical Device Regulation (EU MDR),[93,123,124] whereas in the US, AI-based tools are regulated by the Food and Drug Administration.[125]

To licence an AI-predictive tool in the UK, the MHRA must ensure that it complies with certain 'conformity assessment' standards, which were described by NICE in 2018 and updated in 2021.[126] It is worth mentioning that the NICE framework is designed for AI tools with fixed algorithms (i.e., no change over time) rather than AI tools with adaptive algorithms (i.e., continually and automatically change);[126] the latter are covered by separate standards (including principle 7 of the code of conduct for data-driven health and care technology).[126] Higher-risk AI tools are classified as those that either target vulnerable patient populations, have serious consequences with errors or system failure, are used solely by patients themselves without support from healthcare professionals, or require a change in clinical workflow.[126] For EU approved tools, the tool should be compliant with the general safety and performance requirements stated by the EU MDR.[127,128]

3.2.8. *Maintaining the AI-predictive model*

Maintenance of the AI-predictive model and knowledge management are critical.[129] It may be necessary to update the AI-predictive model as populations, diseases, and treatments change, and include an expiry date.[129] In the UK, the NICE data framework recommends a regression test be done when the model is updated to ensure that any new changes do not have a negative impact on its performance, reliability and functionality.[126] Model developers should also keep users (both clinicians and patients) informed when releasing new versions of the model. In the US, model recertification is needed when AI-predictive models are updated,[106] although the US FDA is currently working on a framework that allows repeated updating of an AI-predictive model without repeated recertification through a change control plan.[130]

3.2.9. *On-going evaluation of the impact of the AI-predictive model.*

The introduction of an AI prediction model into clinical practice can be considered a complex intervention; it usually consists of multiple interacting components including the accuracy of the model predictions, physician and patient understanding and use of these probabilities, expected effectiveness of subsequent actions or interventions, and adherence to these. A new framework has now replaced the UK Medical Research Council's (MRC) guidance for developing and evaluating complex interventions. It focuses on recent developments in methods and the need to optimise the efficiency, use, and impact of research.[131] The

downstream effects on patient outcomes of using an AI prediction model are not always predictable. For example, Kappen et al. described how there was no decrease in the incidence of post-operative nausea and vomiting, despite an increase in the administration of prophylactic antiemetics in the cluster-randomised trial of the AI prediction model (using an *assistive* presentation format).[117] This may indicate that either the predictive performance of the model was insufficient, the impact on physician decision making was too small (e.g., too few prophylactic drugs were administered despite high predicted probabilities), the anti-emetic drugs were not as effective as thought, and/or patients chose not to take them.[117] The collection of additional data (observations and interviews) may be helpful in improving our understanding of these study results.

When designing an impact study before applying to licensing, one needs to also consider whether the complex intervention will have an individual effect on patients or whether it induces a more group-like effect.[117] A prediction model often aims to affect the clinical routine of a physician, which may vary per physician; this could lead to clustering of the effect per physician or per practice (hospital) when the AI model use is compared across providers or practices.[84,93,117] After repeated exposure to the predictions, physicians may also become better at estimating the probability in subsequent similar patients, even when those patients are part of the control group.[84,93,117] This likely dilutes the effectiveness and thus impact of the model use.[109,117] As Kappen et al. highlights, the effects of a learning curve may be minimised, though not completely prevented, by randomisation at a cluster level, e.g. physicians or hospitals.[113,117]

The researcher has provided a road map which clinicians and others developing algorithms can use to develop and evaluate AI-predictive models to inform clinical decision making. The researcher described the nine stages, recognising the challenges that clinicians might face at each stage and practical tips to manage them. A '*blended approach*' should be considered for clinical predictor selection, and the proposed dataset have a good representation of the targeted population where the AI model is intended to be used. Comparing performance measures between the different training, validation and unseen clinical datasets is important for evaluating the risk of over/underfitting and widening the generalisability of the model. The format of the predictive model (assistive or directive) should be carefully chosen and designed. Maintenance of the AI-predictive model is important as populations, diseases, and treatments change. The downstream effects on patient outcome of using an AI prediction model are not always predictable and it is important to evaluate its use in clinical practice using an appropriate study design.

3.3. Chapter Summary

The results presented in this chapter began by describing the different frameworks currently present to guide the development of predictive models and the limitations of these frameworks (Introduction section). Then the researcher described nine key stages for the development and evaluation of AI-predictive models. These key stages start from the clear defining of the outcome of interest, with some examples for how some outcomes were defined (Step one). The next stage was the identification of predictors, with giving more preference to modifiable predictors, which could potentially have higher benefit on the decision-making process. The researcher also introduced the first step in the blended approach, which is related to the selection of different categories of predictors (Step two). In Step three, the researcher explained the different portals in the UK for access of high-quality dataset for models development, the researcher also provided guidance on the important points to focus on when choosing a dataset from the Gateway. Step four and five were more focused on the actual development of the models, throughout the training, validation and testing, the researcher also described the second stage of the blended approach, which is linked to the use of clinical input and machine learning techniques to select features for building the model. Building a model interface and tips for effective presentation of the model interpretations was described in Step six, followed by a description of different regulatory processes in the UK and globally for the licensing of AI-predictive models (step seven), and ending by emphasizing the importance of maintenance (step eight) and ongoing evaluation of these models (step nine).

Chapter 4 will look closer at identifying appropriate clinical variables or predictors related to our outcome of interest. The researcher will also describe how the predictors identification can affect the predictive timeframe and predictive performance of AI-predictive models.

Chapter 4. Identifying the Optimal Set of Predictors Used to Train Machine Learning Algorithms to Predict the Likelihood of Infection and Subsequent Sepsis.

4.1. Introduction

In order to develop and internally validate an AI model to predict the likelihood of post-operative infection in surgical patients (overall aim), it is important to identify the optimal set of predictors used to train machine learning algorithms to predict the likelihood of an infection and subsequent sepsis. In this Chapter, the researcher conducted a systematic review of the literature to identify predictors that had previously been used to develop AI-predictive models for the prediction of sepsis in different care settings. It is important to also understand how each of these developed models defined sepsis and how this definition affected the model performance. In addition, the researcher described the most common predictors used, and how the type of predictors could have an impact on the model performance. These predictors guided the development of the AI-predictive model as described in Chapter 5.

This systematic review has been published in an international peer-reviewed open access journal: Hassan N, Slight R, Weiland D, Vellinga A, Morgan G, Aboushareb F, Slight SP. *Preventing sepsis; how can artificial intelligence inform the clinical decision making process? A systematic review*. International Journal of Medical Informatics. 2021 Jun 1;150:104457., DOI: <https://doi.org/10.1016/j.ijmedinf.2021.104457> (*Appendix B*). The findings from this chapter have also been presented at a national research conference: Great North Pharmacy Research Conference 2021, HSRPP 2021 Conference, and Faculty of Clinical Informatics annual conference 2022 (*Appendix C*).

4.2. Background

Sepsis is a life-threatening condition that is associated with significant morbidity and mortality.[21] Sepsis affects 10% of hospitalized patients, with a 38% rate of mortality in Europe.[85,132] Sepsis also represents a financial burden on healthcare systems, accounting for almost half of the cost of running intensive care.[85] Predicting sepsis can reduce healthcare costs and save patients' lives through preventing the initiation of the multi-organ dysfunction process, reducing Intensive Care Unit (ICU) admissions and improving patients' prognosis.[133]

The use of big data and AI applications, namely machine learning, can predict diseases.[134] Machine learning algorithms have been developed for screening and predicting sepsis,[55,132]

and have showed better predictive abilities compared to current conventional scoring systems.[135] However, these algorithms have got limitations; some have been trained using at least two of SIRS criteria, which can be non-specific and misleading in sepsis prediction.[135,136] Some algorithms have also neglected predictors that influence the susceptibility and exposure of patients to infections.[136]

Clinical decision support systems guide clinicians with their decision making and reduce the likelihood of errors.[137,138] Integrating machine learning algorithms with CDSS could flag high-risk patients and assist clinicians with their care management.[26,135] Predicting the likelihood of sepsis may also help concentrate finite resources, including more aggressive care bundles; higher level care; more regular critical care outreach reviews and observation of vital signs; enhanced nursing; and earlier initiation of broad-spectrum antibiotic therapy in case suspicious symptoms and signs develop.

Patients consenting for surgery could also be provided with more accurate information relating to the risk of post-operative infection and subsequent sepsis. The researcher conducted a systematic review to identify the optimal set of predictors used to train machine learning algorithms to predict the likelihood of an infection and subsequent sepsis, in non-infected patients, and how AI could play a role in guiding clinicians' decision making around sepsis.

4.3. Methods

The review was registered with The International Prospective Register of Systematic Reviews (PROSPERO) (CRD42020158685) and was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.[139] Relevant headings were extracted from PRISMA-P checklist for the preparation of this review.[139]

4.3.1. Eligibility Criteria

Quantitative primary research studies that focused on sepsis prediction (cohort, prospective and retrospective studies) associated with bacterial infection in adult population (>18 years) in all care settings, including data on predictors to develop machine learning algorithms, were eligible for inclusion. The primary search was restricted to the timeframe, 1st January 2000 till the 25th November 2019, to include all articles that may have used sepsis definitions since the 2001 consensus terminology, including sepsis-3 definition.[140] The researcher updated this search on 5th October 2020 to include any articles published over the previous 11 months. Only articles published in the English language were included. Studies published in peer-reviewed journals

or conference proceedings were included. Editorials, commentaries, letters, and opinion articles were excluded.

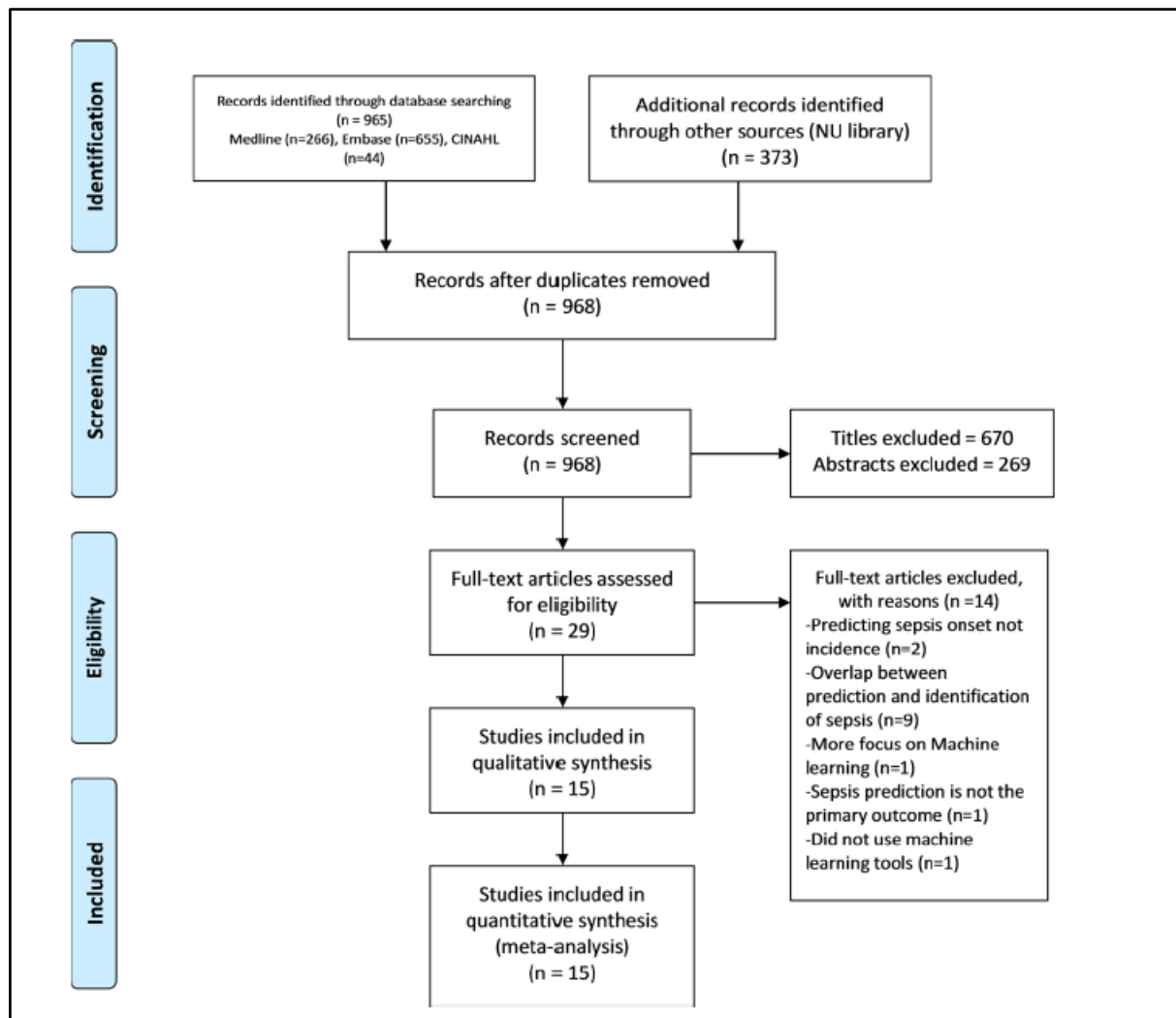
4.3.2. Information sources and search strategy

Three large databases were searched, including Medline (via Ovid), Cumulative Index of Nursing and Allied Health Literature (CINAHL), and Embase (via Ovid). A database of medical related articles held at Newcastle University library was also searched. Appropriate search terms were developed and grouped into three sets, specifically relating to sepsis, predictors, and machine learning and AI. In each set, terms were combined with the “OR” operator, and all sets were then combined with the “AND” operator. The search strategy used, including specific keywords, is described in supplemental material.

4.3.3 Study selection

After removing duplicates using Endnote software (version X9, Clarivate analytics, Philadelphia, United States), titles and abstracts were screened. The researcher excluded manuscripts at the title stage if the title stated clearly that the study was qualitative, on paediatrics, or discussed any type of sepsis other than bacterial sepsis. The same process was carried out in the abstract screening stage. Two independent reviewers (NH and FA) screened all full texts, with any disagreements resolved by discussion with a third reviewer (SPS), if necessary. References of the included articles were reviewed to identify any relevant papers. The reason for excluding any full text article was also documented. Full details on the search strategy can be found in Figure 7.

Figure 7. PRISMA chart for the search strategy and review process for the included studies:



4.3.4 Data collection and analysis

The researcher developed a customised data-extraction sheet which included qualitative data such as: title, authors, year of publication, country, type of publication, objective(s), study design, inclusion and exclusion criteria, data source (i.e., medical records), start date, end date, population under investigation, clinical setting, predictors for infection and/or sepsis, including co-morbidities, sepsis definition, outcome(s) (developed algorithm), imputation of missing data, type of machine learning technique and key conclusions. The data extraction sheet also included quantitative data such as: number of predictors used, sample size for training, sample size for validation, age, prediction timeframe in hours, sensitivity, specificity, positive predictive value, and negative predictive value for the developed algorithms.

A narrative synthesis of all eligible studies was undertaken. Key findings in the included articles were identified through an inductive strategy, which included the simplification of raw data in

the included articles into a comprehensive format. This approach developed a framework with clear associations among the extracted information and allowed arranging the collected data in main groups.

Quality assessment was performed using Newcastle-Ottawa Quality Assessment scale, which was used to evaluate the quality of non-randomized studies as shown in table 1.

Table 1. Newcastle-Ottawa Scale for Quality assessment of the included studies:

Author	Selection				Comparability	Outcome			AHRQ standards
	Representativeness of the exposed cohort	Selection of the non-exposed cohort	Ascertainment of exposure	Demonstration that outcome of interest was not present at start of study		Assessment of outcome	Was follow-up long enough for outcomes to occur	Adequacy of follow-up of cohorts	
Gupta et al.2020 ⁹⁰	1	1	1	1	2	1	1	1	Good Quality
Morrill et al.2020 ¹⁴¹	1	1	1	1	2	1	1	1	Good Quality
Giannini et al.2019 ⁵⁵	1	1	1	1	2	1	1	1	Good Quality
Bloch et al. 2019 ¹¹¹	1	1	1	1	2	1	1	1	Good Quality
Lu et al. 2019 ⁵⁶	1	1	1	1	2	1	1	0	Fair Quality
Scherpf et al. 2019 ¹⁴²	1	0	1	1	2	1	0	0	Poor Quality

Van et al. 2019 ¹⁴³	1	1	1	1	2	1	1	1	Good Quality
Faisal et al. 2018 ¹⁴⁴	1	1	1	1	2	1	1	1	Good Quality
Khojandi et al. 2018 ⁸⁸	1	1	1	1	2	1	1	0	Fair Quality
Nemati et al. 2018 ⁸⁴	1	1	1	1	2	1	1	1	Good Quality
Saqib et al. 2018 ¹⁴⁵	1	1	1	1	2	1	1	1	Good Quality
Danner et al. 2017 ¹⁴⁶	1	1	1	1	2	1	1	0	Fair Quality
Sood et al. 2017 ⁶⁰	1	1	1	1	2	1	1	1	Good Quality
Back et al. 2016 ¹⁴⁷	1	1	1	1	2	1	1	1	Good Quality
Calvert et al. 2016 ¹⁴⁸	1	1	1	1	2	1	0	0	Poor Quality
Wang et al. 2016 ⁵⁴	1	1	1	1	2	1	1	1	Good Quality

Lukaszewski et al. 2008 ⁵⁷	1	0	1	1	2	1	0	0	Poor Quality
---------------------------------------	---	---	---	---	---	---	---	---	--------------

Thresholds for converting the Newcastle-Ottawa scales to AHRQ standards (good, fair, and poor):

Good quality: 3 or 4 stars in selection domain AND 1 or 2 stars in comparability domain AND 2 or 3 stars in outcome/exposure domain

Fair quality: 2 stars in selection domain AND 1 or 2 stars in comparability domain AND 2 or 3 stars in outcome/exposure domain

Poor quality: 0 or 1 star in selection domain OR 0 stars in comparability domain OR 0 or 1 stars in outcome/exposure domain

4.4. Results

The primary search resulted in 1,338 research articles, from Medline (266), Embase (655) CINAHL (44) databases, and Newcastle University electronic library (373). After removing duplicates, 968 articles were eligible for further screening, with 670 and 269 research articles excluded at the title and abstract stages, respectively. Twenty-nine articles were eligible for full text review, of which 15 articles (14 papers and one conference abstract) focused on predicting an infection and subsequent sepsis in non-infected patients in different care settings, and were included in this review.[55-57,60,84,88,111,142-148] The researcher contacted the corresponding author of the conference abstract to obtain the full text article.[88] The updated search identified two additional papers that met the inclusion criteria.[90,141]

Of the seventeen included articles, ten articles were conducted in the US,[54,55,60,84,88,90,143,145,146,148] three articles were conducted in the UK,[57,141,144] and one paper was published from each of the following countries, China,[56] Germany,[142] Israel[111] and South Korea.[147] Regarding the study design, twelve articles had a retrospective design, while the remaining five articles had a prospective design. Ten studies used hospital databases to train and validate their algorithms, three studies used MIMIC III database, and one study used either of a health facts dataset,[88] Cerner Corporations HIPAA-compliant Health Facts database,[90] or PhysioNet/Computing database.[141] MIMIC database is a freely available dataset that includes anonymised data for over forty thousand patients who were admitted to BIDMC.[84,142,145,148] Nemati et al.[84] used two databases, a hospital database to develop their algorithm and the MIMIC II database for external validation of the developed algorithm.[84] Sixteen studies included data from medical inpatients, ICU patients,[55,90] and trauma patients.[56] Sood et al. included data for elective cancer surgery,[60] and Wang et al. included data from community-dwelling adults.[54]

All included studies were classified as prognostic prediction models, according to the TRIPOD criteria.[89] Two studies were classified as Type 3 (externally validated),[88,144] ten studies were classified as Type 1b (internal validation with two different sets),[55,60,84,90,111,141,142,143,146,147] and three studies were not validated.[56,145,148]

4.4.1 Definitions

Eleven studies defined sepsis by the occurrence of two or more of the SIRS criteria in addition to an identified infection for five to 24 hours before the diagnosis of sepsis, and/or Angus classification (using either international classification of the disease (ICD) 9 or 10 codes). One

study defined sepsis with ICD 9 codes: 995.92 (severe sepsis) or 785.52 (septic shock), positive blood cultures, and lactate levels higher than 2.2 mmol/L or a systolic blood pressure less than 90 mmHg, all occurring within a one hour.[55] Two studies confirmed the presence of an infection either through a blood culture or administration of antibiotics.[56,143] Some studies used hypotension and elevated lactate levels as proxy markers for identifying patients with sepsis.[60,84,90,111,144] Only four studies used the most recent Sepsis-3 definition.[55,84,90,141]

4.4.2 Predictors

The researcher identified 194 predictors and grouped these into four categories: patient demographics, vital signs, laboratory investigations and comorbidities. The minimum number of predictors used to develop an algorithm was 2, both of which were vital signs,[141] while the maximum number was 65 predictors.[84] The average number of sepsis predictors was 13, with the most prevalent predictors, including: age, gender, smoking, alcohol intake, heart rate, blood pressure, lactate level, cardiovascular disease, endocrine disease, cancer, chronic kidney disease (estimated Glomerular Filtration Rate eGFR<60ml/min), WBC, liver disease (level not specified in included studies), surgical approach (open or minimally invasive), and pre-operative haematocrit < 30 in surgical patients. Due to the retrospective nature of some of these included studies, missing data was recognised as an important limitation. However, this problem was usually overcome by data imputations, using averaged values or the worst value for a clinical variable in a day.[54,56,142,145] Of these 13 predictors, the researcher classified seven as potentially '*modifiable or controllable*' based on causal inferences found in the literature. These included, for example, smoking, alcohol intake, hypertension, myocardial ischemia, obesity and dyslipidemia. As for surgical patients, the choice of surgical approach (e.g. minimally invasive surgery) and pre-operative hematocrit could also be possible *modifiable* predictors for sepsis.[54-56,60] Table 2 summarises the published predictors for infection and sepsis.

Table 2. Description of sepsis predictors in the included studies:

Study	Non-Modifiable predictors	Predictors which if modified or possibly controlled could potentially reverse sepsis	Number of predictors to train the algorithm
Gupta et al.2020 [90]	Body temperature , mean arterial pressure , systolic blood pressure, glasgow coma scale, respiratory rate, glasgow coma scale, respiratory rate	Glucose	8
Morrill et al.2020 [141]	A ratio between heart rate and systolic blood pressure (shock index)		2
Giannini et al.2019 [55]	Blood typing, absolute basophil count, absolute eosinophil, absolute lymphocyte count, absolute monocyte count, absolute neutrophil count, age, alanine aminotransferase, albumin, alkaline phosphatase, anion gap, arterial blood gases, aspartate aminotransferase, non-invasive DBP, INR, magnesium level, potassium level, sodium level, non-invasive SBP, bilirubin (direct, indirect, total), blood urea nitrogen, calcium level total, carbon dioxide, chloride, creatinine, department, fraction of inspired Oxygen (%), gender, heart rate, height, hematocrit, marital status, mean cell volume, Oxygen saturation (%), Partial thromboplastin time, phosphorus level, platelets, prothrombin time, race, red blood cell count, red blood cell morphology, red cell distribution width, respiration rate, Richmond agitation-sedation scale, temperature (degree Fahrenheit), Urinalysis (urine appearance, urine bacteria, urine blood, urine leukocyte esterase, urine nitrite, urine pH, urine red blood cell, urine specific gravity, urine white blood cells) , venous blood gases, WBC count.	Weight, glucose level, hospital location, hospital service, urine glucose, urine ketones.	20
Bloch et al. 2019[111]	Heart rate, temperature, and mean arterial blood pressures recorded from an arterial line and respiratory rate as recorded from the mechanical ventilator.		4
Lu et al. 2019[56]	Demographic data (age, sex), Vital signs (heart rate, body temperature, respiratory rate, systolic blood pressure, average arterial pressure, and diastolic blood pressure. The severity of injury was determined by the AIS, ISS, and Glasgow Coma Scale (GCS). Laboratory variables (white blood cell count,	drinking, smoking, and comorbidities (including	7

	hemoglobin, hematocrit, platelet, neutrophil ratio, lymphocyte count, lymphocyte ratio, monocyte count, monocyte ratio, and neutrophil count), blood coagulation function, liver function tests, renal function tests, biochemical indicators, and blood gas analysis.	obesity, hypertension, cardiac arrhythmia, and metastatic cancer (after pharmacological or surgical treatment).	
Scherpf et al. 2019[142]	Age, SBP, DBP, pH value, SO ₂ , temperature, heart rate, respiratory rate, PaCO ₂ , WBC.		10
Van et al. 2019[143]	heart rate, (DBP), (SBP),(MAP), temperature, respiratory rate (RR) and (SpO ₂), (WBC) count.		8
Faisal et al. 2018[144]	patient's age, gender, discharge status (alive/dead), admission and discharge date and time, and electronic NEWS, blood test results (albumin, creatinine, hemoglobin, potassium, sodium, urea, and WBC count), and AKI score (0, 1, 2,3).		11
Khojandi et al. 2018[88]	Age group, gender, race, marital status, payer code, WBCs, RR, Temp, HR, and diagnosis information (e.g., ICD-9 code	medications administered (e.g., name of medication, order strength of medication)	9
Nemati et al. 2018[84]	Clinical features: (MAP), Heart Rate (HR), (O ₂ Sat), Systolic (SBP), (DBP), Respiratory Rate, Temperature, Glasgow Coma Scale (GCS), (PaO ₂), Fraction of Inspired O ₂ (FIO ₂). Laboratory: (WBC), Hemoglobin, Hematocrit, Creatinine, Bilirubin and Bilirubin direct, Platelets, (INR), (PTT), (AST), Alkaline Phosphatase, Lactate, Potassium, Calcium, (BUN), Phosphorus, Magnesium, Chloride, (BNP), Troponin, Fibrinogen, CRP, Sedimentation Rate, Ammonia. Laboratory: pH, pCO ₂ , HCO ₃ , Base Excess, SaO ₂ . Demographics/History: Care Unit (Surgical, Cardiac Care, or Neurointensive care), Surgery in the past 12 hours, Wound Class (clean, contaminated, dirty, or infected), Surgical Specialty (Cardiovascular, Neuro, Ortho-Spine, Oncology, Urology, etc.), , Age, , maximum change in SOFA score over the past 6 hours.	Charleston Comorbidity Index (CCI), Mechanical Ventilation (e.g., invasive vs non-invasive, and duration), Glucose, Number of antibiotics in the past 12, 24, and 48 hours.	65
Saqib et al. 2018[145]	WBCs, HR, DBP, SBP, MAP, anion gap, HCO ₃ , SO ₂ , height, temperature, blood urea nitrogen, RR, UOP, SCr, PLT	Glucose, weight	12

Danner et al. 2017[146]	Age, Heart Rate, Respiratory Rate, Temperature, SBP, DBP, MAP, Pulse Oximetry, Heart Rate to SBP ratio		6
Sood et al. 2017[60]	age at surgery, gender, race, neoadjuvant radiotherapy, neoadjuvant chemotherapy, American Society of Anaesthesiologists (ASA) score, operative time, and procedure type	body mass index (BMI), smoking status, comorbidities (hypertension, diabetes mellitus, cardiopulmonary disease (e.g., controlling asthma or COPD), cerebrovascular disease and liver disease (if alcohol-induced), peripheral vascular disease, and chronic kidney disease), preoperative hematocrit (Hct), preoperative serum creatinine, surgical approach (minimally invasive versus open)	
Back et al. 2016 [147]	Age, gender, admission through ER, nursing unit, ICU admission, DBP, SBP, HR, length of stay		
Calvert et al. 2016[148]	SBP, pulse pressure, heart rate, temperature, respiration rate, WBC count, pH, blood oxygen saturation, and age		9
Wang et al. 2016[54]	Age, gender, race, education, income, geographic region, hsCRP,	Smoking, alcohol, comorbidities (AF, cancer, CAD, chronic lung diseases, diabetes, DVT, dyslipidaemia,	13

		hypertension, MI, obesity, PAD, stroke), eGFR<60ml/min, ACR	
Lukaszewski et al. 2008[57]	interleukin-1 (IL-1), IL-6, IL-8, and IL-10, tumor necrosis factor-, FasL, and CCL2 mRNA		6

Abbreviations: DBP: diastolic blood pressure, INR: international normalized ratio, SBP: systolic blood pressure, AIS: Abbreviated Injury Scale, ISS: Injury Severity Score, GCS: Glasgow coma scale, SO₂: oxygen saturation, PaCO₂: partial pressure of carbon dioxide, MAP: mean arterial pressure, SpO₂: partial oxygen pressure, AKI: acute kidney injury, RR: respiratory rate, HR: heart rate, NEWS: national early warning score, ICD: International Classification of Diseases, FiO₂: fraction of inspired oxygen, WBC: white blood cell count, PTT: prothrombin time, AST: aspartate aminotransferase, BUN: blood urea nitrogen, BNP: atrial natriuretic peptide, CRP: C-reactive protein, HCO₃: bicarbonate, CCI: Charleston comorbidity index, SOFA: sequential organ failure assessment, UOP: urine output, ASA: American Society of Anaesthesiologists score, BMI: body mass index, Hct: haematocrit, hsCRP: highly sensitive C-reactive protein, AF: atrial fibrillation, CAD: coronary artery disease, DVT: deep vein thrombosis, MI: myocardial infarction, PAD: peripheral artery disease, eGFR: estimated glomerular filtration rate, ACR: albumin-creatinine ratio.

4.4.3 Machine learning algorithms

Five studies used random forest classifier to train the algorithm,[55,60,88,143,145] two studies used recurrent neural network (RNN),[57,142] and one study used one of the following methods: support vector machines (SMV-RBF),[111] least absolute shrinkage and selection operator (LASSO) technique,[56] modified Weibull-Cox proportional hazards model,[84] tree augmented Bayesian networks,[90] and signature-based regression model.[141] Fourteen studies validated their algorithms against AUC, with three studies reporting an AUC >0.9 indicating excellent discriminative power between infected and non-infected patients,[90,147,148] five studies reporting an AUC of between 0.7 to 0.9, indicating good power.[56,55,141,142,144] Additionally, only one study had an AUC below 0.7,[145] and three studies did not mention their AUC.[54,88,146] The remaining three studies did not attempt any form of algorithm validation.[57,60,143]

The predictive power was described using either specificity,[55,56,84,88,90,111,141,142,143,144,147,148] sensitivity,[55,56,84,88,90,111,141-144,147,148] positive predictive value,[55,57,111,143,144,147] negative predictive value,[57,111,147] or accuracy.[84,88,146] Average sensitivity was 75.7±17.88, with a minimum of 26% and a maximum of 98%, calculated from 13 studies. Average specificity was 63.08±22.01, with a minimum of 30% and a maximum of 98%, calculated from 12 studies. Average positive predictive value was 53.57±24.22 from 6 studies, while average negative predictive value is

77.13±14.74, from three studies. Table 3 described the different strategies used to develop the algorithms and their predictive power based on their working sepsis definition.

Table 3. Machine learning algorithms, predictive power and sepsis definition used:

Study	Name of the Algorithm	Machine learning technique	Predictive power	Sepsis definition
Gupta et al.2020 [90]		tree augmented Bayesian networks	SN:85, SP:76, PPV: ---, NPV:---	Sepsis-3 Definition
Morrill et al.2020 [141]		signature-based regression model	SN:86, SP:33, PPV: ---, NPV:---	Sepsis-3 Definition
Giannini et al.2019 [55]		open source Python programming language and Scikit-learn v0.15. 2 cases were used to train a random-forest classifier.100 estimators (trees) and gini criteria for splits were used.	SN:26, SP:98, PPV: 29, NPV:---	Sepsis was defined based on Sepsis-3 Definition in addition to 3 criteria: 1) International Classification of Diseases, 9th Edition codes (ICD-9) 2) a positive blood culture; 3) a lactate > 2.2 mmol/L or a systolic blood pressure < 90 mmHg, all occurring within a 1 hour. Of note, the time of earliest measured elevated lactate or hypotension was named “sepsis onset” for algorithm’s training.
Bloch et al. 2019[111]		SVM	SN:84, SP:39, PPV: 59, NPV:70.73	The occurrence of at least 2 SIRS criteria within a consecutive 24-hour interval with a diagnosed infection

Lu et al. 2019[56]	Traumatic sepsis score	Univariate logistic regression, LASSO technique was used to select the best subset of features.	SN:64, SP:82, PPV: ---, NPV:---	Suspected or documented infection with, an acute increase in the (SOFA) score by two or more points.
Scherpf et al. 2019[142]	InSight	RNN, the training data was additionally split into training and validation data resulting in training data, validation data and test data for each of the four cross-validation runs	SN:90, SP:30, PPV: ---, NPV:---	the presence of (SIRS) and an infection at the same time.
Van et al. 2019[143]		A random forest classifier using a total of 132 features extracted from a moving time-window.	SN:80, SP:68, PPV: 55, NPV:---	(SIRS) criteria in the presence of the suspicion of infection, indicated by a blood culture and the administration of antibiotics, along with relevant (ICD10) codes.
Faisal et al. 2018[144]	Computer-aided risk of sepsis (CARS) model	RNN	SN:67, SP:88, PPV: 20, NPV:---	(ICD-10) codes identified by an optimized validated method
Khojandi et al. 2018[88]	Algorithm to predict sepsis	A random forest classifier. Scikit-Learn library in Python 2.7 was used for implementation. 700-tree forest were developed by building one tree at a time on a new sub training set, while applying warm-starting technique and aggregating them into one model.	SN:98, SP:88, accuracy: 92	SIRS criteria and ICD-9 codes, especially in pneumonia patients who are at higher risk of sepsis.
Nemati et al. 2018[84]	Artificial Intelligence Sepsis Expert (AISE) algorithm	a modified regularized Weibull-Cox analysis, a type of machine learning approach that results in a more interpretable and generalizable survival model.	SN:85, SP:63, accuracy: 63	Suspected infection with two or more point changes in the SOFA score (tSOFA) from up to 24 hours before to up to 12 hours after the suspecting sepsis.

Saqib et al. 2018[145]		Random forest classifier.	SN:---, SP:---, PPV: ---, NPV:---	Angus criteria (which is an ICD coding system, to identify sepsis)
Danner et al. 2017[146]	Sepsis predictive probability index	Multivariate linear regression was used to analyse the data and create the predictive model utilizing SAS	SN:73.98, accuracy: 80	SIRS criteria and suspected infection.
Sood et al. 2017[60]	Composite sepsis risk score (CSRS)	All model construction and analyses were conducted using the open source Python programming language and Scikit-learn v0.15. 100 estimators (trees) and gini criteria for splits were used. Multivariable adjusted analyses (MVA) were performed to identify the predictors of post-operative sepsis. The score was stratified into low, intermediate, and high risk.	SN:---, SP:---, PPV: ---, NPV:---	Using SIRS criteria and ICD-9 codes.
Back et al. 2016 [147]	Automated Sepsis Risk Assessment System (Auto-SepRAS)	a logistic regression model by selecting risk factors that showed the highest predictive performance	SN:79, SP:82, PPV: 50, NPV:94	ICD-9-CM codes
Calvert et al. 2016[148]	InSight	A sepsis early warning algorithm, InSight, was developed and applied to the prediction of sepsis up to three hours prior to a patient's first five-hour (SIRS) episode.	SN:90, SP:81, PPV: ---, NPV:---	Sustained SIRS episode of at least five hours
Wang et al. 2016[54]	Sepsis risk score (SRS)	a series of multivariable Cox regression models, We performed multiple imputation using chained equations with the Stata "MI" suite, generating 10 imputed datasets and pooling estimates using Rubin's rules	SN:---, SP:---, PPV: ---, NPV:---	A serious infection plus two or more (SIRS) criteria.
Lukaszewski et al. 2008[57]		Neural network analysis	SN: ---, SP:---, PPV: 70, NPV:66.67	The criteria used during the 1991 American College of Chest Physicians/Society of

				Critical Care Medicine Consensus Conference which depends on SIRS criteria.
--	--	--	--	--

Abbreviations: SN: sensitivity, SP: specificity, PPV: positive predictive value, NPV: negative predictive value, SIRS: systemic inflammatory response syndrome.

4.4.4 Prediction timeframe of the algorithms

The prediction timeframe was defined in eight studies as the time window in which the developed algorithm can predict sepsis before its occurrence.[84,88,111,141-143,145,146] The average prediction time from these eight studies was 14.5 hours.[84,88,111,141-143, 145,146] The study that had the highest specificity (88%) and sensitivity (98%) had a prediction time frame of 48 hours, which was the longest time frame among all the included studies.[88] Six studies did not include an hourly timeframe for prediction.[54,55,60,90,111,144] Instead, two of these studies predicted sepsis during hospital stay,[55,56] one study predicted sepsis after emergency room admission,[144] and two studies predicted life-time risk of sepsis.[54,60]

4.5. Discussion

This review identified the predictors used to train machine learning algorithms to predict infection, and subsequent sepsis, in non-infected patients. Seventeen articles were identified from different care settings including ICU, trauma, elective cancer surgery and community. In total, 194 predictors were used in the 17 included articles, with an average of 13 predictors used to develop an algorithm. These predictors were categorised into four major categories: demographic data, vital signs, comorbidities and laboratory investigations. The developed algorithms had an average sensitivity 75.7 ± 17.88 , and average specificity 63.08 ± 22.01 , with an average prediction timeframe of 14.5 hours.[54-57,60,84,88,90,111,141-148] These different sensitivities and specificities were likely to have been influenced by the wide variety of definitions of sepsis that were used in the included papers.[54,55,57,60,111,142,144-148] Three studies defined sepsis through the use of antibiotics or collection of blood cultures.[56,84,143] However, some septic patients might not have had a blood culture taken or the volume of specimen might have been low, which could lead to false-negative results. Only four studies [55,84,90,141] used the most recent definition of sepsis, ‘Sepsis-3’, published in 2016.[140] Limitations of the 2001 definition of sepsis and septic shock include an excessive focus on inflammation, and inadequate specificity and sensitivity of the systemic inflammatory response syndrome (SIRS) criteria,[58] leading to discrepancies in reported incidence and observed mortality.

This systematic review focused on *prognostic* predictive models as described by the TRIPOD framework.[89] These models used predictors to estimate the likelihood of a specific outcome (in my case, infection and subsequent sepsis) at a future time point.[89] The researcher excluded *diagnostic* models,[58,149-150] as these models used predictors to estimate the diagnosis of sepsis in patients with established infection rather than those not currently infected. A recent review used these terms interchangeably, which could possibly have affected the precision of the developed algorithm.[151]

The choice of predictors influenced the predictive power and timeframe of the developed algorithms.[84,88,111,142,143,145,146] The researcher selected seven of these predictors that could be potentially *modified or controlled* to reduce the likelihood of infection and subsequent sepsis, and two additional predictors (surgical approach taken i.e., open or minimally invasive and the correction of pre-operative hematocrit) in surgical patients.[60] Although some variables might be strong predictors for sepsis, the researcher only classified those as potentially '*modifiable or controllable*' if they were found to have had a causal association with sepsis in the literature (described in table 2).[152,153] Due to the retrospective nature of these studies, it is also important to consider issues relating to non-availability of data. Therefore, categorical predictors should be used whenever possible rather than numerical predictors (i.e. using hypotensive and non-hypotensive instead of using the actual value of the blood pressure) to reduce the negative impact of missing data. However, the cut off threshold for conversion of numerical to categorical data needs to be carefully considered to avoid potential inaccuracies.

Vital signs were included in all the studies. They are more descriptive of the changes in the few hours before the occurrence of sepsis, which is important for the development phase of the machine learning algorithm. However, this might be different in the case of algorithms predicting 'lifetime risk' of sepsis,[54,60] and so depending on the context in which the algorithms are intended to be used, different predictors might be more useful. Furthermore, different algorithms could possibly be run at various timepoints before patients develop sepsis; some might focus on whether a patient is going to develop sepsis in the next few hours, whilst others might focus on its development in the next few days or weeks; the former algorithm might be useful for decision making around antibiotic/fluid administration, whilst the latter might be useful in deciding which patients should be monitored in high dependency unit post-operatively.

Additionally, most of the included laboratory investigations were indicative of multi-organ dysfunction, such as kidney function tests (serum creatinine, blood urea nitrogen, urinalysis),

liver function tests (Aspartate aminotransferase-AST- and alanine aminotransferase -ALT-), changes in complete blood count (haemoglobin, haematocrit, leucocytic count, platelet count), and coagulation profile (international normalized ratio and prothrombin time). Although many of these variables can be corrected, the researcher did not consider them as potentially '*modifiable or controllable*' predictors. However, for patients with diabetes or hypertension, correcting blood glucose or blood pressure, respectively, could modify their risk of infection and subsequent sepsis.

This could guide clinicians' decision making around the optimum timing of surgery, for example delaying a procedure until the patient's blood glucose and/or blood pressure are better controlled, so as to reduce the post-operative risk of infection and subsequent sepsis. It could also possibly guide the early use of antibiotics and intravenous fluids in patients with severe sepsis and septic shock (the "golden hour"), which is based upon early recognition, early administration of antibiotics, and early reversal of the shock state. It is important to note that symptoms and signs; examination findings; and non-laboratory investigations were not included as predictors because of the inherent difficulties involved in categorising these data, which are often subjective in nature and heavily textual, for the purposes of machine learning.

Another consideration is the inverse relationship between predictive power (sensitivity and specificity) of the algorithms and the prediction timeframe,[148] which means an increase in the specificity and sensitivity at the expense of a decrease in the predictive timeframe of the developed algorithm.[55,56,148] A short predictive timeframe reduces the time available for clinicians to control or reverse risk factors for sepsis.[55,154] It is still unclear how long this predictive timeframe should be, although it is likely to be at least 1 hour for sepsis.[155]

The researcher identified some limitations in the included studies. Three of the 17 included studies did not validate their algorithms internally.[57,60,143] Additionally, none of the studies mentioned the specific values obtained from the laboratory investigations that were included to develop the algorithms. It would have been helpful to include the reference ranges for these values, with only one study mentioning the values for WBC (<500 cell/L or > 50000 cell/L) that were excluded.[88] Moreover, some studies used the administration of antibiotics or blood cultures (just that it was carried out) as an indication of the presence of infection;[150] it is still possible that some of these patients may not have had a bacterial infection.[56,84,143]

In conclusion, machine learning algorithms offer a promising tool to guide clinicians in predicting sepsis. However, more research is required to optimise the balance between

predictive timeframe and predictive power. Future studies should clearly describe their working definitions of sepsis, noting that the definitions have recently been updated, and how cut off thresholds for the results of laboratory investigations were defined for inclusion and exclusion in the data sets to develop the algorithms. Additionally, clinicians' acceptance of AI applications to guide clinical decision making should be evaluated. More emphasis should be given to the use of modifiable predictors to develop predictive machine learning algorithms. A greater focus on modifiable predictors will guide the development of actionable recommendations based on algorithms' outcomes.

4.6. Chapter Summary

The results presented in this chapter began by describing how the machine learning and big data have been used to inform clinical decisions around sepsis screening and prevention (Background section). In the Methods section the researcher described how the literature was systematically reviewed to identify the optimal set of predictors for training machine learning models to predict the likelihood of an infection and subsequent sepsis. In the results section the researcher demonstrated how the definition of sepsis varied between different studies and the researcher also described the most common 13 predictors that were used in developing previous models which were classified in modifiable and non-modifiable predictors. In the discussion section, the researcher demonstrated how the sepsis definition and predictors type can influence the predictive power and predictive timeframe of machine learning models.

Chapter 5 will describe the development and internal validation of a machine learning model to predict the likelihood of post-operative infection in elective surgical patients.

Chapter 5. The Development of an AI Model to Predict the Likelihood of Post-Operative Infection in Surgical Patients (AIPIS).

5.1. Introduction

In this Chapter, the researcher will describe the training and internal validation of an AI-predictive model to predict post-operative infection in elective abdominal surgery (AIPIS). The researcher will demonstrate and explain the “*blended approach*” which she developed during the process of developing the model. The researcher used the nine key stages described in chapter three to guide the model development and also the choice of predictors was guided by the systematic review described in chapter 4. The findings from this chapter have been internationally presented as a poster in the American Medical Informatics Association Conference 2021 (*Appendix D*) and currently a manuscript is being prepared for submission to Lancet Digital Health Journal.

5.2. Background

The WHO has declared resistant infections a significant global health issue.[156] These infections can trigger an immune system response in the body and lead to life-threatening sepsis.[157] Hospitalisation increases the risk of an individual obtaining a resistant infection and subsequent sepsis.[158] In Europe, about 10% of all hospitalised patients (medical and surgical) develop sepsis, with an estimated mortality rate of 38%.[159] Approximately 19,680 SSIs were reported in Europe (when carrying out a total of 1,255,958 surgical procedures) between 2018 and 2020, with open colon surgery coming top of the list (9.5%, n=119,316).[160] These infections were associated with prolonged post-operative hospital stays, intensive care units admission, additional surgical procedures, treatment and higher mortality.[160] In the UK, sepsis accounts for approximately 48,000 patient deaths annually and incurs an economic cost of 16 billion pounds per year.[23,161] There was 834,000 cases of Healthcare-associated Infections (HCAIs) amongst 13.8 million hospitalised adult inpatients in NHS general and teaching hospitals in England in 2016-17, 28,500 of whom died with infection. These infected cases occupied 7.1 million NHS hospital bed days, with a monetary cost of £2.7 billion to the NHS.[162]

Surgical patients face a higher risk of infection and subsequent sepsis due to the invasive nature of the procedures. The occurrence of post-surgical complications, with infection the most common type of complication, depends on various factors, such as the patient’s age, smoking status, renal and hepatic functions, anaemia, and uncontrolled chronic conditions like diabetes.[163,164] Identifying patients with a higher risk of post-operative infections can

inform clinical decision making pre-surgery and provide an opportunity to intervene early (e.g., controlling blood glucose levels), thus reducing the likelihood of complications occurring.[165] Artificial Intelligence can be utilised to predict the likelihood of post-operative infection in elective surgical patients. As discussed in Chapter 4, several studies have developed AI tools to predict sepsis risk, most of which have focused on community patients, hospitalised medical patients, and critically ill patients.[166] Only one study developed an AI model to predict sepsis following elective cancer surgery in the first 30 post-operative days.[60] However, both medical and surgical models were shown to have limited clinical application and low sensitivity (average: 75.7 ± 17.88) and specificity (average: 63.08 ± 22.01).[167]

5.3. Aim

To develop and internally validate a machine learning model to predict the likelihood of post-operative infection in elective surgical patients to support clinical decision making.

5.3.1. Objectives

1. To clarify the outcome of interest (i.e. post-operative infection);
2. To identify the optimal set of predictors to develop the model, and classify them into modifiable and non-modifiable;
3. To select a relevant dataset to train the model;
4. To train an AI prognostic predictive model using the selected dataset;
5. To internally validate the developed model;
6. To present and interpret the model predictions.

5.4. Study design and methodology

5.4.1. Clarifying the outcome of interest

The machine learning model under development was to predict the likelihood of post-operative infection in elective surgical patients. The researcher sought microbiologist and clinician input on the definitions of post-operative infection and choice of predictors at the very beginning of this work. The researcher defined a bacterial post-operative infection as having abnormal body temperature ($<36^{\circ}\text{C}$ or $>38^{\circ}\text{C}$) AND abnormal total leukocyte count ($<4 \times 10^9/\text{L}$ or $>11 \times 10^9/\text{L}$) AND at least one of the following:

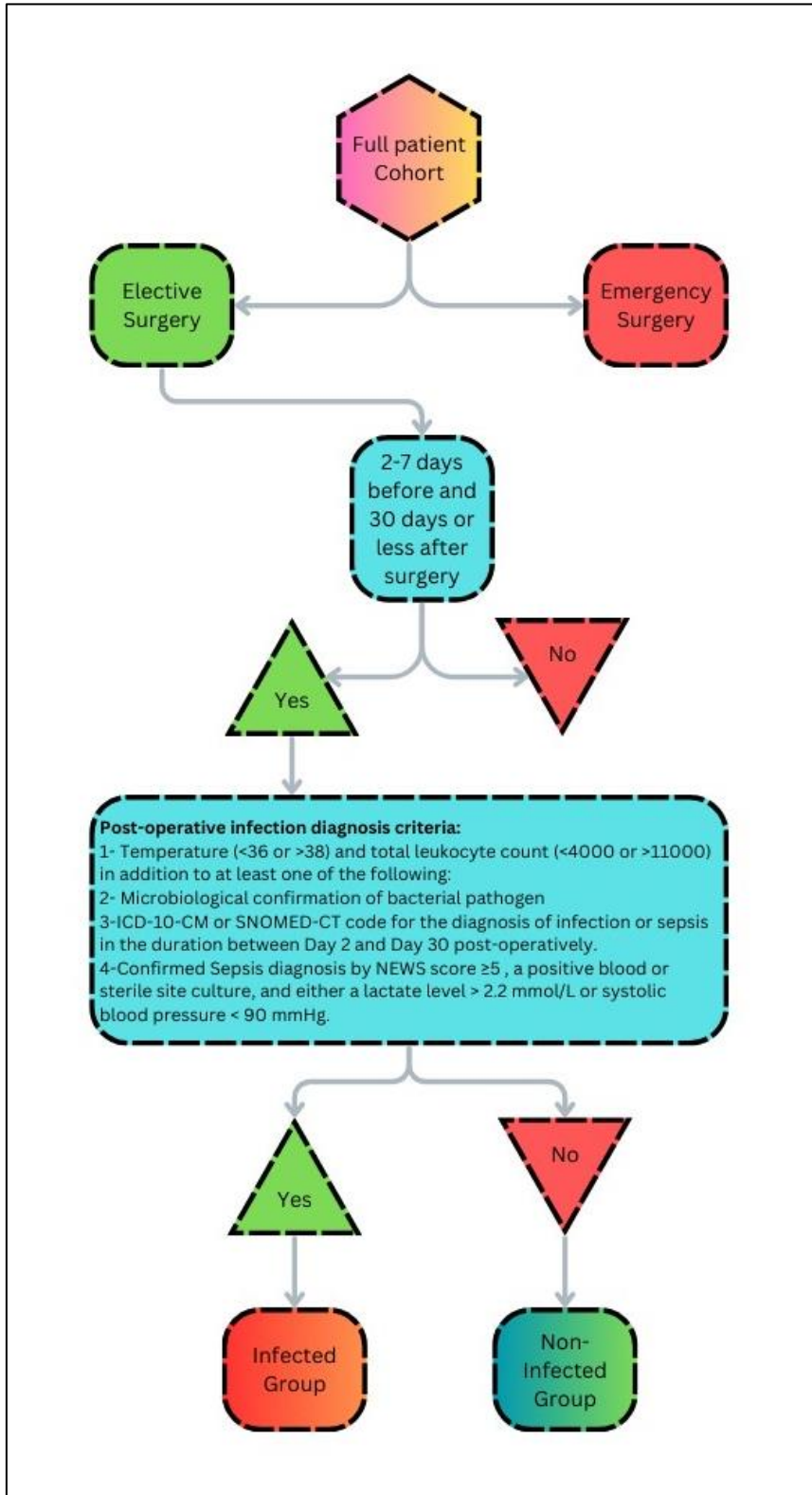
- 1- a positive sterile site culture as defined by the Public Health England Standards for Microbiology Investigations(PHE SMI I),[168]
- 2- confirmation of a particular antigen / antibody in the blood or serum, such as (a) Pneumococcal antigen (urine), (b) Legionella antigen test (urine), (c) atypical pneumonia antibody (blood/serum),
- 3- confirmation of bacterial pathogens detected by Luminex PCR of respiratory tract samples (e.g. Chlamydophila pneumoniae, Mycoplasma pneumoniae, Legionella pneumophila), Meningococcal PCR on blood/EDTA, and 16S rRNA PCR of sterile sites (bacteria)).
- 4- International Classification of Diseases, 10th Revision, Clinical Modification (ICD-10-CM) or Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) code for the diagnosis of infection in the duration between Day 2 and Day 30 post-operatively.

Elective surgical patients who were diagnosed with an infection ***at the time of admission*** e.g., endocarditis (confirmed from anaesthetic pre-op assessment records), or who did not meet the above mentioned criteria, were excluded.

The guidance of the Royal College of Physicians defined sepsis as the patient possessing a positive NEWS score ≥ 5 , a positive blood or sterile site culture, along with either a lactate level ≥ 2.2 mmol/L or a systolic blood pressure < 90 mmHg, or an ICD-10-CM code or SNOMED-CT code for sepsis.[169] If a patient met these Royal College of Physicians criteria and was diagnosed with sepsis, they were classified as part of the "patients with infection" group.

The classification of elective surgical patients

All elective surgical patients who fulfilled the bacterial infection definition detailed above were grouped into ***“patients with infection” (case group)***. All other patients who did not meet the bacterial infection definition were included in the ***“patients without infection” (control group)***, as per Figure 8 below.

Figure 8. Flow chart of the classification of patients in infected and non-infected groups:

5.4.2. Identifying appropriate predictors (feature selection)

As mentioned in Chapter 4, a review of the medical literature was conducted to identify potential predictors to train the machine learning model; 194 clinical predictors were previously used to train machine learning models for sepsis prediction, 13 of which were used across all 17 newly developed models.[87]. It should be noted that all these models included patients in their training datasets that had infections which progressed into sepsis; hence, the researcher used the 13 predictors identified from this review as potential predictors for the development of the model. The 13 predictors were further classified into seven modifiable (i.e., blood glucose level) and six non-modifiable (i.e., age, sex). The variables (n=92) in the sepsis phenotype database covered a wide range of patient-related information, including demographics (age, sex, multiple deprivation index and race), co-morbidities, repeat medications (steroids, insulin, immunosuppressants, and chemotherapy), medical intervention-related factors (type of procedure, incision type, and operative site), cause of admission (main diagnosis that led to hospitalisation), vital observations, microbiological investigations, laboratory investigations, length of ICU/hospital stay, discharge diagnosis, and type and date/time of antibiotics and vasopressors administered. The data was then cleaned and coded numerically. All available pseudo-anonymised variables were extracted from the training database (see next section) for both the "patients with infection" (*case group*) and the "patients without infection" (*control group*). (*Appendix E* describes a manual for detailed data cleaning and analysis steps and machine learning terms glossary).

As mentioned in Chapter 3, the researcher recommended a '*blended approach*' for predictors selection.[170] The developing team (microbiologist, consultant surgeon, and researcher) assessed the 92 variables based on clinical relevance and potential significance in predicting post-operative infections. The team selected 74 variables as potential risk factors, which included both modifiable and non-modifiable, and were filtered again using machine learning algorithms for features selection, which resulted in 19 predictors used to train the model.

The features selection process was carried out using recursive feature elimination with cross-validation (RFECV). This is a machine learning technique used to identify the most important features (i.e., predictors) from the dataset through recursively fitting the model with different groups of features and assessing their performance using cross-validation. Cross-validation assesses the performance of each algorithm on different dataset splits, ensuring that the identified features are not specific to any particular data split and could be replicated on unseen data. Accordingly, the RFECV can identify the optimal number of features that maximise the

predictive ability of the model. RFECV was carried out using five different supervised machine learning algorithms; decision trees, random forest classifiers, logistic regression, gradient boosting classifier and ridge classifier. A voting technique was then used to select the predictors that have been selected by a cutoff point of at least three out of the five algorithms.

5.4.3. Choosing relevant dataset

The Swedish Skaraborg dataset

While the researcher was waiting to obtain access to a UK dataset, she conducted a preliminary analysis using the publicly available anonymised dataset of adult patients admitted to emergency departments with sepsis in Sweden between 2011-12 (n=1572).[171] This Sepsis Skaraborg Database Study was approved by the Regional Ethical Review Board of Gothenburg. Code scripts were written in Python and correlations were assessed between a number of different types of variables.[172] IBM SPSS Statistics 27 was used for descriptive analysis and regression analyses. The purpose of using this dataset was for the researcher to learn programming language and write the python codes, which could then be re-run and/or adjusted using the UK database. The use of this dataset was considered as a staging phase to assure that the used ML algorithms are appropriate and can predict infection with a high performance.

The Deeply Phenotyped Sepsis Patients within Hospital: onset, treatments & outcomes Dataset

The UK HDR Hub for Acute Care (PIONEER), led by the University of Birmingham and University Hospitals Birmingham NHS Foundation Trust, supplied the retrospective pseudonymised data for training this machine learning algorithm. A data request form was submitted and approved by PIONEER. Ethical approval was also obtained from the ethical committee of the HRA under the reference number: 20/EM/0183 (See *Appendix F and G*). Data was extracted from ‘the Deeply phenotyped sepsis patients within hospital: onset, treatments & outcomes’ dataset. This dataset is considered one of the largest sepsis datasets in the UK and has been ranked as a platinum metadata by the health data research UK (HDRUK). The dataset contained 11 Structured Query Language (SQL database tables on admissions, social history, diagnosis, drug administration, lab analysis, microbiology data, observations, prescriptions, prescriptions drug component, site move and surgery. Details of the variables in each table are provided in the Analysis manual in (*Appendix E*)

Pseudo-anonymised data for adult patients (aged 18 years or above) were extracted and included in the model training if they had:

- a) been admitted for elective general abdominal surgery (see below),

- b) data documented for at least 48 hours during their hospital stay (thus day-case elective surgery were excluded),
- c) either undergone ‘clean’ surgery (an incision in which no inflammation was encountered in the surgical procedure) or ‘clean-contaminated’ surgery (an incision through which the respiratory, alimentary or genitourinary tract was entered under controlled conditions but with no contamination encountered),
- d) a length of hospital stay from 2 to 30 days after the operation, and
- e) a pre-operative length of hospital stay of seven days or less.

The decision was made to focus on elective general abdominal surgery patients for a number of reasons. Firstly, the researcher chose ‘elective’ patients as this was the population I was planning to subsequently apply the machine learning model, thus provided me with the opportunity to intervene early and reduce the likelihood of complications occurring. Secondly, the Deeply phenotyped sepsis patients within hospital: onset, treatments & outcomes dataset contained acute care information on elective general abdominal surgery for patients admitted to the University of Birmingham and University Hospitals Birmingham NHS Foundation Trust between 2004 and 2022 (i.e., approximately 18 years of data).[173] Thirdly, all patients within general abdominal surgery had undergone either ‘clean’ or ‘clean-contaminated’ surgery; this facilitated the possibility of testing and applying the developed model to other types of surgery that including class 1 (clean) or class 2 (clean-contaminated) surgical wounds (i.e. orthopaedic surgery, cardiothoracic surgery, etc.).[174]

For patients who had readings with multiple values in a day, it was decided that the average of the highest/lowest three readings would be taken; however, this also depended on the type of the lab test and whether an abnormal value was above or below a certain range. Data sheets were merged to join lab readings to relevant patients, so that each patient had one reading for each available lab investigation/vital sign, detailed processing for each variable is demonstrated in *Appendix E*.

5.4.4. AI model development

Statistical analysis and softwares:

Structured Query Language was used for access and visualisation of the database. Microsoft Excel was used for descriptive analysis of Sepsis Phenotype dataset due to patient confidentiality restrictions and agreements to access and analyse the data through TRE, and

also for data cleaning and merging of tables using PowerPivot and Power Query features. Code scripts for model development were built on an open-source programming language, Python v3.8.2 and visual studio code v1.64.2., and also for the internal validation. *Appendix H* shows all the scripts used to train, validate, and build the model interface.

Model training

As mentioned in Chapter 3, there were four major types of machine learning algorithms; the choice of which will depend on the ‘characters’ of the dataset (i.e., completeness, quality) and the performance measures of the algorithm. The Ensemble model can be used to improve the overall performance of the tool [103]. For AIPIS, the researcher measured the performance with both the best performing model and the ensemble model which included the best three performing model. The performance of the ensemble model was higher, as supported by the literature, hence the ensemble model was used to build the interface.

Management of missing data

The ‘Deeply phenotyped sepsis patients within hospital: onset, treatments & outcomes’ dataset has an impressive data completeness rate of 89% and a low weighted error of only 5%. To handle missing data in the dataset, a complete case analysis was used for patients who had more than 50% of their filtered information missing. This approach involves excluding cases with missing values for any of the selected predictors from the analysis. Out of the eligible patients for inclusion, 16.9% (n=556) were removed by complete case analysis because of having >50% of their filtered variables missing. Descriptive analysis for the cohort removed by complete case analysis (n=556) is demonstrated in *Appendix E*. For the remaining 83.1% of the cohort, missing data was not missing at random, and the decision was made to use K-nearest neighbors (KNN) algorithm for all patients with <50% missing variables. This imputation method uses a machine learning algorithm to impute missing values, based on the information provided from the K nearest data points to the ones with missing value. Python programming language was used to impute for missing data and train the model through Visual Studio Code (VS Code) interface. (*Appendix E* includes definitions for different Python libraries used to train the model, and *Appendix H* contains all the code scripts for model training and validation and building interface).

After importing the dataset into VS code, the researcher divided it into two portions: training the algorithm (80%) and internal validation (20%), randomly balanced between the two classes (infection versus non-infection) are split with the same ratio. As mentioned above, each predictor was then correlated to the outcome of interest (infection diagnosis) using the training

set and the performance measures of the model calculated. This includes the specificity, sensitivity, and the AUC.

The researcher used 12 different algorithms for the model training, including a random forest (RF) classifier, a random forest classifier with grid search, logistic regression, a decision tree, a decision tree with grid search, a gradient boosting classifier, gradient boosting classifier with grid search, a ridge classifier, a ridge classifier with cross validation, support vector classifier, and k-nearest neighbours, definitions for each of these algorithms is explained in the glossary in *Appendix E*. The performance with each of the algorithms was evaluated to determine the best performing model, and the ensemble model was also developed from the best three performing models, based on their performance metrics such as AUC and Youden Index.[105] The AUC measures the distinctive ability of the algorithm to differentiate between infected and non-infected patients. AUC between 0.7 and 0.9 indicates a good discrimination ability while an $AUC > 0.9$ represents a perfect ability to differentiate between two groups.[147] Accordingly, AUC is an overall assessment of a model accuracy. The Youden Index is used to provide an optimal cutoff point for risk categorisation.[175] Youden index has been previously used to categorise the risk into high and low to develop and validate a machine learning sepsis risk stratification model.[147] Youden index is calculated, using sensitivity and specificity to measure the efficiency of the predictive model.

$$\text{Youden index} = [\text{Specificity} + \text{Sensitivity}] - 1$$

The Youden index ranges from zero to one, with one being the highest value of the index. After running the machine learning algorithms, the highest recorded value of the Youden Index was used as the higher cutoff point. Patients with a Youden Index value above this cutoff point were classified as high-risk patients, while those with a value below the cutoff were considered low-risk for infection. [147].

5.4.5. Validating the AI-predictive model

The same algorithms were used for internal validation of the 20% split of data, and performance measures were compared during training and validation to reduce the risk of bias (ROB) and risk of over/underfitting. The PROBAST (Prediction model Risk Of Bias Assessment Tool) guided the development and validation of the model to reduce ROB.[176] More details about PROBAST components and how it was used are provided in the next Chapter (Chapter Six).

5.5. Results

5.5.1. Descriptive analysis

A total of 10,340 patients in the Sepsis phenotype database (accessed through PIONEER research hub), 8,110 of which were elective surgical patients. Of these, 3,272 patients were eligible for inclusion based on the previously mentioned eligibility criteria; 556 (16.9%) of these patients had more than 50% of their clinical variables (i.e., lab data, vital signs and microbiology data) missing after data cleaning and, hence, only 2,716 patients were included. The prevalence of post-operative infection of all types (the output) in this patient group was 38%. The prevalence of SSI in major abdominal surgery is estimated to be between 20 and 40% in the literature.[177] Descriptive analysis of this cohort is provided in Table 4 for categorical variables, and Table 5 for numerical variables. *Appendix E* shows an additional descriptive analysis for emergency surgery patients (the excluded cohort).

Table 4. Descriptive analysis for categorical variables of the 2716 included patients:

		Count (N=2716)
Gender	Male	1447 (53.3%)
	Female	1269 (46.7%)
Ethnicity	White -British	2099 (77.3%)
	Asian/Asian Brit - Pakistani	167 (6.14%)
	Not Known	130(4.8%)
	White - any other White b/g	61 (2.2%)
	White -Irish	56 (2%)
	Any Other Ethnic Group	50 (1.8%)
	Black/black Brit-Caribbean	45 (1.6%)
	Asian/Asian Brit - Indian	17 (0.6%)
	Others (<10 patients)	13 (0.5%)
Index of multiple deprivations	1	480 (17.7%)
	2	509(18.7%)

	3	563 (20.7%)
	4	457 (16.8%)
	5	485 (17.9%)
Microbiology (Cultures within 90 days before operation)		
Positive Blood cultures		4 (0.1%)
Positive other sterile site cultures (i.e. bile)		6(0.2%)
Other positive cultures (sputum, urine,etc.)		429 (15.8%)
Multi-drug resistance (MDR) in blood or sterile site		12 (0.4%)
Infection diagnosis (including those diagnosed with “sepsis”)		1030 (38%)
Sepsis diagnosis		80 (2.9%)
Comorbidities		
Malignancy		1675 (61.7%)
Hypertension		1030 (37.9%)
Cardiac		529 (19.5%)
Diabetes mellitus		509 (18.7%)
Renal		238 (8.8%)
Asthma		227(8.4%)
Hepatic		208 (7.7%)
Obesity		186 (6.8%)
Chronic Obstructive pulmonary disease		93 (3.4%)
Anaemia		82 (3%)
Thrombosis		54 (2%)
Chronic medications		

Steroids	255 (9.4%)
Immune suppressants and chemotherapy	32 (1.2%)
Insulin	217 (8%)
Minimally invasive surgical procedures	
Yes	207 (7.6%)
No	2509 (92.4%)

Table 5. Descriptive analysis for continuous variables of the 2716 included patients:

	Mean±SD	Min	Max
Age (years)	63.7±12.8	19	85
Pre-operative data (within 90 days before operation)			
White Blood Cell Count (WBCs)	8.9±3.8	3.1	16.2
Neutrophils	6.0±2.5	1.8	13.0
Lymphocytes	1.5±0.8	0.3	4.6
Red Blood Cell Count (RBCs)	4.3±0.5	2.2	6.5
Mean Cell Volume (MCV)	88.1±0.1	56.5	112
Mean Cell Haemoglobin (MCH)	29.5±2.3	16.5	38.1
Haematocrit	0.3±0.1	0.1	0.5
Platelets (PLT)*10 ³	230.8±1.6	70	820
Haemoglobin	11±2	7.3	17.2
Bilirubin	41.1±1.7	2	69
Alanine aminotransferase (ALT)	55.4±3.7	5	576
Aspartate aminotransferase (AST)	75.8±1.4	10	928

C-Reactive Protein (CRP)	19.8±0.7	1	569
Blood glucose	8.5±1.5	4.9	16
Serum creatinine	86±0.6	71	157
Operative and Post-operative data			
Operative duration (hrs)	4.2±2.3	<1	22
Pre-operative hospitalisation (days)	1.1±0.9	<1	7
Post-operative hospitalisation (days)	9±5.9	2	30
Length of Intensive Therapy Unit (ITU) stay (days)	1.9±1.1	0	21
Respiratory rate (breath/min)	16.5±1.7	12	23.8
Alert, Voice, Pain, Unresponsive (AVPU)	4±0	4	4
National Early Warning Score (NEWS)	1.9±0.8	0	6
Systolic blood pressure (SBP)	128.5±13.5	98	169
Diastolic blood pressure (DBP)	72.9±7.9	48	103
Mean arterial pressure (MAP)	92.3±9.2	82	130
Temperature	36.6±0.3	35.6	38.7
Blood glucose	9.8±2.2	3.8	28.3
International Normalised Ratio (INR)	1.2±0.2	0.9	3.3
Haemoglobin	10.9±1.6	7.6	16.4
Haematocrit	0.33±0.04	0.2	0.5

Mean cell haemoglobin	29.9±2.5	18.4	39.2
Mean cell volume	89.5±6.3	60.3	113.1
Platelet count*10 ³	279.7±128.8	29.8	648
Red Blood Cell Count (RBCs)	3.5±0.6	2.1	5.7
White Blood Cell Count (WBCs)	11.4±0.1	2.3	80.7
Lymphocytes	2.3±0.1	0.1	36
Neutrophils	10.9±0.2	0.6	96
Neutrophil-lymphocyte ratio	7.1±3.9	1.2	33.5
Erythrocyte sedimentation rate (ESR)	31.9±20.4	2	75
C-reactive protein (CRP)	96.0±2.0	2	325
Alanine transaminase (ALT)	126.2±4.3	9.7	1131
Aspartate aminotransferase (AST)	112.7±7.8	12	711
Total Bilirubin	36.9±58.4	3.2	475.3
Direct Bilirubin	131.3±88.4	12	221
Serum creatinine	70.4±37.6	26.6	203.6
Glomerular filtration rate (GFR)	81.6±13.3	12	90
Lactate	3.1±1.9	0.9	7.9
Creatinine Kinase (CK)	599±157.3	10	3052

5.5.2. Predictors

Nineteen out of 74 predictors were finally selected to train the model, based on their predictive performance (after running RFECV). These variables included demographic data (sex), comorbidities (diabetes, COPD, cardiac, obesity, anemia), microbiology data (positive cultures

within 90 days prior to surgery, multi drug resistance in this culture – resistance to three antibiotics from different classes), Lab investigations within 90 days before operation: WBCs, Lymphocytes, blood glucose, Platelets, hematocrit, mean cell hemoglobin, hemoglobin, total bilirubin, Aspartate aminotransferase, alanine transaminase, CRP.

5.5.3. Model performance

The best performing individual algorithm was the gradient boosting classifier and the five best algorithms were used to develop an ensemble model. The performance measures of the Ensemble and best models are provided in Table 6 below.

Table 6. Performance measures of Ensemble and best models during training and internal validation:

Performance measures	Best model		Ensemble model	
	Training	Internal Validation	Training	Internal validation
Sensitivity	67.9%	90.65%	85.3%	96.9%
Specificity	89.1%	75.3%	74.6%	74.1%
Positive Predictive value	76.8%	58%	64.3%	58.5%
Negative Predictive value	83.8%	95.5%	90.4%	98.4%
Youden index	0.6	0.7	0.6	0.7
AUC	84.3%	83%	88.6%	85.5%

5.5.4. Presenting the model predictions and building the model interface

The prototype of the possible model interface is shown in Figure 9. The user should add the data inputs for the predictors in this interface and press predict, then a statement appears with the risk of infection (Risk of post-operative infection is ...%), which is shown in Figure 10. This interface could potentially be used in the pre-operative setting to guide the surgeon-patient SDM about the risk of post-operative infections. The surgeon can actively show the patient how controlling/modifying any of the inputs (such as correcting anemia, or as part of prehabilitation screening to provide nutritional support to reduce the risk of post-operative infection and subsequent sepsis,[178] which can change their estimated risk of post-operative infection, which can influence their morbidity, mortality (if progressed to sepsis) and their quality of life following the surgery.

The interface was linked to the ensemble model obtained during the training and validation phases of the predictive model development. This ensures that the interface provides accurate risk predictions based on the selected predictors and the model's performance on the training and validation datasets. These predictors included sex, comorbidities (diabetes, COPD, cardiac, obesity, anemia), microbiology data (positive cultures within 90 days prior to surgery, multi drug resistance in this culture – resistance to three antibiotics from different classes), Lab

investigations within 90 days before operation: WBCs, Lymphocytes, blood glucose, Platelets, hematocrit, mean cell hemoglobin, hemoglobin, total bilirubin, Aspartate aminotransferase, alanine transaminase, CRP. Most of these predictors can be modified/controlled during the pre-operative window except for patient's sex. The model interface was designed as a streaming website, making it easily accessible and user-friendly. It requires inputs from the user, whether they are patients or clinicians, to provide the necessary data for predicting the risk of post-operative infection.

Figure 9. A prototype of a possible model interface:

AIPIS: Predicting post-operative infection in surgery

Predictors for post-operative infection:

Gender

☒ Male
☐ Female

Comorbidities

☐ Diabetes Mellitus
☐ Chronic Obstructive Pulmonary Disease
☐ Cardiac diseases
☐ Obesity
☐ Anaemia

Positive Cultures in the last 90 days

☒ Yes
☐ No

Multi Drug Resistance - resistance to 3 antibiotics from 3 different classes

☒ Yes
☐ No

White Blood Cells Count *10⁹/L

0.00 - +

Lymphocytes

0.00 - +

Blood glucose mmol/L

0.00 - +

Platelets count *10⁹/L

0.00 - +

Haematocrite value

0.00 - +

Mean Cell Haemoglobin

0.00 - +

Positive Cultures in the last 90 days

☒ Yes
☐ No

Multi Drug Resistance - resistance to 3 antibiotics from 3 different classes

☒ Yes
☐ No

White Blood Cells Count *10⁹/L

0.00 - +

Lymphocytes

0.00 - +

Blood glucose mmol/L

0.00 - +

Platelets count *10⁹/L

0.00 - +

Haematocrite value

0.00 - +

Mean Cell Haemoglobin

0.00 - +

Haemoglobin g/dl

0.00 - +

Total Bilirubin umol/L

0.00 - +

Alanine aminotransferase U/L

0.00 - +

Aspartate aminotransferase U/L

0.00 - +

C-reactive protein mg/L

0.00 - +

Predict

Figure 10. Model Output and recommendation:

The screenshot shows a web application interface with a browser address bar displaying 'localhost:8302'. The interface contains five input fields for laboratory values, each with a minus and plus button for adjustment:

- 11.99
- Total Bilirubin umol/L: 1.20
- Alanine aminotransferase U/L: 20.00
- Aspartate aminotransferase U/L: 20.00
- C-reactive protein mg/L: 7.00

Below the input fields is a red 'Predict' button. A red box highlights the 'Recommendation' section, which displays a green message: 'Risk of post operative infection is 33.0%'.

Insights from my systematic review, which was conducted to gain a deeper understanding of the perceptions of both clinicians and patients regarding the use of AI tools in clinical settings (see in Chapter Seven) were incorporated in the model interface development. Findings from my qualitative study, which explored the perspectives of clinicians from one English NHS Trust regarding the use of AI-based tools to inform SDM in the clinical environment, were also incorporated (see Chapter Eight).

The involvement of the public and patients (PPIE) in research design is crucial to improve the quality and relevance to clinical care, bringing different perspectives to research. Accordingly, the researcher arranged a 30-minutes meeting with a PPIE group of ten members through VOICE. The meeting was set before data access to get insights from the group on the model development, its benefit to patients and the NHS, and how the model can improve SDM.

The group provided insightful feedback which included:

1. Early involvement of clinicians is necessary to give their perspectives throughout the stages of the tool development, as their specialised knowledge will help improve the application of the tool later in real practice.
2. The development stages of AI tools should take into consideration the dynamic nature of patients' conditions when they are admitted to the hospital, so the tools should spot any change in patients' investigations and modify the risk estimates accordingly.

3. Data anonymisation, protection and security would be a major worry for patients and should be seriously considered by the researcher with data access, processing and development of the tool.
4. At the implementation stage, it should be clarified to the patients that the tool is developed through AI using wordings that are carefully selected and easy to understand, as many patients would be concerned about the idea of AI deciding for their health or medical conditions.
5. General hygiene, sterilisation and aseptic measures in hospitals should be the basic and primary infection prevention measure. Artificial intelligence tools can be an adjunctive to this for infection and subsequent sepsis prevention.
6. Number of patients needed to develop the tool and how they are selected would influence ROB and accuracy of the tool. Including multiple hospitals should also be considered whenever possible, and looking at one type of surgery would also reduce the variation.
7. Risk of infection will vary from one patient to another depending on the type of surgery (site and mode of surgery, i.e. major or minor operation) they will be undergoing, as well as the individualised risk factors in each patient.
8. Number and type of predictors are key factors to consider while developing AI tools, predictors related to patient's lifestyle and quality of life are important to consider in future tools, especially when predicting infection in community.
9. Larger projects should also use national and international data from multiple hospitals, which will allow for the database expansion to include more variables (i.e. considering the incidence of infection with particular microbes), which will better inform infection prediction.
10. Stages that AI tools pass through to assess their accuracy and judge if they are successful enough for implementation in clinical practice.
11. Ownership and patency of the developed AI tool.

5.7. Discussion

This research project aimed to develop a prognostic AI model to predict the likelihood of post-operative infection in one elective surgical setting. The model showed high predictive performance with 85.3% sensitivity, 74.6% specificity, 64.3% PPV, and 90.4% NPV, a Youden

index 0.6 and AUC=88.6%). Du et al. developed a machine learning model to predict post-operative infections in patients admitted to intensive care unit (ICU) and received hydromorphone analgesia.[179] The model included different predictors such as hydromorphone dosage, age, gender, body mass index, WBC, and CRP.[179] The model showed an AUC of 0.835 during training and 0.747 during validation. However, the study excluded high risk patients for post-operative infections, such as those who were obese, and those with impaired renal or hepatic function. Additionally, the model was only developed on 130 retrospective patients, without specifying the type of surgery involved, which makes the generalisation to all types of surgeries challenging.[179] Another study developed a deep learning model that predicted the risk of post-operative infection among elderly population (60 years and above) undergoing elective surgery.[180] The predictors included were gender, American Society of Anaesthesiologists scores, chronic comorbidities, surgical procedures, sedation/local anaesthesia, laparoscopic operation, and severity of surgery. This study included 1,510 patients and their model showed 63.2% sensitivity, 80.5% specificity, and AUC of 0.763. This model was trained using different algorithms with deep learning that included two different neural networks. The authors demonstrated that increasing the number of neural networks was associated with improved sensitivity (from 34.2% to 63.2%). Sensitivity demonstrates the ability of a tool to detect the presence of a condition, when the condition truly exists (true positive). The higher the sensitivity, the better the ability of the tool to identify true positive.[181] The authors also highlighted that selecting predictors more linked to baseline and surgery were associated with improved sensitivity.[180] The study focused only on elderly population who might have different infection predictors than younger age groups.[180]

Another machine learning model was developed to predict the likelihood of SSI following elective colorectal surgery using 275,152 patients from American College of Surgeons National Quality Improvement Program database.[182] Artificial neural network was the best performing model, demonstrating an AUC of 0.769, and predictors of infection included organ-space surgical-site infection present at time of surgery, operative time, oral antibiotic bowel preparation, and surgical approach.[182] Despite the relatively good AUC of the model, the type of predictors included were likely to have less applicability in the pre-operative stage (a key aspect of my model).

Throughout the development of my model, I considered multiple aspects to increase its performance while also considering its clinical utility. My dataset included patients who underwent different types of general abdominal surgery, resulting in a diverse patient cohort with different clinical characteristics. This diversity allows the model to be trained on a wide

range of surgical cases, enhancing its potential for extrapolation to other elective surgical settings. Secondly, the researcher sought expert opinions (senior microbiologist and clinician) on how the researcher defined ‘post-operative infection’. It should be noticed that sepsis diagnosis in the dataset was considered as a positive infection diagnosis for the model training. It was felt that the clear and pre-set criteria used to define post-operative infection in my training model contributed to its higher sensitivity. Furthermore, both modifiable and non-modifiable predictors were chosen to increase its clinical utility. Elective procedures have the option of being postponed if risk factors such as comorbidities like diabetes mellitus, chronic obstructive pulmonary disease, obesity and anaemia could be better controlled. Other potentially controllable features were blood glucose, full blood count with anaemia, high CRP levels which might denote an inflammation. Nevertheless, for the later, this inflammation could be linked to the condition for which the operation is being performed (i.e., biliary related disease), in which case, an elevation in total leucocytic count or CRP, for example, cannot be modified.

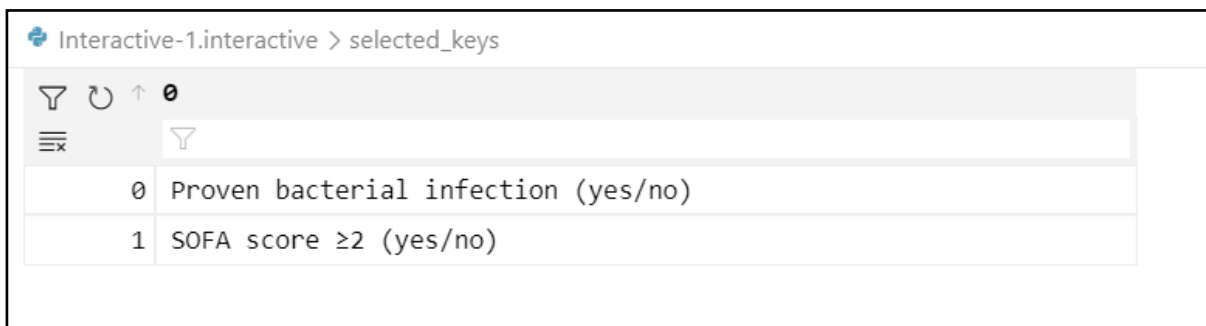
In addition to the models that have been developed to predict post-operative infections, as mentioned above, machine learning models have also been developed to predict sepsis in surgical setting. A model was developed on a patient group who had elective appendectomy,[183] which is considered a minor surgery, with low incidence of post-surgical sepsis and complications, making the extrapolation of the model to major elective surgery with higher risk of infection more challenging. The Sood et al.[60] model showed good performance for post-operative sepsis prediction with an AUC=0.77, however, the dataset was extracted from cancer patients only, who have different clinical characters from other non-cancer patients who might be undergoing different type of major elective surgeries, which limits the clinical use of this model to oncology setting only. A third model was developed to predict the risk of sepsis post elective surgery. [143] However, the model was developed from the MIMIC III database,[143] which includes mainly medical rather than surgical patients. The developers did not specify which type of elective surgery that their model was valid for, which can potentially impact its use in surgical practice.

The blended approach

The development of the model used a blended approach. This was advocated for in Chapter 3 and involves a number of key components. Firstly, a blend of predictors should be selected and include both modifiable and non-modifiable to help increase clinical utility. Secondly, a blend of techniques should be used to identify these predictors. This includes obtaining clinical input alongside the use of machine learning feature selection algorithms. From a computational perspective, only two features (SOFA score and diagnosis of infection) of the developed model

were shown to have significantly high correlation to the output. When these two features proceeded to the training phase, they produced 100% sensitivity and specificity. From a computational perspective, this model performed very well; however, from a clinical perspective, this model was of little use (two features that are non-modifiable and diagnostic rather than prognostic). Thus, it was necessary for the development team (microbiologist, consultant surgeon, and researcher) to filter the features *before* introducing them to the machine learning algorithms (to establish which ones had the highest correlation to the output). Figures 11,12 and 13 demonstrate the changes in the Python script on the Swedish database to add this clinical input and the type of selected predictors to build the model, respectively. Thirdly, a blend of users perspectives (both patients and clinicians) which can inform the developing model and improve the learning healthcare systems (LHS).[184] These systems are typically defined as a group of professionals, stakeholders, or healthcare providers who perform activities and services to enable the continuous learning from routinely delivered care and improvement of provided services.[185] Although, LHS learn from data analysis, they are also driven by social interactions and collaborations which brings multidisciplinary professionals from different specialities to analyse and interpret the data. The input of clinicians can help develop an AI model that is more useful in clinical practice.

Figure 11. Selected features using the RFECV without clinical input:



0	Proven bacterial infection (yes/no)
1	SOFA score ≥ 2 (yes/no)

Figure 12. Changes in Python script to include clinical input:

```

1 import pickle
2
3 #Load the saved pickle file in variables directly (Features without any selection)
4 train_feats, test_feats, out_train, out_test, keys_of_feats = pickle.load(open("Train_Test.p", "rb"))
5 print(keys_of_feats)
6
7 #Add the medical keys in the given list to be selected
8 medical_keys = ["Oxygen saturation (%)", "Body temperature (oC)", "Systolic blood pressure (mmHg)", "Leukocyte particle concentration (x109 cells/L)", "C-reactive protein (
9
10 #Select the medical keys from train and test features
11 #List that will contains the number of columns that have the medical keys
12 selected_idx = []
13 for each_key in medical_keys:
14
15     if each_key in keys_of_feats:
16         # Get the index of the column corresponds to the key
17         idx = list(keys_of_feats).index(each_key)
18         selected_idx.append(idx)
19
20 train_feats_med_sel = train_feats[:, selected_idx]
21 test_feats_med_sel = test_feats[:, selected_idx]
22 print("Number of Medical Features :", len(train_feats_med_sel[0]))
23 pickle.dump( (train_feats_med_sel, test_feats_med_sel, out_train, out_test, medical_keys) , open( "Train_Test_Medical_Selected.p", "wb" ) )
24
25 import pandas as pd
26
27 #Load the saved pickle file in variables directly
28 train_feats_med_sel, test_feats_med_sel, out_train, out_test, keys_of_feats = pickle.load(open("Train_Test_Medical_Selected.p", "rb"))
29 num_selected_feats = 6
30

```

Figure 13. Selected predictors after using a blend of clinical input and machine learning algorithms:

Interactive-1.interactive > selected_keys

0	Age (years)
1	Systolic blood pressure (mmHg)
2	Respiratory rate (breaths/min)
3	Oxygen saturation (%)
4	HR
5	Body temperature (oC)
6	Haemoglobin (g/L)
7	C-reactive protein (mg/L)
8	Procalcitonin (ng/mL)
9	Neutrophil-lymphocyte count ratio

5.6. Limitations of the model development

The researcher acknowledges several limitations to the development and internal validation of the model. These relate to the dataset, data cleaning, predictor selection, clinical application and interface accessibility, each of which is discussed in turn.

The dataset: Despite its reported high quality, the database had a small percentage of weighted error and missing data, which may have affected the accuracy of the resulting model. Additionally, the high prevalence of patients with a history or current diagnosis of cancer in the dataset might have also influenced certain predictors, such as full blood count results, and hence the model's accuracy. To improve future external validation, a more heterogeneous patient cohort (i.e. better representation for ethnic minorities, different surgical specialities..etc.)

should be considered to avoid bias in clinical variables. Increasing the ratio from 1:1 to 1:4 can improve the precision [186]. However, increasing participants in the infected group in my study was challenging due to the retrospective nature of data collection, missing data issues, and specific inclusion criteria for participants.

Data cleaning: The filtering process to select patients based on inclusion criteria resulted in some subjects with all their variables missing. Complete case analysis was used to remove these subjects, which can introduce selection bias. However, this decision was made to ensure more accurate running of machine learning algorithms. The missing data within the selected variables were not missing at random, making traditional multiple imputation methods inapplicable. The use of the K-nearest neighbours machine learning algorithm for data imputation was an alternative approach in this context.

Predictor Selection: As previously explained, the researcher adopted a blended approach to select predictors, which incorporated input from clinicians along with utilising machine learning feature selection algorithms. It is important to note that different clinicians might have varying perspectives on the inclusion of certain predictors chosen by machine learning. These differences in clinical reasoning or concerns about the potential interference of machine learning in predictor selection, could also be presented a limitation.

Clinical application: While the model demonstrated high performance during development and internal validation, its performance would still need to be external validated in the clinical setting. Some sub-groups within the cohort might also have been underrepresented and could be at higher risk of infection, such as patients with low economic/educational level or those from ethnic minority backgrounds. Sub-group analysis for these populations was not possible as the data about educational and income levels were not available; 77.3% of the included cohort identified themselves as white British.[187] This raises the risk of health inequity of using AI tools due to the risk of racial bias, and flags the importance of developing strategies to improve the representation of underrepresented populations.

Interface accessibility: The model's interface which has been developed in the TRE, will need to be converted into a format that is accessible to users. However, this further development would require additional funding and resources.

5.7. Chapter Summary

In this chapter the researcher described how she trained and internally validated an AI-predictive model to predict the risk of post-operative infection in elective abdominal surgery

(AIPIS). The researcher demonstrated how she applied the key stages for AI-predictive models development as described in Chapter 3. The researcher also explained how she structured the blended approach and how it can be applied during the development of AI-predictive models. The researcher also explained the challenges and limitations of the model, highlighting the different sources of ROB in AIPIS which should be considered during the external validation.

In Chapter 6, the researcher will provide an overview on the tools available to assess ROB in AI-predictive models how to identify the different types of bias that developers need to consider when developing and deploying AI models.

Chapter 6. The Potential for Bias During the Development and Internal Validation of AI-based Decision Support Models.

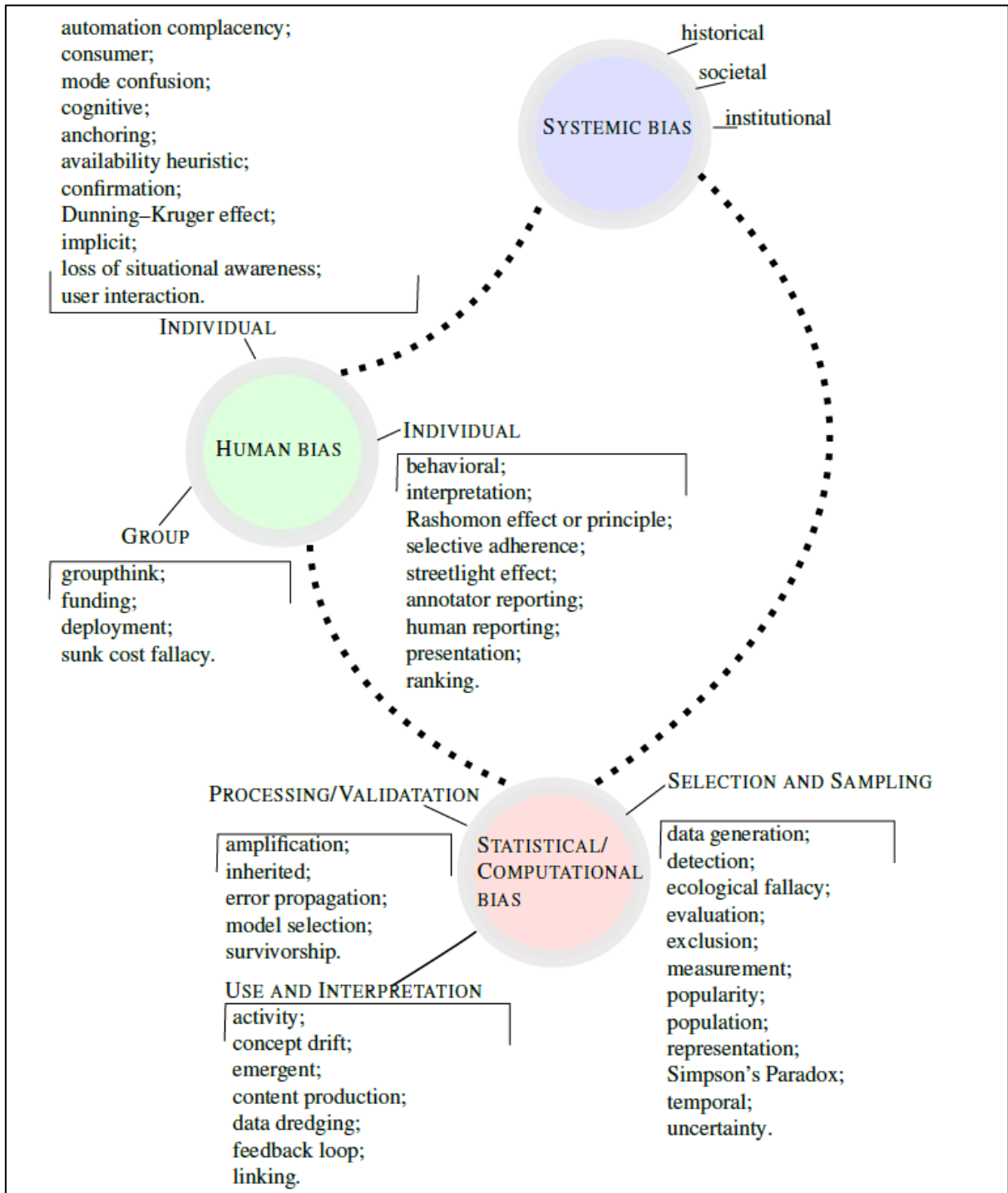
6.1. Introduction

In Chapter three, the researcher described the key stages of developing AI-predictive models, mentioning how bias is one of the key considerations during development of AI-predictive models (sections 3.2.2., 3.2.3, and 3.2.4). In Chapter five, the researcher mentioned the ROB in the limitations of the developed model (see section 5.6). In this chapter, the researcher outlines various types of bias that AI-predictive model developers should be aware of during the development and testing stages, as well as strategies for avoiding them.

6.2. Background

Concerns have been raised regarding the accuracy and potential bias of AI model interpretations, which could negatively impact the clinical care process.[188] Bias refers to a systematic error that produces inaccurate results and undermines the model's validity.[189] One such example could be a model predicting an inaccurate lower risk of morbidity or mortality in a certain patient population (a population that the model was not exposed to during its training). These inaccuracies can lead to a growing sense of mistrust among users, including patients, clinicians, and healthcare stakeholders.[190] There is a wide range of biases documented in the literature; the National Institute of Standards and Technology in the US classified bias in AI technologies into three main categories, which included systemic bias, human bias, and computational and statistical bias.[191] Figure 14 describes the different types of bias under each of these categories.[191]

Figure 14. Different categories of Bias in AI models:[191]



6.3. Different Types of Bias

6.3.1. Bias occurring during development of AI models

Socio-economic and Demographic Bias

Dataset bias emerges in the early stages of AI model development. This bias can manifest in various forms, with gender and race biases being the most prevalent.[192,193] For example, Caliskan et al. developed a machine learning algorithm from internet texts and demonstrated an

association between domestic abuse terms and females, as well as increasingly linked negative expressions to certain races, such as African-Americans, when compared to European-Americans.[194] During the training of AI models, large datasets are utilised, that contain the data of patients who have the condition of interest (i.e., the outcome). Machine learning algorithms generate predictions based on correlations between clinical variables and the outcome, meaning that the accuracy of these predictions relies heavily on the accuracy and diversity within the dataset.[192] Consequently, a biased dataset that disproportionately represents a minority group can lead to bias, commonly referred to as the "*garbage-in, garbage-out*" phenomenon.[195] Obermeyer et al. developed and deployed an AI model in an American healthcare system to inform healthcare decisions.[119] This algorithm displayed a racial bias by assuming that black patients with similar risks were healthier than equally sick white patients.[119] The authors attributed this bias to the fact that, in the training dataset, black patients had less insurance coverage than white patients.[119]

To mitigate the occurrence of dataset bias, three aspects should be considered when selecting the training dataset: data accuracy, validity, and completeness. Data accuracy can be enhanced during the data entry stage, by using objective data collection forms, for example. This has demonstrated a notable improvement in the quality and completeness of data entry.[196] Data validity can also be improved by monitoring data quality throughout the collection and entry process and by encouraging clinicians to utilise standardised medical coding, such as SNOMED and ICD codes, which have shown higher accuracy compared to free text data entry.[197] Incomplete data can amplify ROB due to under-representation of specific populations.

Predictors and Incorporation Bias

Incorporation bias can manifest in three different forms during data preparation and model training.[176] The first form occurs when the variables used to define the outcome of interest are also included as predictors. This dual usage of variables as defining criteria and predictors introduces biased correlations between these predictors and the outcome, compared to other predictors.[198] For example, using SOFA score as a predictor for sepsis, while it is already used for the definition of sepsis. The second form of incorporation bias arises when the correlation between predictors and the outcome is evaluated without prior knowledge of the outcome. Mitigating this form of bias involves masking the outcome from the model development team while assessing the correlation between predictors and the outcome.[199] The third form of incorporation bias pertains to the clinical application of the model, where selected predictors should be available at the time of application.[68] To mitigate incorporation bias, it is essential to exercise clinical judgment in the selection of predictors based on their

correlation with the outcome. Clinical judgment should consider the accuracy and rationale behind the clinical definition of the outcome, the nature of the variables being examined as predictors (modifiable or non-modifiable), and how these variables can enhance the clinical applicability of the tools.[170]

6.3.2. Bias occurring during validation and testing of AI-predictive models

Selective Bias

Selective bias can occur during the development and validation stages of AI models. It arises when a particular patient population is under-represented in the dataset due to biased participant recruitment. For example, a systematic review of 83 different AI models developed for the early detection of ovarian cancer revealed how selective sampling through interrupted recruitment of participants resulting in bias.[200] Data completeness and availability can influence selective inclusion, further contributing to selective bias. This bias can have a negative impact on predictive performance since data are often missing in a selective pattern.[201] Reducing the incidence of selective bias is crucial to enhance the trustworthiness of AI models among users.[202] This can be achieved by addressing the bias associated with underrepresentation in patient recruitment and ensuring that the developed models are used in the same patient population for which they were developed in.[202]

Verification Bias

Verification bias, also known as "work-up" bias, can occur during the testing phase of AI diagnostic predictive models. In order to assess diagnostic accuracy, both a gold standard tool and the diagnostic model are applied to a patient cohort, and the positive diagnoses from both the tool and model compared.[203] Verification bias arises when only a subset of participants undergoes testing with the gold standard tool, while others do not, leading to a bias towards more confirmed diagnoses with the diagnostic model. There are two types of verification bias: partial and differential. Partial verification bias occurs during model testing and can impact accuracy by increasing sensitivity and decreasing specificity, resulting in an underestimation of false negatives when only a subset of patients receives the gold standard tool.[204]

Differential bias, on the other hand, involves the use of two different gold standard tools, which reduces the number of confirmed diagnoses with each tool, leading to a higher bias towards confirmed diagnoses with the diagnostic model. The extent of differential bias depends on the relative accuracy of the two gold standard tests.[204] An example of verification bias can be observed in a study evaluating the accuracy of D-dimer tests for diagnosing deep vein thrombosis (DVT). In this study, Büller et al. describe how patients with a positive D-dimer test underwent a venous ultrasound (first gold standard test), while those with a negative D-

dimer test received clinical follow-up after three months (second gold standard test in clinical practice). Consequently, false negative results for D-dimer may go undiagnosed during the follow-up period due to symptom improvement, leading to an underestimation of false negatives and an overestimation of sensitivity.[205]

To mitigate verification bias, it is crucial to test both the diagnostic predictive model and the gold standard tool on the exact same patient cohort, using only one gold standard test whenever possible.[203] In some cases, it may not be feasible to avoid verification bias due to invasiveness or cost constraints associated with the gold standard tool, such as angiography. In such instances, statistical approaches can be employed to reclassify patients with negative results into the false negative group.[206]

6.3.3. *Bias occurring during the application of AI-predictive models*

Inherent/Latent Bias

Latent bias can manifest during the application of AI models in clinical practice. In this type of bias, the model may become biased over time.[207] Chua et al. developed an AI model that predicted mortality in palliative care, where factors such as socioeconomic or ethnic variations impacted on the accuracy of mortality predictions over time and subsequently influenced clinical decisions.[72] Even if developers choose to exclude such variables from the training datasets, latent bias can still occur if other available variables are correlated with the omitted variables.[207]

Latent bias can particularly arise in a busy clinical environment in the form of automation bias or privilege bias. Automation bias occurs when users, such as clinicians and/or patients, have high expectations of AI models' predictive performance, considering the predictions as unquestionable, and following them "without question" rather than viewing them as "decision-support" tools.[208] The risk of automation bias increases when clinicians face time pressure during consultations or prioritize medico-legal protection.[207] Privilege bias can occur when AI model predictions privilege one population over another. This bias is directly linked to dataset bias, and a prominent example is racial bias, especially if the dataset is derived from a single setting or racial population.[209]

To mitigate latent bias, it is crucial to address it from the early stages of model training by ensuring that the dataset is representative and diverse. Additionally, users should be informed about the existence of latent bias so that they understand that AI model predictions are not infallible due to the potential bias inherent in historical data.[207]

6.4. Tools to assess bias

The PROBAST can be used to identify bias in diagnostic and prognostic predictive models and targets all stages of predictive model development, including training, validation, testing, and maintenance.[176] PROBAST consists of 20 questions, categorised into four sections focusing on participants (patient cohort), predictors, outcome (condition of interest), and data analysis.[176] The first section (patient cohort) assesses the data source(s) used to train the model. Prospective longitudinal studies are associated with low ROB, while the use of data registries and retrospective cohort data increases ROB because the data was not specifically collected for model development.[210] This section also evaluates the appropriateness of the inclusion and exclusion criteria used to prepare the dataset.[176] The second section (predictors) focuses on how predictors were measured and defined. To reduce ROB, predictors should be evaluated without prior knowledge of the positive or negative outcome.[211] Another factor that can lead to ROB is the availability of predictors when the model is running in an active clinical environment. If the developed model includes predictors that might not be available at the time of using the model in practice, it can introduce bias.[212] The third section (outcome) highlights how there should be a pre-determined definition for the outcome and how it is appropriately determined.[170] Ideally, this definition should be standardised through clinical guidelines or an objective tool/score.[213] The outcome should also be determined without prior knowledge of the available predictors. Finally, the duration between capturing predictors and determining the outcome should be appropriate to the severity of the condition (i.e., chronic versus acute condition).[176] The fourth section of PROBAST concerns data analysis and model training. This section focuses on nine different aspects, from the prevalence of the outcome in the training dataset to the way missing data is handled.[118] PROBAST assigns a score to each section using "Yes/No/Unknown", with the overall risk considered low if for example all four sections are rated as low ROB.[176]

The PROBAST-AI is currently under development and will provide guidelines specifically designed for AI-based clinical predictive models.[214] It will be developed through three Delphi survey studies, with the third study currently being performed.[215] PROBAST is not used for studies determining prognostic factors which focuses solely on the identification of the optimal predictors to extract from datasets to train AI-predictive models; instead, the Quality In Prognosis Studies (QUIPS) tool should be used.[216]

The QUIPS tool focuses on six major aspects for the identification of prognostic factors (i.e. predictors) including: study participation, study attrition, prognostic factor measurement, confounding measurement and account, outcome measurement, and analysis and

reporting.[216] The QUIPS uses these six domains to categorise ROB in prognostic factors into low, moderate and high ROB. Starting with the study participation, the sample included should be representative to the wider population to reduce ROB. A study would be judged as having high ROB if the sample size is small, it only focused on one sex or age group, or was very selective in the way participants were recruitment.[216] Study attrition is mainly focused on participants follow-up. If there was a high drop-out rate, this can increase ROB. Prognostic factors measurement domain ensures that the assessment of prognostic factors was similar among all participants. This also requires continuous variables to be reported or at least have clear cut points described, and that any missing data for prognostic factors are appropriately imputed.[217] Regarding the outcome measurement, a consistent method of assessment among all participants is required. This includes a clear definition of the outcome of interest. Study confounders is the fifth major domain in the QUIPS tool, which should be appropriately accounted for in the data analysis. Finally, the statistical analysis should be appropriate for the study design, type of variables, and comparisons, with all primary outcomes stated.[217]

6.5. Governmental Guidance on the detection and mitigation of Bias in AI technologies

The detection and mitigation of bias in AI technology have been under the spotlight in the UK over the last three years. The government launched its consultation ‘Data: a new direction’ on 10 September 2021 to inform its development of proposals to reform the UK’s data protection laws, to secure a pro-growth and trusted data regime as part of the UK’s National Data Strategy.[218] In some cases, access to identifying or personal data may be necessary to mitigate bias, such as reducing societal and demographic among different ethnic groups. This raises legal and ethical considerations for governments, as granting access to such data is a controversial area. Another recent report from Imperial College London that addressed racial and ethnic inequalities in AI technologies also supported the development of appropriate governance systems for healthcare data protection and data exchange, not only within the UK but also throughout Europe, especially post-Brexit.[219] This report described key areas for future research and consideration given to governance systems to potentially mitigate ROB in data-driven technologies. These key areas involve the development of comprehensive frameworks that accurately examine these technologies to ensure an appropriate description of the datasets used in training these tools and their representativeness of different populations, especially minority groups. One industrial developer commented:

“It comes back to regulation. If the FDA says to me, ‘you are about to go build an algorithm that you are going to want to test to predict sepsis, I am going to want to see the validity of your algorithm across these ethnic

groups', then as a developer I would probably start thinking about that day one because I know that is already a regulatory requirement." [219]

Other areas include being open and transparent about measures taken by developers to reduce bias in AI technology, particularly against minority groups. Additionally, it is important to ensure that the tools have been validated and tested across these groups.[219] The report also strongly recommended the future development of a guide for NHS organizations to facilitate the use of AI tools in supporting the decision-making process while auditing the risk of inequity associated with these tools.[219]

6.6. Chapter Summary

Completely eliminating bias from AI systems is challenging and likely to be unrealistic, given that AI algorithms learn from historical data, which is inherently bias.[220] However, developers should strive to minimise bias as much as possible, as discussed earlier. The first step in reducing the impact of bias is to acknowledge its presence. Users and stakeholders should be informed about any unavoidable ROB that may occur at any stage of developing AI-based models, from dataset preparation to clinical application. Users should also be provided with information on how the model was developed and how bias was managed.[191] Governments and national healthcare systems should provide guidance and implement regulations for auditing bias in AI while mitigating its risk, particularly for the most vulnerable groups. In the next chapters, the researcher will focus on users' views of the application of AI tools more broadly to support SDM with the purpose of understanding how to improve usability, accessibility, and trustworthiness of the use of these tools.

Chapter 7. What are Clinicians' and Patients' Perceptions on Using Artificial Intelligence Decision Aids to inform Shared Decision Making?; A systematic review.

7.1. Introduction

In Chapter Six, the researcher described different examples of bias that might occur during the development and evaluation of AI-predictive models. In this Chapter, the researcher conducted a systematic review of the literature to understand clinicians' and patients' perspectives on using AI decision aids to inform SDM at different stages of the patient journey and in various clinical settings. The findings from this chapter won the best oral presentation prize in the Faculty of Clinical Informatics Annual Conference (2021) and has also been presented in the Public Health UK Conference 2021 and published in the *Lancet* : Hassan N, Slight RD, Bimpong K, Weiland D, Vellinga A, Morgan G, Slight SP. *Clinicians' and patients' perceptions of the use of artificial intelligence decision aids to inform shared decision making: a systematic review*. The Lancet. 2021 Nov 1;398:S80. (Appendix I). A manuscript of this Chapter has been submitted to the *Annals of Internal Medicine Journal* currently under review.

7.2. Background

Shared decision making represents a collaborative process through which a clinician supports a patient to make a decision that is right for them.[221] These decisions can occur at different stages of the patient journey such as, for example, when deciding on a suitable treatment option.[67] There are different types of decision aids that are frequently used to assist SDM.[222] Some use in-built AI functionality to predict the likelihood of a disease occurring or progressing,[223] or provide tailored information on the risk of complication(s) after undergoing a particular medical procedure.[224]

It is important to understand patients' and clinicians' perspectives on the use of these AI decision aids.[225] These include the benefits of these tools, such as allowing more time for doctor-patient discussion, and reducing anxiety among surgical patients by providing detailed personalised information about the expected procedure and post-surgical recovery time.[226] However, AI decision aids also have some limitations such as, for example, that patients and/or clinicians may be unable to understand the presented information, or the decision aid itself cannot be integrated with larger electronic healthcare systems.[227] the researcher conducted a systematic review of the literature to understand clinicians' and patients' perspectives on using AI decision aids to support SDM at different stages of the patient journey and in various clinical settings.

7.3. Methods

This review was registered with PROSPERO (CRD42020220320) and aligned with the PRISMA guidelines.[228]

7.3.1. Eligibility Criteria

All quantitative, qualitative and mixed-method studies that included the perspectives of clinicians (including physicians, pharmacists, and nurses) and/or patients on the use of AI decision aids to support SDM were eligible for inclusion. the researcher defined an AI decision aid as any AI-based tool that processed individualised patient information and produced tailored recommendation(s), by considering their preferences, medical history and current clinical conditions.[229] AI-based tools were developed using machine learning algorithms that identified patterns and correlations between different variables and the outcome of interest. Any studies that discussed decision aids without an AI component were excluded. Only articles published in the English language were included, with no restrictions on the medical speciality, clinical setting, or period over which these decision aids were used.

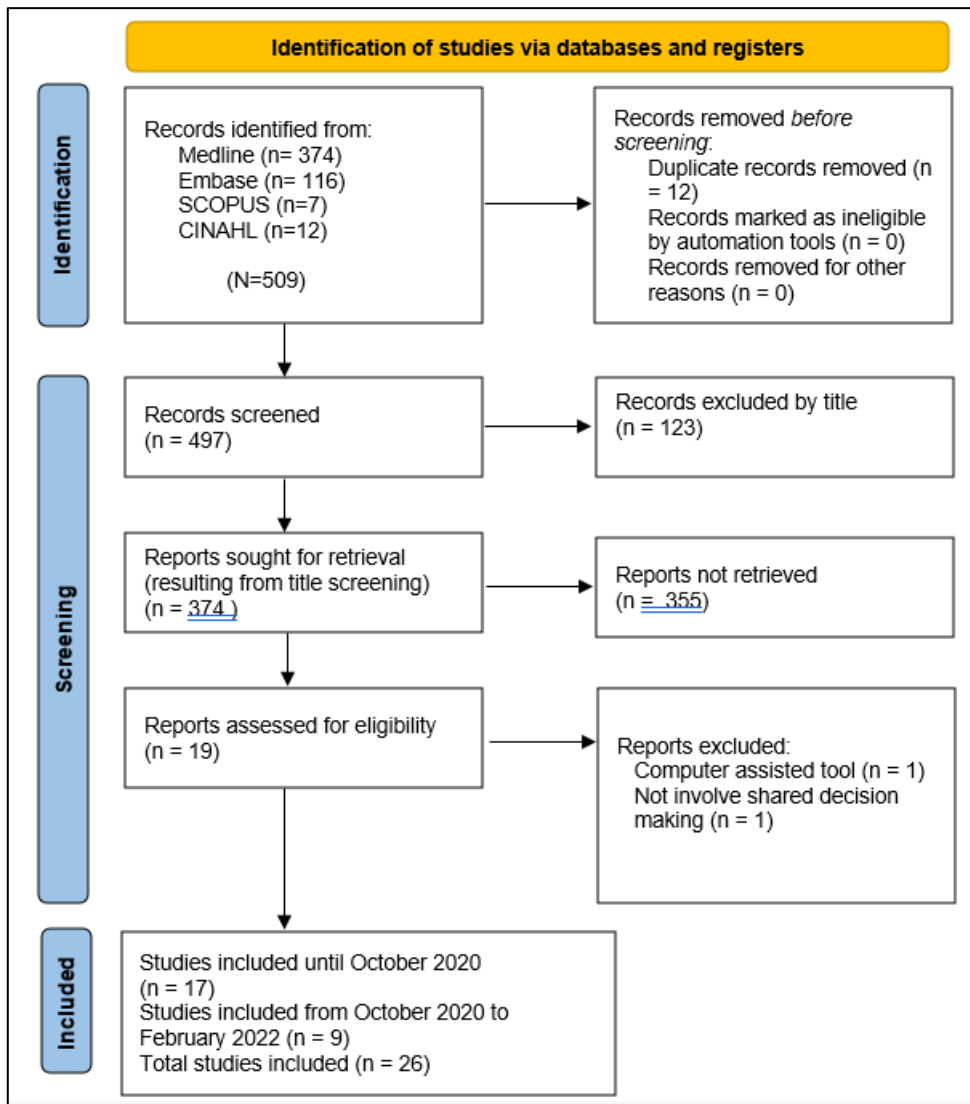
7.3.2. Information sources and search strategy

Four large databases were searched, including MEDLINE and Embase (via Ovid), Cumulative Index to Nursing and Allied Health Literature (CINAHL), and SCOPUS. The search was carried out across two timeframes to include published articles from the date of commencement of each database up until Oct 2020, and updated again from Oct 2020 to February 2022. Relevant keywords relating to perceptions, personnel (clinicians/patients), AI, algorithms/decision aids, and SDM were used and grouped into sets. Each set was combined with an “OR” operator, followed by combining all results with an “AND” operator.

7.3.3. Study selection

Duplicates were removed using Endnote software (version X9, Clarivate Analytics, Philadelphia, US) and manually checked. The corresponding authors of two abstracts were contacted to ask whether there was a complete manuscript available for their studies. Two independent reviewers (NH and KB) screened all titles, abstracts and full texts to determine if they met the inclusion criteria, with any disagreements resolved by discussion with a third reviewer (SPS), if necessary. The reason(s) for excluding any full-text articles were also documented. Inter-rater reliability among the independent reviewers was found to be 100% after discussing disagreements with a third reviewer (SPS). References of the included full-text studies were also screened to identify any potential articles. The search strategy is described in Figure 15.

Figure 15. Search strategy PRISMA flowchart:



7.3.4. Data extraction and analysis

A customised data-extraction sheet was developed to extract specific details from each study, including study title; year of publication; country; publication type; aim; design; type of user (patients/healthcare professionals); methods; ethical approval; population; clinical setting; medical/surgical speciality; name, type and use of the AI decision aids; users' perceptions; inclusion and exclusion criteria; method of recruitment; and key conclusions. Quantitative data on the number of study participants, percent with a particular perception, and/or the percent who were satisfied / dissatisfied was also extracted.

Data extraction followed the ENhancing Transparency in REporting the synthesis of Qualitative research (ENTREQ) guidelines.[230]. A narrative synthesis was carried out using the Research Methods Programme framework (ESRC), which accommodates the inclusion of a wide range of research designs.[231] The extracted data was organised and decision aids grouped by their

main purpose and where in the patient pathway they were used. Clinicians' and patients' perceptions on these different AI decision aids were then compared.

7.3.5. Quality assessment

The quality of included articles was assessed using the Critical Appraisal Skills Programme (CASP) tool for qualitative studies.[232] The tool consists of ten questions, with each "Yes" answer given one point. The total score was calculated. All included studies had a score of seven or higher on the CASP checklist. No studies were excluded at the stage of quality assessment.

7.4. Results

A total of 509 research articles were identified in October 2020, and duplicates (n=12) removed. Articles were screened at the title (n=497), abstract (n=374), and full-text (n=19) stages, with 17 articles eligible for inclusion.[69,70,233-246] This search was updated in February 2022 and an additional nine articles included, giving a final total of 26 articles.[247-255] These included studies had either quantitative, qualitative or mixed methods (i.e., model development and perceptions exploration). Nine qualitative studies used either face-to-face interviews,[234,235,242,248,249] focus groups,[240,243] or a combination of interviews and telephone follow-up calls.[236,252] Fifteen studies used surveys,[69,226,237-239,241,244-247,250,251,253-255], nine of which collected text answers.[69,238,241,244,246,247,250,253,255] Two mixed-method studies used a combination of survey and semi-structured interviews.[233,246]

In terms of the AI decision aids, 12 studies developed and/or evaluated web-based tools,[70,226,236,237,239,240,244,247-251] three studies utilised smart phone and tablet applications,[252-254] and 11 studies evaluated aids that were developed and embedded in larger electronic systems such as Cerner.[69,233-235,238,241-243,245,246,255] Table 7 illustrates the name, type, and purpose of each AI decision aid. Ten studies explored patients' perceptions only,[226,234,236,247,248,250-254] eight studies clinicians' perceptions only,[233,237,238, 239,240,242,243,255] and the remaining eight studies both clinicians' and patients' perceptions.[69,70,235,241, 244-246, 249].

Table 7. Description of AI decision aids:

Study	Name	Type	Medical speciality	Clinical setting	Purpose
Manski-Nankervis et al. [233]	No name specified	Embedded in electronic health system	Infectious diseases	Primary care	To guide the treatment of infections (e.g., antibiotic choice, dose, duration)
Nelson et al. [234]	Direct-to-patient AI tool (DTP AI tool)	web-based	Dermatology	General dermatology clinic	To provide the patient with information on skin cancer screening options
Qu et al. [235]	SMILE	web-based	Dermatology	Lupus clinic	To educate patients about Lupus treatment during a regular clinic visit
Raymond et al. [226]	Surgical Risk Calculator (ACS Calculator)	web-based	Surgery	Pre-operative clinic	To provide patient-specific risks for common adverse perioperative outcomes
Eckman et al. [236]	Atrial Fibrillation Shared Decision Making Tool (AFSDM)	web-based	Cardiology	Arrhythmia clinic	To generate patient-level recommendations for thromboprophylaxis
Ballard et al.[237]	Statin Choice Decision Aid	web-based	Internal medicine (Diabetes)	Mayo Clinic,	To estimate 10-year cardiovascular

	(SCDA) and the Diabetes Medication Choice Decision Aid (DMCDA)			an academic tertiary healthcare centre	risk, the degree of risk reduction with a statin, and the likelihood of adverse events (SCDA) To estimate the impact of different diabetes medications on daily routine, blood sugar control, risk of hypoglycemia, weight change, and cost (DMCDA)
Brown et al.[238]	Clinical prediction models (CPM)	Non-cardiovascular CPMs (different types: Web-based/Software/embedded into systems)	Family medicine	North West England, NHS North Western Deanery	Prognosis of non-cardiovascular long-term conditions
Silvestrin et al.[70]	Women's Health Assessment Tool/ Clinical Decision Support toolkit (WHAT/CDS toolkit)	web-based	Women's health	Two primary care and one gynaecology clinics	To assess depression, vulvovaginal atrophy and provide tailored evidence-based treatment recommendations
Flynn et al.[69]	Computerised decision Aid for Stroke thrombolysis (COMPASS)	iPad based software	Neurology (Stroke)	Stroke units	To predict the patient-specific probability of acute stroke outcomes at three months, with and without thrombolysis

Schroy III et al.[239]	CRC screening decision aids	web-based	Oncology (Colorectal cancer)	Primary care (screening) clinic	To convey key information about colorectal cancer, and the importance of screening
Jimbo et al.[240]	CRC screening decision aids	web-based	Oncology (Colorectal cancer)	Community practice	To inform decision making around the preferred colorectal cancer screening method)
Schackmann et al.[241]	Decision tool for women with BRCA mutation	Embedded into electronic system	Oncology (Breast cancer)	Community practice	To predict the risk of breast cancer and mortality in BRCA1/2 mutation carriers
Cunich et al.[242]	ALProst embedded in Annalisa (AL)	web-based	Oncology (Prostate cancer)	Primary care	To predict prognosis and diagnosis of prostate cancer from prostate antigen test and patient's age.
Harrison et al.[243]	Annalisa, Computer decision aid and paper-based decision aid (3 decision aids)	web-based	Surgery	University of Sydney and Sydney South West Area Health Service	To calculate the risk estimates of surgical treatment of rectal cancer in terms of complications and outcomes
Aoki et al.[244]	u-SHARE	web-based	Neurosurgery	Neurosurgery clinics	To provide favourable treatment options, and critical factors for decision making on unruptured intracranial Aneurysms.
Siminoff et al.[245]	Adjuvant!	embedded into electronic system	Oncology (Breast cancer)	Community Practices	To assist in adjunctive therapy choices for breast cancer.

Thomson et al.[246]	DARTs tool	embedded into electronic system	Neurology (stroke prevention)	General practitioner practices	To support SDM in atrial fibrillation, deciding on whether to take warfarin to prevent stroke.
Berry et al.[247]	Personal Patient Profile- Prostate (P3P)	web-based	Urologic oncology (prostate cancer)	Urology and radiation outpatient clinics	To guide patient-provider communication tailored to race and age.
Jones et al.[248]	RealRisks	web-based	Oncology (Breast cancer)	Community Practices	To promote a woman's understanding of her breast cancer risk and to engage women in planning a tailored chemoprevention plan.
Kim et al.[249]	myAID	Web-based, Australianised version of original US decision aid (UCTO-Ulcerative Colitis Treatment Options)	Gastroenterology (Ulcerative colitis)	Outpatient clinic	To facilitate and support SDM for treatment decisions in ulcerative colitis pharmacological management
Kosch et al.[250]	Evidence-Based Decision Support Tool in Multiple Sclerosis' (EBDiMS)	web-based	Neurology (Multiple Sclerosis)	Multiple Sclerosis Day clinic	To provide an estimation of the individualised long-term prognosis (up to 30 years)
Lau et al.[251]	shouldiscreen.com	web-based	Oncology (Lung Cancer)	Community-based organisations	To provide basic information about low-dose computed tomography screening, education about lung cancer risk factors, and

					a lung cancer risk calculator that computes a personalised risk
Kuppermann et al.[252]	PROCEED	tablet-based	Gynaecology (delivery mode and approach)	prenatal clinics	To guide decision making on trial of labor after caesarean (TOLAC) versus caesarean section.
Weismann et al.[253]	No name specified	hypothetical models to be embedded into electronic records	Respiratory (Chronic Lung Disease)	outpatient pulmonary clinic	To estimate future risk of in-hospital death with advanced chronic lung disease.
Wilson et al.[254]	No name specified	Smart phone application	Internal medicine (Diabetes)	academic medical centre's diabetes clinic	To guide decisions on treatment for type 1 diabetes
Coylewright et al.[255]	HealthDecision	Embedded in electronic health system	Multiple	Outpatient clinic	To calculate and display personalised risk estimates for cardiovascular risk, stroke prevention for atrial fibrillation, fracture prevention in osteoporosis, and breast and lung cancer screening)

The majority of studies (n=22) were conducted in a medical setting,[69,70,233-241,245-251,253-255] and four were done in a surgical setting.[226,243,244,252] All AI decision aids were used to guide SDM process at different stages of the patient journey, including screening and prevention (risk prediction),[234,236,238-240,242,248,251,255] prognosis,[226,237,241,242,244,250,253] and treatment,[69,70,233,235,237,243,245-247,249,252,254] as illustrated in Figure 2. the researcher will discuss each of these different stages in further detail below.

Figure 16. Role of AI tools to guide SDM:

Study	Screening and Prevention	Prognosis (risk assessment/prediction)	Treatment
Aoki et al.[244]			
Ballard et al.[237]			
Berry et al.[247]			
Brown et al.[238]			
Coylewright et al.[255]			
Cunich et al.[242]			
Eckman et al.[236]			
Flynn et al.[69]			
Harrison et al.[243]			
Jimbo et al.[240]			
Jones et al.[248]			
Kim et al.[249]			
Kosch et al.[250]			
Kuppermann et al.[252]			
Lau et al.[251]			
Manski-Nankervis et al.[233]			
Nelson et al.[234]			
Qu et al. [235]			
Raymond et al.[226]			
Schroy III et al.[239]			
Schackmann et al.[241]			
Silvestrin et al.[70]			
Siminoff et al.[245]			
Thomson et al.[246]			
Weismann et al.[253]			
Wilson et al.[254]			

7.4.1. Screening and prevention (risk prediction)

Ten studies evaluated the perceptions of either patients and/or clinicians on AI decision aids for screening (n=6),[234,239,240,242,251,255] and prevention (n=4).[236,238,248,255]

Screening

The *Direct-To-Patient (DTP)* AI tool scanned images of skin lesions submitted by patients before attending their general dermatology clinics in Boston, Massachusetts, US, and estimated their likelihood of having skin cancer.[234] Sixty percent (n=29) of patients felt that *DTP*

increased diagnostic speed, with one patient explaining how *“If the app says, ‘you probably have melanoma—go see your doctor,’ they might actually get in there sooner...so it could be lifesaving”*. [234] However, 85% (n=41) expressed their concerns about the accuracy of the tool when coupled with the lack of a physical examination. One patient highlighted how: *“[AI] is not a substitute for [an] in-person exam. We’re examining a photograph that you get with your phone in variable light. It’s possible that you might get a false-positive or a false-negative.”* [234] A second AI decision aid, *ALProst®*, offered personalised insights on the benefits and risks of prostate cancer screening for an individual patient, based on their prostate-specific antigen (PSA) test result and age. [242] Clinicians described how this information was displayed in an understandable format, which helped patients engage with the tool and gave them a sense of ownership. [242] However, 80% (n=8) of clinicians reported challenges with its use, including how the tool did not incorporate patient-specific cardiovascular risk factors. [242] A second AI decision aid in the area of cancer screening compared various colorectal cancer screening options for patients based on their cost, discomfort, inaccuracy, and requirement for additional tests. This decision aid also obtained the patient’s preferred screening option and shared this with their clinician prior to their consultation. [240] Clinicians perceived an increase in patients’ satisfaction amongst those who completed the tool; however, others expressed their concerns about patients using the tool in the waiting area, in particular patients’ digital literacy, the potential for theft and damage to the device that the patient is using, and a knock-on delay to them attending their appointments. [240]

“[many patients have] no comfort with computers, and I can’t see more than perhaps 15, 20% of our total patient population even being able to access that. And of that, [we would be lucky] if we had 5 or 7% that would actually do it to the level that you are discussing....[F]or a certain group of patients that would be a wonderful tool. I don’t think it would be for our group of patients primarily because we have a large percentage of low functioning patients.” [240]

Prevention

The Atrial Fibrillation Shared Decision making aid recommended the most suitable type of thromboprophylaxis (i.e. warfarin versus DOACs) for individual patients. It also presented information on intracranial and extracranial bleeding risks, which patients found helpful and promoted their adherence to recommended treatments (p-value< 0.0001) after one month. [236] The web-based AI-decision aid, *RealRisks®*, provided women with a personalised risk score of

developing breast cancer over the next 5-years, along with chemo-preventive measures to print out and discuss with their healthcare provider during their consultation.[219] Although all participants (n=21) deemed this tool "*helpful*" and "*acceptable*", 43% (n=9) of patients requested clearer, jargon-free information so as to improve their understanding, and felt that *RealRisks*® put a heavy emphasis on chemoprevention.[248] The aid was also perceived as tailored to Hispanic/Latina individuals, with (43%, n=9) of participants expressing their concerns about the tool's suitability for other ethnicities.[248]

Coylewright et al.[255] examined clinicians' perceptions of an integrated patient decision aid (*iPDA*), which predicted the risk of a patient developing various conditions (atrial fibrillation, cardiovascular disease, fractures with osteoporosis, and lung and breast cancer) based on their demographics, medical history, and preference for preventive measures.[255] Over 80% (n=266) of clinicians used *iPDA* in the last three months to guide their patient-clinician conversations and found it time-saving.[255] However, family physicians shared their concerns about another AI decision aid, which predicted the risk of a patient developing non-cardiac conditions, explaining how the tool could influence over- or under-treatment of patients, mainly due to specific risk factors not being considered.[238] These physicians also felt the tool could increase anxiety amongst patients who had a high mortality or morbidity risk, and so in-turn might have chosen not to receive the information.[238]

7.4.2. Prognosis

Five studies incorporated AI decision aids to predict disease prognosis and/or the likelihood of a particular complication(s).[226,241,244,250,253] The Evidence-Based Decision Support Tool in Multiple Sclerosis (*EBDiMS*) provided information on the long-term prognosis of multiple sclerosis (over 30 years), which participants perceived as "*relevant*" and "*highly understandable*".[250] Ninety-five percent (n=85) of participants would recommend this decision aid to others, with 54% (n=48) highlighting how they would have liked earlier access to the tool itself.[250] A second web-based Surgical Risk Calculator (*ACS Calculator*) predicted post-operative complications, with 90% (n=135) of patients perceiving it as particularly useful during the surgical consenting process.[226] Seventy percent (n=105) of patients were willing to participate in a pre-habilitation programme to lower peri-operative risk after obtaining their risk score, with 38.6% (n=58) considering postponing their surgery until after they had participated in this programme.[226]

The u-SHARE decision aid also provided patients with a personalised risk of post-operative complications following surgery, when compared to watchful waiting.[244] Over half of the

patients (n=8) were comfortable using the tool. However, acceptance appeared to be higher amongst patients than clinicians, with clinicians perceiving the decision aid as a "*structured informed consent tool*".[244] In contrast, oncologists reported higher satisfaction and acceptance with another AI decision aid that provided information about different breast cancer mutations and preventative measures than patients. Oncologists perceived the use of this decision aid improved their interaction and communication with patients during the consultation, with patients finding illustrative graphics and visual aids particularly beneficial.[241]

7.4.3. Treatment

Twelve studies explored patients' and clinicians' perceptions of AI decision aids used for pharmacological or non-pharmacological decisions (i.e. surgical procedure). [69,70,231,233,235,241,243-245,247,250,252] The COMPASS decision aid and Decision Analysis in Routine Treatment Study (DARTs) tool outlined the benefits and risks of thrombolytic therapy.[69,246] While patients found these personalised risk estimates helpful, some had negative feelings around receiving such information.[69] Clinicians noted that patients' health literacy levels, personalities, and preferences influenced their engagement with COMPASS and the SDM process.[69] Some clinicians also expressed reservations about whether DARTs's recommendations incorporated the latest guidance.[246] Similar concerns were raised about the information provided by the Diabetes Medication Choice Decision Aid (DMCDA), which clinicians perceived as somewhat non-relevant and not evidence-based.[237] Only 28% (n=30) of clinicians routinely used the DMCDA tool compared to 80.6% (n=75) who used the Statin Choice Decision Aid (SCDA), a decision aid that also supplied personalised benefits and risks of statin therapy.[237]

The AI decision aid *Annalisa*® presented patients with personalised risk estimates of developing side-effects from drug treatment (medical option) for rectal cancer versus complications from post-surgery (surgical option).[243] Patients felt that *Annalisa*® presented the information clearly and improved their understanding of their clinical condition, in contrast to 25% (n=5) of colorectal surgeons who perceived *Annalisa*® as lacking specific information. Surgeons also raised concerns about the time needed for patients to use the tool and felt it was "*designed for those who speak English*".[243]

Another decision aid, which displayed estimates of a patient's 10-year mortality risk with a surgical intervention versus non-surgical intervention for breast cancer (*Adjuvant*®), was perceived by 48% (n=169) of patients as guiding their decision making on adjuvant therapy and

by 75% (n=43) of clinicians as helping them develop more individualised treatment protocols for their patients.[245] Similarly, the Women's Health Assessment Tool/Clinical Decision Support toolkit (*WHAT*), which offered evidence-based treatments for menopausal symptoms, was perceived by 68% (n=75) of patients as improving the quality of healthcare provided and by 25% (n=3) of clinicians as helping them to “*get right to the point and the patient feels valued. And they feel a part of their healthcare (...) plan*”. [70] GPs (n=7) found another simulated AI decision aid valuable as they could print out the specific guideline on antibiotic treatment and share it with the patient during their consultation. One GP felt that the tool was less valuable and only used it when they were uncertain about the treatment: “*you’re basically looking at the computer while you’re talking to a person in there, and so you look a bit like an idiot, really.*” [233]

Another AI decision aid, which provided clinicians with comprehensive information on the different formulations, costs, doses and side effects of immunosuppressive agents for the treatment of lupus, was found to be useful during consultations as it helped encourage patients to discuss any concerns about their treatment and feel more confident about their treatment plan, potentially improving adherence.[235] However, 48% (n=12) of participants (including physicians, clinic champions and patient advocates) raised concerns about privacy and confidentiality when asked to complete this tool in the patient waiting room.[235]

Finally, AI decision aid in the form of smartphone application alerted type 1 diabetic patients either when their insulin doses were due and/or 30 minutes before a hypoglycaemia event was likely to occur. Over 78% (n=1202) and 64% (n=986) of patients used the tool to guide rapid-acting and long-acting insulin doses, respectively.[254] Although 91% (n=1403) of patients were particularly interested in the ability of the tool to help them avoid hypoglycaemia, 85% (n=1310) asked if the hypoglycaemia alert feature could be further customised based on their blood glucose levels over time.[254]

7.5. Discussion

The researcher explored how patients and clinicians perceived the use of AI decision aids in supporting SDM. Twenty-six articles were included,[69,70,233-255] the majority of which focused on either patients and/or clinicians’ perceptions of using decision aids in the medical setting (n=22). AI-based decision aids were either web-based tools, smart phone applications or integrated within larger electronic systems, and used at different stages of the patient journey, including screening and prevention (risk prediction), prognosis, and treatment. Patients generally found the AI decision aids easy to understand and user-friendly. This helped increase

their engagement with the tool, fostering a sense of ownership and promoting adherence to recommended treatments.[236,242,243,250] Some clinicians perceived AI decision aids facilitated the patient consent process for procedures or treatments,[226,244] and felt patients' interaction with tools was influenced by their health literacy, personalities, and preferences.[69,255] There was a recognised ROB associated with AI decision aids with variations in patients' usage, possibly due to their different levels of health and technological literacy.[69,235,236,248,249] Clinicians expressed concerns about how up-to-date the decision aid information was,[246] and the potential for over- or under-treatment.[234,239,250] The included studies used AI-decision aids mostly for screening purposes, with more perceived challenges reported with their application in treatment and prevention.

A recent scoping review supported these findings reporting patients' positive perceptions on the ease of use of AI tools in general and their potential to enhance efficiency.[256] Another review on the application of AI decision aids specifically to support SDM described how these aids are mainly applied in primary and secondary care settings, rather than tertiary care settings. However, this review included only six studies, and did not explore users' perceptions.[257] In the US, a recent systematic review examined the attitudes of the public on using AI more broadly in healthcare and highlighted similar concerns around data confidentiality and the use and security of their healthcare data, as highlighted in this review.[258]

As AI decision aids become integrated into various aspects of healthcare, it is crucial that they do not perpetuate any existing health inequities for minority ethnic groups. Johnes et al. highlighted how an AI decision aid tested among Hispanic/Latina individuals might lead to racial bias by excluding other ethnic groups.[248] Jimbo et al. also showed that with a web-based tool, concerns were raised by clinicians about the affordability of these devices alongside patients' ability to access the internet and their familiarity with technology. These barriers can contribute to unequal access to healthcare, worse experiences of healthcare, and wider inequities in society.[240]

The suitability of the information displayed by decision aids was also raised,[236,248] with Jones et al. and Eckman et al. suggesting that the display be customised to accommodate the needs of elderly patients and those with visual impairments. Patients found the illustrative graphics and visual aids for a prognostic decision aid particularly beneficial.[241] Patients were able to print out their personalised risk score using the web-based AI-decision aid, *RealRisks*®, and discuss this with their healthcare provider during their consultation; however, this assumes that patients had a suitable device and also had access to a printer. Devices were suggested to

be available for patients in clinics in Jimbo et al. to overcome the barrier of access. However, concerns were raised by the participants about patient privacy and confidentiality in completing the information in waiting rooms,[235] and about the possibility of theft and damage to these devices in another study.[240] In a study by Harrison et al. surgeons could only use a tool if they were proficient in English, excluding those who spoke other languages.[243] From these findings, it is clear that AI decision aids must be designed to prioritise inclusivity, accommodating all patients and clinicians, irrespective of their educational, technological, socioeconomic, or health literacy levels.[69,235,236,248,249]

Predicting the risk or prognosis of a condition generated anxiety for patients in some included studies.[234,238,250] Clinicians' worried about increasing patients' anxiety, especially amongst those with a high risk of mortality or morbidity.[238] Patients proposed that a hybrid approach be used, where patients used the decision aid with their clinician during the consultation so as to help reduce their anxiety.[234] This was described by Nelson et al. as a symbiosis between AI and human clinician.[234] The transparency and explainability of AI models is a central component of the Code of Conduct for Data-Driven Health and Care Technology.[259] It is important to understand the real impact of data-driven technologies on patients' lives, and also have clear expectations regarding the explainability of outputs which are understood by clinicians before, during and after deployment.[219]

Clinicians perceived some AI decision aids e.g. ACS calculator, u-SHARE as assisting with the patient consent process for surgical procedures; it provided a more accurate personalised risk of post-operative complications following surgery, which was felt to improve patients understanding.[226,244] One hundred and twenty-two patients expressed a desire to use the ACS calculator prior to surgery, with fifty eight considering postponing their surgery until after they had participation in the prehabilitation programme to help reduce their post-operative complication risk.[226] Through reduction of post-operative complications, these tools could potentially reduce the length of hospitalisation and subsequently healthcare costs.

This review highlighted how patients' and clinicians' perceptions appeared to vary on the same tool. Although patients often found AI decision aids user-friendly and valuable,[69,244,246,248,249] clinicians expressed reservations about their ease-of-use and whether the information was up-to-date.[69,235,236] AI decisions aids need to be regularly updated once in the clinical setting, both patients and clinicians informed of any updates to guidance so as to increase their trust in the tool.[170] The included studies in this review had some limitations. Some clinicians' perspectives were provided on simulated tools, such as the

tool used to aid antibiotic prescribing for GPs,[233] which may have performed differently when deployed in clinical settings.[233,253] Most of the studies (n=22) included in this review were obtained from the use of tools in the medical setting. None of the included studies conducted a formal ROB assessment, such as the PROBAST tool, when developing their tools.[260] Although the PROBAST tool was more recently published,[260] users need to be made aware of any bias(es) and its impact on outputs. For example, some training datasets might contain data that is unrepresentative or contain socio-historical bias in how the data was entered and collected. Further research is needed to explore how to mitigate the impact of inherent bias in training datasets of AI tools to improve their clinical application and inclusivity to all patient populations. Only studies that included the perspectives of clinicians and/or patients on the use of AI decision aids were included, with the actual development and testing of these decision-aids outside the scope of this review. Evaluating the influence of AI decision aids on SDM through clinical outcomes in both medical and surgical settings also necessitates real-world testing in clinical practice.

7.6. Chapter Summary

In this chapter, the researcher explored the perceptions of clinicians and patients towards using AI decision aids to inform SDM. Patients found AI decision aids to be understandable, user-friendly, and empowering. These tools are likely to have a positive impact on patients' adherence and help inform the shared decision process. However, certain barriers were recognised by both clinicians and patients, most importantly concerns related to bias, anxiety, and privacy. These aspects must be carefully considered during the development and deployment of future AI-based decision aids. In the next Chapter, the researcher conducted a qualitative study the perceptions of clinicians from one English NHS Trust regarding the use of AI-based tools broadly to inform SDM in the clinical environment.

Chapter 8. “This isn't to replace you, this isn't to tell you what to do, this is to help you make decisions, so you are making the decisions with your patients”;

Clinicians' Perceptions Around the Use of AI-based Technologies to Inform Shared Decision Making.

8.1. Introduction

In Chapter Seven, the researcher described how important it is to obtain the perceptions of healthcare professionals towards developing AI tools to inform SDM. In this Chapter, the researcher will describe a qualitative study that she performed using semi-structured interviews with clinicians from different clinical specialties at one English NHS trust to explore their perceptions of using AI-based tools to inform SDM in the clinical environment. The researcher used a reflexive thematic analysis to develop the themes. The findings from this Chapter were presented in the Faculty of Clinical Informatics Conference 2022 and Published in the BMJ Health and Care Informatics Journal: Hassan N, Slight R, Slight S. *Healthcare staff perceptions on using artificial intelligence predictive tools: a qualitative study*. BMJ Health & Care Informatics. 2022:A4-5. (Appendix J).

8.2. Background

The use of AI is a rapidly progressing research area in healthcare. Over the last few years, there has been an exponential increase in the number of different AI technologies developed, such as radiographic scans, robotic surgery, and a range of AI-based predictive models and decision aids.[261] As discussed in Chapter Seven, AI-based technologies can provide information on disease screening, prevention, prognosis, risk prediction and treatment. This information can help optimise the decision making around patient care, improve the efficiency of healthcare services, and better facilitate staff workflow in an already overloaded clinical environment.[261]

The promotion of AI applications in healthcare institutions has inspired healthcare researchers to explore the impact and the performance of different AI tools in clinical practice.[263] However, evaluating the impact of these novel AI-based technologies requires an understanding of the challenges relating to the deployment of these technologies into clinical care, and most importantly, whether the users (both patients and clinicians) perceive potential benefits.[264] The NHS has dedicated £123 million for increased implementation of AI in the national healthcare system,[265] which is described by the NHS Long Term Plan as modernising, and

cost saving technology [266] However, there is a conflict of opinion in the literature amongst users of the healthcare service around the acceptance of these AI technologies into healthcare, with some concerns raised around data privacy and security, and trusting the outputs of these tools.[68] This qualitative study aims to understand the perceptions of clinicians from one English NHS Trust regarding the use of AI-based technologies to inform their SDM. These perceptions can, in turn, potentially inform the development of more user-friendly, better clinically applicable and impactful AI technologies.

8.3. Methods

8.3.1. Participant recruitment

Clinicians were recruited from two hospitals within the Newcastle Upon Tyne Hospital NHS Foundation Trust: the Royal Victoria Infirmary and the Freeman hospitals. A convenient sampling technique was used to select participants. Invitations for participation were sent out to staff (approximately 50) in various specialties through emails and social media platforms such as Twitter. The choice of specialties was based on findings from a systematic review conducted by the research team, which identified the key specialties where AI-based tools had already been deployed in an attempt to support SDM.[68]

8.3.2. Data collection

Semi-structured interviews were conducted with clinicians from different specialties between July 2021 and November 2021. During these interviews, clinicians were asked open-ended questions that allowed them to share their own experiences. Informed consent was obtained, and the interviews were recorded using an audio-recorder. The interviews were conducted via Zoom, utilising a licensed account from Newcastle University. The interview guide (*Appendix K*) was developed after reviewing existing literature on the perceptions of patients and clinicians regarding the use of AI tools in the SDM process.[68] The interview questions focused on three main aspects: (a) exploring the participants' understanding of AI and how AI could assist them in their daily tasks, (b) the significance of communicating risk estimates to patients, and (c) the barriers and facilitators of using AI-based technologies in clinical practice. The Participant Pack included a Participant Information Sheet (*Appendix L*) and a Consent Form (*Appendix M*), and was provided to each clinician who expressed an interest in participating. Prior to each interview, the researcher ensured that the participant had signed the consent form, either during the interview or by returning it via email. The researcher also utilised an interview reflection

log (*Appendix N*), which included additional prompts or questions considered relevant for subsequent interviews.

Transcription of the interviews was promptly carried out after each interview by the researcher, and the transcripts checked for accuracy. Each participant was assigned a unique ID code, combining letters and numbers to indicate their sequential participation and specialty (e.g., 02 Cardio referred to the second participant in the Cardiology department). Any identifiable information was removed during the transcription process to ensure confidentiality.

8.3.3. Data Collection and analysis

Qualitative data collection and analysis followed a reflexive thematic analysis approach, [267] allowing emerging themes to be explored and disconfirming evidence to be sought on the enablers and barriers of using AI-based technologies. Main themes and subthemes were identified using constant comparative analysis. [267] The ‘constant comparison’ technique was used to move backwards and forwards between the data and evolving explanations for the recurring patterns and associations. Themes were discussed among team members, and continually refined and applied systematically to the whole dataset using computerised software Nvivo [QSR International, Melbourne]. Thematic saturation guided the volume of qualitative data collection.

8.4. Results

8.4.1. Descriptive analysis

Ten clinicians from three different professional backgrounds (medical, pharmaceutical and surgical) and nine departments/clinical specialities volunteered to participate in the study. Clinicians had varied experience of clinical practice from between five to twenty years, and half of them got an academic affiliation. Table (8) demonstrates the description of the cohort.

Table 8. Descriptive analysis:

		Count (Percent)
Profession	Physicians	7(70%)
	Surgeons	2(20%)
	Pharmacists	1(10%)
Speciality	Cardiology	2(20%)
	Oncology	1(10%)
	Endocrinology	1(10%)
	Microbiology	1(10%)
	Radiology	1(10%)
	Paediatric Infectious diseases	1(10%)
	Clinical Pharmacy/Informatics	1(10%)
	Head and Neck Surgery	1(10%)
	Transplant Surgery	1(10%)
Years of experience	5 to 10	4(40%)
	10 to 15	2(20%)
	15 to 20	4(40%)
Academic affiliation	Yes	5(50%)
	No	5(50%)
Number of patients seen/week	Less than 50	2(20%)
	50 to 100	8(80%)

Five key themes emerged from the data collected. The themes focused on how the participants understood AI in the general context and in clinical practice, providing some examples of current tools. Additionally, factors influencing the trustworthiness of clinicians in AI were also discussed. The clinicians also expressed their thoughts on the features that they believed can potentially improve the usability of the tools. In addition, to the barriers to the adoption and use of AI, most notably, resistance among clinicians to adopt the tools and overreliance on AI tools. Moreover, the future potential use of AI tools was explored.

8.4.2. Theme 1: Participants' understanding of AI

There was a lack of agreed understanding about the meaning of the term 'AI' amongst clinicians. Some participants described how they did “*not have a huge amount of experience with AI*”.(09-Surgery) Whilst others did not “*use artificial intelligence very much at all at the*

moment”(08-Micro) The participants referred to the role of AI in a general sense, using phrases like *“a big block of data”*(01-Pharma) and how *“advanced technologies and robots”* could be used *“in general to help the human beings, whether to speed up processes or to do a job faster and more efficient”*.(04-Cardio) One participant described how AI could be considered to be the effective use of computers to *“create subset of combined sets [of data] and then start working through those”* (01-Pharma) to develop correlations between data variables and identify patterns or predictors of an outcome.

From a clinical perspective, AI was referred to as being a *“clever tool”* (05-ID) used to *“detect patterns or detect predictors of a particular endpoint”*(01-Pharma), through smart and efficient scanning of multiple variables on EHRs. The data could then be used to alert clinicians about high-risk patients who required immediate attention, for example, *“notifying doctors about something wrong”* (03-Surgery). They also commented on areas in which they could see AI being used in medicine, such as in radiology where they were *“starting to replace some of the human reporting [of radiographic scans] by the artificial intelligence programmes”*(10-Onco).

Examples of Current tools

The participants mentioned different digital tools that they use in their daily practice, perceiving some of them as potentially developed with an AI technology. The tools were classified into two types; the first type related to image recognition feature of AI, describing imaging during coronary interventions which can show coronary lesions with high accuracy *“Using artificial intelligence in some radiological images such as the size of that coronary artery, how big the stents will [need to] be, and the length of the stents”* (02-Cardio), and *“using artificial intelligence to try to report on stress echocardiography [an ultrasound imaging investigation to scan the heart during exercising over a treadmill]”* (04-Cardio). The second type related to diagnostic and prognostic tools. One participant described how diagnostic tools can identify high risk patients for infection:

“We looked at the fever kids tool that [name removed] developed, and that was published a couple of years ago (...) we looked at sort of the tool that has developed at that time, and we looked at all the high risk patients that came through our specialist services to see if the tool that he developed was also applicable to that population, and fortunately, it seems to work rather well.”(05-ID)

Other diagnostic tool that was discussed by one participant was used during acute admission to identify patients who may have a possible blood clot:

“Commonly used in any acute admission units and we do follow the score, so if you score for an above ten, yes, we would actually send the patient to get a CT scan of the long arteries, to see if there's a clot”. (07-Endo)

Another participant referred to a predictive tool that could help provide an estimated risk of the patient developing breast cancer over the coming years and the possible benefits of different therapies.

“In breast cancer, we have some predictive tools, we can just put information in the formula or calculator and that will give us an idea about the future five years, 10 years, 15 years, how's the risk is going to be, and what's the additional benefit of each therapy.” (10-Onco)

8.4.3. Theme 2: Trustworthiness of the tool

Trust in the ability of the tool was described as a “big issue” (01-Pharma, 09-Surgery) and a key consideration for the subsequent adoption of these tools in clinical practice. One participant explained how *“it's about how you involve the [clinical] team in the development of the AI tool, and how they think they're going to act”* (05-ID). The involvement of clinicians in the early development of these tools was perceived as important, as it could help to build their trust in the tool and increase their understanding of the evidence behind the outcomes:

“There's a whole lot of products out there, that will tell us whether the patient's osteoporotic or not, but if we don't have the evidence about outcomes, and whether changing treatments has an impact, then we shouldn't be doing that” (06-Radio)

However, one participant acknowledged how some clinicians might not welcome participation in the development of these tools, fearing that these tools might replace them in the future. One participant described how it is important that clinicians understand how these tools are not being developed to “replace” clinicians, but to aid with their decision making:

“It has to be the sell to a doctor, you need them [clinicians] on board with it, you need to pander to the fragile ego, this isn't to replace you, this isn't to tell you what to do, this is to help you make decisions, so you are making the decisions with your patients, this is to help you do that, it isn't to make decisions for you, it isn't to be listened to, it's an adjunct” (09-Surgery)

One participants felt it was important to know more about the development and testing of the AI tool, so as to understand how relevant the output was to the type of patient that they were treating: *“need to understand if the tools are [were] applicable to the patient in front of us, so if it's not being tested in our 85 year old because they weren't included in the trial design”*.(06-Radio) One participant went further and explained how he would like to know whether the AI tool had been *“developed by using patient samples that are very similar to what is seen in our setting. If that is not the case that would be a barrier”*(05-ID). Ultimately, this was important in order to be able to judge the appropriateness of the tools for patients that he was caring for.

“it needs to be tested for the local population, but, like any intervention in medicine, so you probably need some sort of evidence to show that it works in a setting where you're trying to use it” (06-Radio)

Clinicians also felt it was important that the risk of error within the tool (i.e., its accuracy) was clearly communicated to both clinicians and patients. One participant described how patients might not know this risk of error, and how this risk might not be entirely accurate:

“we're [clinicians are] aware that things can go to different way, but I'm not sure the patients would think of that, so I think it should be clear with the patients, rather than doctors, that this is what we think the risk is going to be, this is a predictive tool, which helps you understand your risks, however, sometimes may deviate and obviously there's always a standard deviation.” (07-Endo)

Validation, maintenance and updating of the developed AI tool were also considered important, and factors that could impact the trustworthiness of the outputs. Clinicians wanted reassurance that AI tools were *“tested to be certain that there is safety”* (04-Cardio) and *“to make sure we're developing tools that are useful, whether that is safety or efficiency or effectiveness”*.(06-Radio) This testing or external validation should be conducted in a clinical setting: *“validate any new tools with our usual practice and to see how much is accurate before implementing the new tool”*.(10-Onco)

Clinicians expressed how AI tools should be continually maintained and updated after deployment. This could likely involve feeding them with real-time patient data, so that *“they are almost like living things”*.(09-Surgery) and was believed to *“make the technology better”*.(05-Micro) Another participant described how the tools being *“up to date is really important, and if you're able to feed in patients and their data, so it's continually updated, it's important”*.(09-Surgery) Participants specifically wanted the tools to be *“developed based on*

solid clinical evidence”,(02-Cardio) and be “*licensed and approved from different licensing bodies, such as CE, NICE, and FDA*”.(06-Radio) This was perceived to increase the trustworthiness of the outputs of such tools.

8.4.4. Theme 3: Usability Features

Participants described some features that they believed could improve the usability of the tools in a clinical environment. These features related to (a) the screen design where data was required to be inputted, (b) the recommendations/outputs presented to the users, and (c) the integration of AI tools within electronic systems. The participants highlighted how the data inputs for the tool should be clear, and “*the questions need to be easy to answer*”(07-Endo) to encourage and facilitate clinicians use of it, especially in a busy clinical environment. Any unclear questions, which may be open to misinterpretation, could lead to clinicians entering inaccurate data, that is likely to influence the accuracy of the output. Furthermore, the data inputs should also be flexible to the user to accommodate possible borderline values: “*for example, sometimes some tools would ask you if the patient is,..., let's say hypertensive and the patient is that borderline, and sometimes you know if you say yes, it changes it completely, so it just needs to be a bit more flexible*”.(07-Endo)

In terms of the tool outputs, one participant described how “*the end result somehow has to be intelligent in itself, so that it's not just noise in the background*”.(08-Micro) Another participant expressed a similar view explaining how “*unnecessary pop-ups create noise in the background, which becomes distracting for clinicians*”.(07-Endo) Intelligent tools should be able to identify true positives and only inform clinicians about real risks; otherwise, clinicians may simply ignore the outputs of these tools. One participant highlighted the intrusive nature of some tools and how more work might be needed to ensure that the interface is usable alongside other decision support systems:

“It [AI tool] had a bit of a bad reputation with prescribers because it's very intrusive, so I think doing some of that more targeted work with a limited range of decision support is where we would want to get to” (01-Pharma)

Some participants suggested that integrating these AI tools into current electronic systems can be “*more effective and efficient*”,(06-Radio) whereas others described how it may “*increase workload*”(06-Radio) if there are multiple different technologies or softwares that you need to navigate through.

“If we start adding layers of complexity, where you've got your scans on your PACS and then you need to go off and get a different software to do an osteoporotic risk calculator, different software to do cardiac

calcium score, it becomes unmanageable. So we need to make sure that it's well integrated, efficient and effective.” (06-Radio)

The tool should be accessible and intuitive to users, requiring little effort to run: *“accessible so that it’s not hidden away in the options menu of a system, so it needs to be running fairly easily in the background, without someone having to do something to make it run in the first place.”(08-Micro)*. The outputs should *“individualise the information to the patient, in a lay person’s language” (04-Cardio)*, Participants also suggested that the AI tools provide related reminders in addition to the outputs (i.e., reminders about subsequent clinical appointment, screening, medication time, etc.), especially if these reminders are on their mobile phones.

“Sometimes patients might find it more attractive if you have this in a very good and attractive way on their mobile phone, so they have these recommendations with them all the time and then this tool acts like a reminder.” (04-Cardio)

Seeking the input of users from an early stage of tool development was perceived as important, because *“you really want to know what they want, what benefit they want to, how they think they're going to act”.* (05-ID)

8.4.5. Theme 4: Barriers to the adoption and use of AI tools in clinical practice

Resistance and Changing Culture

The participants agreed that adaptation to any new innovation and adopting it in to practice can be challenging. The adoption of these tools involved changing workflow like, for example, in radiology *“they will need to review and confirm and sign the reports of each scan”.*(10-Onco). Clinicians would also need to learn how to use the tool and sometimes there can be resistance from those who are from an ‘older generation’, with one participant commenting on how the ‘younger generation’ might welcome new technologies:

“The old generation, not all of them [are] familiar with the technology and it will be difficult for them to re-learn these new things, so I suspect there will be resistance from the senior doctors, and the young fresh doctors, I think, they will welcome these new technologies, because it will make their life easier and also it's handy.”(10-Onco)

In addition, one microbiologist elaborated on how the resistance from senior doctors can impact the adoption of new technologies quite considerably, explaining how: *“if the leaders are not*

comfortable with a new technology, whatever it may be, then it's unlikely to be implemented locally. No matter how good the evidence basis for that technology".(08-Micro)

It is also important to ensure that non-rotational staff who are assigned to a particular hospital ward are trained and supported to use new AI tools, as this was perceived by one participant as potentially creating a 'presence' in these departments and further support to rotational staff should they need it.

"Don't underestimate the effect of junior staff who do a lot of the leg work on the ground, changing over every six months, and if they're not used to working with artificial intelligence tools you are back training up a new cohort of people every six months, basically. Using nurses who have gotten more continued availability and presence in these departments, is going to be beneficial. You need to embed it in culture, but you need to make sure that you have staff involved engaged and supportive of it, who will stay on the unit that you're working on".(05-ID)

Another factor that was perceived about adopting the AI-tools by one participant was the perceived accuracy of the tool and the information that will be shared with the patient. One surgeon explained how the technology might provide a higher percent death rate from a surgical procedure than what he believed the lower percent to be. He would prefer that the patient was not provided with this contradictory information:

"they [surgeons] go: ' yeah Okay, whoever made this little iPad, yeah, they might get a 70% death rate, but in my hands the death rates a lot lower, so tell me what, we'll just turn this iPad off, shall we, and you stick [to] the conversation between you and me, how about that?' And I hear people say it. 'In my hands'. it's different story. " (09-Surgery)

Considering all different aspects of the patient

Another aspect of the AI tool, which was highlighted as important by more than one participant, was its ability to assess the 'first impression' or 'gut feeling' experienced by the clinician on physically meeting with the patient. One participant described how *"the gut feeling of an experienced clinician, tends to be quite a good denominator as to what is going on with the patient"*(05-ID). Another participant explained how *"AI cannot capture the first impression that clinician has, which we all know, is a huge determinate about what someone is going to do"*(05-ID), this can lead to clinicians being *"disempowered"*.(09-Surgery).

“So that's a worry, and for doctors as well, that's disempowerment. The gut feeling, this guy is not going to do well. Computer says he might do well, I don't feel like he's going to do well. How'd you interact that gut feeling with this, it's difficult, what if this [computer] has got a different opinion to you?. How much you rely on this? and how much you rely on you?” (09-Surgery)

This participant went further to say how important it was to understand the uncertainty around a particular patient journey. He described how, even if the AI tool provided a clinician with percent chance of survival of a cancer patient over the next five years, it was not personalised enough to consider the complications of different procedures that they might have to undergo along the way, and whether the patient will be in the percent getting the complication or those without the complication:

“My worry is that it [AI tool] doesn't eliminate uncertainty, you don't get rid of all uncertainty just by providing a huge amount of data for something. So if, for example, you're able to say beyond all reasonable doubt that a patient with a T3N1 laryngeal cancer has a 65% chance of surviving five years. And the big data is telling us that, and the chance of them having a complication is 2% and that's fine and that's a useful piece of information. That's sorted, that's where I feel AI can help us get to, but that still doesn't eliminate uncertainty for that patient, because they don't know if they're in the 65%, and they don't know the effect that those things[complications] will have on them.” (09-Surgery)

Overreliance of the AI technology

Another concern raised by one participant was the overreliance that some doctors might place on AI tools. She highlighted how some more junior members of staff might rely on the tool, possibly becoming misled by the outputs; she felt that this was less likely to happen with more experienced clinicians: *“It might mislead, maybe, some junior doctors or doctors who would not have that much of a clinical sense and would totally depend on it. Maybe that can be a barrier, but I think for someone who's experienced and wise, this tool can only be like an added insight or support or help or maybe another way to look at something.” (07-Endo)*

Clinicians' overreliance on these AI tools was reflected on by one surgeon who worried that some clinicians might direct patients away to another location, the waiting room for example, to use and read additional information provided by the AI tool. This could potentially reduce the amount of time that there was for a direct patient-clinician discussion about the related issue, which he viewed as a possible negative aspect of the use of AI-tools.

“So abusive them, the people going: ‘Right, I’m not going to talk to you, here is all the information, you have a look through that, and just tell me what you think. So, here’s your iPad, you just scroll through, spend as much time as you want it to look through, will sit down in the waiting room, then come back in here and tell what you think of it.’ That’s one thing that worries me.” (09-Surgery)

One participant recounted how *“there are some programmes online where patients can actually put in their symptoms and they can actually get diagnoses.”* (07-Endo) She reflected on this had worried her as patients might consider this diagnosis to be correct; she felt it would be advisable to obtain a clinical assessment alongside the use of the tool, describing how it should be *“clarified to patients that sometimes you need a clinical assessment alongside with these AI-predictive tools”* (07-Endo); these tools were not a substitute for clinical assessment.

Using data with the correct intent

One participant emphasized the importance of communicating to the patients the purpose of using their healthcare data in developing AI tools; *“we really need to focus on that communication about what’s the intent for the data”* (01-Pharma). The participant explained further that only the essential healthcare data should be used to avoid any unnecessary risk of data breach, hence it is specifically challenging *“to share GP information with hospitals and hospital information with GPs, that is a very big real concern”*. (01-Pharma) He also explained that risks to healthcare data security is a common problem to have where *“we do see lots of cases where the technology just doesn’t work, or information is lost or it’s the wrong result and I think those things are very highly reported”* (01-Pharma)

8.4.6. Theme 5: Future potential uses of AI

One participant suggested that AI technology should be *“implemented in a sort of piecemeal fashion, slowly over time, in areas where it’s likely to be safe to use it to begin with, where it’s not so controversial”*. (08-Micro) Another participant commented on how AI could *“move healthcare away from big, centralised hospitals and more into a patient environment”* (01-Pharma). However, this participant also cautioned against this was the risk that clinicians could become more isolated when they *“work more independently on a virtual environment”* (01-Pharma), hence, a *“hybrid approach”* (01-Pharma) incorporating both in-person and virtual might be considered. Participants felt AI based tools could play a role in disease prediction, diagnosis, and screening. One participant highlighted how AI could be useful at predicting deterioration in a patient’s condition early: *“it would be nice to have something that could*

predict by about day five or six before you reach that day, eight to 10 where the deterioration tends to take place, what direction they might travel, you know, are these ones that you can safely send home if they are admitted for just observation” (05-ID). In her opinion, this could give healthcare staff the opportunity to treat them earlier, which would hopefully improve their health-related outcomes: *“because we do know that if they deteriorate, treating them early gives a better result than treating later on” (05-ID)* This participant also described how AI tools could have a role to play in:

“predicting how they [Patients] are going to respond to treatment, or predicting if they might respond better to steroids or immunoglobulin, [...] because at the moment what you get is largely determined by the preference of the doctor or the experience they've had thus far, or are you being treated by Rheumatologists you're more likely to get immunoglobulin, than you are, when you have another general sub specialty.” (05-ID)

Artificial intelligence might also help in the diagnosis of cancers through better detection of abnormalities in radiographic scans. One radiologist described how an AI tool may show up ‘incidental’ findings, i.e., unrelated to the primary problem, from a scan:

“Another big thing in radiology is the incidental findings. Again an opportunity, almost for health screening.”” you could end up with a single scan of the body having multiple different risk assessments and opportunities for intervention.” (06-Radio)

Artificial intelligence could potentially provide better interpretation of these incidental findings, which might *“be quite a powerful tool to facilitate patients to change behaviours”*. (06-Radio) This radiologist described how you could use these results to explain to the patient:

“Listen, there's the impact of smoking on your vessels on your lungs, this gives you a coronary age X, which is 10 years older than your biological age, and that might facilitate change behaviour” (06-Radio)

Other potential use of AI tools to inform diagnosis, was to guide clinicians to request the appropriate clinical test for a patient and interpret the test results.

“Clinicians often don't know what to request, and how to request it and when to request it, and they may not even be aware that they should be doing a test, and so I guess artificial intelligence could happen that way. And then often they don't really know how to interpret a test result either so a lot of times junior

doctors request the test, but they don't actually expect to be able to handle the outcome”(08-Micro)

For screening, AI can be a tool to initially screen data, such as tissues and biopsies, in a pathology lab and help to possibly: *“differentiate between like cancer, non-cancer, at least, initial screening”*. This participant described how AI tools may be able to save both clinicians and patients time by analysing and interpreting the findings from investigations carried out:

“Our clinic is overloaded with unnecessarily investigations done which tend to be normal for example, and the patient is coming just to hear from the doctor how he can interpret these investigations. If we have well validated tools established, then no need for all these busy clinics just you can give access to the patient, to read the report and to be notified, at the same time. Such simple idea can save a lot of time for patient care.”
(10-Onco)

8.5. Discussion

This qualitative study examined clinicians' perceptions regarding the use of various AI-based tools to inform SDM in clinical environments. Some participants provided examples from their practice that highlighted potential uses of AI technology, particularly in areas such as disease prediction and radiographic imaging. Clinicians also raised concerns about trusting the output from these tools and proposed key factors that could be considered to enhance trustworthiness and adoption of these tools in practice. It was felt that clinicians should be involved in the early stages of tool development, ensuring that they understand the clinical reasoning behind their development and how they might be useful in practice. AI tools should be trained on data derived from patient populations similar to those for which the tools are intended to be used. Additionally, the risk of errors should be acknowledged to users. Clinical validity, maintenance, and licensing were also identified as crucial factors for clinicians to confidently adopt and use these tools.

Participants also suggested features that could enhance the user experience, including ease of data input, flexibility, and interpretability of data outputs. Integrating these tools with current electronic systems was seen as a way of facilitating their use and accessibility, though participants cautioned against excessive use of pop-up alerts that could lead to alert fatigue among clinicians. Barriers to clinical applications were discussed, with the most significant one being the need to change behaviors when adopting novel technologies, especially among senior generations. Additionally, conflicts between clinicians' professionalism and the tools'

recommendations could introduce uncertainty, potentially impacting the decision making process. There were also concerns that overreliance on AI tools might result in missing the ‘first impression’ or ‘gut feeling’ experienced by the clinician on physically meeting with the patient, and/or leaving patients to make decisions without clinician intervention. Data privacy and confidentiality were also considered potential threats to the successful application of AI tools in clinical practice. Despite these challenges, clinicians acknowledged the positive potential of these tools for the future, particularly in the areas of disease prediction and prevention, better interpretation of incidental findings in radiographic scans, and faster and more efficient cancer detection.

Trustworthiness towards AI tools was also considered a key factor influencing the successful implementation of AI tools in clinical practice, as revealed through interviews with clinicians. The researcher’s findings highlighted how these related to the tools themselves, such as how they were developed, validated, maintained and licensed. Other trust-related factors were linked to the users and their readiness to change, with the adoption of such novel technology. Similarly, Antes et al. [268] described how the public's trust is closely intertwined with their trust in the healthcare system and the AI tools themselves. In their study, Antes et al. surveyed 936 adults to gauge their perceptions of using AI tools in clinical practice. The aim was to understand the factors influencing these perceptions, with the authors highlighting how individuals need to see the benefits and transparency of these tools. Conversely, trust was negatively associated with concerns about using the tools.[268] Unlike clinicians, who may exhibit resistance towards adopting these tools, one patient in Antes et al. study nicely described how: *“I usually trust a technology until it gives me a reason not to trust it.”* [268]

Another study surveyed 307 members of the public in the US to explore their perceptions towards AI tools in healthcare, with trust towards AI tools emerging as a central theme. It was found to be interconnected with three factors including data privacy concerns, mistrust in the mechanisms of developing AI tools, and social bias that could be associated with these tools.[269] Data privacy concerns primarily revolved around healthcare data access, how this data is used and protected from any unlawful use. These concerns resonate with the data confidentiality concerns raised in our present study, where clinicians identified the risk of data breaches as a significant barrier to the clinical application of AI tools. Furthermore, Esmaeilzadeh [269] highlighted the mistrust in AI mechanisms as users often lack confidence in AI tool recommendations because they lack a clear understanding of how these recommendations are generated. In line with this, our study also revealed that clinicians expressed a desire to be involved in the early stages of AI tool development, aiming to gain a

deeper understanding of the clinical validity of the recommendations generated by these tools. In Esmaeilzadeh study,[269] the public also raised concerns around the risk of societal bias that could affect their trust towards AI tools use, due to the unrepresentative datasets used to develop these tools that might exclude some vulnerable groups (i.e., ethnic minorities). Although societal bias was not raised by the clinicians in our interviews, the clinicians called for the data to represent the population that they would most likely be using it for.

Some clinicians raised concerns about the conflict that could arise between the recommendations of AI tools and their own “gut feeling”, which could potentially misguide the SDM process. This dilemma prompted questions about whether clinicians should rely on their “gut feeling” or adhere to the recommendations of technology. For some clinicians, this raised fears of being replaced by these tools. However, our findings emphasize the importance of clinicians understanding that these tools serve as adjuncts to clinical consultations, intended to support the SDM process rather than replace their clinical judgment. It is imperative to address these concerns and make it clear that AI tools are designed to enhance, not substitute, the role of the clinician in decision making. These concerns were also echoed in a recent review comprising 27 studies that examined stakeholders' attitudes toward the application of AI in clinical practice.[202] Both clinicians and consumers expressed being more comfortable to use AI tools that interpret and synthesize data, rather than those that directly influence their decision making process.[202]

It is important to acknowledge some study limitations. The clinicians in this study were interviewed shortly after the easing of pandemic lockdowns (commencing in July 2021). While this was a challenging time for healthcare systems worldwide, the clinicians recognised that during the COVID-19 pandemic, technology played a pivotal role. They highlighted its effectiveness in patient triage and predicting those in need of mechanical ventilation and intensive care units beds. This not only conserved precious resources but also ensured that they concentrated on treating those who were most in need.

All the clinicians interviewed belonged to the same NHS trust and had exposure to the same electronic systems. This homogeneity may have influenced their perceptions regarding the use of AI tools in their specific healthcare setting, and these findings might not necessarily apply to clinicians in more diverse healthcare environments. Additionally, the recruitment of participants took place during the COVID-19 pandemic and relied completely on voluntary participation. Consequently, the number of participants was limited. However, recruitment continued until thematic saturation was achieved. Finally, the findings from this study help to

improve our understanding of their clinical usability, accessibility, and utility of AI tools in clinical practice.

8.6. Chapter Summary

In this chapter, the researcher described a qualitative study conducted with healthcare professionals from different specialties to understand their perceptions towards the use of AI-based tools to inform SDM in the clinical environment. The researcher demonstrated how clinicians perceive AI in a clinical context and the challenges they encounter during the implementation of AI tools in clinical environments. The researcher also explored their thoughts about the features that they believe would be beneficial to include in these tools and discussed the potential future applications of these tools.

In the next chapter, the researcher will present the key findings from each area of this PhD programme of work, explaining how these findings can potentially guide the clinical application of AI-predictive models in the future, and providing some evidence-based recommendations from NICE and the NHS to support these findings.

Chapter 9. Future Directions and Recommendations.

9.1. Introduction

In this Chapter, the researcher will summarise the key findings from each area of work that was completed during this PhD, highlighting the strengths and weaknesses of each area. The researcher will also describe how these findings could potentially support the future clinical application of the developed model. Additionally, the researcher will provide evidence-based recommendations from the NHS and NICE to support adoption of these tools in the NHS and to improve their clinical utility.

9.2. Key Findings from different areas of this work

In this PhD programme, the researcher developed a framework comprising nine key stages to support the development and evaluation of predictive AI models (Chapter 3).[170] These stages include defining the outcome of interest, identifying predictors, selecting relevant datasets, model training and validation, building user interfaces and data presentations, licensing, updating, maintaining AI-predictive models, and evaluating their impact in clinical practice. This framework acknowledges the challenges that clinicians and researchers may encounter at each stage and provides practical tips for managing them.

The researcher systematically reviewed the literature to identify the optimal predictor set for developing AI-predictive models to predict the likelihood of infection and subsequent infection (Chapter 4).[87] In this work, the researcher described 13 common predictors, categorised as modifiable (e.g., blood pressure, blood glucose, chronic comorbidities) and non-modifiable predictors (e.g., age and sex). Additionally, the researcher demonstrated how different sepsis definitions, and the type of predictors (modifiable or non-modifiable) can influence the predictive power (performance) and predictive timeframe of the models. The researcher also identified a knowledge gap in the development of these models in surgical settings, despite the higher risk of infection.

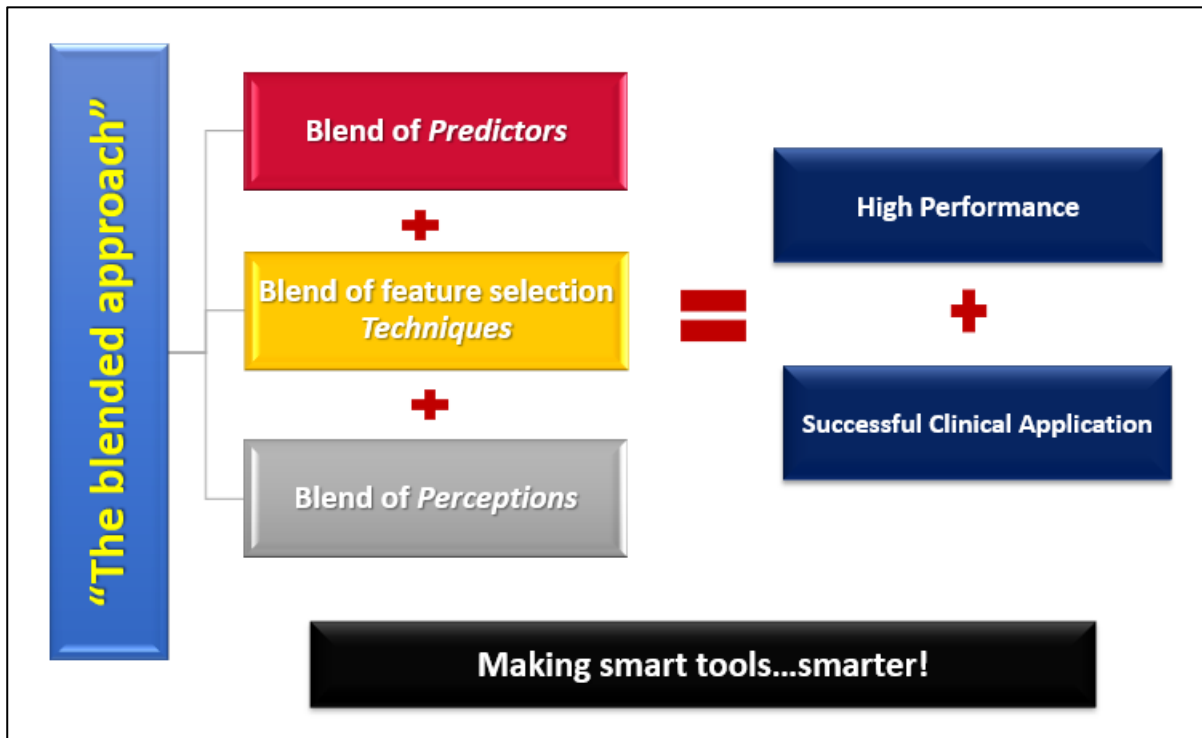
Based on these findings and the developed framework, the researcher addressed this knowledge gap by developing and internally validating an AI-predictive model to assess the risk of post-operative infection in elective abdominal surgery (AIPIS) (Chapter 5). The model utilised 19 predictors grouped into demographics, comorbidities, microbiology data, and lab investigations. The Gradient Boosting Classifier emerged as the best-performing model, with an ensemble model exhibiting high performance during internal validation with 96.9% sensitivity, 74.1% specificity, and an AUC of 85.5%. The researcher also reviewed the literature

for common types of bias encountered during the development and implementation of AI-predictive models. Tools currently available for assessing the ROB in AI-predictive models, such as PROBAST, were also identified and described in Chapter 6.

Alongside developing the model, the researcher explored clinicians' and patients' perspectives of using AI decision aids to inform SDM. The researcher systematically reviewed the literature to understand their perspectives on using these tools at various stages of the patient journey and in different clinical settings (Chapter 7). In this systematic review, patients reported that AI decision aids provided them with a sense of ownership over their health decisions and enhanced adherence to treatment advice. AI decision aids were found to be most usable in disease screening and prevention, prognosis, and treatment. However, both clinicians and patients expressed concerns about data privacy, updating the tools, and recognised a ROB associated with AI decision aids due to the diversity of patient groups using them. The researcher engaged with a Patient and Public Involvement and Engagement (PPIE) group through VOICE to discuss the plan for developing and future implementing our model (Chapter 5, section 5.5.4). The group recognised the potential benefits to patients when this model is integrated into clinical practice and strongly recommended involving clinicians in the early development stages. In response to the PPIE group's input, the researcher also conducted a qualitative study to explore clinicians' thoughts and perceptions regarding the implementation of AI tools in the clinical environment. This study covered the features that they would like to see in these tools, key considerations when developing them, and the enablers and barriers to their implementation in clinical practice (Chapter 8).

Throughout these different but related areas of work, the researcher conceptualised a novel approach, "*the blended approach*," which was viewed to enhance the chances of successful clinical application and improve the predictive performance of these models, thereby better supporting the SDM process. The blended approach comprises of three stages, which were detailed in Chapter 5 and illustrated in Figure (17).

Figure 17. The blended approach to enhance the chances of successful clinical application of AI-predictive models:



Applying a blended approach to our model development demonstrated high performance measures in predicting the likelihood of post-operative infection and subsequent sepsis in surgical setting. The researcher plans to measure its impact of adopting such a tool in a test clinical environment before running into an active clinical environment.

9.3. Strengths and weaknesses of the different areas of work

This PhD programme of work produced an AI-predictive model to predict the likelihood of post-operative infection and subsequent sepsis in elective abdominal surgery. This model addressed an unmet need, with a higher risk of infection in a surgical setting when compared to a medical setting. The researcher adopted a mixed-method approach so as to strength the development of this model and its utility in clinical practice. The qualitative study involved clinicians from different specialities in the NHS that expressed perceptions from different clinical settings, levels of experiences, and patient populations seen. The researcher also gained experience in establishing an academic TRE between Newcastle University and the Pioneer research hub, one of the seven national research hubs in the UK focused on acute and urgent care healthcare data from the University Hospitals Birmingham NHS Foundation Trust.[270]

Nevertheless, the researcher acknowledges several limitations to this work. (a) For the systematic review exploring the optimal predictors set to develop AI-predictive models, some

of the models included were not internally validated, and some studies used the administration of antibiotics or blood cultures (just that it was carried out) as an indication of the presence of an infection, which is a non-specific and inaccurate definition for infection or sepsis. (b) The model development was also limited by different factors linked to the training database, data cleaning, limitation to clinical application and to accessibility to user interface, which were all previously detailed in Chapter 5. (c) The systematic review to understand clinicians' and patients' perspectives of AI decisions included studies that did not carry out a formal bias assessment for the included AI-tools; ROB was recognised and mentioned by both clinicians and patients in these included studies. For the qualitative study, only some clinicians had previous experience of using AI tools in clinical practice, which limited their contributions in a way. Using TRE as described above presented a number of challenges, with further details provided in the next section.

9.4. AI-predictive models development; accessing of databases through TRE

Accessing patient data is a crucial and complex aspect of developing AI models. In the UK, patient data confidentiality is safeguarded by the GDPR, which treats health data as a special category of data. As a result, most data providers grant access to authorised researchers through TREs. These TREs are secure platforms that allow researchers to remotely access a virtual desktop, limiting their access to only the data approved by the data provider. In April 2022, the Secretary of State for Health and Social Care in the UK commissioned a review conducted by Goldacre et al. from the University of Oxford.[271] The review supported the use of TREs to facilitate better, broader, and safer access and analysis of healthcare data.[271]

The review conducted by Goldacre et al. highlighted the potential of adopting TREs to enhance data transparency and minimise the risk of data breaches.[271] These findings have informed the data strategy of the DHSC. The review also recommended the establishment of three types of TREs: a national TRE that allows access to patient data from multiple sites on a national level, local TREs that include patient data from local NHS trusts, and academic TREs that incorporate data from academic research conducted in NHS facilities and are linked to local TREs.[271] Chapter 3 of this thesis provides detailed information on different data sources in the UK and how to access them.

Based on our experience, an academic TRE offers a secure means of accessing patients' data from the NHS within an academic environment, aligning with the recommendations of the Goldacre et al. review. However, through our experience, the researcher also encountered

limitations working with the TRE, which related to: cost, time, risk of error, difficulties extracting analysed outputs, availability of software, and occupational risk:

- **Cost:** It is important to clearly outline from the beginning the costs involved in accessing data through TREs. The Goldacre et al. review also emphasized the importance of securing funding in academic settings to support access to TREs.
- **Time:** The process of setting up a TRE typically takes a couple of months. Any further changes to the TRE may take the technical team a few weeks to 2-3 months to complete, based on our own experience.
- **Risk of Error:** Errors can occur when extracting data from patients records to the TRE. The researcher conducted an initial analysis to detect any errors e.g., extreme values or clinically non-sensible patterns that may be present in the data. These were discussed with the development team (microbiologist and clinician), as they required expert knowledge relating to each variable.
- **Difficulties extracting analysed outputs:** The researcher was required to seek the approval of the data provider / technical team to transfer the analysis outputs outside the TRE. These approvals again took time to obtain. The researcher was also required to develop the coding scripts entirely from scratch (no access to the internet through the TRE). This process typically consumes more time than usual.
- **Availability of software on the TRE:** Not all software was available within the TRE. The researcher needed to undergo additional training on unfamiliar softwares, which took time and came with additional costs.
- **Occupational risk:** Accessing TRE was typically done through a web-based browser, which often had a low resolution. The researcher reported this challenge back to the technical team at Pioneer, who made improvements by allowing access to the TRE via a dedicated remote desktop.

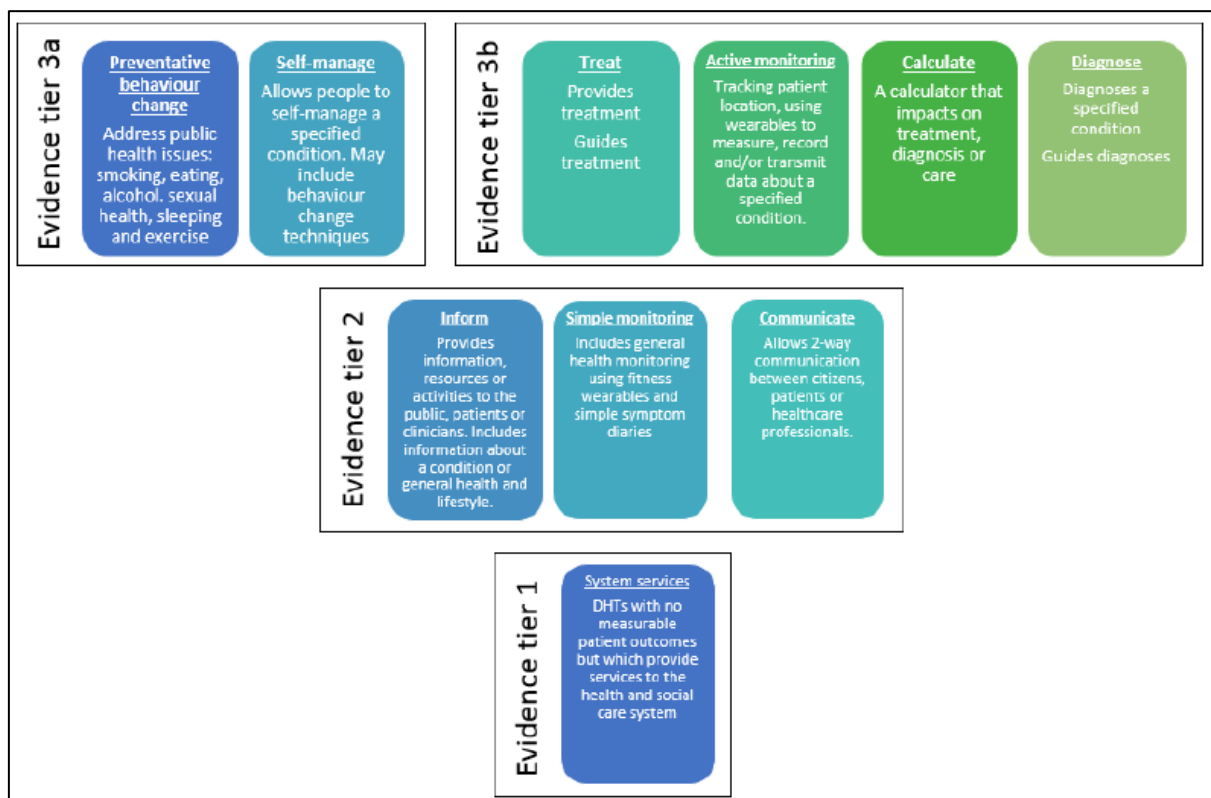
9.5. National Institute for health and Care Excellence (NICE) framework for Digital Health Technologies

The National Institute for health and Care Excellence (NICE) developed an evidence standards framework that guided the development and evaluation of digital health technologies (DHT).[272] Evaluators and decision makers in the health and care system can use these

standards to help identify DHTs that are likely to offer benefits to users and to the health and care system. It was also aligned with regulatory requirements to make it easier to use.

The framework was updated in 2022 to include AI and data-driven technologies with adaptive algorithms. A fixed algorithm is the one that is not connected to real time data, so that it does not change over time. Adaptive algorithms, on the other hand, can automatically change.[273] The framework categorises the DHTs based on their functional classes into three tiers: Tier 1, Tier 2 and Tier 3a and b. The AI-predictive model developed as part of this PhD programme of work could be considered to sit in Tier 3b. Figure (17) is extracted from the NICE framework and demonstrates the description of different DHT tiers based on their function.

Figure 18. Demonstrates different tiers of DHTs based on their function:[272]



The higher the tier of the DHT, the higher the potential risk from the DHT to users.[272] NICE added five contextual questions that should supplement this classification to better estimate the potential risk to users. These questions (or areas) focused on whether the tool will be used in vulnerable populations, such as paediatric populations and vulnerable adults (Question 1),[272] the seriousness of the consequences of DHT failure (Question 2), assistance required for users (Question 3), the inclusion of AI and machine learning (Question 4), and organisational risk, relating to any requirements to adjust the work environment or workflow to adopt the DHT (Question 5). These questions are covered in Table 9 below, alongside the answers relating to our newly developed model.

The British government released a code of conduct for data-driven health and care technology used in the NHS. It is focused on the ethical use of the tool, having a clear function and purpose, assessment of usability and accessibility, ensuring appropriate technical testing, evaluating the clinical safety, data security and transparency. Furthermore, the code also discusses the licensing and regulation of these tools and how to develop the evidence on their clinical, social, behavioural or economic benefits.[273]

Table 9. Risk assessment of the AI model to predict the likelihood of post-operative infections in surgical patients:

Question	Risk assessment
Use by vulnerable population	The model is intended to be used in adults who are undergoing elective general abdominal surgery. It does not include paediatric populations. Additionally, this tool was intended to inform the SDM between patients and clinicians about post-operative infection risk estimation. This decision making process requires patients to have mental capacity to do so.
Seriousness of consequences if the tool failed	This model is a prognostic predictive model. It estimates the likelihood of infection following an elective surgery. Although the model has shown high performance measures during its development and internal validation, it would still need to be externally validated in a real clinical environment.
User supported by a healthcare professional or qualified personnel	This model could be used in preoperative assessment clinics, where patients have clinical consultations with clinicians prior to their surgery. The tool can highlight the post-operative risk and potentially inform the SDM process. There is the opportunity or adjustments to be made (e.g., nutritional) to potentially reduce the risk.
Inclusion of AI or machine learning algorithms	This model was developed using supervised machine learning. The code of conduct for data-driven health and care technology would need to assess the use of the model. This is to ensure that the model operates ethically, provide a clear outcome to the users, to assess its usability and accessibility and to ensure its clinical safety, data protection and transparency.[273]
Assessment of financial and	The tool would need to be tested in a clinical environment and clinicians' perspectives obtained on its use (including its impact on clinical workflow).

organisational risk	
--------------------------------	--

9.6. External Validation

External validation involves re-evaluating the model's performance in a totally unseen dataset. During this stage, the model is tested to predict the outcome of interest and compare this to the clinical diagnosis made by clinicians. External validation is essential for reliable clinical application of AI-predictive models. One example is the Epic Sepsis Model (ESM) which was implemented in practice without external validation and demonstrated worse performance when externally validated compared to the performance reported during training stage.[274] This external validation raised concerns about the nationwide use of this model with poor sepsis prediction, putting patients safety at risk.[274] Our AI-predictive model was developed and internally tested on surgical patients (general abdominal surgery). The future plan is to externally validate this model on patients from other specialities, thereby testing whether it is generalisable to other specialities in which predication of sepsis is important. The research team have started discussions with PIONEER to commence this work (with an amendment/extension to the existing data arrangement). This step will reduce the potential ROB and increase the likelihood of a successful application in clinical practice.

9.7. Exploring how to best present the model predictions/outputs to clinical users

As part of this PhD programme of work, the researcher conducted a systematic review to explore patients' and clinicians' perspectives of using AI decision aids more broadly to improve SDM. The researcher also conducted a qualitative study with clinicians to identify the key features that they would find useful to inform their clinical decision making. The next stage would involve co-producing the user interface with users and clinical informatics experts, developing a suitable prototype, and exploring its integration with current electronic systems. This will include carrying out usability testing and potentially adjust the model for users' needs.

9.8. Designing a future impact study

Impact studies aim to test the clinical impact of a novel tool versus a standard tool / practice. This testing is usually carried out in a silent clinical environment to assess how the novel tool would affect the direct clinical care and the potential workflow. The Effectiveness-implementation Hybrid Framework described three types of designs for testing clinical impact of interventions to improve their public health impact. [275] These types include: (1) testing clinical intervention impact on patients' outcomes while gathering information about

implementation, in parallel; (2) testing an implementation strategy while gathering information about the clinical intervention's effect on patients' outcomes (3) Using a dual testing approach for clinical and implementation strategies. [275] In order to choose the most appropriate design for a future impact study, two steps should be done. Firstly, a normalisation process theory (NPT) analysis could be performed to shape the late-stage translation of an innovation and the delivery of healthcare.[276] The NPT consists of four stages; coherence, engagement, collective action which is what have been done to allow an intervention to take place, and a reflexive monitoring which involves a cost benefit assessment for the intervention. Additionally, different clinical outcomes (e.g., reduction in infection/sepsis rates, ICU admissions, in-hospital mortality and length of hospitalisation) and clinical workflow indicators (i.e., clinicians satisfaction, reduced clinicians fatigue, and patient prioritisation) should be identified. This will also include modelling the scale of potential benefit of this intervention relative to cost. Secondly, the literature should be reviewed to explore the different study designs (e.g., cluster-randomised) that were previously used to evaluate the clinical impact of AI-predictive models in different specialities, and the strengths and limitations of these designs. At this stage, the model should be ready for evaluating its impact in clinical environment (i.e. impact study). This evaluation will guide the real clinical application of the tool in a running clinical environment and highlight any potential weaknesses.

9.9. Licensing and Maintenance

After impact evaluation, the model will be ready for licensing by the MHRA as a medical device as described earlier in Chapter 3. As discussed in Chapter 8, it was perceived as important for users that the model was regularly maintained and upgraded as necessary based on emerging clinical evidence and feedback from users.

9.10. Recommendations

Careful consideration needs to be given to selecting the predictors for the development of any AI model. The researcher highlighted the strengths of having a blend of modifiable and non-modifiable predictors, whilst excluding any variables used to define the outcome of interest, as described in Chapter 6. The researcher would recommend multiple machine learning algorithms should be tried to help with training an ensemble model based on the outcome problem type (i.e., classification, or regression) and the characters of the training dataset, followed by selecting the algorithm with highest performance to build the model. Furthermore, regular feedback from users (i.e., patients, clinicians) and stakeholders should be sought throughout the process of model development.

Bias is a crucial threat for the success of clinical application of AI-predictive models, either prognostic or diagnostic. The researcher has previously highlighted how it is important to select the ‘right’ dataset to train the model and have a clear plan on how to handle missing data. The researcher would recommend using multiple imputation to reduce the ROB associated with missing data. The definition of the predicted outcome should be clear and supported by clinical evidence, as well as the selection of predictors to reduce bias. The researcher had a relatively positive experience using the HDR gateway as a primary and reliable source of data for models training. The gateway includes datasets from all national research hubs in the UK and provides structured information on the data quality and completeness, and how to submit a data request form.

9.11. Impact of COVID on the programme of work

My PhD started six months before the COVID-19 pandemic, which meant spending approximately two years during the pandemic, witnessing three pandemic waves with three lockdowns. The pandemic was a major risk to accomplish this research project, which has always been discussed with my supervisors and all possible scenarios planned for so as to mitigate risks. The pandemic affected my PhD in the following ways:

- **Suspension of research activities in NHS facilities:** At the beginning of the first pandemic wave, all research activities were suspended at Newcastle Upon Tyne Hospitals Foundation Trust (NuTh) due to immense pressures on the NHS at that time. The process was very slow and arduous, and the researcher and supervisory team proposed seeking the necessary approvals at a different site (focus being on Pioneer and accessing the sepsis phenotype database from Pioneer and the researcher started the process of data access approval.
- **Delays with obtaining approvals at the new site:** Applying for data access at the new site took a considerable amount of time (nine months), due to delays during the third wave of the pandemic. The researcher used this time to prepare the code scripts and access a publicly available database (Sepsis Skaraborg Database to run the code and find any inaccuracies. This was an exceptionally stressful time for the researcher who did her best to move things forward in such challenging circumstances.

9.12. Concluding Remarks

Artificial Intelligence is a powerful tool that is currently transforming healthcare, as outlined in the NHS Long Term Plan. AI-predictive models have shown the potential to support SDM by

offering clinicians recommendations that can reduce the predicted risk of a particular condition occurring. However, the successful application of these models in clinical practice is contingent on various factors that should be considered from the early stages of development through to implementation in clinical settings.

By following the nine stages for developing and evaluating AI-predictive models and incorporating our blended approach throughout the model development and testing phases, clinicians can potentially create AI-predictive models with high clinical utility. Involving users in the design and development of these models is crucial. Additionally, testing in a real clinical environment and assessing the clinical impact of these studies is essential for validation before clinical application. The use of the TREs is rapidly expanding; however, further research is needed to address the limitations associated with potentially using TREs.

Chapter 10. Appendices

Appendix A

Open access
Review

BMJ Health & Care Informatics

Road map for clinicians to develop and evaluate AI predictive models to inform clinical decision-making

Nehal Hassan ^{1,2}, Robert Slight,^{2,3} Graham Morgan,⁴ David W Bates,⁵ Suzy Gallier,^{6,7} Elizabeth Sapey ^{6,7}, Sarah Slight^{1,2}

To cite: Hassan N, Slight R, Morgan G, et al. Road map for clinicians to develop and evaluate AI predictive models to inform clinical decision-making. *BMJ Health Care Inform* 2023;30:e100784. doi:10.1136/bmjhci-2023-100784

Received 12 April 2023
Accepted 24 July 2023


Check for updates

© Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹School of Pharmacy, Newcastle University School of Pharmacy, Newcastle Upon Tyne, UK
²Faculty of Medical Sciences, Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, UK
³Freeman Hospital, Newcastle Upon Tyne Hospitals NHS Foundation Trust, Newcastle Upon Tyne, UK
⁴School of Computing, Newcastle University, Newcastle upon Tyne, UK
⁵Department of General Internal Medicine, Harvard Medical School, Boston, Massachusetts, USA
⁶PIONEER Health Data Research Hub, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK
⁷Department of Health Informatics, PIONEER Health Data Research Hub, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK

Correspondence to
Sarah Slight;
Sarah.Slight@newcastle.ac.uk

ABSTRACT

Background Predictive models have been used in clinical care for decades. They can determine the risk of a patient developing a particular condition or complication and inform the shared decision-making process. Developing artificial intelligence (AI) predictive models for use in clinical practice is challenging; even if they have good predictive performance, this does not guarantee that they will be used or enhance decision-making. We describe nine stages of developing and evaluating a predictive AI model, recognising the challenges that clinicians might face at each stage and providing practical tips to help manage them.

Findings The nine stages included clarifying the clinical question or outcome(s) of interest (output), identifying appropriate predictors (features selection), choosing relevant datasets, developing the AI predictive model, validating and testing the developed model, presenting and interpreting the model prediction(s), licensing and maintaining the AI predictive model and evaluating the impact of the AI predictive model. The introduction of an AI prediction model into clinical practice usually consists of multiple interacting components, including the accuracy of the model predictions, physician and patient understanding and use of these probabilities, expected effectiveness of subsequent actions or interventions and adherence to these. Much of the difference in whether benefits are realised relates to whether the predictions are given to clinicians in a timely way that enables them to take an appropriate action.

Conclusion The downstream effects on processes and outcomes of AI prediction models vary widely, and it is essential to evaluate the use in clinical practice using an appropriate study design.

INTRODUCTION

Healthcare systems worldwide generate enormous amounts of patient-related health data, much of which is electronic in developed countries. There is growing interest among clinicians and healthcare staff in how they could use these data to support patient care.¹ Much of medicine is about anticipating and reducing risk, based on current and historical experiences. Predictive analytics in healthcare can help determine the risk of a patient developing a particular condition or complication, which can inform the shared decision-making process between clinicians and patients and improve patient satisfaction with their overall medical care.²⁻⁷ With the new era of artificial intelligence (AI), clinical prediction tools can help personalise treatment and management decisions.

The Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) framework was published to guide developing multivariate predictive models,⁸ outlining what should be reported (eg, data sources, modelling techniques) when written up for publication.⁹ However, a recent systematic review highlighted how these models' reporting has been rather poor since its publication.¹⁰ TRIPOD also only focused on regression-based prediction models (although it can be applied to AI-generated approaches) and highlighted the need for more 'practical methods' for developing models more commonly used in healthcare (ie, supervised learning techniques).¹¹ The Consolidated Standards of Reporting Trials-AI guidelines were published in 2020 to help readers conceive studies with AI interventions; however, there was limited guidance on how these AI predictive models could be developed and usefully applied in clinical practice¹²; clinicians have sought further information on this.¹³ Even if a newly developed AI model has a good predictive performance, this does not guarantee that it will be used in clinical practice or enhance clinical decision-making, let alone improve health outcomes.¹⁴ The quality criteria important for evaluating AI predictive models were described in a recent scoping review; however, little information was provided on how such tools affect the clinical routine of physicians, which may vary per physician.¹⁵

The nine stages for developing and evaluating predictive AI models

Stage 1: clarifying the clinical question or outcome(s) of interest (output).



Hassan N, et al. *BMJ Health Care Inform* 2023;30:e100784. doi:10.1136/bmjhci-2023-100784



Appendix B

International Journal of Medical Informatics 150 (2021) 104457



Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Review article

Preventing sepsis; how can artificial intelligence inform the clinical decision-making process? A systematic review

Nehal Hassan ^a, Robert Slight ^b, Daniel Weiland ^b, Akke Vellinga ^c, Graham Morgan ^d, Fathy Aboushareb ^e, Sarah P. Slight ^{a,*}

^a School of Pharmacy, Newcastle University, King George VI Building, Newcastle upon Tyne, NE1 7RU, UK
^b Newcastle Upon Tyne Hospitals NHS Foundation Trust, Freeman Hospital, High Heaton, Newcastle upon Tyne, NE7 7DN, UK
^c School of Medicine, National University of Ireland Galway, University Road, Galway, H91 TK33, Ireland
^d School of Computing, Newcastle University, Urban Sciences Building, Newcastle upon Tyne, NE4 5TG, UK
^e Northumbria Healthcare NHS Foundation Trust, Rake Lane, North Shields, Tyne and Wear, NE29 8NH, UK



ARTICLE INFO

Keywords:
Sepsis
Prediction
Machine learning
Decision-making
Artificial intelligence

ABSTRACT

Background and objectives: Sepsis is a life-threatening condition that is associated with increased mortality. Artificial intelligence tools can inform clinical decision making by flagging patients at risk of developing infection and subsequent sepsis. This systematic review aims to identify the optimal set of predictors used to train machine learning algorithms to predict the likelihood of an infection and subsequent sepsis.

Methods: This systematic review was registered in PROSPERO database (CRD42020150685). We conducted a systematic literature review across 3 large databases: Medline, Cumulative Index of Nursing and Allied Health Literature, and Embase. Quantitative primary research studies that focused on sepsis prediction associated with bacterial infection in adults in all care settings were eligible for inclusion.

Results: Seventeen articles met our inclusion criteria. We identified 194 predictors that were used to train machine learning algorithms, with 13 predictors used on average across all included studies. The most prevalent predictors included age, gender, smoking, alcohol intake, heart rate, blood pressure, lactate level, cardiovascular disease, endocrine disease, cancer, chronic kidney disease (eGFR < 60 mL/min), white blood cell count, liver dysfunction, surgical approach (open or minimally invasive), and pre-operative haematocrit < 30 %. All included studies used artificial intelligence techniques, with average sensitivity 75.7 ± 17.68, and average specificity 63.08 ± 22.01.

Conclusion: The type of predictors influenced the predictive power and predictive timeframe of the developed machine learning algorithm. Predicting the likelihood of sepsis through artificial intelligence can help concentrate finite resources to those patients who are most at risk. Future studies should focus on developing more sensitive and specific algorithms.

1. Introduction

Sepsis is a life-threatening condition that is associated with significant morbidity and mortality [1]. Sepsis affects 10 % of hospitalized patients, with a 38 % rate of mortality in Europe [2,3]. Sepsis also represents a financial burden on healthcare systems, accounting for almost half of the cost of running intensive care.³ Predicting sepsis can reduce healthcare costs and save patients' lives through preventing the initiation of the multi-organ dysfunction process, reducing ICU

admissions and improving patients' prognosis [4].

The use of big data and artificial intelligence applications, namely machine learning, can predict diseases [5]. Machine learning algorithms have been developed for screening and predicting sepsis [2,6], and have showed better predictive abilities compared to current conventional scoring systems [7]. However, these algorithms have got limitations; some have been trained using at least two of the systemic inflammatory response syndrome (SIRS) criteria, which can be non-specific and misleading in sepsis prediction [7,8]. Some algorithms have also

* Corresponding author.
E-mail addresses: n.a.m.hassan2@newcastle.ac.uk (N. Hassan), bob.slight@nhs.net (R. Slight), dweiland@nhs.net (D. Weiland), akke.vellinga@nuigalway.ie (A. Vellinga), graham.morgan@ncl.ac.uk (G. Morgan), fathy.aboushareb@nhs.net (F. Aboushareb), Sarah.Slight@newcastle.ac.uk (S.P. Slight).

<https://doi.org/10.1016/j.ijmedinf.2021.104457>
Received 23 November 2020; Received in revised form 12 March 2021; Accepted 6 April 2021
Available online 10 April 2021
1386-5056/© 2021 Elsevier B.V. All rights reserved.

Appendix C

118

International Journal of Pharmacy Practice, 2021, Vol. 29, No. S1

2. Crilly P, Kayyali R. A Systematic Review of Randomized Controlled Trials of Telehealth and Digital Technology Use by Community Pharmacists to Improve Public Health. *Pharmacy* 2020;8(3):137. Available from: <https://www.mdpi.com/2226-4787/8/3/137>

PREDICTING INFECTION AND SEPSIS; WHAT PREDICTORS HAVE BEEN USED TO TRAIN MACHINE LEARNING ALGORITHMS? A SYSTEMATIC REVIEW.

N. Hassan¹, R. Slight², D. Weiland², A. Vellinga³, G. Morgan⁴, F. Aboushareb² and S.P. Slight¹, 1. School of Pharmacy, Newcastle University, UK. 2. Newcastle Upon Tyne Hospitals NHS Foundation Trust, Freeman Hospital, Newcastle upon Tyne, UK. 3. School of Medicine, National University of Ireland, Galway, Ireland. 4. School of Computing, Newcastle University, UK

Introduction: Sepsis is a life-threatening condition that is associated with increased mortality. Artificial intelligence tools can inform clinical decision making by flagging patients who may be at risk of developing infection and subsequent sepsis and assist clinicians with their care management.

Aim: To identify the optimal set of predictors used to train machine learning algorithms to predict the likelihood of an infection and subsequent sepsis and inform clinical decision making.

Methods: This systematic review was registered in PROSPERO database (CRD42020158685). We searched 3 large databases: Medline, Cumulative Index of Nursing and Allied Health Literature, and Embase, using appropriate search terms. We included quantitative primary research studies that focused on sepsis prediction associated with bacterial infection in adult population (>18 years) in all care settings, which included data on predictors to develop machine learning algorithms. The timeframe of the search was 1st January 2000 till the 25th November 2019. Data extraction was performed using a data extraction sheet, and a narrative synthesis of eligible studies was undertaken. Narrative analysis was used to arrange the data into key areas, and compare and contrast between the content of included studies. Quality assessment was performed using Newcastle-Ottawa Quality Assessment scale, which was used to evaluate the quality of non-randomized studies. Bias was not assessed due to the non-randomised nature of the included studies.

Results: Fifteen articles met our inclusion criteria (Figure 1). We identified 194 predictors that were used to train machine learning algorithms to predict infection and subsequent sepsis, with 13 predictors used on average across all included studies. The most significant predictors included age, gender, smoking, alcohol intake, heart rate, blood pressure, lactate level, cardiovascular disease, endocrine disease, cancer, chronic kidney disease (eGFR<60ml/min), white blood cell count, liver dysfunction, surgical approach (open or minimally invasive), and pre-operative haematocrit < 30%. These predictors were used for the development of all the algorithms in the fifteen articles. All included studies used artificial intelligence techniques to predict the likelihood of sepsis, with average sensitivity 77.5±19.27, and average specificity 69.45±21.25.

Conclusion: The type of predictors used were found to influence the predictive power and predictive timeframe of

the developed machine learning algorithm. Two strengths of our review were that we included studies published since the first definition of sepsis was published in 2001, and identified factors that can improve the predictive ability of algorithms. However, we note that the included studies had some limitations, with three studies not validating the models that they developed, and many tools limited by either their reduced specificity or sensitivity or both. This work has important implications for practice, as predicting the likelihood of sepsis can help inform the management of patients and concentrate finite resources to those patients who are most at risk. Producing a set of predictors can also guide future studies in developing more sensitive and specific algorithms with increased predictive time window to allow for preventive clinical measures.

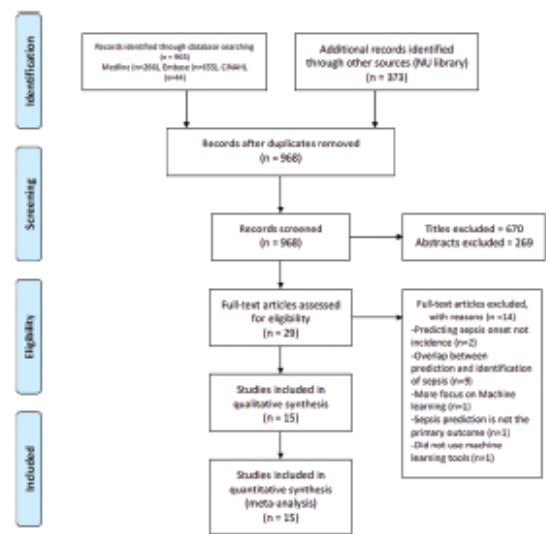


Figure 1: PRISMA diagram.

DESIGNING A CONTINUOUS DATA-DRIVEN FEEDBACK AND LEARNING INITIATIVE TO IMPROVE ELECTRONIC PRESCRIBING: AN INTERDISCIPLINARY QUALITY IMPROVEMENT STUDY

M. McLeod¹, S. Farah¹, K. Macaulay², T. Sheth³, M. Patel³, A. Ghafour³, M. Denning⁴, A. Mulla⁵, J. Kerai³, A. Chu⁶, D. Patel² and B. Franklin^{1,7}, 1. Centre for Medication Safety and Service Quality, Imperial College Healthcare NHS Trust, London, UK. 2. Improvement Team, Imperial College Healthcare NHS Trust, London, UK. 3. Pharmacy Department, Imperial College Healthcare NHS Trust, London, UK. 4. Department of Surgery and Cancer, Imperial College London, London, UK. 5. Imperial Business Intelligence (Data Warehouse), Imperial College Healthcare NHS Trust, London, UK. 6. Department of Renal Medicine, Imperial College Healthcare NHS Trust, London, UK. 7. UCL School of Pharmacy, London, UK.

Introduction: The WHO Global Patient Safety Challenge aims to reduce severe avoidable medication-related harm by 50% by 2023[1]. Research suggests that providing timely, trusted feedback that incorporates relevant action can improve practice. However, a key barrier is lack of prescribing

Appendix D

AI-based Sepsis prediction in Surgical Practice; Lessons learnt from Model Development.

Nehal Hassan, MSc¹, Robert Slight, PhD^{1,2}, Daniel Weiland, FRCPath², Graham Morgan, PhD³, Akke Vellinga, PhD⁴, Hazem Mamdouh, BSc⁵, Sarah Slight, PhD^{1,2}

¹School of Pharmacy, Newcastle University, Newcastle Upon Tyne, UK, ²Newcastle Upon Tyne Hospitals NHS Foundation Trust, Freeman Hospital, Newcastle Upon Tyne, UK, ³School of Computing, Newcastle University Newcastle Upon Tyne, UK, ⁴University College Dublin, Dublin, Ireland, ⁵Research Development International (RDI), Cairo, Egypt.

Introduction: Sepsis is a leading cause of in-hospital mortality, costing more than 24 billion US dollars annually.¹ Although there are artificial intelligence (AI) models for early detection and prediction of sepsis, these models are limited by their sensitivity and specificity when applied in clinical practice. Our aim was to develop an AI model that can predict the likelihood of infection and subsequent sepsis in patients, using a blend of techniques that could potentially increase its applicability in clinical practice.

Methods: We conducted a systematic literature review to identify clinical predictors previously used in developed sepsis prediction models.² We used the publicly available, *Sepsis Skaraborg study* dataset of adult hospitalized patients (n=1572), which had 23 variables including demographic information, vital signs, blood culture results (positive/negative), sequential organ failure assessment score, systemic inflammatory response syndrome score, confirmed bacterial infection. Code scripts for model development were built on Python v3.8.2 and visual studio code v1.64.2. Patients were categorized into being either septic or non-septic (Sepsis-3 definition). Features (predictors) selection was performed using a blend of machine learning methods (feature elimination) and clinician input on the clinical meaningfulness and relevance of different modifiable (i.e., blood glucose level) and non-modifiable predictors. Recursive feature elimination with cross-validation was then re-run on these predictors. The dataset was divided into training (80%) and internal validation (20%) set. The model was trained using different types of classifier, supervised machine learning methods (i.e., random forest classifier, decision trees classifier) followed by training of an ensemble model. Performance measures were calculated. We also conducted semi-structured interviews with clinicians to seek their input on how the model interface should be designed and outcomes should be presented. Interview questions were informed by a systematic review on patients and clinicians perspectives towards using AI tools.³ Ethical approval was obtained (20/EM/0183, IRAS 280077).

Results: We identified nine predictors for the model training (age, systolic blood pressure, respiratory rate, oxygen saturation, heart rate, body temperature, hemoglobin, C-reactive protein, procalcitonin, Neutrophil-Lymphocyte count ratio), seven of which were modifiable, thus allowing clinicians to make recommendations on how to reduce an individual's estimated sepsis risk. The performance measures for ensemble model in the training phase were (50% sensitivity, 88% specificity, 70.5% PPV, and 77.2% NPV, and AUROC=75.5%), while in internal validation phase (41% sensitivity, 91.8% specificity, 75.4% PPV, 72% NPV, and AUROC=72.6%), which shows low risk of overfitting. Ten clinicians participated in one-hour interviews via Zoom teleconferencing between April and November 2021. Clinicians suggested a number of different ways to improve the model's clinical application in general, which included ensuring the tool that was "quick to use" (P7), used language that was "simple and handy and easy to understand" (P10) and "give us an immediate answer" (P10).

Conclusion: This study demonstrated opportunities for the application of AI predictive tools in clinical practice. Adopting a blended approach throughout the development of AI predictive model could improve the chances of successful clinical application without compromising performance measures or risk of over/underfitting.

References:

1. Rudd KE, Johnson SC, Agesa KM, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *The Lancet*. 2020 Jan 18;395(10219):200–11.
2. Hassan N, Slight R, Weiland D, et al. Preventing sepsis; how can artificial intelligence inform the clinical decision-making process? A systematic review. *International Journal of Medical Informatics*. 2021 Jun 1;150:104457.
3. Hassan N, Slight RD, Bimpong K, et al. Clinicians' and patients' perceptions of the use of artificial intelligence decision aids to inform shared decision making: a systematic review. *The Lancet*. 2021 Nov 1;398:S80.

Appendix E

Machine learning Glossary and Data Analysis Manual for Chapter 5.

1. Basics and Definitions:

1.1. *Artificial intelligence:*

The ability of machines to perform functions that are normally carried out by humans, such as voice recognition, image recognition, and decision making. These machines could be an automated cashier in a supermarket, a mobile camera, a robot, or a software on a hospital system. The physical shape of the machine depends on its purpose. A medical example would be robotic surgery, where the physical structure and appearance of a robot would be necessary for a physical surgical purpose (i.e. doing a particular intervention during an operation, i.e. C-arm robots used in angiography), while an application (tool) integrated in the electronic medical records to guide decision making would be more convenient and cost-effective to be in the form of software rather than a physical appearance of a robot.

1.2. *Machine learning:*

A sub-speciality of AI, and it is the method by which machines are trained to be intelligent systems.

1.2.1. *Machine learning models versus algorithms:*

Although the terms “models” and “algorithms” are usually used interchangeably in the computer science literature, they are not the same. Algorithms are the codes written in a programming language and are run on the training, validation or testing dataset. While models are the output obtained from running the algorithm. In our infection example, random forest would be one of the algorithms that was trained on the dataset, while the tool that we got for infection prediction, is the model.

1.2.2. *Types of machine learning algorithms:*

There are four major machine learning algorithms: Supervised learning, Unsupervised learning, Semi-supervised and Reinforcement learning. Most recently, deep learning has been introduced, which is a sub-classification from supervised learning. Supervised learning is most widely used for predictive models in different specialities, which we will focus on.

1.2.3. *Supervised learning algorithms:*

These are group of algorithms used when the output is known. For infection prediction, if we have a dataset where we have an output of infection diagnosis (i.e. we know infected and non-infected patients), and we use this output to train the model, this will be supervised learning.

Unsupervised learning is where the dataset doesn't have an output and the machine has to learn it by itself. If we have a dataset to predict infection and we don't know who were the infected and the non-infected patients, and we used an algorithm to develop the output, this will be unsupervised learning.

Supervised learning is sub-classified into two types; regression algorithms and classification algorithms. In a dataset with variables X_{0-10} and output Y , each X will correlate to Y to a particular level. Supervised learning algorithms can be used to train a predictive model for Y , based on the level of correlations between each of X and Y . If Y is numerical (continuous) variable, regression algorithms should be used to train the model. If Y is a discrete/binary variable (i.e. infected, non-infected), classification algorithms should be used.

An example, if a model is developed to predict the systolic blood pressure value based on other vital signs, in this case, a regression algorithm should be used. While a model developed to predict the status of blood pressure (hypotensive, hypertensive, normotensive) based on vital signs, a classification model should be used. It is worth to mention that X s could be solely numerical, solely discrete or a mix of both numerical and discrete. A model can be one dimensional (i.e. uses only one X), two-dimensional (uses 2 X s), multidimensional (uses a finite number of X s), or it can be developed from infinite number of X s, in this case support vector machine (SVM) algorithm can be used. Support vector machine is a type of supervised learning that can work for both regression and classification types, but it requires high computational processing power and more resources.

In our infection model, the output is infection diagnosis, which is a discrete output (binomial), meaning that we will use supervised classification algorithms. There are multiple types of classification algorithms that we have used to develop our model, including logistic regression, decision trees, random forest, gradient booster, ridge, and perceptron, which will be explained in detail in the feature selection section.

2. Dataset preparation:

2.1.Terminology:

- 2.1.1. **Data frame:** This is the dataset spreadsheet that is running on python, containing columns and rows, in our example, we have only one data frame.
- 2.1.2. **Keys:** These are the names of the clinical variables or features (names of columns present in row 1 in the dataset). Gender, age, respiratory rate are keys. These are different from a Key column in SQL that was explained earlier.

- 2.1.3. **Output:** this is the variable that the model will be trained on to predict. In our example, infection diagnosis was used as output, and sepsis diagnosis was defined using SEPSIS-3 definition. This is because SEPSIS-3 definition is the most specific to multi-organ dysfunction that is characteristic to sepsis. Infection definition was detailed in the methods section.
- 2.1.4. **Python script:** this is the python interface where commands (codes) are written. Please note for all the scripts' figures in chapter five, green hashtag text in the scripts are only for clarifications and not a part of the codes.
- 2.1.5. **Python terminal:** this is part of the python console where codes written in the script are executed.
- 2.1.6. **Python Libraries:** These are packages in Python that contain ready-made codes to perform different functions (mathematical operations), the suitable package should be imported to the python script before running the relevant function.
- 2.1.7. **Pandas Library:** This is one of the python libraries which contains codes for data analysis and manipulation.
- 2.1.8. **Loop function:** One of the functions used in Python to iterate over the inputs of a key (i.e. allows reading each cell in a key in order to be further used to identify missing values and manage them), it repeats the process so it reads a cell in an excel sheet after another one by one.
- 2.1.9. **Not a number (NaN):** it is an expression that substitutes missing values in a way that is readable by Python. Model training cannot be run unless all missing data are converted to NaN values. NaN can be substituted with predicted values through the development of a model to predict missing values. This was not needed in our case as the missing data did not exceed 10% in the whole data frame. NaN should be checked in the data preparation stage before starting the model training as the model accuracy will be affected by empty cells.

3. Feature elimination:

3.1. Terminology:

- 3.1.1. **Feature elimination:** this is the process of removing features (variables) that are least correlated to the output using machine learning algorithms. In this process, different machine learning algorithms are used to select features with highest correlation to predict the output and eliminate features with least

correlation. Feature elimination helps eliminate unnecessary features which if included will represent a cost and need of unnecessarily higher computational power, in addition to reduction in the model accuracy. This step is followed by a voting technique where the most common highlight correlated features from all the algorithms involved are then selected. The number of algorithms used depends on the type of features.

3.1.2. Recursive feature elimination (RFE): this is a feature selection algorithm that is used in python to select the features with highest correlation with the output and eliminate features with the least correlation (hence the name feature elimination). This algorithm requires the programmer to decide on the number of features needed to be selected (i.e. how many features you want your predictive model to include at the end) which can be decided from the medical literature or according to the working environment where the model will be used (clinicians will not be keen to use a tool where they have to enter 50 entries, for example), and the type of algorithms that you will use to predict the correlations between features and output (i.e. decision trees, random forest..etc.). RFE works for both regression and classification outputs.

3.1.3. Recursive feature elimination with cross validation (RFECV): this algorithm is more accurate and more commonly used than RFE as it reduces the risk of overfitting. It is similar to RFE in the concept of feature elimination, but with two differences, firstly, RFECV evaluates the correlations produced from different algorithms between the features and the output through subclassifying the data frame into multiple subsets to validate the correlations (i.e. the cross validation), which means that correlation patterns should be almost similar in all the data frame subsets so that it guarantees that the correlations are reproducible and hence less risk of overfitting. In RFE, the correlations are predicted on the data frame as one whole set. Secondly, RFECV automatically decides the number of features that will be selected. The algorithm internally set a number of selected features and check the accuracy of this number on the validation. Accordingly, it starts to change this number and check accuracy on multiple validation subset until reaching a certain number that has the best accuracy and that is the best number of selected feature with respect to the cross-validation technique.

- 3.1.4. Overfitting:** it means that the performance measures (sensitivity, specificity, AUC, positive predictive values, negative predictive values) appear very good during the model training, which is then significantly reduced in the validation and testing data frames. There are some techniques and ML algorithms that should be used to reduce the risk of overfitting, which will be detailed at a later stage.
- 3.1.5. Logistic regression algorithm:** This a machine learning supervised classifier algorithm, which is known as linear classifier. It predicts the best fitting linear relationship between the features and output through weighing the features in relation to the variables.
- 3.1.6. Decision tree algorithm:** This a machine learning supervised classifier algorithm which can provide classification based on the number of selected features. The larger (or more precisely, the deeper) the decision tree, the greater number of features is needed, and the better fitting the model will be. To clarify, if the model is to predict infection, so the first leaf of the tree will be temperature, if high, then infected, if low then no infection, adding another feature (or leaf) to this tree would improve the accuracy of the prediction, such as WBCs. In this case if temperature is high and WBCs is high then infection. This example shows how the depth of the tree can improve the accuracy of the prediction.
- 3.1.7. Random forest algorithm:** multiple and deeper decision trees form a forest. This is the random forest supervised classifier algorithm, which is usually expected to improve the performance of decision trees algorithms and reduce the risk of overfitting.
- 3.1.8. Gradient boosting algorithm:** this is known as ensemble classifier algorithm, which is a supervised, advanced type of decision trees and random forest. It depends on adding decision trees in a sequential manner by one by one, creating a complex and huge ensemble of decision trees with large number of probabilities to boost the performance of the model. These decision trees are then fit together using gradient descent optimization algorithm (creating slope between different points to minimise the value between them) and hence called gradient boosting classifier. [side note: it's important to save, and preferably backup any important files on personal laptops with low processing power before running gradient boosting classifiers on python as

laptops might crash while running, as this algorithm requires relatively high computational power].

3.1.9. Ridge classifier algorithm: it is an advanced supervised, regression classifier algorithm, which divides the data points into multiple subsets, thus creating better fitting regression curves between datapoints, compared to those developed by logistic regression algorithms, which treats the dataset as one large subset, creating only one best fitting line passing through all the data points.

3.1.10. Perceptron algorithm: it is a supervised linear classifier algorithm. It is classified as the simplest type of neural networks as it deals with the cases rows in the data as single neurons and starts to predict the output (infection versus non-infection), with each row, the algorithm calculates the percentage of error in the prediction and considers it in the following row, this how it is trained to predict an output. So, this type of algorithms learns to separate between two sides of the output using a boundary. However, this type of algorithm requires larger sample size of the data frame compared to the previously mentioned ones.

4. Data Analysis Manual

4.1. Step 1: Data Cleaning

4.1.1. Data access

The data access was granted after approving a data request form to PIONEER research hub. The data was accessed through a TRE which was set up by PIONEER in the form of a virtual desktop. More details on TREs are explained in chapter two of this thesis. This virtual desktop has the database uploaded in the form of Structured Query Language (SQL) database tables, in addition to all softwares added to the data request form for data processing and model development, which were for the purpose of this model: Python, SQL, Excel, and Anaconda.

4.1.2. Exploring data tables on SQL:

SQL is a software that uses a query programming language to manage and visualise big datasets. The database was uploaded on Microsoft SQL Server Management Studio, Microsoft for Windows, United states. The patients data was uploaded into 11 SQL tables from the data provider. The variables in each of the 11 tables are described in the table 1.

Table 1. All the variables available on the SQL database:

Name of the table	Variables included	Description
Admission	Age, gender, ethnicity, index of multiple deprivation (IMD), IMD quintile, date of admission, date of discharge, length of hospital stay, length of ITU stay (if admitted to ITU).	Demographics and admission information
Assessments (social history)	Assessment, source value (i.e. status), source value expanded, date recorded.	Social history (alcohol consumption through AUDIT-C score and smoking status).
Diagnosis	Diagnosis type, diagnosis order, diagnosis source code (i.e. ICD-10) code, diagnosis source code expanded, diagnosis name, coding system, time of recording the diagnosis.	Diagnosis on admission, final diagnosis, and any comorbidities.

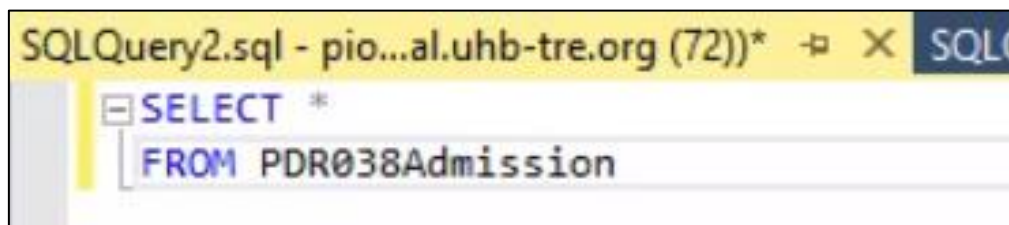
Drug administration	Time prescribed, time administered, administration status, form, route, quantity, quantity_unit, concentration, concentration_unit, infusion_rate, infusion_rate_unit, volume remaining, volume remaining unit, own medication	Quantity of drug administered and status of administration
Lab analysis	Lab_analysis, sample type, time of sample collection, time sample received, time result released, result, unit	All laboratory investigations done during hospitalisation (pre-and post-operative).
Microbiology	Time collected, time displayed, specimen types, organisms, antibiotic, result.	Information on positive cultures only
Observation	Observation, time performed, result, result expanded, unit.	Vital signs, AVPU, and NEWS score.
Prescription	Prescription type, written time, start time, mode, frequency, route, alternative route	Route of administration, and if the medication is scheduled or when needed.
Prescription drug component	Drug, form, alternative form, quantity, quantity, unit, alternative quantity, alternative quantity unit, concentration, concentration unit, alternative concentration, alternative concentration unit, infusion rate, min infusion rate, max infusion rate.	Name of the drug, concentration and dosage form
Site move	Site move time, move order, site source, site	Patient transfer between wards during hospitalisation

Surgery	Procedure ICD-10 code, procedure_code_expanded, procedure order, theatre, time_into_theatre, time_knife_to_skin, time_into_recovery,	Intra-operative information.
---------	--	------------------------------

All the tables have one fixed column known as “the database key”. This key column created by data admin has a unique identifying code (mix of numbers and letters) for each patient, which can be used to link information from the 11 tables to the relevant patient. In this database, the key column was named “care_spell_id, which was used to link information for the eligible patients in the following steps.

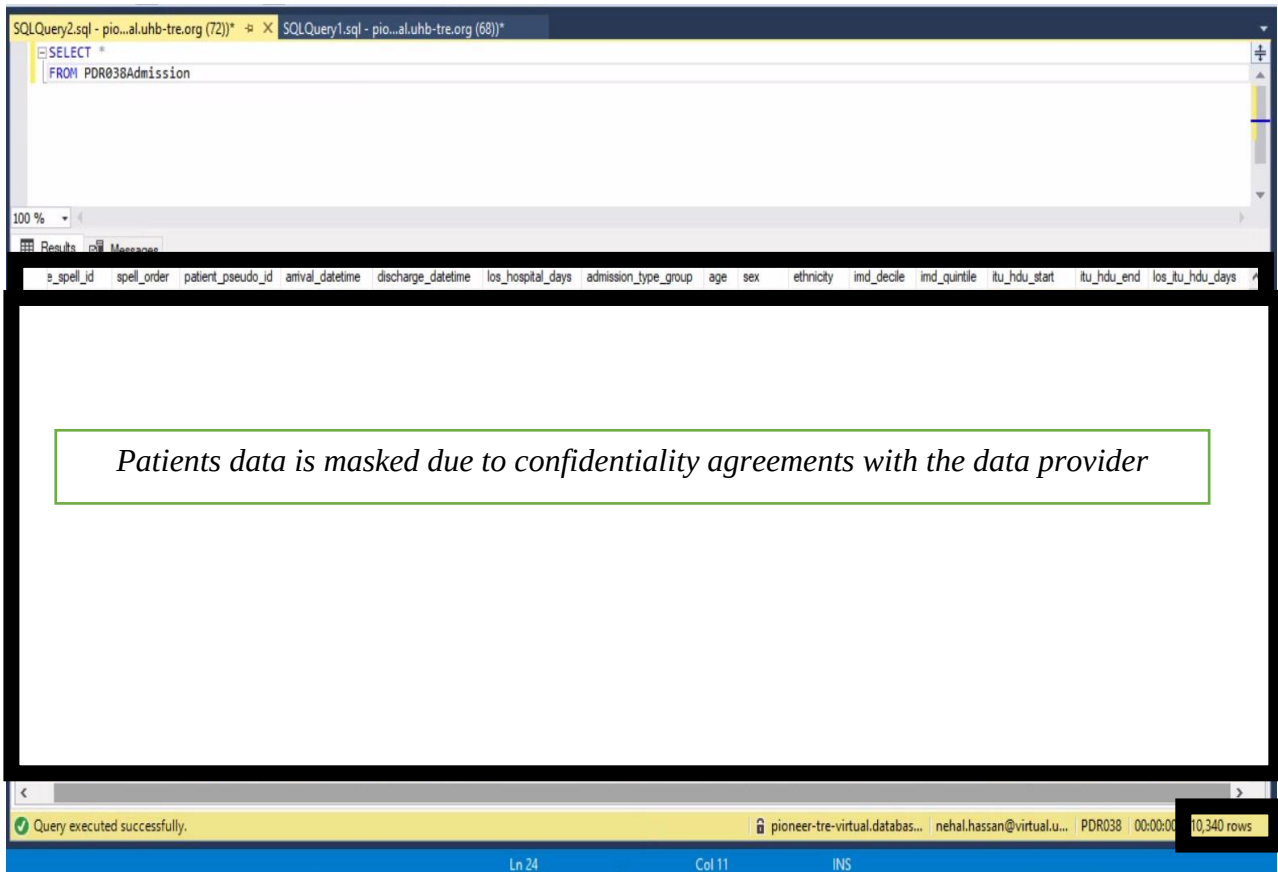
Exploring data on SQL requires running queries. For example, to access the admission table, a SELECT query should be written as shown in figure 1:

Figure 1: Query to visualise a table on SQL database:



A “*” after the SELECT statement denotes to all the variables in this particular table, instead, if only particular variables are needed, the exact name (i.e. case sensitive) of the variables as they appear on the table should be typed in the SELECT statement and separated by a comma (,). In the FROM statement, the exact name of the table should be typed, as the name saved on the database.

The next step is running the query by selecting “execute” from the tool bar. The output of the above query is shown below in Figure 2.

Figure 2: Output for executing query in figure 1:

The output appears on the lower part of the query, the names of the columns (i.e variables) appear on the top, and then the patients data is shown in the form of rows (masked in figure 2 due to patient information confidentiality). On the bottom right side on the yellow bar, the number of rows is displayed.

Similarly, the exploration of all the 11 tables can be performed in the same way. However, the authorised researchers (i.e. data users) do not have writing access to the database. Hence, any data manipulation or creation of new tables from the available database is carried out on another software. In our case, this was Microsoft Excel 2016, Microsoft for Windows, United States, which was also accessed through the TRE.

4.1.3. Data management and filtration:

Eligible patients were extracted from the database based on the inclusion criteria mentioned above. This extraction was carried out on multiple steps. The first step was to create “flying columns”. These are columns created by the data users through running some calculation queries, but they cannot be saved in the original tables of the database (hence, flying). An example is including patients who had pre-operative stay of $\leq$ seven days, and post-operative stay of < 30 days. The admission table only provides the date of arrival and date of discharge, while the surgery table provides the date of surgery (i.e. time_knife_to_skin). Accordingly, to

calculate pre and post-operative hospital stay, the admission and surgery tables are joined using “JOIN” clause, and then the difference between the dates is calculated using “DATEDIFF” clause. Additionally, only elective patients are eligible, hence a “WHERE” clause should be used to filter the results to elective patients only. Furthermore, general abdominal surgeries were eligible for inclusion, which have also been added to a WHERE clause using ICD-10 codes for each procedure. The executed queries are shown in figure 3.

Figure 3: Query for extracting patients based on the inclusion criteria:

```

SELECT surgery.care_spell_id, procedure_code, procedure_code_expn, time_knife_to_skin, time_in_to_recovery, admission.care_spell_id, arrival_datetime,
los_hospital_days, discharge_datetime, admission_type_group, age, sex, ethnicity, imd_decile, imd_quintile, los_itu_hdu_days,
DATEDIFF (hour, time_knife_to_skin, time_in_to_recovery) AS operative_duration,
DATEDIFF (day, time_knife_to_skin, discharge_datetime) AS duration_op_to_discharge,
DATEDIFF (day, arrival_datetime, time_knife_to_skin) AS preop_duration
FROM dbo.PDR038Admission as admission
LEFT JOIN dbo.PDR038Surgery as surgery
ON admission.care_spell_id = surgery.care_spell_id
WHERE admission_type_group = 'ELECTIVE' AND los_hospital_days > 2 AND procedure_code IN ('J185', 'J183', 'Y752', 'J563', 'J021', 'J023', 'J022',
'J032', 'T309', 'G331', 'J562', 'J339', 'Y502', 'J182', 'J189', 'J292', 'Y508', 'J181', 'J374', 'J275', 'J031', 'J026', 'J311', 'J028', 'J027', 'J188', 'J249', 'T259',
'H073', 'T243', 'J239', 'J078', 'J099', 'T258', 'J029', 'J035', 'Z301', 'Z302', 'J015', 'J238', 'J268', 'J269', 'T253', 'T279', 'J073', 'T248', 'J058', 'Y751', 'J168',
'T203', 'J079', 'H119', 'T209', 'J291', 'T278', 'J233', 'H113', 'J184', 'H079', 'J092', 'J161', 'J053', 'T252', 'H154', 'J038', 'T202', 'J219', 'H118', 'J169', 'H074',
'G514', 'J093', 'T268', 'J178', 'H111', 'H114', 'H062', 'H061', 'J088', 'H078', 'J211', 'T272', 'J208', 'J024', 'J089', 'H159')

```

The resulting table has been extracted into an excel sheet, where “filter column” feature was used to limit the “preop_duration” to a maximum of seven days and limit the duration_op_to_discharge to a maximum of 30 days.

For each of the 11 tables, flying columns were created to be either used as filters for inclusion criteria or potential predictors for infection or sepsis during the model training, which required some clinical input to best define the predictors. These flying columns were created after multiple discussions with the data administrators to understand how data entry and medical coding were performed. Taking these variables one by one:

- ***Cause of admission:***

The diagnoses coded as the “admitting diagnosis” were selected as the cause of admission.

- ***Surgical procedure:***

The database included information from gastroenterology and hepatology departments, hence we decided to include general abdominal surgeries due to the high risk of post-surgical infection and sepsis (surgical site and non-surgical site related). Any mentioning of laparoscopic or laparotomy procedure was considered as a minimally invasive procedure in a separate column (Yes/No). It is worth to mention that almost 60% of the patients on the dataset had malignancy, where it is estimated from the medical literature that risk of post-surgical infection in this surgical population is up to 30%.

- **Chronic medications**

A list of all medication on the database was explored using the query demonstrated in figure 4.

Figure 4: Query for the description of medications in the database:

The screenshot shows a SQL query window with the following query:

```
SELECT drug, COUNT (*) AS meds
FROM PDR038PrescriptionDrugComponent
GROUP BY drug
ORDER BY COUNT (*) desc
```

The results are displayed in a table with two columns: 'drug' and 'meds'. The table lists 14 medications and their corresponding counts, ordered from highest to lowest count.

drug	meds
1 Paracetamol	25240
2 Sodium Chloride	17579
3 Tramadol	13982
4 Cyclizine	11553
5 Codeine Phosphate	10215
6 Hartmann's Solution	9944
7 Enoxaparin	9382
8 AES (anti-embols...	8878
9 Morphine Sulfate ...	8745
10 Morphine Sulfate I...	8734
11 Ondansetron	8571
12 Lansoprazole	8198
13 Glucose (5%)	7852
14 Co-amoxiclav	6958

The status bar at the bottom indicates: "Query executed successfully." and "1,734 rows".

There were 1734 different types of medications were administered during hospitalisation. These medications were categorised into groups according to their pharmacological classes into antibiotics, vasopressors, insulins, corticosteroids, immunosuppressants..etc.

- **Investigations (Laboratory tests):**

All the lab tests were uploaded to Lab_analysis table, the available labs were clinically screened for their relevance and significance, with consideration for our published systematic review on sepsis predictors. A list of labs (N= 21) was selected from all the available labs (N= 68). The selected labs were: Lymphocytes, Neutrophils, Neutrophil/lymphocyte ratio (as a flying column), Red cell count (RBCs), Platelets, Haematocrit, Mean cell volume (MCV), Mean cell haemoglobin (MCH), Blood glucose, Haemoglobin, WBCs, Serum creatinine, estimated glomerular filtration rate (GFR), Bilirubin, International normalised ratio (INR), Alanine transferase (ALT), Aspartate transferase (AST), CRP, Erythrocyte sedimentation rate (ESR), Lactate, COVID-19 test (all were negative so was excluded later), Creatinine kinase.

For each lab investigation, Powerpivot feature of Microsoft excel was used to calculate the average readings 90 days pre-operatively using pivot-unpivot columns, and merging data sheets to join lab readings to relevant patients, so that each patient has one pre-operative reading for each available lab investigation. Additionally, post-operative labs were also analysed to guide the diagnosis of infection.

- **Observations (i.e. vital signs)**

Similar process with the lab investigations were replicated for the observations from the observations table. Eight out of 13 observations were filtered based on clinical significance and relevance to be included in the cleaned data sheet for model training. The filtered observations included: oxygen saturation, respiratory rate, temperature, systolic blood pressure, diastolic blood pressure, mean arterial pressure, NEWS score (National Early Warning Score), and AVPU (Alert, Voice, Pain, Unresponsive) score. The post-operative vital signs and NEWS score guided the diagnosis of infection, as explained in the methods section.

- **Microbiology data:**

Data on Microbiology table were only for patients who had a culture withdrawn and came back positive. It did not include patients with negative results or those who haven't had any cultures requested. A flying column for multidrug resistant strains was created, which was defined as resistance to antibiotics from three different classes. Positive cultures from sterile sites were added to a flying column and separated from other positive cultures. Post-operative positive cultures guided the diagnosis of post-operative infection, while pre-operative positive cultures, 90 days before the operation was used as potential predictor for infection.

The following stage was to join the variables tables into three sets (i.e. large tables) so that they can be connected to the eligible patients using the “Power Pivot” feature of Microsoft Excel. Each of the tables had a serial number which was created by the researcher (i.e. data user) and a key column created by the data admin (care_spell_id). This column was created to facilitate the merge between different tables and for easier identification of any errors. These two columns were used to join the tables together. The variables set are described in table 2:

Table 2. Variable sets:

Set	Joined tables	Categorical or Ordinal Variables	Numerical variables
Demographics	Admission, Diagnosis, Surgery, Assessment	Serial number, care_spell_id, cause of admission, type of procedure, Minimally invasive procedure (Y/N), sex, ethnicity, smoking, IMD ,IMD	Operative duration (hrs), pre-operative duration, post-operative duration, length of hospital stay,

		decile, comorbidities including malignancy, cardiovascular, hypertension, diabetes, asthma, COPD, renal, hepatic, anaemia, thrombosis, obesity, .	ITU stay, age, Audit C.
Medications	Drug administration, Prescription, Prescription drug component.	Serial number, care_spell_id, Chronic medications including insulin use, immunosuppressants and chemotherapy, corticosteroids use.	
Investigations and Infection and Sepsis diagnosis	Lab analysis, microbiology, observations, Drug administration, Prescription, Prescription drug component	Serial number, care_spell_id, prophylaxis antibiotics, week 1 use of antibiotics, week 2 use of antibiotics, week 3 use of antibiotics, week 4 use of antibiotics, week 1 use of vasopressors, week 2 use of vasopressors, week 3 use of vasopressors, week 4 use of vasopressors, ICD-10 coded diagnosis of infection, ICD-10 coded diagnosis of sepsis, positive sterile site cultures, other	Blood glucose, INR, haemoglobin, haematocrite, mean cell volume, platelets, red blood cell count, white blood cell count, lymphocytes, neutrophils, Erythrocyte sedimentation rate, C-reactive protein, Liver enzymes (ALT, AST), bilirubin, lactate, creatinine

		positive cultures, multi drug resistant organism (resistance to 3 antibiotics from 3 different classes)	kinase, systolic blood pressure, diastolic blood pressure, mean arterial pressure, temperature, oxygen saturation, AVPU score, NEWS score.
--	--	---	--

4.1.4. *Joining all the data in one table for training.*

Now the three sets are ready in the form of three large tables that should be joined into one table. Each table was revised for the accuracy of the information. Each patient had only one row in each of the three tables. The three tables were then joined simultaneously using combine queries (merge) function in PowerPivot in excel. The resulting final table contained all the information required for model training. Each patient had their data in one row. The total number of patient in this sheet was 3272 patients. However, there was some patients with missing data (more than 50% of variables). Consequently, complete case analysis method was used to exclude patients who had all their variables missing. This exclusion resulted in 2716 patients, eligible to include and ready for model training. Missing data within these was imputed using K-nearest neighbours algorithm, with K=5.

4.2. *Step 2: Dataset visualisation:*

Data visualisation and understanding of variables is the first step for model development after data cleaning. This understanding can be developed through descriptive analysis to ensure the uniformity and completeness of data and to carry further measures before starting the training of the model. Descriptive analysis can be performed using different statistical packages such as R, SAS, SPSS or Microsoft Excel. For the purpose of descriptive analysis, Microsoft excel was used. This is because of data security regulation that hindered extracting the excel sheet with the dataset from the TRE and the absence of any software which can be used for descriptive analysis on the TRE by the time of data analysis. Results of descriptive analysis are shown below. The dataset used for model training included 2716 infected and non-infected surgical patients who were admitted to the hospital. Descriptive analysis of the included cohort (elective

surgery patients eligible for inclusion) is described in chapter Five. Additional descriptive analysis for the excluded cohort (emergency surgery) is provided in table 3 and 4 below.

Table 3. Descriptive analysis of emergency surgery (excluded) cohort:

		Count (N=2148)
Gender	Male	1217 (56.6%)
	Female	918 (42.7%)
	Missing	13 (0.6)
Ethnicity	White -British	1479 (68.9%)
	Asian/Asian Brit - Pakistani	171 (8%)
	Not Known	105 (4.9%)
	White - any other White b/g	45 (2.1%)
	White -Irish	64 (3%)
	Any Other Ethnic Group	53 (2.5%)
	Black/black Brit-Caribbean	95 (4.4%)
	Asian/Asian Brit - Indian	28 (1.3%)
	Others (<10 patients)	4 (0.2%)
	Missing	104 (4.8%)
Index of multiple deprivations	1	372 (17.3%)
	2	390 (18.2%)
	3	384 (17.9%)
	4	402 (18.7%)
	5	377 (17.6%)
	Missing	223 (10.4%)
Microbiology (Cultures within 90 days before operation)		

Positive Blood cultures	0 (0%)
Positive other sterile site cultures (i.e. bile)	2 (0.1%)
Other positive cultures (sputum, urine,etc.)	9 (0.4%)
MDR in blood or sterile site	2 (0.1%)
Comorbidities	
Malignancy	1284 (59.8%)
Hypertension	984 (45.8%)
Cardiac	583 (27.1%)
Diabetes mellitus	341 (15.9%)
Renal	284 (13.2%)
Asthma	236 (11%)
Hepatic	198 (9.2%)
Obesity	167 (7.8%)
Chronic Obstructive pulmonary disease	58 (2.7%)
Anaemia	76 (3.5%)
Thrombosis	29 (1.4%)
Chronic medications	
Steroids	126 (5.9%)
Immune suppressants and chemotherapy	11 (0.5%)
Insulin	103 (4.8%)
Minimally invasive surgical procedures	
Yes	96 (4.5%)
No	2052 (95.5%)

Table 4. Descriptive analysis for continuous variables of the emergency surgery (excluded) cohort:

	Mean±SD	Min	Max
Age (years)	53.6±16.1	18	85
Pre-operative data (within 90 days before operation)			
White Blood Cell Count (WBCs)	9.2±2.4	4.2	32.4
Neutrophils	8.3±1.6	2.6	18
Lymphocytes	2.8±0.7	0.6	6.8
Red Blood Cell Count (RBCs)	4.3±0.3	2.4	6.1
Mean Cell Volume (MCV)	92±0.4	62.3	120
Mean Cell Haemoglobin (MCH)	28.3±1.4	14.3	41.2
Haematocrit	0.4±0.1	0.1	0.7
Platelets (PLT)*10 ³	249.2±2.8	23	684
Haemoglobin	10.1±1.3	6.3	14.2
Bilirubin	40.2±1.4	4.3	82
Alanine aminotransferase (ALT)	64.3±12.3	12	1023
Aspartate aminotransferase (AST)	83.6±5.8	16	1197
C-Reactive Protein (CRP)	22.3±1.8	12	894
Blood glucose	7.2±0.9	4.4	19
Serum creatinine	102±3.6	79	165
Operative and Post-operative data			
Operative duration (hrs)	5.7±2.8	2	22

Post-operative hospitalisation (days)	11.9±6.5	3	370
Length of Intensive Therapy Unit (ITU) stay (days)	1.6±0.1	0	27
Respiratory rate (breath/min)	18.3±1.2	14	25.6
Alert, Voice, Pain, Unresponsive (AVPU)	3±0.3	1	4
National Early Warning Score (NEWS)	2.7±0.6	1	8
Systolic blood pressure (SBP)	125.2±10.3	84	189
Diastolic blood pressure (DBP)	81.3±4.2	52	112
Mean arterial pressure (MAP)	89.3±3.6	62	128
Temperature	36.9±0.2	35.4	39.6
Blood glucose	8.4±2.3	3.9	26.8
International Normalised Ratio (INR)	1.3±0.1	1	4.2
Haemoglobin	10.4±1.2	6.6	15.2
Haematocrit	0.31±0.02	0.28	0.43
Mean cell haemoglobin	26.5±1.9	17.9	36.2
Mean cell volume	76.2±5.4	64.1	98.2
Platelet count*10 ³	251.2±12.6	18	1085.9
Red Blood Cell Count (RBCs)	3.2±0.2	2.3	4.9
White Blood Cell Count (WBCs)	12.3±0.3	3.1	34.3
Lymphocytes	2.3±0.2	0.2	42

Neutrophils	11.3±0.1	0.9	72
Neutrophil-lymphocyte ratio	4.9±0.1	3.3	12.2
Erythrocyte sedimentation rate (ESR)	43.9±8.4	10	84
C-reactive protein (CRP)	102.0±3.0	13	468
Alanine transaminase (ALT)	32.2±2.3	11.6	984
Aspartate aminotransferase (AST)	112.7±7.8	12	711
Total Bilirubin	36.9±58.4	3.2	475.3
Direct Bilirubin	131.3±88.4	12	221
Serum creatinine	70.4±37.6	26.6	245
Glomerular filtration rate (GFR)	83.2±11.3	6	128
Lactate	4.1±1.3	1.2	7.6
Creatinine Kinase (CK)	568±34.3	29	1284

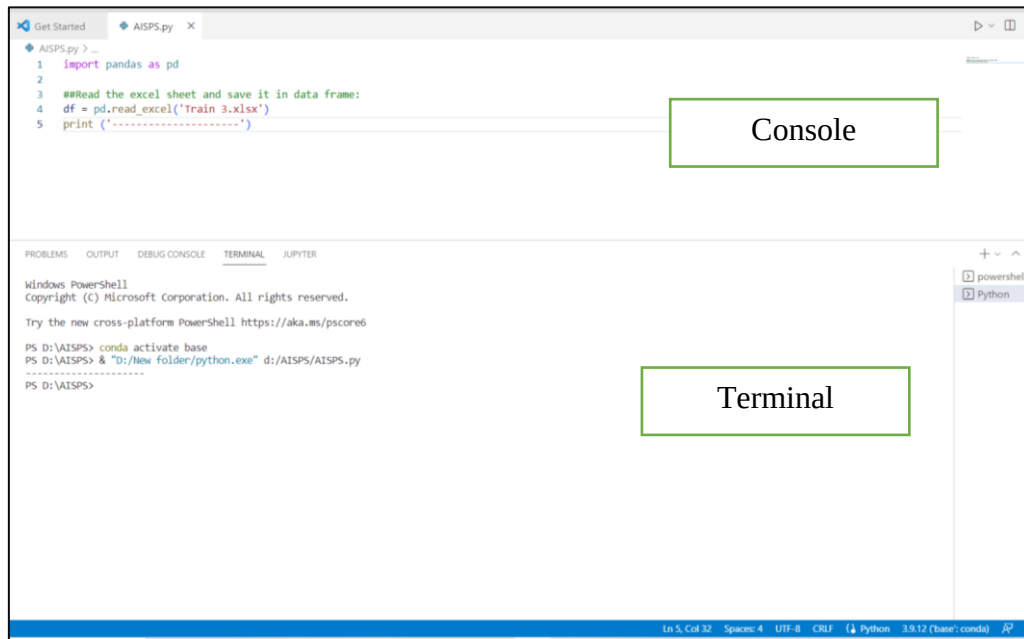
4.3. Step 3: Dataset preparation:

Predictive models development can be performed using different programming languages. For this model, I used Python with a source-code editor visual studio code for its wide use and applicability, easiness of the language and its free-of-charge availability.

4.3.1. Opening the data frame and data visualisation in python:

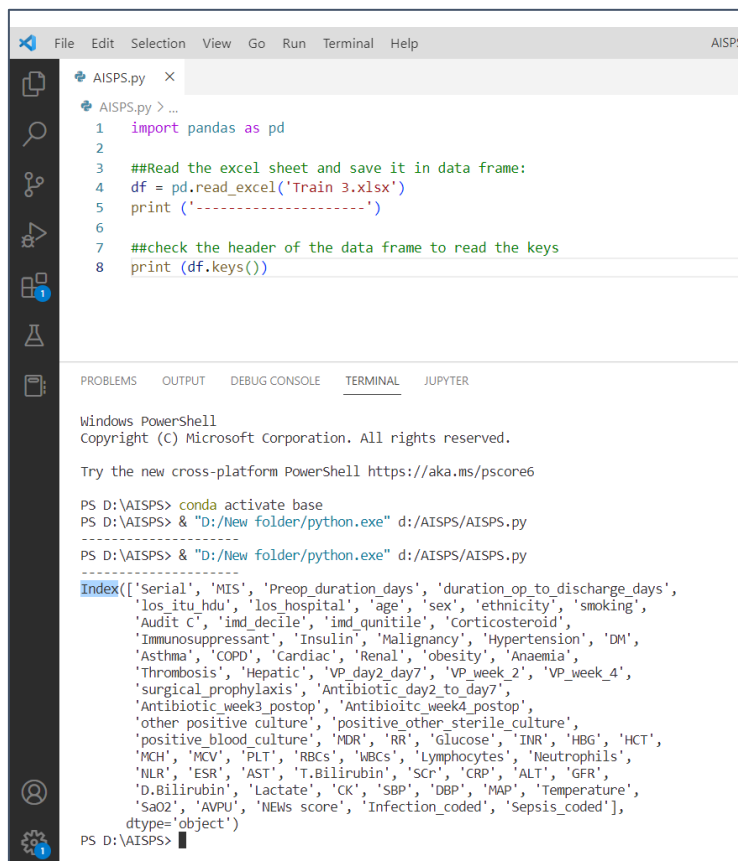
In our example, the data frame was saved as an excel spread sheet. It has to be opened in python to start the data preparation. If the library is not yet installed (using python for the first time), install the library using the command “pip install pandas” in the terminal. Based on the python version used, “xlsx” may need to be installed if not already present, which is responsible for reading excel sheets in python. Press “run” for commands in the console and press “enter” for commands in the terminal (installing packages). The following figure (5) shows how to read an excel sheet and save it as a data frame readable in python:

Figure 5: opening excel sheets in Python:



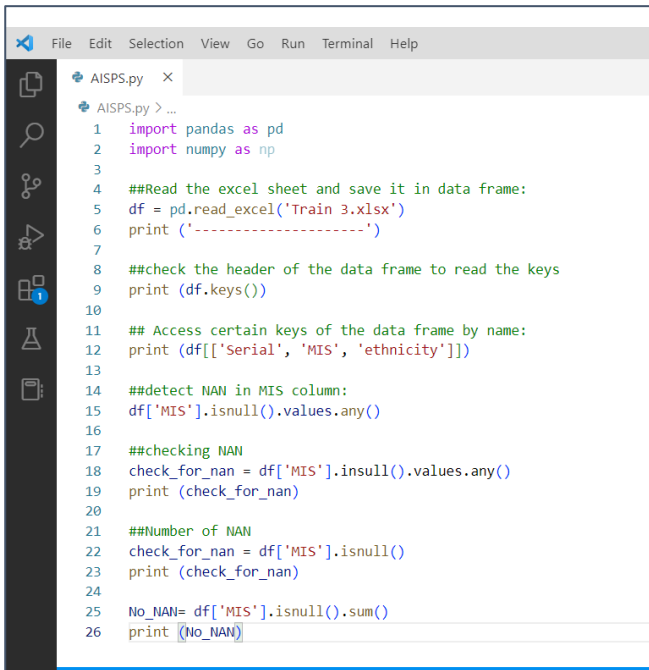
After saving the data frame, the headers of the data frame should be read to know the keys available in the data frame. The command is written in the console and the keys (variables) will appear in the terminal, as shown in figure 6.

Figure 6: Reading keys in the data frame:



To identify missing data, check for NaN in order to identify the amount of missing data and substitute them, an example that is shown here is for the column “MIS (minimal invasive surgery)”, which was coded with (1), if the surgery incision was minimally invasive. In the figure 11, NaN is checked in the MIS column, and the missing values appear as Nan in figure 7, from figure 8, it’s obvious that the number of NaNs is 731, which means that 731 patients had a conventional incision rather than a minimally invasive incision.

Figure 7: Codes to detect NaN:

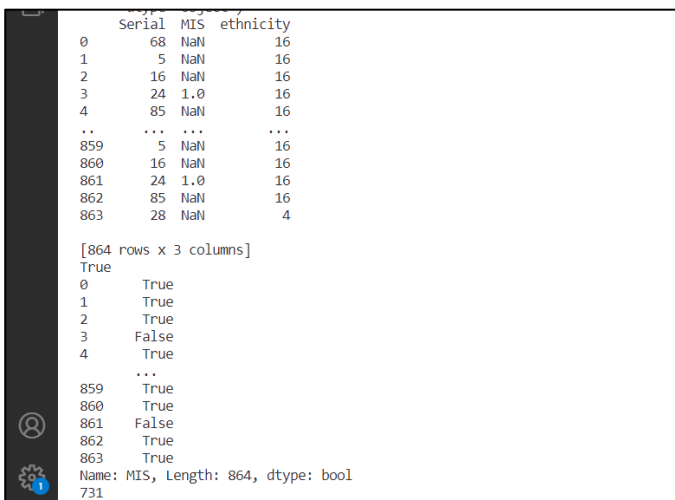


```

AISPS.py
1 import pandas as pd
2 import numpy as np
3
4 ##Read the excel sheet and save it in data frame:
5 df = pd.read_excel('Train 3.xlsx')
6 print ('-----')
7
8 ##check the header of the data frame to read the keys
9 print (df.keys())
10
11 ## Access certain keys of the data frame by name:
12 print (df[['Serial', 'MIS', 'ethnicity']])
13
14 ##detect NaN in MIS column:
15 df['MIS'].isnull().values.any()
16
17 ##checking NaN
18 check_for_nan = df['MIS'].isnull().values.any()
19 print (check_for_nan)
20
21 ##Number of NaN
22 check_for_nan = df['MIS'].isnull()
23 print (check_for_nan)
24
25 No_NAN= df['MIS'].isnull().sum()
26 print (No_NAN)

```

Figure 8: Terminal results showing Nan values (appearing as “True”) and their number:



```

Serial MIS ethnicity
0      68 NaN      16
1       5 NaN      16
2      16 NaN      16
3      24 1.0      16
4      85 NaN      16
...    ...    ...
859     5 NaN      16
860    16 NaN      16
861    24 1.0      16
862    85 NaN      16
863    28 NaN       4

[864 rows x 3 columns]
True
0      True
1      True
2      True
3     False
4      True
...
859    True
860    True
861    False
862    True
863    True
Name: MIS, Length: 864, dtype: bool
731

```

4.3.2. Mapping all variables into numerical values:

For Python to train the model on the dataset, the dataset has to be in numerical values. Accordingly, any words or categorical entries have to be mapped into numerical values before splitting the dataset into train and test sets. Steps to write the codes are:

1. Import necessary libraries and functions (Figure 9:lines 28 to 31 in the script)
2. Keep the original copy of excel sheet as a backup for any errors.
3. Substitute empty fields with zero.
4. Transform the text values into numerical values.
5. Loop on each column to repeat step 3 and 4 on each column (Figure 10).
6. Save the produced file in a “new data sheet” that will be used from now on.

Here you don't expect something to appear on the terminal, but you will find the new excel sheet (“new data sheet”) in your directory. Codes for mapping of variables are described below in Figure 9 and 10, green hashtag text in the script are for clarifications and not part of the code.

Figure 9: Codes for mapping variables into numerical values.

```

28 import pandas as pd
29 import numpy as np
30 from pandas.api.types import is_numeric_dtype
31 from pandas.api.types import is_string_dtype
32
33 ## Read the excel sheet and save it in a data frame
34 dr = pd.read_excel('Train 3.xlsx')
35
36 # Copy the current dataframe into the new one to keep the original unchanged:
37 df_copy = df.copy()
38

```

Figure 10: Codes for mapping variables into numerical values (cont.).

```

39 # Handling the empty fields
40 # Transform the text values into to numerical values
41 # Loop over columns of the dataframe
42 for key in df_copy:
43     col = df_copy[key]
44     # Check if the column contains text only
45     if is_string_dtype(col):
46         # Column values are strings (character based)
47         # Handling the empty values in the text values with Unknown category
48         col = col.fillna("Unknown")
49         # to convert the column to numerical we can use .astype('category')
50         # it categorize the string values into index inheritly
51         # for example: unique column values are: female, male and Unkown
52         # .astype('category') will map all female values to 0 while male values to 1 and Unkown to 2
53         # codes to the category based on the occurrence in the data what comes first takes a value first
54         col = col.astype("category")
55         # col here is still strings by categorical to transform it into numerical values we can use .cat.codes
56         # This command will transform the column from categorical to integer number
57         col = col.cat.codes
58     # Check if the column contains numbers only
59     elif is_numeric_dtype(col):
60         # Column values are numerical
61         # Handling the empty values in the column with zero number
62         col = col.fillna(0)
63
64     # Assign the column in the copied data frame
65     df_copy[key] = col
66
67 file_name = 'new data sheet.xlsx'
68 # saving the excel
69 df_copy.to_excel(file_name)
70 print (".....")

```

4.3.3. Missing data imputation:

Missing data can limit the accuracy of the model. Complete case analysis was initially used to remove all cases (i.e., patients) who had all their variables missing, leaving 2716 patients for inclusion. There are two common types used for missing data imputation with big data. The first method is multiple imputation with chain equation (MICE), however this method requires the satisfaction of missing at random condition (MAR). The MAR condition guarantees that the imputation can use relevant available information in other variables to predict the missing. In our database the MAR condition was not satisfied. Accordingly, we used the K-nearest neighbours (KNN) algorithm. This is a machine learning algorithm that can predict missing data regardless to being missing at random. The optimum number of Ks (neighbours) was 5, which provided the highest imputation accuracy. Tables 5 and 6 show descriptive analysis for categorical (table 5) and numerical variables (table 6) of the excluded cohort using complete case analysis

Table 5. Descriptive analysis of excluded cohort by complete case analysis (N=556):

		Count (N=556)	Missing
Gender	Male	322 (58%)	0 (0%)
	Female	234 (42%)	
Ethnicity	White -British	421 (75.7%)	0 (0%)
	Asian/Asian Brit - Pakistani	62 (11.2%)	
	Not Known	10 (1.8%)	
	White - any other White b/g	10 (1.8%)	
	White -Irish	9 (1.6%)	
	Any Other Ethnic Group	12 (2.1%)	
	Black/black Brit-Caribbean	7 (1.3%)	
	Asian/Asian Brit - Indian	7 (1.3%)	
	Others (<10 patients)	6 (1%)	
Index of multiple deprivations	1	133 (24%)	24 (4.4%)
	2	84 (15.1%)	
	3	117 (21.1%)	
	4	90 (16.2%)	
	5	106 (19.1%)	
Microbiology (Cultures within 90 days before operation)			
Positive Blood cultures		1 (<1%)	NA (could be missing or negative result)
Positive other sterile site cultures (i.e. bile)		0 (0%)	
Other positive cultures (sputum, urine,etc.)		3 (<1%)	
Multi-drug resistance (MDR) in blood or sterile site		0 (0%)	

Infection diagnosis (including those diagnosed with “sepsis”)	211 (38%)	
Sepsis diagnosis	16 (2.9%)	
Comorbidities		
Malignancy	56 (10%)	NA (Could be missing or not having the comorbidity)
Hypertension	156 (28%)	
Cardiac	83 (15%)	
Diabetes mellitus	90 (16.2%)	
Renal	18 (3.2%)	
Asthma	41 (7.4%)	
Hepatic	34 (6.2%)	
Obesity	32 (5.7%)	
Chronic Obstructive pulmonary disease	13 (2.4%)	
Anaemia	10 (1.8%)	
Thrombosis	54 (<1%)	
Chronic medications		
Steroids	11 (2%)	NA (Could be missing or not on the medication)
Immune suppressants and chemotherapy	5 (<1%)	
Insulin	18 (3.2%)	
Minimally invasive surgical procedures		
Yes	18 (3.2%)	0 (0%)
No	538 (96.8%)	

Table 6. Descriptive analysis for continuous variables of excluded cohort by complete case analysis (N=556):

	Mean±SD (N=556)	Min	Max	Missing values (N/%)
Age (years)	64.9±10.1	22	83	23 (4.1%)
Pre-operative data (within 90 days before operation)				
White Blood Cell Count (WBCs)	8.2±1.3	3.4	14.3	497 (89.4%)
Neutrophils	4.5±1.1	1.8	13.0	497 (89.4%)
Lymphocytes	1.2±0.6	0.3	3.9	497 (89.4%)
Red Blood Cell Count (RBCs)	4.2±0.3	1.9	4.8	497 (89.4%)
Mean Cell Volume (MCV)	85.2±0.8	51.3	92	497 (89.4%)
Mean Cell Haemoglobin (MCH)	27.6±3.9	14.3	34.2	497 (89.4%)
Haematocrit	0.5±0.1	0.2	0.8	497 (89.4%)
Platelets (PLT)*10 ³	248±3.8	239	268	497 (89.4%)
Haemoglobin	10.7±1.3	7.1	15.2	497 (89.4%)
Bilirubin	32.1±1.2	24	36	508 (91.4%)
Alanine aminotransferase (ALT)	46±2.8	26	68	508 (91.4%)
Aspartate aminotransferase (AST)	64±2.6	56	82	508 (91.4%)
C-Reactive Protein (CRP)	13.8±1.3	8	287	532 (95.7%)
Blood glucose	6.9±1.8	4.7	9.2	527 (94.8%)
Serum creatinine	82±1.4	76	136	513 (92.3%)

Operative and Post-operative data				
Operative duration (hrs)	3.8±1.3	<1	6	384 (69.1%)
Pre-operative hospitalisation (days)	1.3±0.2	1	5	384 (69.1%)
Post-operative hospitalisation (days)	5±1.6	2	8	384 (69.1%)
Length of Intensive Therapy Unit (ITU) stay (days)	2±1.1	1	6	NA (missing or might not have been in ICU)
Respiratory rate (breath/min)	14.8±1.3	12	26	532 (95.7%)
Alert, Voice, Pain, Unresponsive (AVPU)	4±0	4	4	542 (97.5%)
National Early Warning Score (NEWS)	1.3±0.2	0	3	542 (97.5%)
Systolic blood pressure (SBP)	121.2±6.5	94	150	532 (95.7%)
Diastolic blood pressure (DBP)	70.6±3.8	62	102	532 (95.7%)
Mean arterial pressure (MAP)	90±3.2	81	102	532 (95.7%)
Temperature	36.9±0.2	36.2	38.0	532 (95.7%)
Blood glucose	8.7±1.1	4.6	11.2	458 (82.4%)
International Normalised Ratio (INR)	1.1±0.1	0.9	2.5	533 (95.9%)
Haemoglobin	10.8±1.3	8	13.3	521 (93.7%)
Haematocrit	0.4±0.03	0.2	0.6	521 (93.7%)
Mean cell haemoglobin	27.6±1.3	22.4	37.2	521 (93.7%)

Mean cell volume	84.3±4.2	63.2	103.5	521 (93.7%)
Platelet count*10 ³	284.3±10.4	203	321	521 (93.7%)
Red Blood Cell Count (RBCs)	2.2±0.3	2.1	4.8	521 (93.7%)
White Blood Cell Count (WBCs)	12.3±0.3	3.9	18.6	521 (93.7%)
Lymphocytes	2.8±0.2	2.1	10.3	521 (93.7%)
Neutrophils	9.3±0.2	1.3	14.3	521 (93.7%)
Neutrophil-lymphocyte ratio	3.3±1.6	0.6	1.38	521 (93.7%)
Erythrocyte sedimentation rate (ESR)	36.4±5.6	4	54	542 (97.5%)
C-reactive protein (CRP)	84.6±3.8	4	278	542 (97.5%)
Alanine transaminase (ALT)	86±2.8	23	97	513 (92.3%)
Aspartate aminotransferase (AST)	135±4.5	68	287	513 (92.3%)
Total Bilirubin	18.6±8.5	2.9	35.6	549 (98.7%)
Direct Bilirubin	9.3±4.6	7	18.9	549 (98.7%)
Serum creatinine	79.2±3.7	70	154	482(86.7%)
Glomerular filtration rate (GFR)	104±5.8	54	129	496 (89.2%)
Lactate	0	0	0	556 (100%)
Creatinine Kinase (CK)	0	0	0	556 (100%)

Figure 11, 12 and 13 demonstrate the codes and libraries to run the KNN imputation.

Figure 11: KNN imputation algorithm codes:

```

KNN_trial3.py > ...
1  import pandas as pd
2  import numpy as np
3  from numpy import isnan
4  from pandas import read_excel
5  from numpy import mean
6  from numpy import std
7  import matplotlib
8  import pickle
9  from sklearn.impute import KNNImputer
10 from sklearn.ensemble import RandomForestClassifier
11 from sklearn.model_selection import cross_val_score
12 from sklearn.model_selection import RepeatedStratifiedKFold
13 from sklearn.pipeline import Pipeline
14 from sklearn.metrics import mean_squared_error
15
16 def select_input_output_keys(dataframe, inp_keys, out_keys):
17     """
18     Description:
19     Select the keys of input features and output target from the dataframe
20     returning two arrays of the values for the input and output
21     the returning arrays will be of shape length of dataframe x length of keys
22     """
23     Input Parameters:
24     dataframe: (pandas object) dataframe containing the given keys that will be split into input and output
25     inp_keys: (list) contains the keys and the sequence that will be retrieved for input features
26     out_keys: (list) contains the keys and the sequence that will be retrieved for output targets
27
28     Returns:
29     inp_feats: (array) values of the keys columns with the same sequence as inp_keys
30     out_targets: (array) values of the keys columns with the same sequence as out_keys
31     """

```

Figure 12: KNN imputation algorithm codes (cont.):

```

KNN_trial3.py > ...
32 values_df = dataframe.values
33 keys_df = dataframe.keys().values
34 # Get true value on the indices of the input keys
35 inp_indices = [x in inp_keys for x in keys_df]
36 inp_feats = values_df[:,inp_indices]
37 # Get true value on the indices of the output keys
38 out_indices = [x in out_keys for x in keys_df]
39 # make the predictions 0 and 1 instead of 1 and 2 where 0 is for no and 1 is for yes
40 out_targets = values_df[:,out_indices]
41 out_targets = out_targets.astype('int')
42
43 return inp_feats, out_targets
44
45
46 def Get_Error(df1, df2, cols):
47     euc_dist = np.linalg.norm(df1[cols].values - df2[cols].values,
48                             axis=0)
49     mse_list = []
50     for col in cols:
51         mse_list.append(mean_squared_error(df1[col].values, df2[col].values))
52     return euc_dist, mse_list
53
54 imputed_keys = ['WBCs', 'Lymphocytes', 'Neutrophils', 'BG', 'RBCs', 'PLT', 'HCT', 'MCV', 'MCH', 'HGB', 'GFR', 'SCR', 'T.bilirubin', 'ALT', 'AST', 'CRP']
55
56 ## Load Dataset
57 dataframe = pd.read_excel(r'Z:\SQL exports\Model training preop\Trials for missing data 4.3.23\excels\misstrial2.xls', na_values='')
58
59 ## Summarise the first few rows
60 print(dataframe.head())
61

```

Figure 13: KNN imputation algorithm codes (cont.)

```

63 Before_imputation,_ = select_input_output_keys(dataframe,imputed_keys,[])
64
65 print('Data before performing imputation\n')
66 imputer = KNNImputer(n_neighbors=5, weights='uniform', metric='nan_euclidean')
67 imputer.fit(Before_imputation)
68 After_imputation = imputer.fit_transform(Before_imputation)
69 print('\n\nAfter performing imputation\n', After_imputation)
70
71 ##convert array into a dataframe and save excel sheet
72 New_dataframe = pd.DataFrame(After_imputation)
73 New_dataframe.columns = imputed_keys
74 New_dataframe.to_excel(r'Z:\SQL exports\Model training preop\Trials for missing data 4.3.23\imp.xls', index=False)
75 pickle.dump(imputer, open(r'Z:\SQL exports\Model training preop\Trials for missing data 4.3.23\knn_imputer_without_infection.p', 'wb'))
76
77
78 ## Load Dataset
79
80 imputer = pickle.load(open(r'Z:\SQL exports\Model training preop\Trials for missing data 4.3.23\knn_imputer_without_infection.p', 'rb'))
81 complete_df = pd.read_excel(r'Z:\SQL exports\Model training preop\Trials for missing data 4.3.23\excels\misstrial2.xls', na_values='')
82 missing_df = pd.read_excel(r'Z:\SQL exports\Model training preop\Trials for missing data 4.3.23\excels\misstrial2.xls', na_values='')
83
84 miss_data_list,_ = select_input_output_keys(missing_df,imputed_keys,[])
85 complete_data_list,_ = select_input_output_keys(complete_df,imputed_keys,[])
86
87 imputed_df = pd.DataFrame(imputer.transform(miss_data_list))
88 complete_df = pd.DataFrame(complete_data_list)
89 complete_df.columns = imputed_keys
90 imputed_df.columns = complete_df.columns
91 ##calculate mean square root error for each variable:
92 euc_distance, mse = Get_Error(complete_df, imputed_df, complete_df.columns)
93 print(euc_distance)
94 print(mse)
95

```

4.4. Step 4: Features selection (elimination):

Now after preparing the data, the following step is selecting the features of high correlation with the output. This correlation can be either assessed by statistical methods using regression analysis or using machine learning methods through feature elimination algorithm. For the present model, I used a “medical feature selection” approach which involved using clinical input to make an initial filtration of the 92 variables, in order to select all the clinically relevant variables, that can also be obtained in a clinical environment, to improve the usability and clinical application of the model. 74 variables were eligible for the second feature selection stage using machine learning. Five machine learning algorithms were used for the selection of features that will build the model. These algorithms are Logistic regression, Random forest classifiers, decision trees, gradient booster classifier, and Ridge classifier. Recursive feature elimination method with cross validation (RFECV) was used. In the present method, both methods were tried and compared. It has been shown that the use of feature elimination provided more clinically relevant features compared to statistical regression. Additionally, machine learning algorithm can use a “cross validation” method that provides more precise estimates that depends on predictions, compared to statistical regression which depends mainly on data inferences. Accordingly, machine learning feature elimination technique was the mathematical approach used to decide on the features that will be taken forward to train the model. Features selected by the statistical regression are described in table 7.

4.5. Splitting dataset into train and test sets

4.5.1. Splitting dataset and saving pickle files:

In order to select the features that will be used to train the model, the data frame is split into training and testing datasets and saving them into pickle files. Figure 14 shows the codes to split the dataset and save it into pickle files. Steps are as follows:

- 1- Import the installed library “SK learn”
- 2- Import the function from the library to split the files (line 3)
- 3- Identify the training and testing outputs and features, and select the percent needed for the test dataset. In our example, we divided the dataset into 80 to 20%, as shown in line 6 in the script.
- 4- Import pickle function to save pickle files. (Line 9)
- 5- Save pickle files (line 11)
- 6- The pickle file name will be “Train_Test.p” (important to memorise to be recalled later).

Figure 14: Data splitting and saving as pickle files:

```

1  import sklearn
2  # Import the library responsible for splitting the train and test data
3  from sklearn.model_selection import train_test_split
4
5  # Split the features and corresponding the output by 20% of the data
6  train_feats, test_feats, out_train, out_test = train_test_split(input_feats, output, test_size=0.2)
7
8  # Import the package responsible for saving python objects
9  import pickle
10 # Save the data as pickle files to be loaded directly in the next tasks
11 pickle.dump((train_feats, test_feats, out_train, out_test, keys_of_feats), open("Train_Test.p", "wb"))

```

4.5.2. Feature selection using RFECV:

Now after splitting the dataset into training and test, we will import the pickle for training dataset and start the features selection process using the RFECV method, to identify the predictors that will be used to develop the model. Figure 15, 16 and 17 demonstrate the codes used for features selection. Steps to run the RFECV:

- 1- Load the pickle file for training
- 2- Import RFECV from the SKlearn library.
- 3- Import each of the five algorithms (logistic regression, decision trees, random forest, gradient booster classifier, and ridge classifier).
- 4- Rank the selected features based on their correlation to the output for each model (lines 16, 18 in the regression model and similar ones in the rest of models).

Figure 15: Features selection using RFECV:

```

1  import pickle
2
3  #Load the saved pickle file in variables directly
4  train_feats, test_feats, out_train, out_test, keys_of_feats = pickle.load(open("Train_Test.p", "rb"))
5
6  from sklearn.feature_selection import RFECV
7
8  from sklearn.linear_model import LogisticRegression
9
10 ## Logistic Regression with parameters to make it converge ;)
11 model = LogisticRegression(penalty='l1', solver='liblinear')
12 rfecv = RFECV(model, step=1)
13 fit = rfecv.fit(train_feats, out_train)
14 logistic_ranking = fit.ranking_
15 ## Get the key names that have the highest ranked score for logistic regression
16 ranked_keys_logistic = keys_of_feats[logistic_ranking==1]
17
18
19 from sklearn.tree import DecisionTreeClassifier
20
21 ## Decision Tree classifier
22 model = DecisionTreeClassifier()
23 rfecv = RFECV(model, step=1)
24 fit = rfecv.fit(train_feats, out_train)
25 decision_ranking = fit.ranking_
26 ## Get the key names that have the highest ranked score for decision tree
27 ranked_keys_decision = keys_of_feats[decision_ranking==1]
28

```

Figure 16: Features selection using RFECV (cont.):

```

29
30 from sklearn.ensemble import RandomForestClassifier
31
32 ## Random Forest classifier
33 model = RandomForestClassifier()
34 rfecv = RFECV(model, step=1)
35 fit = rfecv.fit(train_feats, out_train)
36 random_ranking = fit.ranking_
37 ## Get the key names that have the highest ranked score for decision tree
38 ranked_keys_random = keys_of_feats[random_ranking==1]
39
40 from sklearn.linear_model import RidgeClassifier
41
42 ## Ridge Classifier
43 model = RidgeClassifier()
44 rfecv = RFECV(model, step=1)
45 fit = rfecv.fit(train_feats, out_train)
46 ridge_ranking = fit.ranking_
47 ## Get the key names that have the highest ranked score for ridge
48 ranked_keys_ridge = keys_of_feats[ridge_ranking==1]
49
50 ##Gradient boosting classifier
51 from sklearn.ensemble import GradientBoostingClassifier
52
53 model = GradientBoostingClassifier()
54 rfecv = RFECV(model, step=1)
55 fit = rfecv.fit(train_feats, out_train)
56 gradient_ranking = fit.ranking_
57 ## Get the key names that have the highest ranked score for decision tree
58 ranked_keys_gradient = keys_of_feats[gradient_ranking==1]
59
60

```

Figure 17: Features selection using RFECV (cont.):

```

60
61 from sklearn.linear_model import Perceptron
62
63 ## Perceptron
64 model = Perceptron()
65 rfecv = RFECV(model, step=1)
66 fit = rfecv.fit(train_feats, out_train)
67 perceptron_ranking = fit.ranking_
68 ## Get the key names that have the highest ranked score for perceptron
69 ranked_keys_perceptron = keys_of_feats[perceptron_ranking==1]
70

```

4.5.3. Voting for the features selection:

After the RFECV is processed using the five models, a voting technique is used to choose the features that were highly correlated in more than model. The cut off number of models should be decided. This is the number of models that selected a particular features to be included as model predictors. For example, if the cut off vote is 3, it means that a feature should be selected at least with three models as significantly correlated to the output, in order to appear in the voted list of selected features.

Steps:

- 1- Voting function to collect votes for the highest ranked features.

- 2- Looping on each feature in each model to determine the features that were selected (+=1).
- 3- Repeat for every model of the five models.
- 4- Determine the cut off number of models for votes. In this example, 3 cut off votes was chosen. It means that a feature has to be selected by at least half of the used models to be included in the final voted list of features.
- 5- Codes from line 101 to 110 are to visualise the features as names and indices-their order in the data frame.
- 6- Codes from 112 to 117 are to save a new pickle file with the chosen features that will be used from now on for the model training and testing.

Figures 18 and 19 demonstrate the codes for voting technique.

Figure 18. Codes for voting technique:

```

72 # Collect votes for the highest ranked keys of the fit models
73 voted_dict = dict()
74 # Loop on each key for the key list
75 for each_item in keys_of_feats:
76     # Initialize the vote of the key by zero
77     voted_dict[each_item] = 0
78     # Increment the vote of the key by 1 if it is mentioned in the ranked list
79     # Key mentioned in logistic regression ranking
80     if each_item in ranked_keys_logistic:
81         voted_dict[each_item] += 1
82     # Key mentioned in decision tree ranking
83     if each_item in ranked_keys_decision:
84         voted_dict[each_item] += 1
85     # Key mentioned in random forest ranking
86     if each_item in ranked_keys_random:
87         voted_dict[each_item] += 1
88     # Key mentioned in Gradient boosting classifier
89     if each_item in ranked_keys_gradient:
90         voted_dict[each_item] += 1
91     # Key mentioned in Ridge
92     if each_item in ranked_keys_ridge:
93         voted_dict[each_item] += 1
94     # Key mentioned in Perceptron
95     if each_item in ranked_keys_perceptron:
96         voted_dict[each_item] += 1
97

```

Figure 19. Codes for voting technique (cont.):

```

97
98 # select the highest voted keys from train and test feats
99 # determine which vote value will be accepted
100 cut_off_vote = 3
101 # Initialize a list to import the selected keys index corresponds to the features
102 selected_keys_indices = []
103 selected_keys = []
104 keys_of_feats = list(keys_of_feats)
105 # Loop on the voted dictionary
106 for each_key in voted_dict:
107     # Check if the value of the vote equal or exceed our threshold then add it to the selected list
108     if voted_dict[each_key] >= cut_off_vote:
109         selected_keys_indices.append(keys_of_feats.index(each_key))
110         selected_keys.append(each_key)
111
112 # Get the selected features for train and test data
113 selected_train = train_feats[:,selected_keys_indices]
114 selected_test = test_feats[:,selected_keys_indices]
115
116 pickle.dump( (selected_train, selected_test, out_train, out_test,selected_keys) , open( "Train_Test_Selected.p", "wb" ) )
117 print (".....")
118

```

The features selected with these codes were: demographic data (gender), comorbidities (diabetes, COPD, cardiac, obesity, anemia), microbiology data (positive cultures within 90 days before operation, multi drug resistance in this culture – resistance to three antibiotics from different classes), Lab investigations within 90 days before operation: WBCs, Lymphocytes, blood glucose, Platelets, hematocrit, mean cell hemoglobin, hemoglobin, total bilirubin, Aspartate aminotransferase, alanine transaminase, CRP).

It is worth to mention that the number of features on the final selected list can be controlled by the number of cut-off votes. The higher the number of cut-off votes, the lower the number of selected features after voting techniques, and the stronger the correlation between the features and the output. Usually an odd number of cut off votes is expected to result in higher performance measures. In this example, the cut-off votes were 3 out of 5 models.

4.6. Step 5: Training of the predictive model:

Now after features selection and saving a pickle file with the selected features, the pickle training file is ready to train the predictive model. Twelve different algorithms were used for the model training, including decision trees, a random forest (RF) classifier, a random forest classifier with grid search, logistic regression, a decision tree, a decision tree with grid search, a gradient boosting classifier, gradient boosting classifier with grid search, a ridge classifier, a ridge classifier with cross validation, Support vector classifier, and k-nearest neighbours. After training, the algorithm with the best performance can be chosen for the model development. Figure 20 demonstrate the codes for model development. Steps:

- 1- Continuing on the RFECV codes, start with one of the algorithms and repeat the following steps for each algorithm.
- 2- Import the algorithm from the relevant library (for logistic regression, Line 146)
- 3- Train the model using the selected features in the saved pickle file and the output (line 151)
- 4- Test the trained model on the 20% of the data frame that is used for testing (line 153)
- 5- Write commands to show performance of the model (classification report) (Line 155)
- 6- Write commands to show the confusion matrix (observed versus predicted)
- 7- Repeat steps from 2 to 6 for each of the remaining 11 algorithms. An example of codes written for one algorithm is shown in figure 20. The full codes for the algorithms is in appendix F.

Figure 20. Codes for training models:

```

64
65 def Train_Logistic_Regression(train_feats, train_targets, valid_feats, valid_targets ):
66     ...
67     Description:
68     Using Logistic Regression for binary classification returning the model and valid accuracies
69     -----
70     Input Parameters:
71     train_feats: (array) training features to be fitting the model on it
72     train_targets: (array) target for the training features
73     valid_feats: (array) valid features that will be used in reporting accuracy
74     valid_targets: (array) target for validation data
75
76     Returns:
77     model: (object) model object after being fit
78     acc_param: (dict) dictionary contain the TP,TN,FP,FN of the validation dataset
79     ...
80     model = LogisticRegression(solver='liblinear', penalty='l1', max_iter=300)
81     model.fit(train_feats, train_targets)
82     valid_pred = model.predict(valid_feats)
83     conf_mat = confusion_matrix(valid_targets, valid_pred)
84     valid_result = get_acc_from_conf(conf_mat)
85     print (valid_result)
86     roc = roc_auc_score(valid_targets, valid_pred)
87     result = get_acc_from_conf(conf_mat)
88     result ["AUCROC"]=roc
89     print(result)
90
91
92     return model, valid_result
93

```

After running the codes, the performance measures and confusion matrix will appear in the terminal for each model. The Gradient boosting classifier was the best performing model. An ensemble model was then developed using the best performing 3 models.

4.7. Step 6: Building user interface:

After developing the model, it has to be available in a user-friendly form. The model was given a name: Artificial intelligence for sepsis prediction in surgery (**AIPIS**), which will be used in the rest of this chapter. Codes to build the user interface are shown below in figure 21, 22, 23 and 24. However, these codes in this section are only for demonstration and cannot be run as such. In appendix F, the researcher has cleaned all the scripts mentioned in this section and all previous sections and collected them into five collective large scripts to run sequentially to produce the user interface that is shown in figure 22,23 and 24. The aim of these collective scripts is to minimise the incidence of errors in codes writing, ease the process of recalling functions for AIPIS training and to structure more organised scripts to refer to for any update or maintenance for AIPIS in the future.

Steps:

- 1- If using Anaconda interface for the first time, Open Anaconda software from the windows icon → environment → base → open terminal → run “pip install streamlit” → press enter (as shown in figure 21)
- 2- Return back to visual studio interface for python. In the terminal, choose command prompt from the right side from the “+” sign.
- 3- In the terminal type: streamlit run file name.py and press enter.
- 4- A link will appear on the terminal, press on the link.

Figure 21. installing streamlit package:

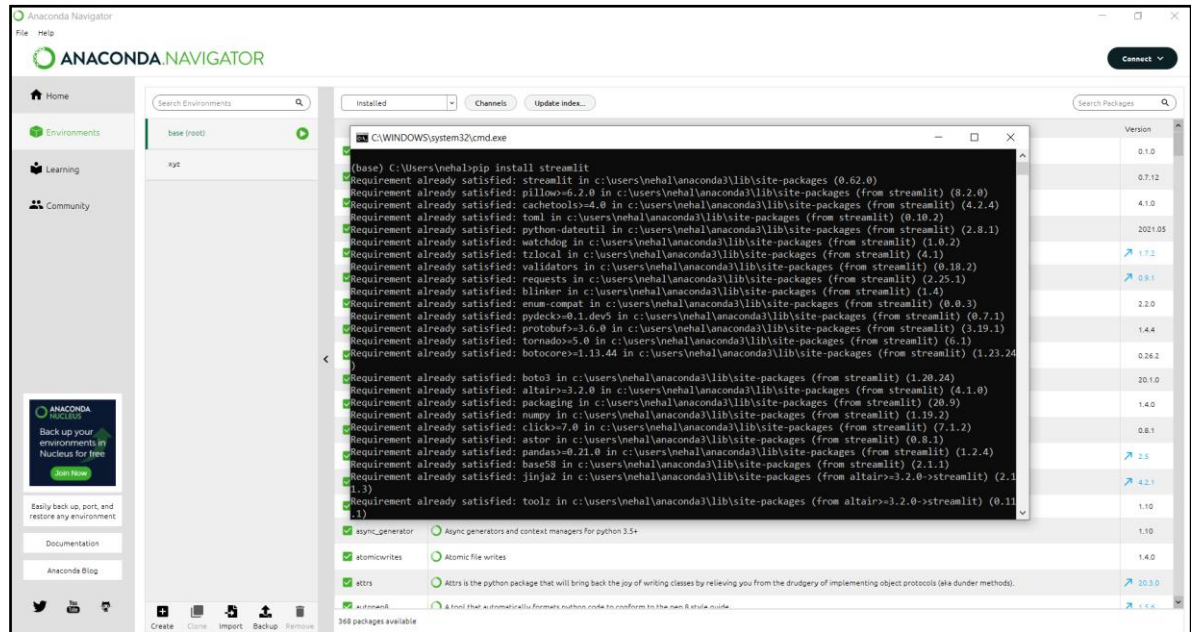


Figure 22. Codes for building user interface:

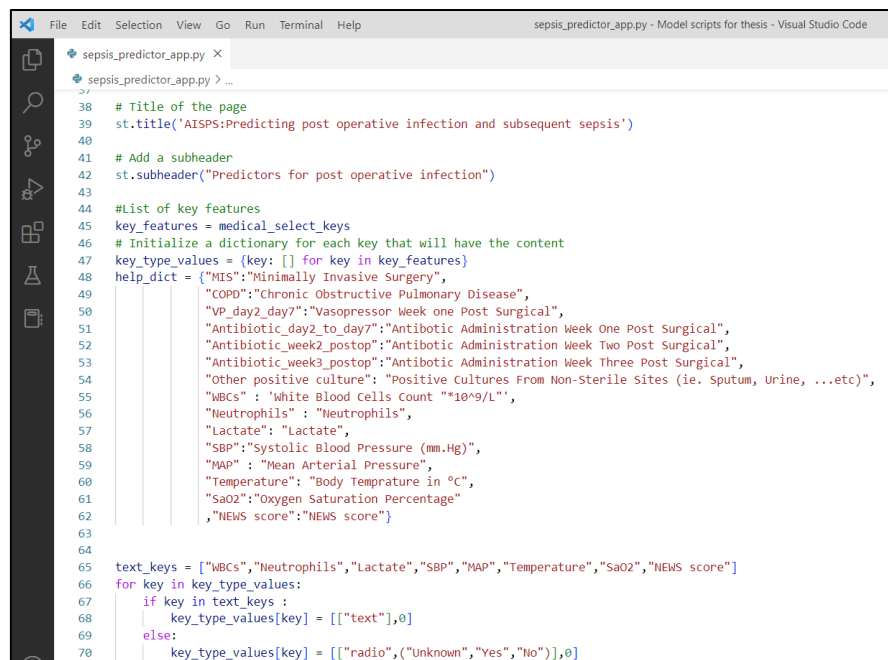
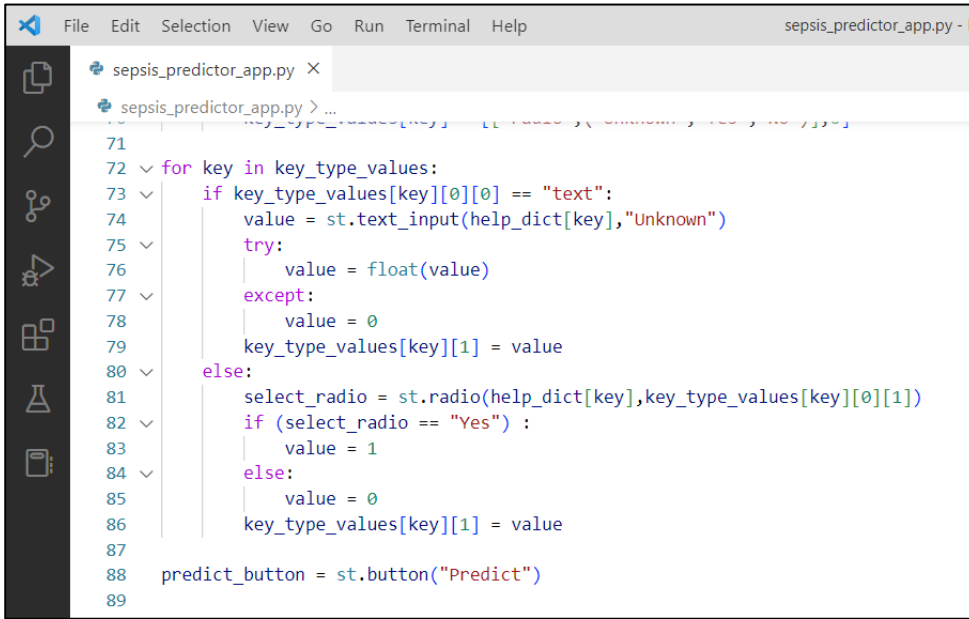


Figure 23. Codes for building user interface (cont.):

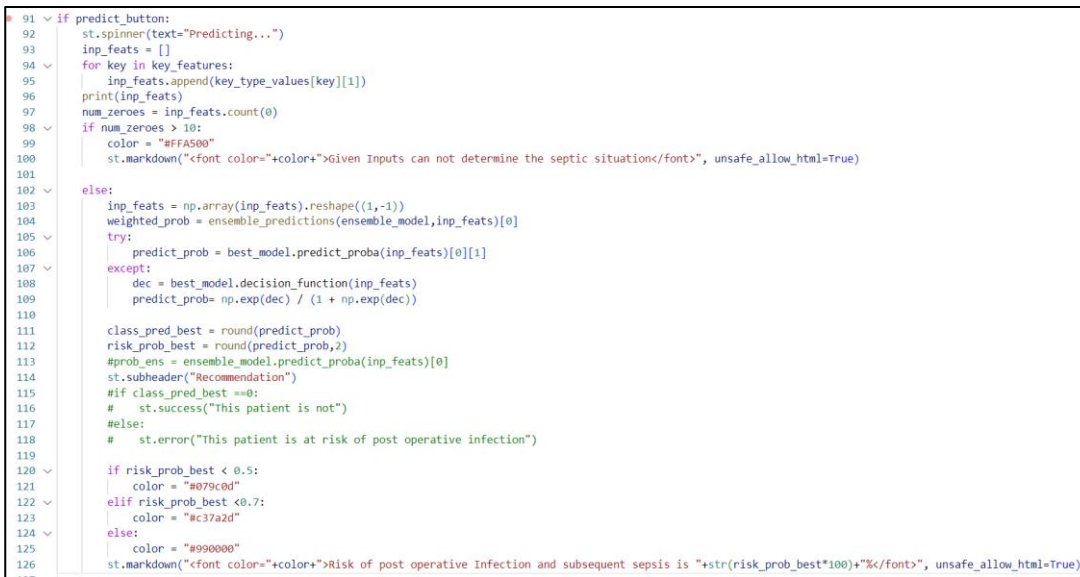


```

71
72 for key in key_type_values:
73     if key_type_values[key][0][0] == "text":
74         value = st.text_input(help_dict[key], "Unknown")
75         try:
76             value = float(value)
77         except:
78             value = 0
79         key_type_values[key][1] = value
80     else:
81         select_radio = st.radio(help_dict[key], key_type_values[key][0][1])
82         if (select_radio == "Yes") :
83             value = 1
84         else:
85             value = 0
86         key_type_values[key][1] = value
87
88 predict_button = st.button("Predict")
89

```

Figure 24. Codes for building user interface (cont.):



```

91 if predict_button:
92     st.spinner(text="Predicting...")
93     inp_feats = []
94     for key in key_features:
95         inp_feats.append(key_type_values[key][1])
96     print(inp_feats)
97     num_zeroes = inp_feats.count(0)
98     if num_zeroes > 10:
99         color = "#FFA500"
100         st.markdown("<font color='"+color+"'>Given Inputs can not determine the septic situation</font>", unsafe_allow_html=True)
101     else:
102         inp_feats = np.array(inp_feats).reshape((1,-1))
103         weighted_prob = ensemble_predictions(ensemble_model,inp_feats)[0]
104         try:
105             predict_prob = best_model.predict_proba(inp_feats)[0][1]
106         except:
107             dec = best_model.decision_function(inp_feats)
108             predict_prob = np.exp(dec) / (1 + np.exp(dec))
109
110         class_pred_best = round(predict_prob)
111         risk_prob_best = round(predict_prob,2)
112         #prob_ens = ensemble_model.predict_proba(inp_feats)[0]
113         st.subheader("Recommendation")
114         #if class_pred_best ==0:
115         #    st.success("This patient is not")
116         #else:
117         #    st.error("This patient is at risk of post operative infection")
118
119         if risk_prob_best < 0.5:
120             color = "#079c0d"
121         elif risk_prob_best < 0.7:
122             color = "#c37a2d"
123         else:
124             color = "#990000"
125         st.markdown("<font color='"+color+"'>Risk of post operative Infection and subsequent sepsis is "+str(risk_prob_best*100)+"%</font>", unsafe_allow_html=True)
126

```

In figure 24, these are the codes for what happens after the user presses “predict”, the word “predicting...” starts to appear till prediction is complete, which is coded in lines 91 and 92. Lines 93 to 112 demonstrate looping on the user entries processed by the developed final algorithm. Lines from 97 to 102 are codes for the result predicted, so if based on ensemble model the patient is “0”, it means that the patient is unlikely to get post-operative infection and subsequent sepsis, anything else, will be “high likelihood of infection”. Lines from 120 to 125 represents the colour code, infection will appear in red, non-infection will appear in green. The cut off points (0.5 and 0.7), are set based on Youden index calculation.

Appendix F

Favorable opinion letter


**Health Research
Authority**
East Midlands - Leicester Central Research Ethics Committee
 The Old Chapel
 Royal Standard Place
 Nottingham
 NG1 6FS
 Telephone: 0207 972 2568
 Fax:

Please note: This is the favourable opinion of the REC only and does not allow you to start your study at NHS sites in England until you receive HRA Approval

03 August 2020

Dr Sarah Slight
 School of Pharmacy, Newcastle University
 King George VI Building
 Newcastle Upon Tyne
 NE1 7RU

Dear Dr Slight

Study title:	Sepsis Prevention in Surgery: Informing the Clinical Process through Artificial Intelligence
REC reference:	20/EM/0183
IRAS project ID:	280077

Thank you for your letter of 31 July 2020, responding to the Proportionate Review Sub-Committee's request for changes to the documentation for the above study.

The revised documentation has been reviewed and approved on behalf of the PR sub-committee.

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form, protocol and supporting documentation as revised.

Conditions of the favourable opinion

The REC favourable opinion is subject to the following conditions being met prior to the start of the study.

Confirmation of Capacity and Capability (in England, Northern Ireland and Wales) or NHS management permission (in Scotland) should be sought from all NHS organisations involved in the study in accordance with NHS research governance arrangements. Each NHS organisation must confirm through the signing of agreements and/or other documents that it has given permission for the research to proceed (except where explicitly specified otherwise).

Guidance on applying for HRA and HCRW Approval (England and Wales)/ NHS permission for research is available in the Integrated Research Application System.

For non-NHS sites, site management permission should be obtained in accordance with the procedures of the relevant host organisation.

Sponsors are not required to notify the Committee of management permissions from host organisations.

Registration of Clinical Trials

It is a condition of the REC favourable opinion that **all clinical trials are registered** on a publicly accessible database. For this purpose, 'clinical trials' are defined as the first four project categories in IRAS project filter question 2. Registration is a legal requirement for clinical trials of investigational medicinal products (CTIMPs), except for phase I trials in healthy volunteers (these must still register as a condition of the REC favourable opinion).

Registration should take place as early as possible and within six weeks of recruiting the first research participant at the latest. Failure to register is a breach of these approval conditions, unless a deferral has been agreed by or on behalf of the Research Ethics Committee (see here for more information on requesting a deferral:

<https://www.hra.nhs.uk/planning-and-improving-research/research-planning/research-registration-research-project-identifiers/>

As set out in the UK Policy Framework, research sponsors are responsible for making information about research publicly available before it starts e.g. by registering the research project on a publicly accessible register. Further guidance on registration is available at: <https://www.hra.nhs.uk/planning-and-improving-research/research-planning/transparency-responsibilities/>

You should notify the REC of the registration details. We routinely audit applications for compliance with these conditions.

Publication of Your Research Summary

We will publish your research summary for the above study on the research summaries section of our website, together with your contact details, no earlier than three months from the date of this favourable opinion letter. Should you wish to provide a substitute contact point, make a request to defer, or require further information, please visit:

<https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/research-sum>

maries/

For research studies related to COVID-19, we are fast-tracking the publication of research summaries. During this public health emergency, it is vital that everyone can promptly identify all relevant research related to COVID-19 that is taking place globally. If you have not already done so, please register your study on a public registry as soon as possible and provide the HRA with the registration detail, which will be posted alongside other information relating to your project. We are also asking sponsors not to request deferral of publication of research summary for any projects relating to COVID-19. In addition, to facilitate finding and extracting studies related to COVID-19 from public databases, please enter the WHO official acronym for the coronavirus disease (COVID-19) in the full title of your study. Approved COVID-19 studies can be found at: <https://www.hra.nhs.uk/covid-19-research/approved-covid-19-research/>

It is the responsibility of the sponsor to ensure that all the conditions are complied with before the start of the study or its initiation at a particular site (as applicable).

After ethical review: Reporting requirements

The attached document "After ethical review – guidance for researchers" gives detailed guidance on reporting requirements for studies with a favourable opinion, including:

- Notifying substantial amendments
- Adding new sites and investigators
- Notification of serious breaches of the protocol
- Progress and safety reports
- Notifying the end of the study, including early termination of the study
- Final report

The latest guidance on these topics can be found at <https://www.hra.nhs.uk/approvals-amendments/managing-your-approval/>.

Ethical review of research sites

The favourable opinion applies to all NHS/HSC sites taking part in the study, subject to management permission being obtained from the NHS/HSC R&D office prior to the start of the study (see "Conditions of the favourable opinion" above).

Approved documents

The documents reviewed and approved by the Committee are:

<i>Document</i>	<i>Version</i>	<i>Date</i>
IRAS Application Form [IRAS_Form_30062020]		30 June 2020
IRAS Application Form XML file [IRAS_Form_30062020]		30 June 2020
IRAS Checklist XML [Checklist_30062020]		30 June 2020
Letter from sponsor [Letter from sponsor]	Version 1	22 June 2020
Other [Summary CV - RS]		
Other [Responses to REC]	Version 1	31 July 2020

Research protocol or project proposal [Research protocol]	Version 2 (UPDATED)	31 July 2020
Summary CV for Chief Investigator (CI) [CV for Chief investigator]	Version 1	23 June 2020
Summary CV for student [CV for Student]	Version 1	22 June 2020
Summary CV for supervisor (student research) [CV for Supervisor]	Version 1	23 June 2020

Statement of compliance

The Committee is constituted in accordance with the Governance Arrangements for Research Ethics Committees and complies fully with the Standard Operating Procedures for Research Ethics Committees in the UK.

User Feedback

The Health Research Authority is continually striving to provide a high quality service to all applicants and sponsors. You are invited to give your view of the service you have received and the application procedure. If you wish to make your views known please use the feedback form available on the HRA website:

<http://www.hra.nhs.uk/about-the-hra/governance/quality-assurance/>

HRA Learning

We are pleased to welcome researchers and research staff to our HRA Learning Events and online learning opportunities– see details at:

<https://www.hra.nhs.uk/planning-and-improving-research/learning/>

IRAS project ID: 280077 correspondence	Please quote this number on all
---	---------------------------------

With the Committee's best wishes for the success of this project.

Yours sincerely



Pp Mrs Lynne Fryatt
Chair

Email: leicestercentral.rec@hra.nhs.uk

Enclosures: "After ethical review – guidance for researchers"

Copy to: Mrs Kay Howes

Lead Nation:
England: approvals@hra.nhs.uk

Appendix G

HRA approval



Dr Sarah Slight
School of Pharmacy, Newcastle University
King George VI Building
Newcastle Upon Tyne
NE1 7RU

Email: approvals@hra.nhs.uk
HCRW.approvals@wales.nhs.uk

03 August 2020

Dear Dr Slight

**HRA and Health and Care
Research Wales (HCRW)
Approval Letter**

Study title:	Sepsis Prevention in Surgery: Informing the Clinical Process through Artificial Intelligence
IRAS project ID:	280077
REC reference:	20/EM/0183
Sponsor	Newcastle University

I am pleased to confirm that [HRA and Health and Care Research Wales \(HCRW\) Approval](#) has been given for the above referenced study, on the basis described in the application form, protocol, supporting documentation and any clarifications received. You should not expect to receive anything further relating to this application.

Please now work with participating NHS organisations to confirm capacity and capability, in line with the instructions provided in the "Information to support study set up" section towards the end of this letter.

How should I work with participating NHS/HSC organisations in Northern Ireland and Scotland?

HRA and HCRW Approval does not apply to NHS/HSC organisations within Northern Ireland and Scotland.

If you indicated in your IRAS form that you do have participating organisations in either of these devolved administrations, the final document set and the study wide governance report (including this letter) have been sent to the coordinating centre of each participating nation. The relevant national coordinating function/s will contact you as appropriate.

Please see [IRAS Help](#) for information on working with NHS/HSC organisations in Northern Ireland and Scotland.

How should I work with participating non-NHS organisations?

HRA and HCRW Approval does not apply to non-NHS organisations. You should work with your non-NHS organisations to [obtain local agreement](#) in accordance with their procedures.

What are my notification responsibilities during the study?

The standard conditions document "[After Ethical Review – guidance for sponsors and investigators](#)", issued with your REC favourable opinion, gives detailed guidance on reporting expectations for studies, including:

- Registration of research
- Notifying amendments
- Notifying the end of the study

The [HRA website](#) also provides guidance on these topics, and is updated in the light of changes in reporting expectations or procedures.

Who should I contact for further information?

Please do not hesitate to contact me for assistance with this application. My contact details are below.

Your IRAS project ID is 280077. Please quote this on all correspondence.

Yours sincerely,
Barbara Cuddon

Approvals Specialist

Email: approvals@hra.nhs.uk

Copy to: Mrs Kay Howes

Appendix H

The final Scripts used on Visual Studio code to train and validate the model.

Run sequentially in this order and put all in one folder with the relevant excel sheets before running)

Defining Machine learning algorithms

```
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.ticker as ticker
import pandas as pd
from ast import import increment_lineno
from matplotlib.ticker import import NullFormatter
from sklearn.neighbors import KNeighborsClassifier
from sklearn import metrics
from sklearn.linear_model import perceptron
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.preprocessing import StandardScaler
from sklearn import datasets
from sklearn import metrics
from sklearn.model_selection import train_test_split
from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.linear_model import LogisticRegressionCV
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import RidgeClassifier, RidgeClassifierCV
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.metrics import confusion_matrix

def get_acc_from_conf(conf_mat):
    """
    Description:
    Calculate the accuracy parameters from the confusion matrix for all classes
    reporting sensitivity, specificity, positive predictive value
    negative predictive value, false positive rate, false negative rate, false discovery rate
    overall accuracy
    ---
    Input Parameters:
        conf_mat: (array) confusion matrix of the predictions

    Returns
        result: (dict) containing the accuracy parameters for the given confusion matrix
    """
    FP = conf_mat.sum(axis=0) - np.diag(conf_mat)
```

```

FP = FP[1]
FN = conf_mat.sum(axis=1) - np.diag(conf_mat)
FN = FN[1]
TP = np.diag(conf_mat)
TP = TP[1]
TN = conf_mat.sum() - (FP + FN + TP)
SENS = TP/(TP+FN)
SPEC = TN/(TN+FP)
PPV = TP/(TP+FP)
NPV = TN/(TN+FN)
FPR = FP/(FP+TN)
FNR = FN/(TP+FN)
FDR = FP/(TP+FP)
ACC = (TP+TN)/(TP+FP+FN+TN)
YODEN = (SENS+SPEC)-1
result = { "FP":FP, "FN":FN, "TP":TP, "TN":TN,
           "SENS":SENS, "SPEC":SPEC, "PPV":PPV,
           "NPV":NPV, "FPR":FPR, "FNR":FNR, "FDR":FDR,
           "ACC":ACC, "YODEN":YODEN }
return result
print(result)

```

```
def Train_Logistic_Regression(train_feats, train_targets, valid_feats, valid_targets):
```

```
'''
```

Description:

Using Logistic Regression for binary classification returning the model and valid accuracies

```
----
```

Input Parameters:

train_feats: (array) training features to be fitting the model on it
train_targets: (array) target for the training features
valid_feats: (array) valid features that will be used in reporting accuracy
valid_targets: (array) target for validation data

Returns:

model: (object) model object after being fit
acc_param: (dict) dictionary contain the TP,TN,FP,FN of the validation dataset

```
'''
```

```

model = LogisticRegression(solver='liblinear', penalty='l1', max_iter=300)
model.fit(train_feats, train_targets)
valid_pred = model.predict(valid_feats)
conf_mat = confusion_matrix(valid_targets, valid_pred)
valid_result = get_acc_from_conf(conf_mat)
print (valid_result)
roc = roc_auc_score(valid_targets, valid_pred)
result = get_acc_from_conf(conf_mat)
result ["AUCROC"]=roc
print(result)

```

```
return model, valid_result
```

```

def Train_KNN(train_feats, train_targets, valid_feats, valid_targets ):
    """
    Description:
    Using Logistic Regression for binary classification returning the model and valid accuracies
    ----
    Input Parameters:
        train_feats: (array) training features to be fitting the model on it
        train_targets: (array) target for the training features
        valid_feats: (array) valid features that will be used in reporting accuracy
        valid_targets: (array) target for validation data

    Returns:
        model: (object) model object after being fit
        acc_param: (dict) dictionary contain the TP,TN,FP,FN of the validation dataset
    """
    k = 6
    model = KNeighborsClassifier(n_neighbors= k).fit(train_feats,train_targets)
    model.fit(train_feats, train_targets)
    valid_pred = model.predict(valid_feats)
    print("Accuracy of model",metrics.accuracy_score(valid_targets, valid_pred))
    conf_mat = confusion_matrix(valid_targets, valid_pred)
    valid_result = get_acc_from_conf(conf_mat)
    print (valid_result)
    roc = roc_auc_score(valid_targets, valid_pred)
    result = get_acc_from_conf(conf_mat)
    result ["AUCROC"]=roc
    print(result)

    return model, valid_result

def Train_SVC(train_feats, train_targets, valid_feats, valid_targets ):
    """
    Description:
    Using Logistic Regression for binary classification returning the model and valid accuracies
    ----
    Input Parameters:
        train_feats: (array) training features to be fitting the model on it
        train_targets: (array) target for the training features
        valid_feats: (array) valid features that will be used in reporting accuracy
        valid_targets: (array) target for validation data

    Returns:
        model: (object) model object after being fit
        acc_param: (dict) dictionary contain the TP,TN,FP,FN of the validation dataset
    """
    sc = StandardScaler()
    sc.fit(train_feats)
    train_feats_std = sc.transform(train_feats)

```

```

valid_feats_std = sc.transform(valid_feats)
model = SVC(C=1.0, random_state=1, kernel='linear')
model.fit(train_feats_std, train_targets)
valid_pred = model.predict(valid_feats_std)
print("Accuracy score %.3f" %metrics.accuracy_score(valid_targets, valid_pred))
conf_mat = confusion_matrix(valid_targets, valid_pred)
print(conf_mat)
valid_result = get_acc_from_conf(conf_mat)
print (valid_result)
roc = roc_auc_score(valid_targets, valid_pred)
result = get_acc_from_conf(conf_mat)
result ["AUCROC"]=roc
print(result)

return model, valid_result

```

```
def Train_Perceptron(train_feats, train_targets, valid_feats, valid_targets ):
```

```
'''
```

Description:

Using Logistic Regression for binary classification returning the model and valid accuracies

```
----
```

Input Parameters:

train_feats: (array) training features to be fitting the model on it

train_targets: (array) target for the training features

valid_feats: (array) valid features that will be used in reporting accuracy

valid_targets: (array) target for validation data

Returns:

model: (object) model object after being fit

acc_param: (dict) dictionary contain the TP,TN,FP,FN of the validation dataset

```
'''
```

```

model = Perceptron()
model.fit(train_feats, train_targets)
valid_pred = model.predict(valid_feats)
conf_mat = confusion_matrix(valid_targets, valid_pred)
print(conf_mat)
valid_result = get_acc_from_conf(conf_mat)
print (valid_result)
roc = roc_auc_score(valid_targets, valid_pred)
result = get_acc_from_conf(conf_mat)
result ["AUCROC"]=roc
print(result)

```

```
return model, valid_result
```

```
def Train_Logistic_RegressionCV(train_feats, train_targets, valid_feats, valid_targets ):
```

```

'''
Description:
Using Logistic Regression Cross Validation for binary classification returning the model and valid
accuracies
----
Input Parameters:
    train_feats: (array) training features to be fitting the model on it
    train_targets: (array) target for the training features
    valid_feats: (array) valid features that will be used in reporting accuracy
    valid_targets: (array) target for validation data

Returns:
    model: (object) model object after being fit
    acc_param: (dict) dictionary contain the TP,TN,FP,FN of the validation dataset
'''
model = LogisticRegressionCV(cv=6,solver='liblinear', penalty='l1', max_iter=300)
model.fit(train_feats, train_targets)
valid_pred = model.predict(valid_feats)
conf_mat = confusion_matrix(valid_targets, valid_pred)
valid_result = get_acc_from_conf(conf_mat)
print (valid_result)
roc = roc_auc_score(valid_targets, valid_pred)
result = get_acc_from_conf(conf_mat)
result ["AUCROC"]=roc
print(result)

```

```

return model, valid_result

```

```

def Train_Decision_Tree(train_feats, train_targets, valid_feats, valid_targets ):
'''
Description:
Using Decision Tree for binary classification returning the model and valid accuracies
----
Input Parameters:
    train_feats: (array) training features to be fitting the model on it
    train_targets: (array) target for the training features
    valid_feats: (array) valid features that will be used in reporting accuracy
    valid_targets: (array) target for validation data

Returns:
    model: (object) model object after being fit
    acc_param: (dict) dictionary contain the TP,TN,FP,FN of the validation dataset
'''
model = DecisionTreeClassifier()
model.fit(train_feats, train_targets)
valid_pred = model.predict(valid_feats)
conf_mat = confusion_matrix(valid_targets, valid_pred)
valid_result = get_acc_from_conf(conf_mat)
print (valid_result)

```

```

roc = roc_auc_score(valid_targets, valid_pred)
result = get_acc_from_conf(conf_mat)
result ["AUCROC"]=roc
print(result)

```

```

return model, valid_result

```

```

def Train_GridSearch_Decision_Tree(train_feats, train_targets, valid_feats, valid_targets ):

```

```

'''

```

Description:

Using Grid Search Decision Tree for binary classification returning the model and valid accuracies

It can be used to find the best depth for decision tree using cross validation

```

----

```

Input Parameters:

train_feats: (array) training features to be fitting the model on it

train_targets: (array) target for the training features

valid_feats: (array) valid features that will be used in reporting accuracy

valid_targets: (array) target for validation data

Returns:

model: (object) model object after being fit

acc_param: (dict) dictionary contain the TP,TN,FP,FN of the validation dataset

```

'''

```

```

parameters = {'max_depth':range(3,20)}

```

```

clf = GridSearchCV(DecisionTreeClassifier(), parameters, n_jobs=8)

```

```

clf.fit(X=train_feats, y=train_targets)

```

```

model = clf.best_estimator_

```

```

valid_pred = model.predict(valid_feats)

```

```

conf_mat = confusion_matrix(valid_targets, valid_pred)

```

```

valid_result = get_acc_from_conf(conf_mat)

```

```

print (valid_result)

```

```

roc = roc_auc_score(valid_targets, valid_pred)

```

```

result = get_acc_from_conf(conf_mat)

```

```

result ["AUCROC"]=roc

```

```

print(result)

```

```

return model, valid_result

```

```

def Train_Random_Forest(train_feats, train_targets, valid_feats, valid_targets ):

```

```

'''

```

Description:

Using Random Forest for binary classification returning the model and valid accuracies

```

----

```

Input Parameters:

train_feats: (array) training features to be fitting the model on it

train_targets: (array) target for the training features

valid_feats: (array) valid features that will be used in reporting accuracy

valid_targets: (array) target for validation data

Returns:

model: (object) model object after being fit

```

    acc_param: (dict) dictionary contain the TP,TN,FP,FN of the validation dataset
'''
model = RandomForestClassifier()
model.fit(train_feats, train_targets)
valid_pred = model.predict(valid_feats)
conf_mat = confusion_matrix(valid_targets, valid_pred)
valid_result = get_acc_from_conf(conf_mat)
print (valid_result)
roc = roc_auc_score(valid_targets, valid_pred)
result = get_acc_from_conf(conf_mat)
result ["AUCROC"]=roc
print(result)

return model, valid_result

def Train_GridSearch_RandomForest(train_feats, train_targets, valid_feats, valid_targets ):
'''
Description:
Using Grid Search Random Forest  returning the model and valid accuracies
----
Input Parameters:
    train_feats: (array) training features to be fitting the model on it
    train_targets: (array) target for the training features
    valid_feats: (array) valid features that will be used in reporting accuracy
    valid_targets: (array) target for validation data

Returns:
    model: (object) model object after being fit
    acc_param: (dict) dictionary contain the TP,TN,FP,FN of the validation dataset
'''
n_trees = [10, 50, 100, 500, 1000,2000]
depths = [i for i in range(1,8)] + [None]
parameters = {'max_samples':np.linspace(0.1, 0.9, num=8),'max_depth':depths,'n_estimators':n_trees}
clf = GridSearchCV(RandomForestClassifier(), parameters, n_jobs=-1)
clf.fit(X=train_feats, y=train_targets)
model = clf.best_estimator_
valid_pred = model.predict(valid_feats)
conf_mat = confusion_matrix(valid_targets, valid_pred)
valid_result = get_acc_from_conf(conf_mat)
print (valid_result)
roc = roc_auc_score(valid_targets, valid_pred)
result = get_acc_from_conf(conf_mat)
result ["AUCROC"]=roc
print(result)

return model, valid_result

def Train_Gradient_Boost(train_feats, train_targets, valid_feats, valid_targets ):
'''
Description:

```

Using Gradient Boost for binary classification returning the model and valid accuracies

Input Parameters:

train_feats: (array) training features to be fitting the model on it
train_targets: (array) target for the training features
valid_feats: (array) valid features that will be used in reporting accuracy
valid_targets: (array) target for validation data

Returns:

model: (object) model object after being fit
acc_param: (dict) dictionary contain the TP,TN,FP,FN of the validation dataset

'''

```
model = GradientBoostingClassifier()
model.fit(train_feats, train_targets)
valid_pred = model.predict(valid_feats)
conf_mat = confusion_matrix(valid_targets, valid_pred)
valid_result = get_acc_from_conf(conf_mat)
print(valid_result)
roc = roc_auc_score(valid_targets, valid_pred)
result = get_acc_from_conf(conf_mat)
result["AUCROC"]=roc
print(result)
return model, valid_result
```

```
def Train_GridSearch_Gradient_Boost(train_feats, train_targets, valid_feats, valid_targets):
```

'''

Description:

Using Grid Search Gradient Boost for binary classification returning the model and valid accuracies

Input Parameters:

train_feats: (array) training features to be fitting the model on it
train_targets: (array) target for the training features
valid_feats: (array) valid features that will be used in reporting accuracy
valid_targets: (array) target for validation data

Returns:

model: (object) model object after being fit
acc_param: (dict) dictionary contain the TP,TN,FP,FN of the validation dataset

'''

```
parameters = { "loss":["deviance"],
               "learning_rate": [0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2],
               "min_samples_split": np.linspace(0.1, 0.5, 12),
               "min_samples_leaf": np.linspace(0.1, 0.5, 12),
               "max_depth":[3,5,8],
               "max_features":["log2","sqrt"],
               "criterion": ["friedman_mse"],
               "subsample":[0.5, 0.618, 0.8, 0.85, 0.9, 0.95, 1.0],
               "n_estimators":[10]
             }
clf = GridSearchCV(GradientBoostingClassifier(), parameters, n_jobs=-1)
```

```

clf.fit(X=train_feats, y=train_targets)
model = clf.best_estimator_
valid_pred = model.predict(valid_feats)
conf_mat = confusion_matrix(valid_targets, valid_pred)
valid_result = get_acc_from_conf(conf_mat)
print (valid_result)
roc = roc_auc_score(valid_targets, valid_pred)
result = get_acc_from_conf(conf_mat)
result ["AUCROC"]=roc
print(result)
return model, valid_result

```

```
def Train_Ridge(train_feats, train_targets, valid_feats, valid_targets):
```

```
'''
```

Description:

Using Ridge Classifier for binary classification returning the model and valid accuracies

```
----
```

Input Parameters:

train_feats: (array) training features to be fitting the model on it
train_targets: (array) target for the training features
valid_feats: (array) valid features that will be used in reporting accuracy
valid_targets: (array) target for validation data

Returns:

model: (object) model object after being fit
acc_param: (dict) dictionary contain the TP,TN,FP,FN of the validation dataset

```
'''
```

```

model = RidgeClassifier()
model.fit(train_feats, train_targets)
valid_pred = model.predict(valid_feats)
conf_mat = confusion_matrix(valid_targets, valid_pred)
valid_result = get_acc_from_conf(conf_mat)
print (valid_result)
roc = roc_auc_score(valid_targets, valid_pred)
result = get_acc_from_conf(conf_mat)
result ["AUCROC"]=roc
print(result)

```

```
return model, valid_result
```

```
def Train_RidgeCV(train_feats, train_targets, valid_feats, valid_targets):
```

```
'''
```

Description:

Using Ridge Classifier Cross Validation for binary classification returning the model and valid accuracies

```
----
```

Input Parameters:

train_feats: (array) training features to be fitting the model on it
train_targets: (array) target for the training features
valid_feats: (array) valid features that will be used in reporting accuracy

`valid_targets: (array)` target for validation data

Returns:

`model: (object)` model object after being fit

`acc_param: (dict)` dictionary contain the TP,TN,FP,FN of the validation dataset

'''

```
model = RidgeClassifierCV(alphas=[1e-3, 1e-2, 1e-1, 1])
```

```
model.fit(train_feats, train_targets)
```

```
valid_pred = model.predict(valid_feats)
```

```
conf_mat = confusion_matrix(valid_targets, valid_pred)
```

```
valid_result = get_acc_from_conf(conf_mat)
```

```
print (valid_result)
```

```
roc = roc_auc_score(valid_targets, valid_pred)
```

```
result = get_acc_from_conf(conf_mat)
```

```
result ["AUCROC"]=roc
```

```
print(result)
```

```
return model, valid_result
```

Defining feature selection techniques

```
from sklearn.feature_selection import RFE
from sklearn.feature_selection import RFECV
```

```
from sklearn.linear_model import LogisticRegression
from sklearn.linear_model import RidgeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import GradientBoostingClassifier
```

```
def RFE_Logistic_Regression(inp_feats, out_targets, num_selected_feats):
```

```
    """
```

Description:

Apply Recursive feature elimination using logistic regression

```
    ----
```

Input Parameters:

inp_feats: (array) input features that needed to be ranked for elimination

out_targets: (array) output targets must be the same length of the input to get the elimination based on

num_selected_feats: (int) must be less than the feature keys and it is the number of the highest ranked

keys

Returns:

feats_rank: (list) rank for each index for the input features

```
    """
```

```
    model = LogisticRegression(solver='liblinear', penalty='l1', max_iter=300)
```

```
    rfe = RFE(model, n_features_to_select=num_selected_feats, step=1)
```

```
    fit = rfe.fit(inp_feats, out_targets)
```

```
    ranking = fit.ranking_
```

```
    return ranking
```

```
def RFECV_Logistic_Regression(inp_feats, out_targets):
```

```
    """
```

Description:

Apply Recursive feature elimination Cross Validation using logistic regression

that eliminates the features from CV perspective

```
    ----
```

Input Parameters:

inp_feats: (array) input features that needed to be ranked for elimination

out_targets: (array) output targets must be the same length of the input to get the elimination based on

Returns:

feats_rank: (list) rank for each index for the input features

```
    """
```

```
    model = LogisticRegression(solver='liblinear', penalty='l1', max_iter=300)
```

```
    rfe = RFECV(model, step=1)
```

```
    fit = rfe.fit(inp_feats, out_targets)
```

```
    ranking = fit.ranking_
```

```
    return ranking
```

```

def RFE_Ridge_Classifer(inp_feats, out_targets, num_selected_feats):
    """
    Description:
    Apply Recursive feature elimination using ridge classifier
    ----
    Input Parameters:
        inp_feats: (array) input features that needed to be ranked for elimination
        out_targets: (array) output targets must be the same length of the input to get the elimination based on
        num_selected_feats: (int) must be less than the feature keys and it is the number of the highest ranked
keys

    Returns:
        feats_rank: (list) rank for each index for the input features
    """
    model = RidgeClassifier()
    rfe = RFE(model, n_features_to_select=num_selected_feats, step=1)
    fit = rfe.fit(inp_feats, out_targets)
    ranking = fit.ranking_
    return ranking

def RFECV_Ridge_Classifer(inp_feats, out_targets):
    """
    Description:
    Apply Recursive feature elimination Cross Validation using ridge classifier
    that eliminates the features from CV perspective
    ----
    Input Parameters:
        inp_feats: (array) input features that needed to be ranked for elimination
        out_targets: (array) output targets must be the same length of the input to get the elimination based on

    Returns:
        feats_rank: (list) rank for each index for the input features
    """
    model = RidgeClassifier()
    rfe = RFECV(model, step=1)
    fit = rfe.fit(inp_feats, out_targets)
    ranking = fit.ranking_
    return ranking

def RFE_Random_Forest(inp_feats, out_targets, num_selected_feats):
    """
    Description:
    Apply Recursive feature elimination using random forest
    ----
    Input Parameters:
        inp_feats: (array) input features that needed to be ranked for elimination
        out_targets: (array) output targets must be the same length of the input to get the elimination based on
        num_selected_feats: (int) must be less than the feature keys and it is the number of the highest ranked
keys

```

Returns:

feats_rank: (list) rank for each index for the input features

'''

```
model = RandomForestClassifier()
rfe = RFE(model, n_features_to_select=num_selected_feats,step=1)
fit = rfe.fit(inp_feats, out_targets)
ranking = fit.ranking_
return ranking
```

```
def RFECV_Random_Forest(inp_feats, out_targets):
```

'''

Description:

Apply Recursive feature elimination Cross Validation using random forest that eliminates the features from CV perspective

Input Parameters:

inp_feats: (array) input features that needed to be ranked for elimination

out_targets: (array) output targets must be the same length of the input to get the elimination based on

Returns:

feats_rank: (list) rank for each index for the input features

'''

```
model = RandomForestClassifier()
rfe = RFECV(model,step=1)
fit = rfe.fit(inp_feats, out_targets)
ranking = fit.ranking_
return ranking
```

```
def RFE_Gradient_Boosting(inp_feats, out_targets, num_selected_feats):
```

'''

Description:

Apply Recursive feature elimination using Gradient Boosting Classifier

Input Parameters:

inp_feats: (array) input features that needed to be ranked for elimination

out_targets: (array) output targets must be the same length of the input to get the elimination based on

num_selected_feats: (int) must be less than the feature keys and it is the number of the highest ranked

keys

Returns:

feats_rank: (list) rank for each index for the input features

'''

```
model = GradientBoostingClassifier()
rfe = RFE(model, n_features_to_select=num_selected_feats,step=1)
fit = rfe.fit(inp_feats, out_targets)
ranking = fit.ranking_
return ranking
```

```
def RFECV_Gradient_Boosting(inp_feats, out_targets):
```

'''

Description:

Apply Recursive feature elimination Cross Validation using Gradient Boosting Classifier that eliminates the features from CV perspective

Input Parameters:

inp_feats: (array) input features that needed to be ranked for elimination

out_targets: (array) output targets must be the same length of the input to get the elimination based on

Returns:

feats_rank: (list) rank for each index for the input features

'''

```
model = GradientBoostingClassifier()
```

```
rfe = RFECV(model,step=1)
```

```
fit = rfe.fit(inp_feats, out_targets)
```

```
ranking = fit.ranking_
```

```
return ranking
```

Missing data imputation using K-nearest neighbours algorithm:

```

import pandas as pd
import numpy as np
from numpy import isnan
from pandas import read_excel
from numpy import mean
from numpy import std
import matplotlib
import pickle
from sklearn.impute import KNNImputer
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import RepeatedStratifiedKFold
from sklearn.pipeline import Pipeline
from sklearn.metrics import mean_squared_error

def select_input_output_keys(dataframe, inp_keys, out_keys):
    """
    Description:
    Select the keys of input features and output target from the dataframe
    returning two arrays of the values for the input and output
    the returning arrays will be of shape length of dataframe x length of keys
    ---
    Input Parameters:
        dataframe: (pandas object) dataframe containing the given keys that will be split into input and output
        inp_keys: (list) contains the keys and the sequence that will be retrieved for input features
        out_keys: (list) contains the keys and the sequence that will be retrieved for output targets

    Returns:
        inp_feats: (array) values of the keys columns with the same sequence as inp_keys
        out_targets: (array) values of the keys columns with the same sequence as out_keys
    """
    values_df = dataframe.values
    keys_df = dataframe.keys().values
    # Get true value on the indecies of the input keys
    inp_indecies = [x in inp_keys for x in keys_df]
    inp_feats = values_df[:,inp_indecies]
    # Get true value on the indecies of the output keys
    out_indecies = [x in out_keys for x in keys_df]
    # make the predictions 0 and 1 instead of 1 and 2 where 0 is for no and 1 is for yes
    out_targets = values_df[:,out_indecies]
    out_targets = out_targets.astype('int')

    return inp_feats, out_targets

def Get_Error (df1, df2, cols):
    euc_dist = np.linalg.norm(df1[cols].values - df2[cols].values,
                              axis=0)
    mse_list = []
    for col in cols:
        mse_list.append (mean_squared_error(df1[col].values ,df2[col].values))

```

```

return euc_dist, mse_list

imputed_keys =
['WBCs','Lymphocytes','Neutrophils','BG','RBCs','PLT','HCT','MCV','MCH','HGB','GFR','SCr','T.bilirubin',
'ALT','AST','CRP']

## Load Dataset
dataframe = pd.read_excel(r"Z:\SQL exports\Model training preop\Trials for missing data
4.3.23\excels\miss2747.xls", na_values="")

## Summarise the first few rows
print(dataframe.head())

Before_imputation,_ = select_input_output_keys(dataframe,imputed_keys,[])

print ('Data before performing imputation\n')
imputer = KNNImputer(n_neighbors=5, weights='uniform', metric='nan_euclidean')
imputer.fit (Before_imputation)
After_imputation = imputer.fit_transform(Before_imputation)
print("\n\nAfter performing imputation\n", After_imputation)

##convert array into a dataframe and save excel sheet
New_dataframe = pd.DataFrame (After_imputation)
New_dataframe.columns = imputed_keys
New_dataframe.to_excel(r'Z:\SQL exports\Model training preop\Trials for missing data 4.3.23\imp.xls',
index=False)
pickle.dump (imputer, open('Z:\SQL exports\Model training preop\Trials for missing data
4.3.23\kNN_Imputer_Without_Infection.p', 'wb'))

## Load Dataset

imputer = pickle.load(open('Z:\SQL exports\Model training preop\Trials for missing data
4.3.23\kNN_Imputer_Without_Infection.p', 'rb'))
complete_df = pd.read_excel (r"Z:\SQL exports\Model training preop\Trials for missing data
4.3.23\excels\misstrial.xls", na_values="")
missing_df = pd.read_excel (r"Z:\SQL exports\Model training preop\Trials for missing data
4.3.23\excels\misstrial2.xls", na_values="")

miss_datalist,_ = select_input_output_keys(missing_df,imputed_keys,[])
complete_datalist,_ = select_input_output_keys(complete_df,imputed_keys,[])

imputed_df = pd.DataFrame(imputer.transform (miss_datalist))
complete_df = pd.DataFrame(complete_datalist)
complete_df.columns = imputed_keys
imputed_df.columns = complete_df.columns
##calculate mean square root error for each variable:
euc_distance, mse = Get_Error (complete_df, imputed_df, complete_df.columns)
print (euc_distance)

```

print (mse)

Model training

```

import pandas as pd

import pickle
from Feature_Elimination_Techniques import *
from ML_Techniques import *
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score
from matplotlib import pyplot

def read_excel_generate_numeric_values(excel_sheet_path):
    """
    Description:
    Read an excel sheet to be in a data frame
    check each column values to ensure it is always mapped to a numeric values
    mapping any unknown or NAN values into 0
    ----
    Input Parameters:
        excel_sheet_path: (string) path directory to the excel sheet

    Returns:
        numeric_df: (pandas object) dataframe while converting all values to numeric ones
    """
    df = pd.read_excel(excel_sheet_path)
    numeric_df = df.copy()

    for key in numeric_df:
        col = numeric_df[key]
        # fill any NAN value by zero
        col = col.fillna(0)
        # fill empty cells with zero
        col.loc[col.isnull()] = 0
        col.loc[col == ""] = 0
        col.loc[col == ' '] = 0
        # replace any Yes value by 1, any No value by 2 and others by 0
        ##col.replace({"Yes": 1, "No": 2, "Cannot be determined":0 }, inplace=True)
        # replace any Male value by 1, any Female value by 2
        ##col.replace({"Male": 1, "Female": 2}, inplace=True)
        numeric_df[key] = col

    return numeric_df

def select_input_output_keys(dataframe, inp_keys, out_keys):
    """
    Description:
    Select the keys of input features and output target from the dataframe
    returning two arrays of the values for the input and output
    the returning arrays will be of shape length of dataframe x length of keys

```

Input Parameters:

dataframe: (pandas object) dataframe containing the given keys that will be split into input and output
 inp_keys: (list) contains the keys and the sequence that will be retrieved for input features
 out_keys: (list) contains the keys and the sequence that will be retrieved for output targets

Returns:

inp_feats: (array) values of the keys columns with the same sequence as inp_keys
 out_targets: (array) values of the keys columns with the same sequence as out_keys

'''

```
values_df = dataframe.values
keys_df = dataframe.keys().values
# Get true value on the indecies of the input keys
inp_indecies = [x in inp_keys for x in keys_df]
inp_feats = values_df[:,inp_indecies]
# Get true value on the indecies of the output keys
out_indecies = [x in out_keys for x in keys_df]
# make the predictions 0 and 1 instead of 1 and 2 where 0 is for no and 1 is for yes
out_targets = values_df[:,out_indecies]
out_targets = out_targets.astype('int')

return inp_feats, out_targets
```

```
def eliminate_feats(train_feats, train_targets, source_keys, vote_cutoff, mode="RFE", num_feats=39):
```

'''

Description:

Eliminate the least correlated features with the target using different models
 Voting ensemble technique is used to get the highest correlated features

Input Parameters:

train_feats: (array) input features needed to used in training model
 train_targets: (array) output targets that will be used in the elimination technique
 source_keys: (list) keys of the features that will be eliminated
 mode: (string) either RFE or RFECV determining either to use a manual num of feats or not
 num_feats: (int) determine the number of features that will be used in RFE. IT WILL NOT BE USED

IN RFECV mode

vote_cutoff: (int) voting threshold for the ranking, if the key rank have this cut off or higher will be selected

Returns:

selected_keys: (list) selected keys after models ensembling and elimination whatever it is resultant from RFE or RFECV

'''

```
if mode not in ["RFE", "RFECV"]:
    return source_keys
if mode == "RFE":
    lr_ranking = RFE_Logistic_Regression(train_feats, train_targets, num_feats)
    ridge_ranking = RFE_Ridge_Classifier(train_feats, train_targets, num_feats)
    gbc_ranking = RFE_Gradient_Boosting(train_feats, train_targets, num_feats)
    rf_ranking = RFE_Random_Forest(train_feats, train_targets, num_feats)
else:
```

```

lr_ranking = RFECV_Logistic_Regression(train_feats,train_targets)
ridge_ranking = RFECV_Ridge_Classifer(train_feats,train_targets)
gbc_ranking = RFECV_Gradient_Boosting(train_feats,train_targets)
rf_ranking = RFECV_Random_Forest(train_feats,train_targets)

```

```

voted_dict = dict()
for idx,each_item in enumerate(source_keys):
    voted_dict[each_item] = 0
    if lr_ranking[idx] == 1:
        voted_dict[each_item] += 1
    if ridge_ranking[idx] == 1:
        voted_dict[each_item] += 1
    if gbc_ranking[idx] == 1:
        voted_dict[each_item] += 1
    if rf_ranking[idx] == 1:
        voted_dict[each_item] += 1

```

```

selected_keys = list()
for each_key in voted_dict:
    if voted_dict[each_key] >= vote_cutoff:
        selected_keys.append(each_key)
return selected_keys

```

```
def filter_features_by_keys(inp_feats, source_keys, target_keys):
```

```
'''
```

Description:

Filter the input features to select the target keys only knowing the key indexes from source keys

```
---
```

Input Parameters:

inp_feats: (array) input features relative to the source keys

source_keys: (list) keys of the input features

target_keys: (list) list of the selected keys that will be remained in the feats

Returns:

selected_feats: (array) selected feats from target keys

selected_keys: (list) same as target keys but sorted by the same order as feats

```
'''
```

```

selected_keys = [x for x in source_keys if x in target_keys]
source_keys_selected = [x in target_keys for x in source_keys]
selected_feats = inp_feats[:,source_keys_selected]

```

```
return selected_feats,selected_keys
```

```
def ensemble_predictions(models, test_feats):
```

```
'''
```

Description:

Get the benefits of the given model to predict more robust result

the ensembling technique is weighted average technique for probability

```
---
```

Input Parameters:

models: (list) list of model objects and its weight that will be used in ensembling
 test_feats: (array) input features for the given models that will be used in predictions

Returns:

weighted_average_prob: (array) resultant probabilities from the weighted predictions

'''

prediction_list = []

weights = []

for each_model,weight in models:

try:

predict_prob = each_model.predict_proba(test_feats)[:,-1]

except:

dec = each_model.decision_function(test_feats)

predict_prob= np.exp(dec) / (1 + np.exp(dec))

prediction_list.append(predict_prob)

weights.append(weight)

weighted_average_prob = np.average(np.array(prediction_list),axis=0,weights=weights)

return weighted_average_prob

def train_models (train_feats, train_targets, valid_feats, valid_targets, mode="ensemble", n_best=5):

'''

Description:

Training different models using the training data and choose to get the best models or ensembling models

Input Parameters:

train_feats: (array) input features needed to be used in training model

train_targets: (array) target of the input features

valid_feats: (array) input features needed to be used in evaluating model

valid_targets: (array) actual targets of the valid data

mode: (string) either ensemble or best to return the criteria that will be used in predicting

Returns:

model: (model object) either a model object in case of best mode or the list of best models

result: (dict) dictionary of the evaluation of the model on the devset

predict_prob: (array) prediction probabilities of the selected mode of the valid data

'''

model_list = []

print ("Training Logistic Regression ...")

model_list.append(Train_Logistic_Regression(train_feats,train_targets,valid_feats,valid_targets))

print ("Training K-Nearest Neighbors ...")

model_list.append(Train_KNN(train_feats, train_targets, valid_feats, valid_targets))

print ("Training Support Vector Classifier ...")

model_list.append(Train_SVC(train_feats, train_targets, valid_feats, valid_targets))

```

print ("Training Logistic Regression CV ...")
model_list.append(Train_Logistic_RegressionCV(train_feats,train_targets,valid_feats,valid_targets))

print ("Training Decision Tree ...")
model_list.append(Train_Decision_Tree(train_feats,train_targets,valid_feats,valid_targets))

print ("Training Grid Search Decision Tree ...")
model_list.append(Train_GridSearch_Decision_Tree(train_feats,train_targets,valid_feats,valid_targets))

print ("Training Random Forest ...")
model_list.append(Train_Random_Forest(train_feats,train_targets,valid_feats,valid_targets))

print ("Training GridSearch RandomForest ...")
model_list.append(Train_GridSearch_RandomForest(train_feats,train_targets,valid_feats,valid_targets))

print ("Training Gradient Boosting Classifier ...")
model_list.append(Train_Gradient_Boost(train_feats,train_targets,valid_feats,valid_targets))

print ("Training GridSearch Gradient Boosting Classifier ...")
model_list.append(Train_GridSearch_Gradient_Boost(train_feats,train_targets,valid_feats,valid_targets))
)

print ("Training Ridge Classifier ...")
model_list.append(Train_Ridge(train_feats,train_targets,valid_feats,valid_targets))

print ("Training Ridge Classifier CV ...")
model_list.append(Train_RidgeCV(train_feats,train_targets,valid_feats,valid_targets))

accuracies = [m[1][1]"ACC" for m in model_list]
best_indices = (-np.array(accuracies)).argsort()[:n_best]
best_acc = [accuracies[i] for i in best_indices]
model_weights = best_acc/best_acc[-1]
best_models_ens = [(model_list[i][0],model_weights[w]) for w,i in enumerate(best_indices)]
predict_prob_ens = ensemble_predictions(best_models_ens,valid_feats)
predict_targets_ens = np.array(predict_prob_ens>=0.5).astype(int)
conf_mat = confusion_matrix(valid_targets,predict_targets_ens)
roc = roc_auc_score(valid_targets, predict_prob_ens)
result_ens = get_acc_from_conf(conf_mat)
result_ens["AUCROC"]=roc

index = np.argmax(accuracies)
model = model_list[index][0]
result = model_list[index][1]
try:
    predict_prob = model.predict_proba(valid_feats)[:,-1]
except:
    dec = model.decision_function(valid_feats)
    predict_prob= np.exp(dec) / (1 + np.exp(dec))
roc = roc_auc_score(valid_targets, predict_prob)
result["AUCROC"] = roc

```

```

return model, result, predict_prob, best_models_ens, result_ens, predict_prob_ens

train_df = read_excel_generate_numeric_values(r"Z:\SQL exports\Model training preop\Trials for missing
data 4.3.23\ready.xls")

medical_keys = train_df.drop(['Infection', 'Serial'], axis = 1)

out_targets = ["Infection"]

medical_feats, targets = select_input_output_keys(train_df, medical_keys, out_targets)

train_med_feats, valid_med_feats, train_targets, valid_targets = train_test_split(medical_feats, targets,
test_size=0.2)

medical_select_keys =
eliminate_feats(train_med_feats, train_targets, medical_keys, 2, mode="RFE", num_feats=12)
print("Training Keys are :")
print(medical_select_keys)
print("-----")
train_sel_med_feats, medical_select_keys =
filter_features_by_keys(train_med_feats, medical_keys, medical_select_keys)
valid_sel_med_feats, medical_select_keys =
filter_features_by_keys(valid_med_feats, medical_keys, medical_select_keys)

best_model, best_result, best_predict_prob, ensemble_model, ensemble_result, ensemble_predict_prob =
train_models(train_sel_med_feats, train_targets, valid_sel_med_feats, valid_targets, mode="ensemble")
print ("Results of Ensemble Models:")
print (ensemble_result)
print ("-----")
print ("Results of Best Model:")
print (best_result)
print ("-----")

ensemble_fpr, ensemble_tpr, _ = roc_curve(valid_targets, ensemble_predict_prob)
best_fpr, best_tpr, _ = roc_curve(valid_targets, best_predict_prob)

pyplot.figure(0)
pyplot.plot(ensemble_fpr, ensemble_tpr, marker='.', label='Ensemble Models')
pyplot.xlabel('False Positive Rate')
pyplot.ylabel('True Positive Rate')
pyplot.legend()
pyplot.savefig("Ensemble_ROC_Curve2.png")

pyplot.figure(1)
pyplot.plot(best_fpr, best_tpr, marker='.', label='Best Model')
pyplot.xlabel('False Positive Rate')
pyplot.ylabel('True Positive Rate')
pyplot.legend()

```

```
pyplot.savefig("Best_ROC_Curve2.png")

pickle.dump((ensemble_model,ensemble_result,medical_select_keys),open("Ensemble_Models.p","wb"))
pickle.dump((best_model,best_result,medical_select_keys),open("Best_Model.p","wb"))
pickle.dump((train_sel_med_feats,train_targets,valid_sel_med_feats,valid_targets,medical_select_keys),open("Train_Valid_Sel_Data.p","wb"))

print (train_df)
```

Model Internal Validation

```

import pandas as pd
import pickle
from Feature_Elimination_Techniques import *
from ML_Techniques import *
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score
from matplotlib import pyplot

def read_excel_generate_numeric_values(excel_sheet_path):
    """
    Description:
    Read an excel sheet to be in a data frame
    check each column values to ensure it is always mapped to a numeric values
    mapping any unknown or NAN values into 0
    ----
    Input Parameters:
        excel_sheet_path: (string) path directory to the excel sheet

    Returns:
        numeric_df: (pandas object) dataframe while converting all values to numeric ones
    """
    df = pd.read_excel(excel_sheet_path, na_values="")
    numeric_df = df.copy()
    """
    for key in numeric_df:
        col = numeric_df[key]
        # fill any NAN value by zero
        col = col.fillna(0)
        # fill empty cells with zero
        col.loc[col.isnull()] = 0
        col.loc[col == ""] = 0
        col.loc[col == ' '] = 0
        # replace any Yes value by 1, any No value by 2 and others by 0
        ##col.replace({"Yes": 1, "No": 2, "Cannot be determined":0 }, inplace=True)
        # replace any Male value by 1, any Female value by 2
        ##col.replace({"Male": 1, "Female": 2}, inplace=True)
        numeric_df[key] = col
    """
    return numeric_df

def select_input_output_keys(dataframe, inp_keys, out_keys):
    """
    Description:
    Select the keys of input features and output target from the dataframe
    returning two arrays of the values for the input and output
    the returning arrays will be of shape length of dataframe x length of keys

```

Input Parameters:

dataframe: (pandas object) dataframe containing the given keys that will be split into input and output
 inp_keys: (list) contains the keys and the sequence that will be retrieved for input features
 out_keys: (list) contains the keys and the sequence that will be retrieved for output targets

Returns:

inp_feats: (array) values of the keys columns with the same sequence as inp_keys
 out_targets: (array) values of the keys columns with the same sequence as out_keys

'''

```
values_df = dataframe.values
keys_df = dataframe.keys().values
# Get true value on the indecies of the input keys
inp_indecies = [x in inp_keys for x in keys_df]
inp_feats = values_df[:,inp_indecies]
# Get true value on the indecies of the output keys
out_indecies = [x in out_keys for x in keys_df]
# make the predictions 0 and 1 instead of 1 and 2 where 0 is for no and 1 is for yes
out_targets = values_df[:,out_indecies]
out_targets = out_targets.astype('int')

return inp_feats, out_targets
```

```
def filter_features_by_keys(inp_feats, source_keys, target_keys):
```

'''

Description:

Filter the input features to select the target keys only knowing the key indexes from source keys

Input Parameters:

inp_feats: (array) input features relative to the source keys
 source_keys: (list) keys of the input features
 target_keys: (list) list of the selected keys that will be remained in the feats

Returns:

selected_feats: (array) selected feats from target keys
 selected_keys: (list) same as target keys but sorted by the same order as feats

'''

```
selected_keys = [x for x in source_keys if x in target_keys]
source_keys_selected = [x in target_keys for x in source_keys]
selected_feats = inp_feats[:,source_keys_selected]
```

```
return selected_feats,selected_keys
```

```
def ensemble_predictions(models, test_feats):
```

'''

Description:

Get the benefits of the given model to predict more robust result
 the ensembling technique is weighted average technique for probability

Input Parameters:

models: (list) list of model objects and its weight that will be used in ensembling
 test_feats: (array) input features for the given models that will be used in predictions

Returns:

weighted_average_prob: (array) resultant probabilities from the weighted predictions

'''

prediction_list = []

weights = []

for each_model,weight in models:

try:

predict_prob = each_model.predict_proba(test_feats)[:,-1]

except:

dec = each_model.decision_function(test_feats)

predict_prob= np.exp(dec) / (1 + np.exp(dec))

prediction_list.append(predict_prob)

weights.append(weight)

weighted_average_prob = np.average(np.array(prediction_list),axis=0,weights=weights)

return weighted_average_prob

test_df = read_excel_generate_numeric_values("Z:\SQL exports\Model training preop\Trials for missing data 4.3.23\Testing.xls")

ensemble_model,ensemble_result,medical_select_keys = pickle.load(open("Ensemble_Models.p","rb"))

best_model,best_result,medical_select_keys = pickle.load(open("Best_Model.p","rb"))

imputer = pickle.load(open('imputer.p', 'rb'))

imputed_keys =

['WBCs','Lymphocytes','Neutrophils','BG','RBCs','PLT','HCT','MCV','MCH','HGB','GFR','SCr','T.bilirubin',
 'ALT','AST','CRP','Infection']

data_miss, _ = select_input_output_keys(test_df,imputed_keys,[])

imputed_df = pd.DataFrame(imputer.transform (data_miss))

imputed_df.columns = imputed_keys

for each_key in imputed_keys:

test_df[each_key] = imputed_df[each_key]

out_targets = ["Infection"]

inp_feats, targets = select_input_output_keys(test_df,medical_select_keys,out_targets)

ensemble_prob = ensemble_predictions(ensemble_model,inp_feats)

predict_targets_ens = np.array(ensemble_prob>=0.5).astype(int)

conf_mat = confusion_matrix(targets,predict_targets_ens)

print(conf_mat)

```

roc = roc_auc_score(targets, predict_targets_ens)
result_ens = get_acc_from_conf(conf_mat)
result_ens["AUCROC"]=roc
print ("Result of Ensemble Model:")
print(result_ens)
print ("-----")

try:
    best_prob = best_model.predict_proba(inp_feats)[: ,1]
except:
    dec = best_model.decision_function(inp_feats)
    best_prob= np.exp(dec) / (1 + np.exp(dec))

predict_targets_best = np.array(best_prob>=0.5).astype(int)
conf_mat = confusion_matrix(targets,predict_targets_best)
print(conf_mat)
roc = roc_auc_score(targets, predict_targets_best)
result_best = get_acc_from_conf(conf_mat)
result_best["AUCROC"]=roc

print ("Result of Best Model:")
print(result_best)
print ("-----")

ensemble_fpr, ensemble_tpr, _ = roc_curve(targets, ensemble_prob)
best_fpr, best_tpr, _ = roc_curve(targets, best_prob)

pyplot.figure(0)
pyplot.plot(ensemble_fpr, ensemble_tpr, marker='.', label='Ensemble Models')
pyplot.xlabel('False Positive Rate')
pyplot.ylabel('True Positive Rate')
pyplot.legend()
pyplot.savefig("Test_ROC_Curve_Ensemble.png")

pyplot.figure(1)
pyplot.plot(best_fpr, best_tpr, marker='.', label='Best Model')
pyplot.xlabel('False Positive Rate')
pyplot.ylabel('True Positive Rate')
pyplot.legend()
pyplot.savefig("Test_ROC_Curve_Best.png")

```

Model Interface:

```

import streamlit as st
import numpy as np
import pandas as pd
import pickle

def select_input_output_keys(dataframe, inp_keys, out_keys):
    """
    Description:
    Select the keys of input features and output target from the dataframe
    returning two arrays of the values for the input and output
    the returning arrays will be of shape length of dataframe x length of keys
    ---
    Input Parameters:
        dataframe: (pandas object) dataframe containing the given keys that will be split into input and output
        inp_keys: (list) contains the keys and the sequence that will be retrieved for input features
        out_keys: (list) contains the keys and the sequence that will be retrieved for output targets

    Returns:
        inp_feats: (array) values of the keys columns with the same sequence as inp_keys
        out_targets: (array) values of the keys columns with the same sequence as out_keys
    """
    values_df = dataframe.values
    keys_df = dataframe.keys().values
    # Get true value on the indecies of the input keys
    inp_indecies = [x in inp_keys for x in keys_df]
    inp_feats = values_df[:,inp_indecies]
    # Get true value on the indecies of the output keys
    out_indecies = [x in out_keys for x in keys_df]
    # make the predictions 0 and 1 instead of 1 and 2 where 0 is for no and 1 is for yes
    out_targets = values_df[:,out_indecies]
    out_targets = out_targets.astype('int')

    return inp_feats, out_targets

imputer = pickle.load(open('imputer.p', 'rb'))
def impute_predictors(inp_feats):
    impute_keys =
    ['WBCs','Lymphocytes','Neutrophils','BG','RBCs','PLT','HCT','MCV','MCH','HGB','GFR','SCr','T.bilirubin',
    'ALT','AST','CRP','Infection']
    numerical_keys = ["WBCs",
    "Lymphocytes","BG","PLT","HCT","MCH","HGB","T.bilirubin","ALT","AST","CRP"]
    miss_data = np.empty((1,len(impute_keys)))
    miss_data[:] = np.nan
    for idx in range(0,len(impute_keys)):
        for idx_k in range(0,len(numerical_keys)):
            if numerical_keys[idx_k] == impute_keys[idx]:
                if inp_feats[idx_k+8] != 0:
                    miss_data[0,idx] = inp_feats[idx_k+8]
                break

```

```

print (miss_data)
imputed_df = pd.DataFrame(imputer.transform (miss_data))
imputed_df.columns=impute_keys
imputed_data, _ = select_input_output_keys(imputed_df,numerical_keys,[])
inp_feats[8:]=imputed_data[0,:]
return inp_feats

def ensemble_predictions(models, test_feats):
    """
    Description:
    Get the benefits of the given model to predict more robust result
    the ensembling technique is weighted average technique for probability
    ---
    Input Parameters:
        models: (list) list of model objects and its weight that will be used in ensembling
        test_feats: (array) input features for the given models that will be used in predictions

    Returns:
        weighted_average_prob: (array) resultant probabilities from the weighted predictions
    """
    prediction_list = []
    weights = []
    for each_model,weight in models:
        try:
            predict_prob = each_model.predict_proba(test_feats)[:,-1]
        except:
            dec = each_model.decision_function(test_feats)
            predict_prob= np.exp(dec) / (1 + np.exp(dec))
        prediction_list.append(predict_prob)
        weights.append(weight)

    weighted_average_prob = np.average(np.array(prediction_list),axis=0,weights=weights)

    return weighted_average_prob

ensemble_model,ensemble_result,medical_select_keys = pickle.load(open("Ensemble_Models.p","rb"))
best_model,best_result,medical_select_keys = pickle.load(open("Best_Model.p","rb"))

# Title of the page
st.title('AIPIS: Predicting post operative infection in surgery')

# Add a subheader
st.subheader("Predictors for post operative infection:")

#List of key features
key_features = medical_select_keys
# Initialize a dictionary for each key that will have the content
key_type_values = {key: [] for key in key_features}

```

```

help_dict = {"sex":"Gender","COPD":"Chronic Obstructive Pulmonary Disease","DM": "Diabetes Mellitus",
"Cardiac":"Cardiac diseases","Obesity":"Obesity","Anaemia":"Anaemia","Other_positive_cultures":"Positive Cultures in the last 90 days",
"MDR ":"Multi Drug Resistance - resistance to 3 antibiotics from 3 different classes","WBCs":"White Blood Cells Count *10^9/L",
"Lymphocytes":"Lymphocytes","BG":"Blood glucose mmol/L","PLT":"Platelets count *10^9/L",
"HCT":"Haematocrite value","MCH":"Mean Cell Haemoglobin","HGB":"Haemoglobin g/dl","T.bilirubin":"Total Bilirubin umol/L","ALT":"Alanine aminotransferase U/L","AST":"Aspartate aminotransferase U/L","CRP":"C-reactive protein mg/L"}

```

```

text_keys = ["WBCs",
"Lymphocytes","BG","PLT","HCT","MCH","HGB","T.bilirubin","ALT","AST","CRP"]
check_box_items = ["DM","COPD","Cardiac","Obesity","Anaemia"]

```

```

for key in key_type_values:

```

```

    if key in text_keys :
        key_type_values[key] = [{"text"},0]
    elif key == "sex":
        key_type_values[key] = [{"radio"},("Male","Female"),0]
    elif key in check_box_items:
        key_type_values[key] = [{"checkbox"},0]
    else:
        key_type_values[key] = [{"radio"},("Yes","No"),0]

```

```

for key in key_type_values:

```

```

    if key_type_values[key][0][0] == "text":
        value = st.number_input(help_dict[key],min_value=0.00)
        try:
            value = float(value)
        except:
            value = 0
        key_type_values[key][1] = value
    elif key_type_values[key][0][0] == "checkbox":
        box_checked = st.checkbox(help_dict[key])
        value = 0
        if box_checked:
            value = 1
        key_type_values[key][1] = value
    elif key == "sex":
        select_radio = st.radio(help_dict[key],key_type_values[key][0][1])
        if (select_radio == "Male"):
            value = 0
        else:
            value = 1
        key_type_values[key][1] = value
        st.text("Comorbidities")
    else:
        select_radio = st.radio(help_dict[key],key_type_values[key][0][1])
        if (select_radio == "Yes"):

```

```

        value = 0
    else:
        value = 1
    key_type_values[key][1] = value

predict_button = st.button("Predict")

if predict_button:
    st.spinner(text="Predicting...")
    inp_feats = []
    for key in key_features:
        inp_feats.append(key_type_values[key][1])
    print(inp_feats)
    inp_feats = impute_predictors(inp_feats)
    print(inp_feats)

    inp_feats = np.array(inp_feats).reshape((1,-1))
    weighted_prob = ensemble_predictions(ensemble_model,inp_feats)[0]

    risk_prob_ensemble = round(weighted_prob,2)
    st.subheader("Recommendation:")
    #if class_pred_best ==0:
    #    st.success("Likely to be Non-Septic")
    # else:
    #    st.error("Likely to be Septic")

    if risk_prob_ensemble < 0.5:
        color = "#079c0d"
    elif risk_prob_ensemble <0.7:
        color = "#c37a2d"
    else:
        color = "#990000"
    st.markdown("<font color="+color+">Risk of post operative infection is
"+str(risk_prob_ensemble*100)+"%</font>", unsafe_allow_html=True)

```

Appendix I

Meeting Abstracts

Clinicians' and patients' perceptions of the use of artificial intelligence decision aids to inform shared decision making: a systematic review

Nehal Hassan, Robert D Slight, Kweku Bimpong, Daniel Weiland, Akke Vellinga, Graham Morgan, Sarah P Slight

Abstract

Background Artificial intelligence (AI) decision aids can offer individualised and tailored information to patients and clinicians to inform shared decision making (SDM). We aimed to explore the perspectives of clinicians and patients on the use of AI decision aids to inform SDM in different clinical settings.

Methods We did a systematic review by searching four large databases (MEDLINE, CINAHL, SCOPUS, and Embase), using appropriate MeSH terms and key words ("perceptions", "clinicians", "patients", "artificial intelligence", "decision-aids", "shared decision-making"), for articles published in English from database inception until Oct 5, 2020. All quantitative, qualitative, mixed-method studies that discussed the perspectives of clinicians or patients from any medical speciality, clinical setting, and timeperiod were included and reviewed by two independent reviewers (inter-rater reliability was 100%). Studies discussing decision aids without an AI component were excluded. We assessed quality and bias using the Critical Appraisal Skills Programme (CASP) checklist. We did a narrative synthesis based on the Research Methods Programme framework developed by the UK Economic and Social Research Council. This review was registered with PROSPERO (number CRD42020220320) and adhered to the PRISMA guidelines.

Findings 509 articles were identified, with 17 articles eligible for inclusion. All included studies received a score of 7 or higher on the CASP checklist. AI decision aids were used to support SDM at various stages of the patient journey, including prevention, screening, prognosis, diagnosis, and treatment. It was perceived by patients and clinicians that these decision aids helped to promote patient engagement and communication with their clinicians, increase patient confidence and compliance, inform the consenting process, present risk estimates in an interactive and individualised manner, and improve patient satisfaction around their overall clinical care. The main barriers to using such aids were around patients' variability in technology literacy, and incomplete or missing information that could potentially affect the findings.

Interpretation AI decision aids were perceived to be useful tools to inform the SDM process for both patients and clinicians at different stages of the patient journey. However, some AI decision aids were used in a simulated research environment, and so clinicians' and patients' perspectives on these might differ if used in the clinical setting.

Funding Newcastle University Overseas Research Scholarship.

Copyright © 2021 Published by Elsevier Ltd. All rights reserved.

Contributors

RDS, SPS, and NH conceived this review. NH and KB conducted the literature search and extracted the data. NH, RDS, and SPS led the writing of this manuscript, with all other coauthors (DW, GM, and AV) commenting on subsequent drafts. All authors gave their approval for the final version to be published.

Declaration of interests

We declare no competing interests.

Acknowledgments

NH was awarded a Newcastle University Overseas Research Scholarship, which covers tuition fees; this project is a part of a doctorate degree.

Published Online
November 26, 2021
School of Pharmacy
(N Hassan MSc,
K Bimpong MPharm,
Sarah P Slight PhD), and School
of Computing (G Morgan PhD),
Newcastle University,
Newcastle upon Tyne, UK;
Newcastle Upon Tyne Hospitals
NHS Foundation Trust,
Freeman Hospital, Newcastle
upon Tyne, UK (R D Slight PhD),
D Weiland FRCPATH; School of
Medicine, National University
of Ireland Galway, Galway,
Ireland (A Vellinga PhD)

Correspondence to:
Dr Sarah P Slight, School of
Pharmacy, Newcastle University,
Newcastle upon Tyne
NE1 7RU, UK
sarah.slight@newcastle.ac.uk

Appendix J

is a growing need for a digitally literate workforce. However, there are no national standards on DH education for UK medical students. Consequently, this study sought to assess the current provisions, perceptions and challenges regarding DH education in the undergraduate medical curriculum.

Methods An anonymous cross-sectional online survey was developed following a literature search and by collecting iterative feedback from both researchers and external collaborators. The survey consisted of questions in 6 areas: (a) understanding of DH; (b) existing provision of DH education; (c) interest in DH education; (d) preferred means of delivering and assessing DH education; (e) impact of the COVID-19 pandemic on DH; and (f) demographic information.

The survey was administered via Qualtrics from March to October 2021, and disseminated to UK medical students via university mailing lists, social media and student representatives. Quantitative and qualitative data were collected pertaining to demographics, attitudes, preferences, and current provisions regarding DH education. Qualitative responses underwent thematic analysis. For quantitative analysis, R (version 3.5.0) and R Studio (version 1.1a) were used.

Results 514 complete responses were received from 39 UK medical schools in 2021. 57.2% of respondents were female, with a mean age of 22.9 ± 3.2 . 65.8% of students considered DH 'extremely important' to future clinical practice, particularly the domains of electronic patient records, telehealth and smartphone applications. However, only 18.1% felt aware of the DH competencies required in clinical medicine. 70.2% of students reported receiving some DH education, with the highest proportion being in the form of lectures or seminars (30.5%, $n=157$), e-learning modules (28.6%, $n=147$) and ad hoc teaching during clinical placements (22.8%, $n=117$). However, only 25.7% felt satisfied with these provisions. Themes for student satisfaction related to a practical teaching approach, delivery of content appropriate for their training stage and coverage of topics in student interest. Conversely, student dissatisfaction originated from inadequate teaching, and subsequent fears of falling behind. 56.1% preferred DH education to be mandatory rather than elective, ideally

7 HEALTHCARE STAFF PERCEPTIONS ON USING ARTIFICIAL INTELLIGENCE PREDICTIVE TOOLS: A QUALITATIVE STUDY

Nehal Hassan, Robert Slight, Sarah Slight. School of Pharmacy, Newcastle University

10.1136/bmjhci-2022-FCIASC.7

Objective Artificial intelligence (AI) predictive tools can help inform the clinical decision-making process by, for example, detecting early signs of patient deterioration or predicting the likelihood of a patient developing a particular disease or complications postsurgery. However, it is unclear how acceptable or useful clinicians find these tools in practice. This project aims to explore healthcare staff' perceptions on the benefits and challenges of using AI tools to inform clinical decision-making in practice.

Methods Healthcare staff (physicians, pharmacists and nurses) working in different departments at one large teaching hospital in the North East were invited to participate in semi-structured interviews. Interviews were conducted between August and November 2021 by zoom videoconferencing, with questions focused on what AI predictive tools they currently use, how they guide daily tasks around diagnosis, management, prevention, prognosis and screening, and what challenges they face with their use. All transcribed files were checked for accuracy. Thematic saturation guided the volume of qualitative data collection. Qualitative data analysis and development of themes was performed for each interview using Nvivo 12 software. Ethical approval was obtained (20/EM/0183, IRAS 280077).

Results Ten healthcare staff were interviewed (physicians ($n=7$), pharmacists ($n=1$), surgeons ($n=2$)) from different medical specialities (e.g., Oncology, Endocrinology, Cardiology, Head and Neck, and transplant surgery). Five themes emerged, including the meaning of the term AI, the usefulness of AI predictive tools in informing clinical decision-making, features that healthcare staff found helpful, and challenges

Appendix K



**Clinicians' perceptions on the use of AI tools to inform clinical decision making.
Participant Consent Form**

Participant ID:

Date of Interview:

Thank you for agreeing to participate in this research project. I am Nehal Hassan, a PhD student based at Newcastle University and I would like to understand clinicians' perceptions on the use of predictive tools, which have or could have an AI component, in clinical practice. Is there anything you would like to ask me before we start? Are you still happy for this interview to be audio and video recorded? (Confirm receipt of signed consent form)

Questions:

1. Can you begin by telling me a little bit about yourself? What speciality do you work in? How many years of experience have you in this speciality?
2. What is your understanding of the term Artificial Intelligence or AI?

Prompt: Examples - Can predict an individual's risk of developing a complication post-surgery? Can detect abnormalities in images and scans (radiology)?

3. How might AI be able to help you in your daily tasks?

Prompt: Are there any tasks which involve you predicting the risk of something? For example, the risk of complications from a particular procedure or from prescribing a particular medicine?

4. Do you usually talk to your patients about risk? For example, when consenting for a particular operation?
5. How do you think AI might be able to help you with your clinical decision making?

Prompt: Personalising the risk for patients?

6. Can you think of any barriers to use AI predictive models or decision aids in your practice?



7. Thinking back over all the things we discussed today, is there anything else you would like to add before we finish? Might I be able to contact you again, if I had any follow-up questions?

Thank you for giving your time for this interview. It is really appreciated.

It would be helpful for us to know, just for our records, whether you have an academic affiliation? And approximately how many patients you see daily (or weekly)?

Reflection Log:

Participant ID:

Date of Interview:

1. How I think the interview went?
2. Reflection on my personal rapport with the participant
3. Important points that the participant raised
4. Aspects that might be useful to explore in the analysis
5. Additional questions to ask in subsequent interviews
6. Any other notes

Appendix L



Clinicians' perceptions on the use of AI tools to inform clinical decision making.

Participant Information Sheet

Names of Investigators: Ms Nehal Hassan, Dr Robert Slight, Dr Sarah Slight

Ethical Approval Ref: 20/EM/0183

IRAS Project ID: 280077

Invitation paragraph

You have been invited to take part in a research study. Before you decide whether or not to take part, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully.

Background

The use of big data and artificial intelligence (AI) can help predict the likelihood of diseases before they occur. Predicting the likelihood of sepsis, for example, can reduce healthcare costs and save patients' lives through preventing the initiation of the multi-organ dysfunction process, reducing ICU admissions and improving patients' prognosis. We would like to understand clinicians' perceptions on the use of predictive tools, which have or could have an AI component, in clinical practice.

Who has reviewed this study?

This study has been reviewed and approved by the NHS Health Research Authority research and ethics committee.

Who funds/sponsors the research?

This research is funded and sponsored by Newcastle University (Newcastle University overseas research scholarship - NUORS) and is being conducted as part of a PhD project.

What will taking part involve?

Participation will involve taking part in an interview (which could either be face-to-face, over the phone or via call-based software) to explore your understanding of AI and how AI might be able to help you in your daily tasks, the importance of communicating risk estimates to patients, and the barriers and facilitators of using AI predictive models. The interview is expected to last approximately 30-40 minutes. The interview will be recorded and typed up word for word (transcribed). Once transcribed, the recording will be erased.

Why have I been chosen to take part?

Any clinician currently working at Newcastle Upon Tyne Hospitals NHS Foundation Trust at any professional level is welcome to participate. We are keen to invite clinicians from a wide variety of clinical specialities and working at different professional levels.



Will my taking part in this study be kept confidential?

All information provided will be kept confidential. Any information reported will not enable you to be recognised. You will not automatically be expected to take part in any future research. Consent forms will be securely stored within Newcastle university in a filing cabinet within a locked office, or on a password-protected University computer, and held securely in accordance with regulations. Any personal data will be destroyed as soon as is practical and reasonable.

What will happen to the results of this research study?

The researchers plan to submit the findings of this study to medical journals for publication. If you wish to be given a summary of the results, we would be happy to provide this to you.

What if something goes wrong?

In case you have a complaint, you can approach the Chief Investigator: Dr Sarah Slight (contact details below). Independent advice is also available from Newcastle Upon Tyne Hospitals NHS Foundation Trust Research & Development Department via phone: +44 (0191) 282 5959 or email: trust.r&d@nuth.nhs.uk.

How will we use information about you?

We will need to use information from you for this research project. This information will include only your opinions. People will use this information to do the research or to check your records to make sure that the research is being done properly. We will keep all information about you safe and secure. Once we have finished the study, we will keep some of the data so we can check the results. We will write our reports in a way that no-one can work out that you took part in the study.

What are your choices about how your information is used?

You can stop being a part of the study at any time, without giving a reason, but we will keep information/audio-recordings about you that we already have. We need to manage your records in specific ways for the research to be reliable. This means that we won't be able to let you see or change the data we hold about you.

Where can you find out more about how your information is used?

You can find out more about how we use your information:

- at www.hra.nhs.uk/information-about-patients/
- by asking one of the research team via email (contact information below).

Who can I contact for more information about this study?

PhD student: Ms. Nehal Hassan, School of Pharmacy, King George VI Building, Newcastle University, Newcastle Upon Tyne, NE1 7RU. Email: n.a.m.hassan2@newcastle.ac.uk.

Supervisor/Chief Investigator: Dr Sarah Slight, School of Pharmacy, King George VI Building, Newcastle University, Newcastle Upon Tyne, NE1 7RU. Email: Sarah.Slight@ncl.ac.uk.

Thank you very much for considering taking part in this research study.

Clinicians' perceptions on the use of AI tools to inform clinical decision making.
IRAS Project ID 280077 / Participant Information Sheet -Version 1.1 /18 July 2021

Appendix M



**Clinicians' perceptions on the use of AI tools to inform clinical decision making.
Participant Consent Form**

Ethical Approval Ref: 20/EM/0183
IRAS Project ID: 280077

Names of Investigators: Ms Nehal Hassan, Dr Robert Slight, Dr Sarah Slight

Name of Participant: _____

Please initial box

1. I confirm that I have read and understand the participant information sheet version number, dated for the above study and have had the opportunity to ask questions. ☐
2. I understand that my participation is voluntary and that I am free to withdraw at any time, without giving a reason. ☐
3. I understand that should I withdraw, the information collected so far cannot be erased, and that this information may still be used in the project analysis. ☐
4. I understand that notes may be taken during the recorded interview, and I will be asked a few questions. The interview (which could either be face-to-face, over the phone or via call-based software) will be transcribed, and any anonymous direct quotes may be used in the study reports and publications. ☐
5. I understand that all information supplied will be kept confidential. Any information reported or published will not enable me to be recognised. ☐
6. I understand that relevant sections of the data collected during the study may be looked at by individuals from: the study team, the sponsor (Newcastle University), and from regulatory authorities, where it is relevant to taking part in this research. ☐
7. I agree to take part in the above study. ☐

Name of Participant

Date

Signature

Researcher taking consent

Date

Signature

For more information about the study, or to ask further questions, feel free to contact the lead researcher: Nehal Hassan, n.a.m.hassan2@newcastle.ac.uk
2 copies: 1 for participant and 1 for the project notes

Appendix N



**Clinicians' perceptions on the use of AI tools to inform clinical decision making.
Participant Consent Form**

Reflection Log

Participant ID:

Date of Interview:

1. How I think the interview went?
2. Reflection on my personal rapport with the participant
3. Important points that the participant raised.
4. Aspects that might be useful to explore in the analysis.
5. Additional questions to ask in subsequent interviews.
6. Any other notes.

Chapter 11. References.

1. Antimicrobial resistance fact sheets. <https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance> [Accessed on 14th January 2022]
2. Knight GM, Costelloe C, Deeny SR, Moore LS, Hopkins S, Johnson AP, Robotham JV, Holmes AH. Quantifying where human acquisition of antibiotic resistance occurs: a mathematical modelling study. *BMC medicine*. 2018 Dec;16(1):1-1.
3. Murray CJ, Ikuta KS, Sharara F, Swetschinski L, Aguilar GR, Gray A, Han C, Bisignano C, Rao P, Wool E, Johnson SC. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*. 2022 Feb 12;399(10325):629-55.
4. Kumar M, Silori R, Mazumder P, Shrivastava V, Loge F, Barceló D, Mahlkecht J. Wars and Pandemics: AMR Accelerators of the 21st Century?. *Environmental Science & Technology Letters*. 2023 Mar 26;10(4):289-91.
5. The review of antimicrobial resistance: https://amr-review.org/sites/default/files/AMR%20Review%20Paper%20%20Tackling%20a%20crisis%20for%20the%20health%20and%20wealth%20of%20nations_1.pdf [Accessed on 19th May 2023]
6. Medina MJ, Legido-Quigley H, Hsu LY. Antimicrobial resistance in one health. *Global Health Security: Recognizing Vulnerabilities, Creating Opportunities*. 2020:209-29.
7. House of Commons, Health and Social Care Committee eleventh report <https://publications.parliament.uk/pa/cm201719/cmselect/cmhealth/962/962.pdf> [Accessed on 19th May 2023]
8. Llewelyn MJ, Budgell EP, Laskawiec-Szkonter M, Cross EL, Alexander R, Bond S, Coles P, Conlon-Bingham G, Dymond S, Evans M, Fok R. Antibiotic review kit for hospitals (ARK-Hospital): a stepped-wedge cluster-randomised controlled trial. *The Lancet Infectious Diseases*. 2023 Feb 1;23(2):207-21.
9. Michael CA, Dominey-Howes D, Labbate M. The antimicrobial resistance crisis: causes, consequences, and management. *Frontiers in public health*. 2014 Sep 16;2:145.
10. ECDC: fact sheets for experts. <https://www.ecdc.europa.eu/en/antimicrobial-resistance/facts/factsheets/experts> [Accessed on 19th May 2023]
11. Why is it so hard to develop new antibiotics? <https://wellcome.org/news/why-is-it-so-hard-develop-new-antibiotics> [Accessed on 9th September 2023]
12. Ahiawodzi PD, Kelly K, Massengill A, Thompson DK. Risk factors for sepsis morbidity in a rural hospital population: A case-control study. *American journal of infection control*. 2018 Sep 1;46(9):1041-6.
13. Annual Report of the Chief Medical Officer, Volume Two, 2011, Infections and the Rise of Antimicrobial Resistance; Department of Health and Social Care: London, UK, 2012.

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/138331/CMO_Annual_Report_Volume_2_2011.pdf [Accessed on 19th May 2023]

14. Dadgostar P. Antimicrobial resistance: implications and costs. *Infection and drug resistance*. 2019 Dec 20;3903-10.

15. Edwardson S, Cairns C. Nosocomial infections in the ICU. *Anaesthesia & Intensive Care Medicine*. 2019 Jan 1;20(1):14-8.

16. National risk register

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/952959/6.6920_CO_CCS_s_National_Risk_Register_2020_11-1-21-FINAL.pdf [Accessed on 19th May 2023]

17. Guy R, Higgins H, Rudman J, Fountain H, Bennet KF, Hopkins KL, Demirjian A, Gerver SM, Mirfenderesky M, Henderson KL. Antimicrobial Resistance in England 2017 to 2021 (ESPAUR Report 2021–22). *In Medical Sciences Forum* 2023 Mar 24 (Vol. 15, No. 1, p. 3). MDPI.

18. Horner C, Costelloe CE. Exploring temporal trends and risk factors for resistance in *Escherichia coli*-causing bacteraemia in England between 2013 and 2018: an ecological study. *Europe*;9:11.

19. Naylor NR, Pouwels KB, Hope R, Green N, Henderson KL, Knight GM, Atun R, Robotham JV, Deeny SR. The health and cost burden of antibiotic resistant and susceptible *Escherichia coli* bacteraemia in the English hospital setting: A national retrospective cohort study. *PloS one*. 2019 Sep 10;14(9):e0221944.

20. Surveillance of surgical site infections in NHS hospitals in England April 2021 to March 2022

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1123846/SSI-annual-report-2021-to-2022.pdf [Accessed on 19th May 2023]

21. Bhattacharjee P, Edelson DP, Churpek MM. Identifying patients with sepsis on the hospital wards. *Chest*. 2017 Apr 1;151(4):898-907

22. Alberto L, Marshall AP, Walker R, Aitken LM. Screening for sepsis in general hospitalized patients: a systematic review. *Journal of Hospital Infection*. 2017 Aug 1;96(4):305-15.

23. UK Sepsis Trust <https://sepsistrust.org/about/about-sepsis/references-and-sources/> [Accessed on 17th September 2023]

24. The sepsis manual. (<https://sepsistrust.org/wp-content/uploads/2020/01/5th-Edition-manual-080120.pdf>) [Accessed on 8th September 2023]

25. Improving outcomes for patients with sepsis. A cross-system action plan

<https://www.england.nhs.uk/wp-content/uploads/2015/08/Sepsis-Action-Plan-23.12.15-v1.pdf> [Accessed on 8th September 2023]

26. Fathi M, Markazi-Moghaddam N, Ramezankhani A. A systematic review on risk factors associated with sepsis in patients admitted to intensive care units. *Australian Critical Care*. 2019 Mar 1;32(2):155-64
27. Yang WS, Chang YC, Chang CH, Wu LC, Wang JL, Lin HH. The association between body mass index and the risk of hospitalization and mortality due to infection: a prospective cohort study. *InOpen forum infectious diseases* 2021 Jan (Vol. 8, No. 1, p. ofaa545). US: Oxford University Press.
28. Wang HE, Baddley J, Griffin RL, Judd S, Howard G, Donnelly JP, Safford MM. Physical inactivity and long-term rates of community-acquired sepsis. *Preventive medicine*. 2014 Aug 1;65:58-64.
29. Tackling antimicrobial resistance 2019–2024, The UK’s five-year national action plan https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1070263/UK_AMR_5_year_national_action_plan.pdf [Accessed on 11th September 2023]
30. Bini SA. Artificial intelligence, machine learning, deep learning, and cognitive computing: what do these terms mean and how will they impact health care?. *The Journal of arthroplasty*. 2018 Aug 1;33(8):2358-61.
31. Shailaja K, Seetharamulu B, Jabbar MA. Machine learning in healthcare: A review. In 2018 Second international conference on electronics, communication and aerospace technology (ICECA) 2018 Mar 29 (pp. 910-914). IEEE.
32. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, Allison T, Arnaout O, Abbosh C, Dunn IF, Mak RH. Artificial intelligence in cancer imaging: clinical challenges and applications. *CA: a cancer journal for clinicians*. 2019 Mar;69(2):127-57.
33. Mello-Thoms C, Mello CA. Clinical applications of artificial intelligence in radiology. *The British Journal of Radiology*. 2023 Apr;96:20221031.
34. Borges AF, Laurindo FJ, Spínola MM, Gonçalves RF, Mattos CA. The strategic use of artificial intelligence in the digital era: Systematic literature review and future research directions. *International Journal of Information Management*. 2021 Apr 1;57:102225.
35. Kiaii B. Robotic Cardiac Valvular Surgery. In *Evidence-Based Practice in Perioperative Cardiac Anesthesia and Surgery* 2021 (pp. 485-489). Springer, Cham.
36. Shared decision making NHS summary guide: <https://www.england.nhs.uk/wp-content/uploads/2019/01/shared-decision-making-summary-guide-v1.pdf> [Accessed on 11th September 2023]
37. Gutacker N, Street A. Use of large-scale HRQoL datasets to generate individualised predictions and inform patients about the likely benefit of surgery. *Quality of life research*. 2017 Sep;26:2497-505.

38. Hassan AM, Rajesh A, Asaad M, Nelson JA, Coert JH, Mehrara BJ, Butler CE. Artificial intelligence and machine learning in prediction of surgical complications: Current state, applications, and implications. *The American Surgeon*. 2023 Jan;89(1):25-30.
39. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. *Cancer Inf*. 2007;2:59-77.
40. Miller RA, Pople HE, Myers JD. Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. *N Engl J Med*. 1982;307(8):468-476.
41. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: Promises and perils. *Ann Surg*. 2018;268(1):70-76.
42. Hassan AM, Biaggi AP, Asaad M, et al. Development and assessment of machine learning models for individualized risk assessment of mastectomy skin flap necrosis. *Ann Surg*. 2022; *Published ahead of print*. [10.1097SLA.0000000000005386](https://doi.org/10.1097SLA.0000000000005386)
43. Hassan A, Lu S, Asaad M, Offodile AC, Sidey-Gibbons C. Novel machine learning approach for prediction of hernia recurrence, surgical complications, and 30-day readmission following abdominal wall reconstruction. *J Am Coll Surg*. 2022;234(5):918-927.
44. Soguero-Ruiz C, Fei WM, Jenssen R, et al. Data-driven temporal prediction of surgical site infection. *AMIA Annu Symp Proc*. 2015;2015:1164-1173.
45. Lee CK, Hofer I, Gabel E, Baldi P, Cannesson M. Development and validation of a deep neural network model for prediction of postoperative in-hospital mortality. *Anesthesiology*. 2018;129(4):649-662.
46. Fei Y, Hu J, Li WQ, Wang W, Zong GQ. Artificial neural networks predict the incidence of portosplenomesenteric venous thrombosis in patients with acute pancreatitis. *J Thromb Haemostasis*. 2017;15(3):439-445.
47. Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, Hall MK. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. *Academic emergency medicine*. 2016 Mar;23(3):269-78.
48. Silver DL. Machine lifelong learning: Challenges and benefits for artificial general intelligence. Paper presented at: International conference on artificial general intelligence; 2011
49. Gepperth A, Hammer B. Incremental learning algorithms and applications. Paper presented at: European symposium on artificial neural networks (ESANN); 2016.
50. Ij H. Statistics versus machine learning. *Nat Methods*. 2018 Apr;15(4):233.
51. Lette J, Colletti BW, Cerino M, Mcnamara D, Eybalin MC, Levasseur A, Natnel S. Artificial intelligence versus logistic regression statistical modelling to predict cardiac complications after noncardiac surgery. *Clinical cardiology*. 1994 Nov;17(11):609-14.

52. Salati M, Migliorelli L, Moccia S, Andolfi M, Roncon A, Guiducci GM, Xiume F, Tiberi M, Frontoni E, Refai M. A machine learning approach for postoperative outcome prediction: Surgical data science application in a thoracic surgery setting. *World Journal of Surgery*. 2021 May;45:1585-94.
53. Hopkins BS, Mazmudar A, Driscoll C, Svet M, Goergen J, Kelsten M, Shlobin NA, Kesavabhotla K, Smith ZA, Dahdaleh NS. Using artificial intelligence (AI) to predict postoperative surgical site infection: a retrospective cohort of 4046 posterior spinal fusions. *Clinical neurology and neurosurgery*. 2020 May 1;192:105718.
54. Wang HE, Donnelly JP, Griffin R, Levitan EB, Shapiro NI, Howard G, Safford MM. Derivation of a Novel Risk Prediction Scores for Community-Acquired Sepsis and Severe Sepsis. *Critical care medicine*. 2016 Jul;44(7):1285.
55. Giannini HM, Ginestra JC, Chivers C, Draugelis M, Hanish A, Schweickert WD, Fuchs BD, Meadows L, Lynch M, Donnelly PJ, Pavan K. A machine learning algorithm to predict severe sepsis and septic shock: Development, implementation and impact on clinical practice. *Critical care medicine*. 2019 Nov;47(11):1485.
56. Lu HX, Du J, Wen DL, Sun JH, Chen MJ, Zhang AQ, Jiang JX. Development and validation of a novel predictive score for sepsis risk among trauma patients. *World Journal of Emergency Surgery*. 2019 Dec;14(1):1-8.
57. Lukaszewski RA, Yates AM, Jackson MC, Swingler K, Scherer JM, Simpson AJ, Sadler P, McQuillan P, Titball RW, Brooks TJ, Pearce MJ. Presymptomatic prediction of sepsis in intensive care unit patients. *Clinical and Vaccine Immunology*. 2008 Jul;15(7):1089-94.
58. Nachimuthu SK, Haug PJ. Early detection of sepsis in the emergency department using Dynamic Bayesian Networks. In *AMIA Annual Symposium Proceedings 2012 (Vol. 2012, p. 653)*. *American Medical Informatics Association*
59. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, Shimabukuro D, Chettipally U, Feldman MD, Barton C, Wales DJ. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. *JMIR medical informatics*. 2016 Sep 30;4(3):e5909.
60. Sood A, Abdollah F, Sammon JD, Arora N, Weeks M, Peabody JO, Menon M, Trinh QD. Postoperative sepsis prediction in patients undergoing major cancer surgery. *Journal of Surgical Research*. 2017 Mar 1;209:60-9.
61. Yao RQ, Jin X, Wang GW, Yu Y, Wu GS, Zhu YB, Li L, Li YX, Zhao PY, Zhu SY, Xia ZF. A machine learning-based prediction of hospital mortality in patients with postoperative sepsis. *Frontiers in medicine*. 2020 Aug 11;7:445.
62. Bunn C, Kulshrestha S, Boyda J, Balasubramanian N, Birch S, Karabayir I, Baker M, Luchette F, Modave F, Akbilgic O. Application of machine learning to the prediction of postoperative sepsis after appendectomy. *Surgery*. 2021 Mar 1;169(3):671-7.

63. Ninh A, Wood K, Bui AH, Leitman IM. Risk factors and outcomes for sepsis after appendectomy in adults. *Surgical Infections*. 2019 Dec 1;20(8):601-6.
64. Emergency Surgery <https://uemssurg.org/surgicalspecialties/emergency-surgery/> [Accessed on 8th September 2023]
65. De Simone B, Sartelli M, Coccolini F, Ball CG, Brambillasca P, Chiarugi M, Campanile FC, Nita G, Corbella D, Leppaniemi A, Boschini E. Intraoperative surgical site infection control and prevention: a position paper and future addendum to WSES intra-abdominal infections guidelines. *World journal of emergency surgery*. 2020 Dec;15(1):1-23.
66. Fontana G, Ghafur S, Torne L, Goodman J, Darzi A. Ensuring that the NHS realises fair financial value from its data. *The Lancet Digital Health*. 2020 Jan 1;2(1):e10-2. [https://doi.org/10.1016/S2589-7500\(19\)30225-0](https://doi.org/10.1016/S2589-7500(19)30225-0)
67. Elwyn G, Frosch DL, Kobrin S. Implementing shared decision making: consider all the consequences. *Implementation Science*. 2015 Dec 1;11(1):114.
68. Hassan N, Slight RD, Bimpong K, Weiland D, Vellinga A, Morgan G, Slight SP. Clinicians' and patients' perceptions of the use of artificial intelligence decision aids to inform shared decision making: a systematic review. *The Lancet*. 2021 1st November;398:S80. [https://doi.org/10.1016/S0140-6736\(21\)02623-4](https://doi.org/10.1016/S0140-6736(21)02623-4)
69. Flynn D, Nesbitt DJ, Ford GA, McMeekin P, Rodgers H, Price C, Kray C, Thomson RG. Development of a computerised decision aid for thrombolysis in acute stroke care. *BMC medical informatics and decision making*. 2015 1st December;15(1):6. DOI: <https://doi.org/10.1186/s12911-014-0127-1>
70. Silvestrin TM, Steenrod AW, Coyne KS, Gross DE, Esinduy CB, Kodsí AB, Slifka GJ, Abraham L, Araiza AL, Bushmakín AG, Luo X. An approach to improve the care of mid-life women through the implementation of a Women's Health Assessment Tool/Clinical Decision Support toolkit. *Women's Health*. 2016 Sep;12(5):456-64. <https://doi.org/10.1177/1745505716664742>
71. Motorny S, Sarnikar S, Noteboom C. Design of an Intelligent Patient Decision aid Based on Individual Decision making Styles and Information Need Preferences. *Information Systems Frontiers*. 2021 Mar 23;1-6. <https://doi.org/10.1007/s10796-021-10125-9>
72. Chua IS, Ritchie CS, Bates DW. Enhancing serious illness communication using artificial intelligence. *npj Digital Medicine*. 2022 Jan 27;5(1):1-4. <https://doi.org/10.1038/s41746-022-00556-2>
73. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Journal of British Surgery*. 2015 Feb;102(3):148-58. <https://doi.org/10.1002/bjs.9736>
74. Najafabadi AH, Ramspek CL, Dekker FW, Heus P, Hooft L, Moons KG, Peul WC, Collins GS, Steyerberg EW, van Diepen M. TRIPOD statement: a preliminary pre-post analysis of reporting and methods of prediction models. *BMJ open*. 2020 Sep 1;10(9):e041537. doi: 10.1136/bmjopen-2020-041537

75. Andaur Navarro CL, Damen JA, Takada T, Nijman SW, Dhiman P, Ma J, Collins GS, Bajpai R, Riley RD, Moons KG, Hooft L. Completeness of reporting of clinical prediction models developed using supervised machine learning: A systematic review. *arXiv. org.* 2021 1st July. <https://doi.org/10.1186/s12874-021-01469-6>
76. Collins GS, Moons KG. Reporting of artificial intelligence prediction models. *The Lancet.* 2019 20th April;393(10181):1577-9. [https://doi.org/10.1016/S0140-6736\(19\)30037-6](https://doi.org/10.1016/S0140-6736(19)30037-6)
77. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *bmj.* 2020 9th September;370. doi: <https://doi.org/10.1136/bmj.m3164>
78. Hercheui M, Mech G. Factors Affecting The Adoption Of Artificial Intelligence In Healthcare. *Global Journal of Business Research.* 2021;15(1):77-88.
79. Kappen TH, van Klei WA, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, Moons KG. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. *Diagnostic and prognostic research.* 2018 Dec;2(1):1-1.
80. De Hond AA, Leeuwenberg AM, Hooft L, Kant IM, Nijman SW, van Os HJ, Aardoom JJ, Debray T, Schuit E, van Smeden M, Reitsma JB. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *npj Digital Medicine.* 2022 Jan 10;5(1):1-3. <https://doi.org/10.1038/s41746-021-00549-7>
81. Romero-Brufau S, Wyatt KD, Boyum P, Mickelson M, Moore M, Cognetta-Rieke C. A lesson in implementation: a pre-post study of providers' experience with artificial intelligence-based clinical decision support. *International journal of medical informatics.* 2020 1st May;137:104072. <https://doi.org/10.1016/j.ijmedinf.2019.104072>
82. Nanji KC, Garabedian PM, Shaikh SD, Langlieb ME, Boxwala A, Gordon WJ, Bates DW. Development of a Perioperative Medication-Related Clinical Decision Support Tool to Prevent Medication Errors: An Analysis of User Feedback. *Applied Clinical Informatics.* 2021 Oct;12(05):984-95. DOI: 10.1055/s-0041-1736339
83. Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific reports.* 2020 Jul 20;10(1):1-2. <https://doi.org/10.1038/s41598-020-68771-z>
84. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Critical care medicine.* 2018 Apr;46(4):547.
85. Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, Rubenfeld G, Kahn JM, Shankar-Hari M, Singer M, Deutschman CS. Assessment of clinical criteria for sepsis: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *Jama.* 2016 Feb 23;315(8):762-74.
86. Marik PE, Taeb AM. SIRS, qSOFA and new sepsis definition. *Journal of thoracic disease.* 2017 Apr;9(4):943. doi: 10.21037/jtd.2017.03.125

87. Hassan N, Slight R, Weiland D, Vellinga A, Morgan G, Aboushareb F, Slight SP. Preventing sepsis; how can artificial intelligence inform the clinical decision making process? A systematic review. *International Journal of Medical Informatics*. 2021 Apr 10;104:457.
88. Khojandi A, Tansakul V, Li X, Koszalinski RS, Paiva W. Prediction of sepsis and in-hospital mortality using electronic health records. *Methods of information in medicine*. 2018 Sep;57(04):185-93.
89. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of internal medicine*. 2015 Jan 6;162(1):W1-73.
90. Gupta A, Liu T, Shepherd S. Clinical decision support system to assess the risk of sepsis using tree augmented Bayesian networks and electronic medical record data. *Health informatics journal*. 2020 Jun;26(2):841-61. <https://doi.org/10.1177/1460458219852872>
91. Sidey-Gibbons JA, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC medical research methodology*. 2019 Dec;19(1):1-8. <https://doi.org/10.1186/s12874-019-0681-4>
92. D'Ascenzo F, De Filippo O, Gallone G, Mittone G, Deriu MA, Iannaccone M, Ariza-Solé A, Liebetrau C, Manzano-Fernández S, Quadri G, Kinnaird T. Machine learning-based prediction of adverse events following an acute coronary syndrome (PRAISE): a modelling study of pooled datasets. *The Lancet*. 2021 16th January;397(10270):199-207. [https://doi.org/10.1016/S0140-6736\(20\)32519-8](https://doi.org/10.1016/S0140-6736(20)32519-8)
93. Beckers R, Kwade Z, Zanca F. The EU medical device regulation: Implications for artificial intelligence-based medical device software in medical physics. *Physica Medica*. 2021 1st March;83:1-8.
94. Reilly G, Varma S. Health Data Research Innovation Gateway. *ITNOW*. 2021;63(2):60-3. <https://doi.org/10.1093/itnow/bwab061>
95. Health Data Research Innovation Gateway. <https://www.hdruk.ac.uk/access-to-health-data/health-data-research-innovation-gateway/> [Accessed on 18th March 2022]
96. Gallier S, Price G, Pandya H, McCarmack G, James C, Ruane B, Forty L, Crosby BL, Atkin C, Evans R, Dunn KW. Infrastructure and operating processes of PIONEER, the HDR-UK Data Hub in Acute Care and the workings of the Data Trust Committee: a protocol paper. *BMJ health & care informatics*. 2021;28(1). doi: 10.1136/bmjhci-2020-100294
97. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*. 2021 Jul 13;54(6):1-35. <https://doi.org/10.1145/3457607>
98. Clifford, G.D.; Scott, D.J.; Villarroel, M. User Guide and Documentation for the MIMIC II Database; MIMIC-II Database Version; Physionet.org: Cambridge, MA, USA, 2009; Volume 2

99. Liang Y, Abbott D, Howard N, Lim K, Ward R, Elgendi M. How effective is pulse arrival time for evaluating blood pressure? Challenges and recommendations from a study using the MIMIC database. *Journal of clinical medicine*. 2019 Mar;8(3):337. <https://doi.org/10.3390/jcm8030337>
100. Symons JD, Ashrafian H, Dunscombe R, Darzi A. From EHR to PHR: let's get the record straight. *BMJ open*. 2019 Sep 1;9(9):e029582. <http://dx.doi.org/10.1136/bmjopen-2019-029582>
101. Budhdeo S, Weerasuriya CK, Zhang J, Thomas JP, Mariam NB, Sharma N, Cosco TD, Harrison OT, Banerjee A. Interoperability in NHS Acute Trusts within England: a Situation and Capability Analysis using Freedom of Information requests. *medRxiv*. 2021 1st January. <https://doi.org/10.1101/2021.10.23.21265418>
102. Johnson AE, Pollard TJ, Shen L, Li-Wei HL, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. *Scientific data*. 2016 May 24;3(1):1-9. <https://doi.org/10.1038/sdata.2016.35>
103. Sarker IH. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*. 2021 May;2(3):1-21. <https://doi.org/10.1007/s42979-021-00592->
104. Huang JC, Ko KM, Shu MH, Hsu BM. Application and comparison of several machine learning algorithms and their integration models in regression problems. *Neural Computing and Applications*. 2020 May;32(10):5461-9. <https://doi.org/10.1007/s00521-019-04644-5>
105. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, Topol EJ, Ioannidis JP, Collins GS, Maruthappu M. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *Bmj*. 2020 25th March;368. doi: <https://doi.org/10.1136/bmj.m689>
106. Oshiro TM, Perez PS, Baranauskas JA. How many trees in a random forest?. In *International workshop on machine learning and data mining in pattern recognition 2012 Jul 13 (pp. 154-168)*. Springer, Berlin, Heidelberg.
107. Sarker IH, Colman A, Han J, Khan AI, Abushark YB, Salah K. Behavdt: a behavioral decision tree learning to build user-centric context-aware predictive model. *Mobile Networks and Applications*. 2020 Jun;25(3):1151-61. <https://doi.org/10.1007/s11036-019-01443-z>
108. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *New England Journal of Medicine*. 2019 4th April;380(14):1347-58. DOI: 10.1056/NEJMra1814259
109. Battineni G, Sagaro GG, Chinatalapudi N, Amenta F. Applications of machine learning predictive models in the chronic disease diagnosis. *Journal of personalised medicine*. 2020 Jun;10(2):21. <https://doi.org/10.3390/jpm10020021>
110. Downloading Python: <https://www.python.org/downloads/> [Accessed on 7th May 2022]
111. Bloch E, Rotem T, Cohen J, Singer P, Aperstein Y. Machine learning models for analysis of vital signs dynamics: a case for sepsis onset prediction. *Journal of healthcare engineering*. 2019 Nov 3;2019.

- 112.What Do We Do About the Biases in AI? <https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai> [Accessed on 14th June 2022].
- 113.Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where?. *Clinical Kidney Journal*. 2021 Jan;14(1):49-58. <https://doi.org/10.1093/ckj/sfaa188>
- 114.Tsvetanova A, Sperrin M, Peek N, Buchan I, Hyland S, Martin GP. Missing data was handled inconsistently in UK prediction models: a review of method used. *Journal of Clinical Epidemiology*. 2021 Dec 1;140:149-58. <https://doi.org/10.1016/j.jclinepi.2021.09.008>
- 115.Michie S, Johnston M. Changing clinical behaviour by making guidelines specific. *BMJ*. 2004;328:343–5. doi: <https://doi.org/10.1136/bmj.328.7435.343>
- 116.Lobach D, Sanders GD, Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux R, Samsa G, Hasselblad V, Williams JW, Wing L, Musty M, Kendrick AS. Enabling health care decision making through clinical decision support and knowledge management. *Evid Rep Technol Assess (Full Rep)*. 2012(203):1–784.
- 117.Kappen TH, van Loon K, Kappen MAM, van Wolfswinkel L, Vergouwe Y, van Klei WA, et al. Barriers and facilitators perceived by physicians when using prediction models in practice. *J Clin Epidemiol*. 2016;70:136–45. <https://doi.org/10.1016/j.jclinepi.2015.09.008>
- 118.Bates DW, Auerbach A, Schulam P, Wright A, Saria S. Reporting and implementing interventions involving machine learning and artificial intelligence. *Annals of internal medicine*. 2020 Jun 2;172(11_Supplement):S137-44. <https://doi.org/10.7326/M19-0872>
- 119.Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019 Oct 25;366(6464):447-53. DOI: 10.1126/science.aax2342
- 120.Kennedy G, Gallego B. Clinical prediction rules: a systematic review of healthcare provider opinions and preferences. *International journal of medical informatics*. 2019 Mar 1;123:1-0. <https://doi.org/10.1016/j.ijmedinf.2018.12.003>
- 121.National Health Service. Accelerating AI in Health and Care: Results from a State of the Nation Survey. London, United Kingdom: Department of Health and Social Service; 2018
- 122.Guidance: Medical device stand-alone software including apps (including IVDMDs) v1.08 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/999908/Software_flow_chart_Ed_1-08b-IVD.pdf [Accessed on 5th January 2022]
- 123.Scheibner J, Sleight J, Ienca M, Vayena E. Benefits, challenges, and contributors to success for national eHealth systems implementation: a scoping review. *Journal of the American Medical Informatics Association*. 2021 Sep;28(9):2039-49. <https://doi.org/10.1093/jamia/ocab096>
- 124.Benjamens S, Dhunoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ digital medicine*. 2020 Sep 11;3(1):1-8. <https://doi.org/10.1038/s41746-020-00324-0>

125. Larson DB, Harvey H, Rubin DL, Irani N, Justin RT, Langlotz CP. Regulatory frameworks for development and evaluation of artificial intelligence–based diagnostic imaging algorithms: summary and recommendations. *Journal of the American College of Radiology*. 2021 1st March;18(3):413-24. <https://doi.org/10.1016/j.jacr.2020.09.060>
126. Evidence standards framework for digital health technologies:
<https://www.nice.org.uk/corporate/ecd7/resources/evidence-standards-framework-for-digital-health-technologies-pdf-1124017457605> [Accessed on 2nd May 2022]
127. Pesapane F, Volonté C, Codari M, Sardanelli F. Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. *Insights into imaging*. 2018 Oct;9(5):745-53. <https://doi.org/10.1007/s13244-018-0645-y>
128. Petersen E, Potdevin Y, Mohammadi E, Zidowitz S, Breyer S, Nowotka D, Henn S, Pechmann L, Leucker M, Rostalski P, Herzog C. Responsible and Regulatory Conform Machine Learning for Medicine: A Survey of Technical Challenges and Solutions. *arXiv preprint arXiv:2107.09546*. 2021 20th July. DOI: 10.1109/ACCESS.2022.3178382
129. Solomonides AE, Koski E, Atabaki SM, Weinberg S, McGreevey III JD, Kannry JL, Petersen C, Lehmann CU. Defining AMIA’s artificial intelligence principles. *Journal of the American Medical Informatics Association*. 2022 Apr;29(4):585-91. <https://doi.org/10.1093/jamia/ocac006>
130. Food and Drug Administration. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). <https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf> [Accessed on 10th June 2022]
131. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, Boyd KA, Craig N, French DP, McIntosh E, Petticrew M. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. *bmj*. 2021 Sep 30;374.
132. L. Alberto, A.P. Marshall, R. Walker, L.M. Aitken, Screening for sepsis in general hospitalized patients: a systematic review, *J. Hosp. Infect.* 96 (August (4)) (2017) 305–315.
133. Sartelli M, Kluger Y, Ansaloni L, Hardcastle TC, Rello J, Watkins RR, Bassetti M, Giamarellou E, Coccolini F, Abu-Zidan FM, Adesunkanmi AK. Raising concerns about the Sepsis-3 definitions. *World Journal of Emergency Surgery*. 2018 Dec;13(1):1-9.
134. Frey AL, Karran M, Jimenez RC, Baxter J, Adeogun M, Bose N, Chan D, Crawford J, Dagum P, Everson R, Hinds C. Harnessing the potential of digital technologies for the early detection of neurodegenerative diseases (EDoN).
135. Islam MM, Nasrin T, Walther BA, Wu CC, Yang HC, Li YC. Prediction of sepsis patients using machine learning approach: a meta-analysis. *Computer methods and programs in biomedicine*. 2019 Mar 1;170:1-9.
136. Seetharaman S, Wilson C, Landrum M, Qasba S, Katz M, Ladikos N, Harris JE, Galiatsatos P, Yousem DM, Knight AM, Pearse DB. Does use of electronic alerts for systemic inflammatory response syndrome (SIRS) to identify patients with sepsis improve mortality?. *The American Journal of Medicine*. 2019 Jul 1;132(7):862-8.

137. Beam AL, Kohane IS. Big data and machine learning in health care. *Jama*. 2018 Apr 3;319(13):1317-8.
138. Slight SP, Tolley CL, Bates DW, Fraser R, Bigirimurame T, Kasim A, Balaskonis K, Narrie S, Heed A, Orav EJ, Watson NW. Medication errors and adverse drug events in a UK hospital during the optimisation of electronic prescriptions: a prospective observational study. *The Lancet Digital Health*. 2019 Dec 1;1(8):e403-12.
139. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group*. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*. 2009 Aug 18;151(4):264-9.
140. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche JD, Coopersmith CM, Hotchkiss RS. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *Jama*. 2016 Feb 23;315(8):801-10.
141. Morrill JH, Kormilitzin A, Nevado-Holgado AJ, Swaminathan S, Howison SD, Lyons TJ. Utilization of the Signature Method to Identify the Early Onset of Sepsis From Multivariate Physiological Time Series in Critical Care Monitoring. *Critical Care Medicine*. 2020 Oct 30;48(10):e976-81.
142. Scherpf M, Gräßer F, Malberg H, Zaunseder S. Predicting sepsis with a recurrent neural network using the MIMIC III database. *Computers in biology and medicine*. 2019 Oct 1;113:103395.
143. van Wyk F, Khojandi A, Mohammed A, Begoli E, Davis RL, Kamaleswaran R. A minimal set of physiomarkers in continuous high frequency data streams predict adult sepsis onset earlier. *International journal of medical informatics*. 2019 Feb 1;122:55-62.
144. Faisal M, Scally A, Richardson D, Beatson K, Howes R, Speed K, Mohammed MA. Development and external validation of an automated computer-aided risk score for predicting sepsis in emergency medical admissions using the patient's first electronically recorded vital signs and blood test results. *Critical care medicine*. 2018 Apr 1;46(4):612-8.
145. Saqib M, Sha Y, Wang MD. Early prediction of sepsis in EMR records using traditional ML techniques and deep learning LSTM networks. In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2018 Jul 18 (pp. 4038-4041). *IEEE*.
146. Danner OK, Hendren S, Santiago E, Nye B, Abraham P. Physiologically-based, predictive analytics using the heart-rate-to-systolic-ratio significantly improves the timeliness and accuracy of sepsis prediction compared to SIRS. *The American Journal of Surgery*. 2017 Apr 1;213(4):617-21.
147. Back JS, Jin Y, Jin T, Lee SM. Development and validation of an automated sepsis risk assessment system. *Research in nursing & health*. 2016 Oct;39(5):317-27.
148. Calvert JS, Price DA, Chettipally UK, Barton CW, Feldman MD, Hoffman JL, Jay M, Das R. A computational approach to early sepsis detection. *Computers in biology and medicine*. 2016 Jul 1;74:69-73.

149. Mehta S, Gill SE. Improving clinical outcomes in sepsis and multiple organ dysfunction through precision medicine. *Journal of Thoracic Disease*. 2019 Jan;11(1):21.
150. Liu R, Greenstein JL, Granite SJ, Fackler JC, Bembea MM, Sarma SV, Winslow RL. Data-driven discovery of a novel sepsis pre-shock state predicts impending septic shock in the ICU. *Scientific reports*. 2019 Apr 16;9(1):6145.
151. Fleuren LM, Klausch TL, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, Swart EL, Girbes AR, Thorat P, Ercole A, Hoogendoorn M. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive care medicine*. 2020 Mar;46:383-400.
152. Lederer DJ, Bell SC, Branson RD, Chalmers JD, Marshall R, Maslove DM, Ost DE, Punjabi NM, Schatz M, Smyth AR, Stewart PW. Control of confounding and reporting of results in causal inference studies. Guidance for authors from editors of respiratory, sleep, and critical care journals. *Annals of the American Thoracic Society*. 2019 Jan;16(1):22-8.
153. Leisman DE, Harhay MO, Lederer DJ, Abramson M, Adjei AA, Bakker J, Ballas ZK, Barreiro E, Bell SC, Bellomo R, Bernstein JA. Development and reporting of prediction models: guidance for authors from editors of respiratory, sleep, and critical care journals. *Critical care medicine*. 2020 May;48(5):623.
154. Ginestra JC, Giannini HM, Schweickert WD, Meadows L, Lynch MJ, Pavan K, Chivers CJ, Draugelis M, Donnelly PJ, Fuchs BD, Umscheid CA. Clinician perception of a machine learning-based early warning system designed to predict severe sepsis and septic shock. *Critical care medicine*. 2019 Nov;47(11):1477.
155. Wheeler DS. Is the “golden age” of the “golden hour” in sepsis over?. *Crit Care*. 2015 Dec 1;19(1):447.
156. Tacconelli E, Carrara E, Savoldi A, Harbarth S, Mendelson M, Monnet DL, Pulcini C, Kahlmeter G, Kluytmans J, Carmeli Y, Ouellette M. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *The Lancet Infectious Diseases*. 2018 Mar 1;18(3):318-27.
157. Henry KE, Kornfield R, Sridharan A, Linton RC, Groh C, Wang T, Wu A, Mutlu B, Saria S. Human-machine teaming is key to AI adoption: clinicians’ experiences with a deployed machine learning system. *npj Digital Medicine*. 2022 Jul 21;5(1):1-6.
158. Liang YW, Zhu YF, Zhang R, Zhang M, Ye XL, Wei JR. Incidence, prognosis, and risk factors of sepsis-induced cardiomyopathy. *World Journal of Clinical Cases*. 2021 Nov 11;9(31):9452.
159. Tanner J, Khan D, Aplin C, Ball J, Thomas M, Bankart J. Post-discharge surveillance to identify colorectal surgical site infection rates and related costs. *J Hosp Infect*. 2009;72(3):243-250..
160. Annual Epidemiological Report for 2018–2020 Healthcare-associated infections: surgical site infections <https://www.ecdc.europa.eu/sites/default/files/documents/Healthcare-associated%20infections%20-%20surgical%20site%20infections%202018-2020.pdf> [Accessed on 15th September 2023]

161. York Health Economics Consortium (YHEC)-Feb 2017. <https://www.york.ac.uk/news-and-events/news/2017/research/sepsis-economy-report/> [Accessed on 14th January 2022]
162. Guest JF, Keating T, Gould D, Wigglesworth N. Modelling the annual NHS costs and outcomes attributable to healthcare-associated infections in England. *BMJ open*. 2020 Jan 1;10(1):e033367.
163. Savage JW, Anderson PA. An update on modifiable factors to reduce the risk of surgical site infections. *Spine J*. 2013;13(9):1017–29.
164. Páramo-Zunzunegui J, Alonso-García M, Rodríguez-Villar D, Drewniak-Jakubowska J, Calvo-Espino P, Cuberes-Montserrat R, Rodríguez-Caravaca G, Durán-Poveda M. Incidence of surgical infection and risk factors in colorectal surgery-A prospective cohort study. *Cirugía y cirujanos*. 2021 Apr;89(2):156-62..
165. Martínez-Escribano C, Arteaga Moreno F, Pérez-López M, Cunha-Pérez C, Belenguer-Varea Á, Cuesta Peredo D, Blanco González FJ, Tarazona-Santabalbina FJ. Malnutrition and increased risk of adverse outcomes in elderly patients undergoing elective colorectal Cancer surgery: A case-control study nested in a cohort. *Nutrients*. 2022 Jan 3;14(1):207.
166. Gillis C, Wischmeyer PE. Pre-operative nutrition and the elective surgical patient: why, how and what?. *Anaesthesia*. 2019 Jan;74:27-35.
167. El-Rashidy N, Abuhmed T, Alarabi L, El-Bakry HM, Abdelrazek S, Ali F, El-Sappagh S. Sepsis prediction in intensive care unit based on genetic feature optimisation and stacked deep ensemble learning. *Neural Computing and Applications*. 2022 Mar;34(5):3603-32.
168. UK Standards for Microbiology Investigations : http://www.apsi.it/public/ufiles/smi/b37_8_en_141104.pdf [Accessed on 2nd August 2023]
169. Tom Evans. Diagnosis and management of sepsis. *Clinical Medicine* Apr 2018, 18 (2) 146-149; DOI: 10.7861/clinmedicine.18-2-146
170. Hassan N, Slight R, Morgan G, Bates DW, Gallier S, Sapey E, Slight S. Road map for clinicians to develop and evaluate AI predictive models to inform clinical decision making. *BMJ Health & Care Informatics*. 2023 Aug 1;30(1):e100784.
171. Ljungström L, Pernestig AK, Jacobsson G, Andersson R, Usener B, Tilevik D. Diagnostic accuracy of procalcitonin, neutrophil-lymphocyte count ratio, C-reactive protein, and lactate in patients with suspected bacterial sepsis. *PloS one*. 2017 Jul 20;12(7):e0181704.
172. <https://zenodo.org/record/823967> [Accessed on 25th July 2023]
173. Deeply phenotyped sepsis patients within hospital: onset, treatments & outcomes <https://www.pioneerdatahub.co.uk/dataset/deeply-phenotyped-sepsis-patients-within-hospital-onset-treatments-outcomes/> [Accessed on 3rd August 2023]
174. Herman TF, Bordoni B. Wound Classification. [Updated 2023 Aug 11]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK554456/> [Accessed on 3rd August 2023]
175. Yin J, Tian L. Joint confidence region estimation for area under ROC curve and Youden index. *Statistics in medicine*. 2014 Mar 15;33(6):985-1000.

176. Moons KG, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Annals of internal medicine*. 2019 Jan 1;170(1):W1-33.
177. Collaborative G. Determining the worldwide epidemiology of surgical site infections after gastrointestinal resection surgery: protocol for a multicentre, international, prospective cohort study (GlobalSurg 2). *BMJ open*. 2017 Jul 1;7(7):e012150.
178. West MA, Wischmeyer PE, Grocott MP. Prehabilitation and nutritional support to improve perioperative outcomes. *Current Anesthesiology Reports*. 2017 Dec;7:340-9.
179. Du Y, Shi H, Yang X, Wu W. Machine learning for infection risk prediction in postoperative patients with non-mechanical ventilation and intravenous neurotargeted drugs. *Frontiers in Neurology*. 2022;13.
180. Li P, Wang Y, Li H, Cheng B, Wu S, Ye H, Ma D, Fang X. Prediction of postoperative infection in elderly using deep learning-based analysis: an observational cohort study. *Aging Clinical and Experimental Research*. 2023 Jan 4:1-9.
181. Parikh R, Mathai A, Parikh S, Sekhar GC, Thomas R. Understanding and using sensitivity, specificity and predictive values. *Indian journal of ophthalmology*. 2008 Jan;56(1):45.
182. Chen KA, Joisa CU, Stem J, Guillem JG, Gomez SM, Kapadia MR. Improved Prediction of Surgical Site Infection after Colorectal Surgery Using Machine Learning. *Diseases of the Colon & Rectum*. 2022 Nov 30:10-97.
183. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PloS one*. 2017 Apr 6;12(4):e0174708.
184. Hardie T, Horton T, Thornton-Lee N, Home J, Pereira P. Developing Learning Health Systems in the, UK: Priorities for Action. The Health Foundation. 2022 Sep 1;10.
185. Foley T, Horwitz L, Zahran R. Realising the potential of learning health systems. The Learning Healthcare Project. 2021 May.
186. Rigby AS, ROBINSONOE MB. Statistical methods in epidemiology. IV. Confounding and the matched pairs odds ratio. *Disability and Rehabilitation*. 2000 Jan 1;22(6):259-65.
187. O'Brien N, Van Dael J, Clarke J, Gardner C, O'Shaughnessy J, Darzi A, Ghafur S. Addressing racial and ethnic inequities in data-driven health technologies.
188. Duggal R, Brindle I, Bagenal J. Digital healthcare: regulating the revolution. *BMJ*. 2018 Jan 15;360.
189. d'Eon G, d'Eon J, Wright JR, Leyton-Brown K. The spotlight: A general method for discovering systematic errors in deep learning models. In 2022 ACM Conference on Fairness, Accountability, and Transparency 2022 Jun 21 (pp. 1962-1981).
190. Hassan N, Slight R, Slight S. Healthcare staff perceptions on using artificial intelligence predictive tools: a qualitative study. *BMJ Health & Care Informatics*. 2022:A4-5.

191. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence
<https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf> [Accessed on 15th January 2023]
192. Schönberger D. Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. *International Journal of Law and Information Technology*. 2019 Jun 1;27(2):171-203.
193. Khan MS, Bates D, Kovacheva VP. The Quest for Equitable Health Care: The Potential for Artificial Intelligence. *NEJM Catalyst Innovations in Care Delivery*. 2021 Dec 22;2(6).
194. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. *Science*. 2017 Apr 14;356(6334):183-6.
195. Kilkenny MF, Robinson KM. Data quality: “Garbage in–garbage out”. *Health Information Management Journal*. 2018 Sep;47(3):103-5.
196. Shamsa A, Jang AJ, McGee TM. Documentation of instrumental vaginal deliveries: The benefits of a pro forma. *Health Information Management Journal*. 2016 Dec;45(3):116-20.
197. Flood MM, McDonald SJ, Pollock WE, Davey MA. Data accuracy in the Victorian perinatal data collection: results of a validation study of 2011 data. *Health Information Management Journal*. 2017 Sep;46(3):113-26.
198. Speiser JL, Miller ME, Tooze J, Ip E. A comparison of random forest variable selection methods for classification prediction modeling. *Expert systems with applications*. 2019 Nov 15;134:93-101.
199. Ray S. A quick review of machine learning algorithms. In 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon) 2019 Feb 14 (pp. 35-39). IEEE.
200. Geomini P, Kruitwagen R, Bremer GL, Cnossen J, Mol BW. The accuracy of risk scores in predicting ovarian malignancy: a systematic review. *Obstetrics & Gynecology*. 2009 Feb 1;113(2):384-94.
201. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, Reitsma JB, Collins GS. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS medicine*. 2014 Oct 14;11(10):e1001744.
202. Scott IA, Carter SM, Coiera E. Exploring stakeholder attitudes towards AI in clinical practice. *BMJ Health & Care Informatics*. 2021;28(1).
203. O’Sullivan JW, Banerjee A, Heneghan C, Pluddemann A. Verification bias. *BMJ evidence-based medicine*. 2018 Feb 27.
204. Whiting PF, Rutjes AW, Westwood ME, Mallett S, QUADAS-2 Steering Group. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *Journal of clinical epidemiology*. 2013 Oct 1;66(10):1093-104.
205. Büller HR, ten Cate-Hoek AJ, Hoes AW, Joore MA, Moons KG, Oudega R, Prins MH, Stoffers HE, Toll DB, van der Velde EF, van Weert HC. Safely ruling out deep venous thrombosis in primary care. *Annals of Internal Medicine*. 2009 Feb 17;150(4):229-35.

206. Umemneku Chikere CM, Wilson K, Graziadio S, Vale L, Allen AJ. Diagnostic test evaluation methodology: a systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard—an update. *PLoS One*. 2019 Oct 11;14(10):e0223832.
207. DeCamp M, Lindvall C. Latent bias and the implementation of artificial intelligence in medicine. *Journal of the American Medical Informatics Association*. 2020 Dec;27(12):2020-3.
208. Jones-Jang SM, Park YJ. How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. *Journal of Computer-Mediated Communication*. 2023 Jan;28(1):zmac029.
209. Makhni S, Chin MH, Fahrenbach J, Rojas JC. Equity challenges for artificial intelligence algorithms in health care. *Chest*. 2022 May 1;161(5):1343-6.
210. Riley RD, Ensor J, Snell KI, Debray TP, Altman DG, Moons KG, Collins GS. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *bmj*. 2016 Jun 22;353.
211. Navarro CL, Damen JA, Takada T, Nijman SW, Dhiman P, Ma J, Collins GS, Bajpai R, Riley RD, Moons KG, Hooft L. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *bmj*. 2021 Oct 20;375.
212. Crossnohere NL, Elsaid M, Paskett J, Bose-Brill S, Bridges JF. Guidelines for Artificial Intelligence in Medicine: Literature Review and Content Analysis of Frameworks. *Journal of Medical Internet Research*. 2022 Aug 25;24(8):e36823.
213. Park SH, Han K, Jang HY, Park JE, Lee JG, Kim DW, Choi J. Methods for clinical evaluation of artificial intelligence algorithms for medical diagnosis. *Radiology*. 2023 Jan;306(1):20-31.
214. Collins GS, Dhiman P, Navarro CL, Ma J, Hooft L, Reitsma JB, Logullo P, Beam AL, Peng L, Van Calster B, van Smeden M. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ open*. 2021 Jul 1;11(7):e048008.
215. PROBAST+AI Delphi survey <https://osf.io/w3cfe/> [Accessed on 16th September 2023]
216. Hayden JA, van der Windt DA, Cartwright JL, Côté P, Bombardier C. Assessing bias in studies of prognostic factors. *Annals of internal medicine*. 2013 Feb 19;158(4):280-6.
217. Hayden JA, Côté P, Bombardier C. Evaluation of the quality of prognosis studies in systematic reviews. *Annals of internal medicine*. 2006 Mar 21;144(6):427-37.
218. Data: a new direction - government response to consultation <https://www.gov.uk/government/consultations/data-a-new-direction/outcome/data-a-new-direction-government-response-to-consultation#introduction> [Accessed on 19th September 2023]
219. Addressing Racial and Ethnic Inequities in Data-driven Health Technologies https://spiral.imperial.ac.uk/bitstream/10044/1/94902/2/Imperial_IGHI_AddressingRacialandEthnicInequities%20_Report.pdf [Accessed on 17th August 2023]

220. Roselli D, Matthews J, Talagala N. Managing bias in AI. In Companion Proceedings of The 2019 World Wide Web Conference 2019 May 13 (pp. 539-544).
221. What is shared decision making?(and what it is not). Academic Emergency Medicine. 2016 Dec;23(12):1320-4.
222. Stiggelbout AM, Pieterse AH, De Haes JC. Shared decision making: concepts, evidence, and practice. Patient education and counseling. 2015 Oct 1;98(10):1172-9.
223. Kopecky KE, Urbach D, Schwarze ML. Risk calculators and decision aids are not enough for shared decision making. JAMA surgery. 2019 Jan 1;154(1):3-4.
224. McCannon JB, O'donnell WJ, Thompson BT, El-Jawahri A, Chang Y, Ananian L, Bajwa EK, Currier PF, Parikh M, Temel JS, Cooper Z. Augmenting communication and decision making in the intensive care unit with a cardiopulmonary resuscitation video decision support tool: a temporal intervention study. Journal of palliative medicine. 2012 Dec 1;15(12):1382-7.
225. Nundy S, Montgomery T, Wachter RM. Promoting trust between patients and physicians in the era of artificial intelligence. Jama. 2019 Aug 13;322(6):497-8.
226. Raymond BL, Wanderer JP, Hawkins AT, Geiger TM, Ehrenfeld JM, Stokes JW, McEvoy MD. Use of the American College of Surgeons National Surgical Quality Improvement Program surgical risk calculator during pre-operative risk discussion: the patient perspective. Anesthesia & Analgesia. 2019 Apr 1;128(4):643-50.
227. Gilkey MB, Marcus JL, Garrell JM, Powell VE, Maloney KM, Krakower DS. Using HIV risk prediction tools to identify candidates for pre-exposure prophylaxis: Perspectives from patients and primary care providers. AIDS patient care and STDs. 2019 Aug 1;33(8):372-8.
228. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. International journal of surgery. 2021 Apr 1;88:105906.
229. Will I. An introduction to patient decision aids. Bmj. 2013 Jul 27;347:27.
230. Tong A, Flemming K, McInnes E, Oliver S, Craig J. Enhancing transparency in reporting the synthesis of qualitative research: ENTREQ. BMC medical research methodology. 2012 Dec;12(1):1-8.
231. Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M, Britten N, Roen K, Duffy S. Guidance on the conduct of narrative synthesis in systematic reviews. A product from the ESRC methods programme Version. 2006 Apr 1;1:b92.
232. Critical Appraisal Skills Programme (2018). CASP (Qualitative) Checklist. (online)
233. Manski-Nankervis JA, Biezen R, Thursky K, Boyle D, Clark M, Lo S, Buising K. Developing a Clinical Decision Support Tool for Appropriate Antibiotic Prescribing in Australian General Practice: A Simulation Study. Medical Decision Making. 2020 Jun 8:0272989X20926136.
234. Nelson CA, Pérez-Chada LM, Creadore A, Li SJ, Lo K, Manjaly P, Pournamdari AB, Tkachenko E, Barbieri JS, Ko JM, Menon AV. Patient perspectives on the use of artificial

- intelligence for skin cancer screening: a qualitative study. *JAMA dermatology*. 2020 May 1;156(5):501-12.
235. Qu H, Hu X, Singh JA. Factors influencing implementation of a computerised, individualised, culturally tailored lupus decision aid in lupus clinics: a qualitative semi-structured interview study. *Clinical Rheumatology*. 2019 Oct 1;38(10):2793-801.
 236. Eckman MH, Costea A, Attari M, Munjal J, Wise RE, Knochelmann C, Flaherty ML, Baker P, Ireton R, Harnett BM, Leonard AC. Shared decision making tool for thromboprophylaxis in atrial fibrillation—A feasibility study. *American heart journal*. 2018 May 1;199:13-21.
 237. Ballard AY, Kessler M, Scheitel M, Montori VM, Chaudhry R. Exploring differences in the use of the statin choice decision aid and diabetes medication choice decision aid in primary care. *BMC medical informatics and decision making*. 2017 Dec;17(1):1-6.
 238. Brown B, Cheraghi-Sohi S, Jaki T, Su TL, Buchan I, Sperrin M. Understanding clinical prediction models as ‘innovations’: a mixed methods study in UK family practice. *BMC medical informatics and decision making*. 2016 Dec 1;16(1):106.
 239. Schroy III PC, Mylvaganam S, Davidson P. Provider perspectives on the utility of a colorectal cancer screening decision aid for facilitating shared decision making. *Health Expectations*. 2014 Feb;17(1):27-35.
 240. Jimbo M, Shultz CG, Nease DE, Feters MD, Power D, Ruffin IV MT. Perceived barriers and facilitators of using a Web-based interactive decision aid for colorectal cancer screening in community practice settings: findings from focus groups with primary care clinicians and medical office staff. *Journal of Medical Internet Research*. 2013;15(12):e286.
 241. Schackmann EA, Munoz DF, Mills MA, Plevritis SK, Kurian AW. Feasibility evaluation of an online tool to guide decisions for BRCA1/2 mutation carriers. *Familial cancer*. 2013 Mar 1;12(1):65-73.
 242. Cunich M, Salkeld G, Dowie J, Henderson J, Bayram C, Britt H, Howard K. Integrating evidence and individual preferences using a web-based multi-criteria decision analytic tool. *The Patient: Patient-Centered Outcomes Research*. 2011 Sep 1;4(3):153-62.
 243. Harrison JD, Masya L, Butow P, Solomon M, Young J, Salkeld G, Whelan T. Implementing patient decision support tools: moving beyond academia?. *Patient education and counseling*. 2009 Jul 1;76(1):120-5.
 244. Aoki N, Sakai M, Nakayama T, Fukuhara S, Ohta S, Kikuchi N, Oishi M, Kiuchi T, Nozaki K, Hashimoto N. u-SHARE: web-based decision support/risk communication tool for healthcare consumers with unruptured intracranial aneurysms. In *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems 2007* (p. 1012). IOS Press.
 245. Siminoff LA, Gordon NH, Silverman P, Budd T, Ravdin PM. A decision aid to assist in adjuvant therapy choices for breast cancer. *Psycho-Oncology: Journal of the Psychological, Social and Behavioral Dimensions of Cancer*. 2006 Nov;15(11):1001-13.
 246. Thomson R, Robinson A, Greenaway J, Lowe P. Development and description of a decision analysis-based decision support tool for stroke prevention in atrial fibrillation. *BMJ Quality & Safety*. 2002 Mar 1;11(1):25-31.

247. Berry DL, Hong F, Blonquist TM, Halpenny B, Xiong N, Filson CP, Master VA, Sanda MG, Chang P, Chien GW, Jones RA. Decision regret, adverse outcomes, and treatment choice in men with localised prostate cancer: Results from a multi-site randomised trial. *In Urologic Oncology: Seminars and Original Investigations* 2021 Jan 19. Elsevier.
248. Jones T, Guzman A, Silverman T, Freeman K, Kukafka R, Crew K. Perceptions of Racially and Ethnically Diverse Women at High Risk of Breast Cancer Regarding the Use of a Web-Based Decision Aid for Chemoprevention: Qualitative Study Nested Within a Randomized Controlled Trial. *Journal of Medical Internet Research*. 2021 Jun 8;23(6):e23839.
249. Kim AH, Girgis A, De Cruz P, Siegel CA, Karimi N, Ruban SO, Sechi AJ, Ng WS, Andrews JM, Connor SJ. Development and Feasibility of a Web-Based Decision Aid for Patients With Ulcerative Colitis: Qualitative Pilot Study. *Journal of medical Internet research*. 2021;23(2):e15946.
250. Kosch R, Schiffmann I, Daumer M, Lederer C, Scalfari A, Galea I, Scheiderbauer J, Rahn A, Heesen C. Long-term prognostic counselling in people with multiple sclerosis using an online analytical processing tool. *Multiple Sclerosis Journal*. 2021 Aug;27(9):1442-50.
251. Lau YK, Bhattarai H, Caverly TJ, Hung PY, Jimenez-Mendoza E, Patel MR, Coté ML, Arenberg DA, Meza R. Lung Cancer Screening Knowledge, Perceptions, and Decision Making Among African Americans in Detroit, Michigan. *American Journal of Preventive Medicine*. 2021 Jan 1;60(1):e1-8.
252. Kuppermann M, Kaimal AJ, Blat C, Gonzalez J, Thiet MP, Bermingham Y, Altshuler AL, Bryant AS, Bacchetti P, Grobman WA. Effect of a patient-centered decision support tool on rates of trial of labor after previous cesarean delivery: the PROCEED randomised clinical trial. *Jama*. 2020 Jun 2;323(21):2151-9.
253. Weissman GE, Yadav KN, Srinivasan T, Szymanski S, Capulong F, Madden V, Courtright KR, Hart JL, Asch DA, Ratcliffe SJ, Schapira MM. Preferences for Predictive Model Characteristics among People Living with Chronic Lung Disease: A Discrete Choice Experiment. *Medical Decision Making*. 2020 Jul;40(5):633-43.
254. Wilson LM, Tyler N, Jacobs PG, Gabo V, Senf B, Reddy R, Castle JR. Patient input for design of a decision support smartphone application for type 1 diabetes. *Journal of diabetes science and technology*. 2020 Nov;14(6):1081-7.
255. Coylewright M, Keevil JG, Xu K, Dodge SE, Frosch D, Field ME. Pragmatic study of clinician use of a personalised patient decision aid integrated into the electronic health record: an 8-year experience. *Telemedicine and e-Health*. 2020 May
256. Chew HS, Achananuparp P. Perceptions and needs of artificial intelligence in health care to increase adoption: scoping review. *Journal of medical Internet research*. 2022 Jan 14;24(1):e32939.
257. Abbasgholizadeh Rahimi S, Cwintal M, Huang Y, Ghadiri P, Grad R, Poenaru D, Gore G, Zomahoun HT, Légaré F, Pluye P. Application of artificial intelligence in shared decision making: scoping review. *JMIR Medical Informatics*. 2022 Aug 9;10(8):e36199.
258. Beets B, Newman TP, Howell EL, Bao L, Yang S. Surveying Public Perceptions of Artificial Intelligence in Health Care in the United States: Systematic Review. *Journal of Medical Internet Research*. 2023 Apr 4;25:e40337.

259. A guide to good practice for digital and data-driven health technologies. Accessed from: <https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology>. [Accessed on 17th August 2023].
260. de Jong Y, Ramspek CL, Zoccali C, Jager KJ, Dekker FW, van Diepen M. Appraising prediction research: a guide and meta-review on bias and applicability assessment using the Prediction model Risk Of Bias ASsessment Tool (PROBAST). *Nephrology*. 2021 Dec;26(12):939-47.
261. AlZaabi A, AlMaskari S, AalAbdulsalam A. Are physicians and medical students ready for artificial intelligence applications in healthcare?. *Digital Health*. 2023 Jan;9:20552076231152167.
262. Goodman K, Zandi D, Reis A, Vayena E. Balancing risks and benefits of artificial intelligence in the health sector. *Bulletin of the World Health Organization*. 2020 Apr 4;98(4):230.
263. Tran BX, Vu GT, Ha GH, Vuong QH, Ho MT, Vuong TT, La VP, Ho MT, Nghiem KC, Nguyen HL, Latkin CA. Global evolution of research in artificial intelligence in health and medicine: a bibliometric study. *Journal of clinical medicine*. 2019 Mar 14;8(3):360.
264. Ali O, Abdelbaki W, Shrestha A, Elbasi E, Alryalat MA, Dwivedi YK. A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities. *Journal of Innovation & Knowledge*. 2023 Jan 1;8(1):100333.
265. The Artificial Intelligence in Health and Care Award <https://transform.england.nhs.uk/ai-lab/ai-lab-programmes/ai-health-and-care-award/> [Accessed on 13th September 2023]
266. The NHS Long Term Plan: <https://www.longtermplan.nhs.uk/wp-content/uploads/2019/08/nhs-long-term-plan-version-1.2.pdf> [Accessed on 13th September 2023]
267. Slight SP, Howard R, Ghaleb M, Barber N, Franklin BD, Avery AJ. The causes of prescribing errors in English general practices: a qualitative study. *British Journal of General Practice*. 2013 Oct 1;63(615):e713-20.
268. Antes AL, Burrous S, Sisk BA, Schuelke MJ, Keune JD, DuBois JM. Exploring perceptions of healthcare technologies enabled by artificial intelligence: an online, scenario-based survey. *BMC medical informatics and decision making*. 2021 Dec;21(1):1-5.
269. Esmailzadeh P. Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives. *BMC medical informatics and decision making*. 2020 Dec;20(1):1-9.
270. PIONEER - Our Hub for Acute Care. <https://www.hdruk.ac.uk/helping-with-health-data/health-data-research-hubs/pioneer/> [Accessed on 14th August 2023]
271. Better, Broader, Safer: Using Health Data for Research Analysis. Accessed from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1067053/goldacre-review-using-health-data-for-research-and-analysis.pdf [Accessed on 19th May 2023]

272. Evidence Standards Framework for Digital Health Technologies (March 2019)
<https://www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/evidence-standards-framework/digital-evidence-standards-framework.pdf> [Accessed on 7th June 2023]
273. The code of conduct for data-driven health and care technology; A guide to good practice for digital and data-driven health technologies
<https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology#have-a-clear-value-proposition> [Accessed on 7th June 2023]
274. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, Pestruie J, Phillips M, Konye J, Penzoza C, Ghous M. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*. 2021 Aug 1;181(8):1065-70.
275. Curran GM, Bauer M, Mittman B, Pyne JM, Stetler C. Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. *Medical care*. 2012 Mar;50(3):217.
276. May CR, Albers B, Bracher M, Finch TL, Gilbert A, Girling M, Greenwood K, MacFarlane A, Mair FS, May CM, Murray E. Translational framework for implementation evaluation and research: a normalisation process theory coding manual for qualitative research and instrument development. *Implementation Science*. 2022 Dec;17(1):1-5.